**­­­Improving the Forecast Accuracy of Protected Data Using Time Series Features**

Cameron D. Bale[[1]](#footnote-1)\*, Matthew J. Schneider1, and Jinwook Lee1

1: Drexel University, 3141 Chestnut St., Philadelphia, PA 19104

\**Corresponding Author:*

Cameron D. Bale

3141 Chestnut St., Philadelphia, PA 19104

Office 734

Email: cdb327@drexel.edu

**Abstract**

Time series features are useful for classification, clustering, forecast accuracy prediction, model selection, and forecast combinations. We consider the case of a data owner protecting time series data directly for sharing with forecasters. Current research shows that existing data privacy methods degrade forecast accuracy to unusable levels (Gonçalves et al. 2021a). To overcome this problem, we develop a matrix-based privacy method that swaps time series values when their respective time series are similar on a feature space. Our machine learning-based feature selection methodology significantly improves the forecast accuracy of protected data compared to competitor methods. We show that the most useful features for predicting forecast accuracy are not necessarily the most useful for swapping time series values.  The proposed methodology applies to other use cases where time series features are important, such as replacing missing values, outlier removal, or nowcasting.

Keywords: data privacy; time series features; forecast accuracy; machine learning

1. **Introduction**

Despite the past few years of worldwide data protection legislation, the number of reported data breaches has more than doubled.[[2]](#footnote-2) Fundamental drivers of data breaches include the increased availability of data, parties in data sharing collaborations, and threats from adversaries. Of these threats, it has been estimated that 74% of all data breaches involved human error or deliberate misuse, and 19% of all data breaches were internal involving full-time employees, independent contractors, interns, or other staff. (Verizon Data Breach Investigations Report (2023)). The overarching message is that although privacy can be assured legally and contractually, the data still gets out.

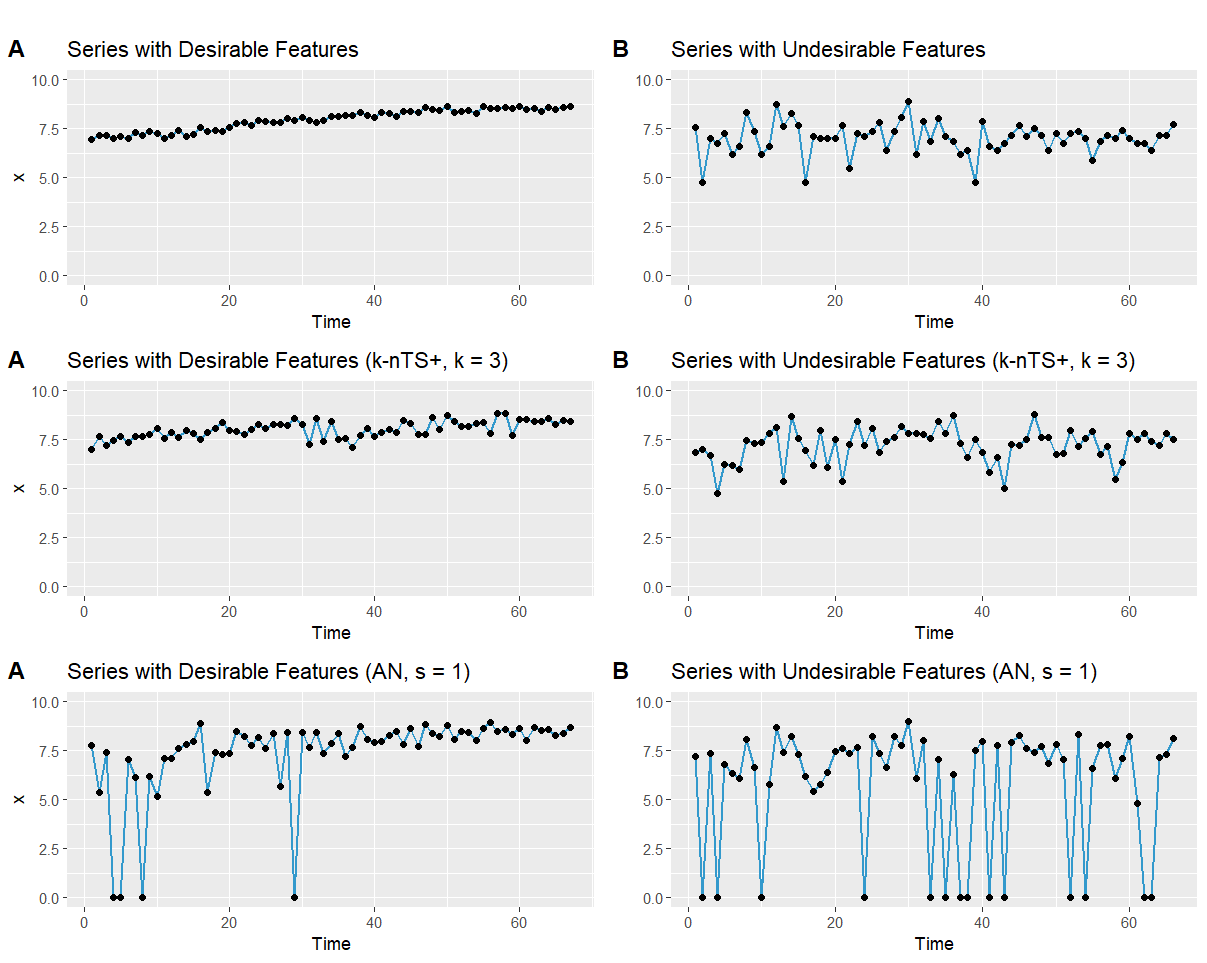
To support good privacy practices, data owners may protect time series data by altering the original data set. This approach assumes the data eventually gets out and protects against the worst-case scenario of a hacker or internal employee breaching the data. Typically, data protection has two goals. The first goal is to limit the ability of an adversary to identify data subjects (identification disclosure, *i.e.*, revealing the true identity of a time series). The second goal of data protection is to limit the ability of an adversary to learn sensitive information about data subjects (attribute disclosure, *e.g.,*, learning sensitive values after an identification is made). In tandem, these goals limit an adversary’s increased knowledge about specific data subjects and have been the primary focus of researchers using differential privacy (Dwork XXXX).

However, an often-overlooked goal and the focus of this paper is maintaining the usefulness of protected data, which has been treated as an afterthought in the computer science and statistical communities (Blanco-Justicia et. al. 2022). For publishing standards, the importance of the theoretical underpinnings of differential privacy dominates the importance of practical problems with real applications (Drechsler 2023). The forecasting literature also found that the tradeoff between the usefulness of protected data and differential privacy levels was poor. Gonçalves et al. (2021a) showed that even protected data with a weak differential privacy level of (it is recommended that lie in the range 0.01 to ln 3 (Dwork, 2011; Blanco-Justicia et al., 2022) reduced forecast accuracy to unusable levels for some data owners.

To achieve good forecast accuracy and comply with privacy practices, Gonçalves et al. (2021c) propose a data market where sensitive data is shared only with a market operator, who utilizes the data to sell forecasts to interested buyers. Assuming a trustworthy market operator, this approach provides privacy since sellers only have access to their own time series and buyers only have access to forecasts from the market operator. In practice, however, data owners may be hesitant to share their data over privacy concerns with the market operator.Sharing forecasts also reduces the usefulness of time series data and has limitations including the assumption of the forecasting model (*e.g.*, SES, VAR, or LGBM) and the inability of a forecaster to calculate time series features (*e.g.*, spectral entropy, strength of trend, or max variance shift). Other researchers (Goncalves et al. 2021b) hare more information such as forecasting model parameters,, which has the advantage of sharing non-personal data which obviates the need to implement data security measures or anonymize personal data according to the General Data Protection Regulation (GDPR),[[3]](#footnote-3) but this assumes the data generating process and hides the underlying data distribution. To bypass these assumptions, our paper introduces a privacy solution that alters the original time series data directly. We propose k-nearest time series+ (k-nTS+) swapping which generates protected data that preserves those time series features needed for good forecast accuracy.

Consider two monthly time series in Figure 1, one with the best of features (series A) and one with the worst of features (series B). Series A has a strong positive trend, low spectral entropy indicating good forecastability, and small shifts in variance. Series B has a negligible trend, high spectral entropy indicating poor forecastability, and large shifts in variance. Any forecaster interested in good forecast accuracy would prefer series A over series B. Furthermore, the only time series features that may be of interest to a forecaster for series B may be the mean and variance.

**Figure 1: Time Series with Desirable (A) and Undesirable (B) Features**



Now, suppose both time series included personal data requiring protection such as monthly household energy consumption, retail transactions, or movements from smart devices (Boone et. al. 2019). How should a data owner alter the time series values in series A so that an adversary cannot re-identify series A and a forecaster can still make good forecasts? The method we propose in this paper involves swapping the past values of series A with similar time series on the feature space. One idea may be to swap the values of series A randomly with time series that have a similar spectral entropy and strength of trend. However, this method may prove ineffective for maintaining forecast accuracy because similarity based on a low spectral entropy and high strength of trend could select time series with strong seasonalities and negative trends. Hence, we propose a machine learning based method which initially selects features using RReliefF (Robnik-Sikonja & Kononenko, 2003) and consolidates the remaining correlated features with a random forest-based recursive feature elimination (RFE) algorithm. Our k-nTS+ swapping method can also be applied to the tasks of outlier replacement (cite Talagal), missing data replacement (Twumasi & Twumasi, 2022), and nowcasting ( Barbaglia et al., 2022). It is important to note that the randomization from the swapping mechanism can be removed for tasks that do not require privacy.

Our contributions to the literature are three-fold. First, we enable forecasters to share a protected version of a time series dataset with good forecast accuracy and reasonable privacy ( GDPR Recital 26). Compared to the prior literature and benchmark methods, our k-nTS+ method performs well on the tradeoff between privacy and accuracy. The method is flexible since it does not assume a forecasting model and gives the user the option to customize the protected data for specific forecast horizons, accuracy metrics, or models. To the best of our knowledge, this paper is the first to produce protected time series data with both acceptable privacy and usable forecast accuracy.

Second, we use a machine-learning based feature selection process to incorporate the usefulness of data (forecast accuracy) into theprotection process. By doing so, welimit the degradation in forecast accuracy in our empirical application to around 14% instead of 40% without the machine learning process. This advantage follows a theme in recent research (Schneider et. al., 2018) which maintained marketing metrics within 10-15% of the original by including a marketing loss function in their privacy method.

Third, we show that the most useful features for predicting forecast accuracy (Makridakis et al., 2018; Spiliotis et al., 2020) are not necessarily the most useful for swapping time series values. Specifically, we find that variance, max variance shift, spike, mean, and max level shift are the most important features to consider for swapping values from similar time series. We analyze how these features change vis-à-vis accuracy and show which forecasting models are more robust to changes in time series features after privacy protection. We also examine whether the magnitude, direction, or volatility of these adjusted forecasts vary the effects of data protection on forecast accuracy.

The rest of the paper is as follows. In Section 2, we review the relevant literature. Section 3 describes the *k*-nTS swapping method and proposes the k-nTS+ swapping method with a feedback loop of the time series features. Section 4 presents the empirical application using M3 Competition data and hierarchical retail data from the M5 competition. Finally, section 5 concludes.

1. **Literature Review**
   1. *Time Series Features*

Time series features are commonly used for classification (Fulcher & Jones, 2014), clustering (Bandara et al., 2018), forecast accuracy prediction (Makridakis et al., 2018; Spiliotis et al., 2020), and model selection and forecast combination (Montero-Manso et al., 2020; Qi et al., 2022; Talagala et al., 2022; Li et al., 2022; Kang et al., 2022). Table 1 illustrates a subset of features used in the literature along with values of these features from series A and series B in Figure 1.

**Table 1: Time series feature descriptions, ranges, and values from series A (desirable) and series B (undesirable).**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Feature** | **Description** | **Value Range** | **Series A** | **Series B** |
| *Spectral Entropy* | Signal-to-noise ratio of the time series. | [0, 1] | 0.07 | 1.00 |
| *Hurst* | Long-range dependence (self-similarity) of a time series. | [0, 1] | 0.99 | 0.50 |
| *Skewness* | Symmetry of the distribution of time series values. |  | -0.41 | -0.57 |
| *Kurtosis* | Weight of the tails of the distribution of time series values. |  | -1.24 | 1.16 |
| *Error ACF* | First autocorrelation coefficient of the error component of the decomposed series. | [-1, 1] | -0.09 | -0.19 |
| *Trend* | Strength of the trend. | [0, 1] | 0.97 | 0.12 |
| *Seasonality* | Strength of the seasonality. | [0, 1] | 0.16 | 0.23 |
| *Mean* | Mean of the time series. |  | 7.96 | 7.01 |
| *Variance* | Variance of the time series. |  | 0.29 | 0.65 |
| *Spike* | Variance of the leave-one-out variances of the remainder component of the decomposed series. |  | 0.0000 | 0.0001 |
| *Max Variance Shift* | Largest variance shift between two consecutive sliding windows. |  | 0.05 | 1.10 |
| *Max Level Shift* | Largest mean shift between two consecutive sliding windows. |  | 0.57 | 0.70 |

In examining which features improve forecast accuracy, Bandara et al. (2018) clustered similar time series based on eighteen interpretable features, including the mean, variance, and strength of seasonality to improve the accuracy of recurrent neural networks (RNNs) between 2 and 11%. Furthermore, the initial results from the M4 competition suggested that the randomness and linearity of time series were the most important determinants of forecast accuracy and that seasonal time series (typically less noisy) are easier to forecast (Makridakis et al., 2018). In a follow-up study, Spiliotis et al. (2020) used multiple linear regression to confirm the importance of randomness, linearity, and seasonal strength in predicting mean absolute scaled error (MASE) values of the ETS, ARIMA, Theta, and Naïve 2 (random walk applied to seasonally adjusted data) models from the M4 competition. They found that increasing the frequency, kurtosis, linearity, and seasonal strength of time series improved forecast accuracy, but increasing skewness, self-similarity, and randomness degraded forecast accuracy.

The predictive power of time series features has been leveraged for other tasks such as forecasting model selection. Qi et al. (2022)found that forecasts using the strength of trend and seasonality for exponential smoothing model selection had lower errors across multiple forecast accuracy metrics than information-based selection methods for the majority of forecast horizons. Petropoulos & Siemsen (2022) created a representativeness metric that selects models with trend and seasonality components when the respective signals of these components are strong. For most data frequencies, their approach lowered MASE on the M, M3, and M4 competition data and selected the best forecasting model approximately 3% more often than the other selection methods.Leveraging time series features can also improve forecast combination accuracy. Montero-Manso et. al. (2020) used 42 time series features on historical data with FFORMA (Feature-based FORecast Model Averaging) to tune a weighted combination of forecasts. They applied their method to the M4 competition data and found that their point estimates and prediction intervals outperformed all but one method in the M4 competition. Talagala et al. (2022) applied a meta-learning algorithm based on Bayesian multivariate surface regression to 37 features, including spectral entropy and the Hurst exponent, to predict the model combination that would yield the minimum forecast error for the M4 competition data. This approach achieved forecast accuracy on par with the top M4 competition methods with less computational cost. Li et al. (2022) used features such as the first ACF value to propose an interpretable Bayesian forecast combination framework with time-varying weights. In experiments using the M3 competition data, this method reduced the MASE by approximately 1.1% relative to the next-best forecast combination method. Interestingly, using the diversity of forecasts as a single feature achieves similar or better forecast accuracy compared to using dozens of features from historical data (Kang et al., 2022; Li et al., 2022).

* 1. *Privacy Methods*

Time series data are either stored in a single data set (centralized) or spread across multiple data owners and/or data sets (decentralized).In the decentralized scenario (*e.g.*, in the renewable energy sector), multi-party computation and federated learning enable privacy-preserving collaborative forecasting to ensure accurate forecasts while protecting sensitive data (Gonçalves et al., 2021a; Gonçalves et al., 2021b; Sommer et al., 2021). For example, data owners can sell time series data to a market operator who then sells forecasts of the time series data to multiple buyers. This approach has the advantage of creating a market of economic incentives for data sharing while limiting data transfer and protecting privacy. However, there are still privacy risks including potential data breaches with the transfer and storage of the time series data to the market operator. Goncalves et al. (2021c) modeled a data market where data owners are compensated for sharing their time series data and purchase only forecasts based on the data from other parties. However, the original time series were still shared with a central party which discourages data owners from sharing time series due to possibility of a data breach. Other privacy-preserving solutions for collaborative forecasting include secure multi-party computation, decomposition-based methods, and data transformation techniques (see Gonçalves et al. 2021a).

In a centralized scenario, the data owner anonymizes the time series data by directly altering the values within the data set to increase the privacy level of the protected data set. Gonçalves et al. (2021a) showed that differential privacy reduces the forecast accuracy of VAR models under very high values of the privacy parameter (weak privacy protection). Others have also studied the application of differential privacy to time series (Imtiaz et al., 2020; Liyue Fan & Li Xiong, 2014). Luo et al. (2018) simulated data integrity attacks and found that multiplicative noise reduces forecast accuracy by over 21% when only half the data points are altered. Their results likely understate the reduction in forecast accuracy from privacy methods because only half the data points were altered.

Other privacy methods include generalization where the structure of the original data set is changed. Data records can be aggregated or combined to make every record (or time series) identical to at least other records (or time series). For example, daily time series data can be aggregated to weekly time series data (temporal aggregation), or each time series can be averaged with its most similar time series ( anonymity). Using anonymity (weak privacy), Nin & Torra (2009) evaluated the change in forecast accuracy for simple exponential smoothing, double exponential smoothing, linear regression, multiple linear regression, and polynomial regression. They found an overall reduction in forecast accuracy but did not provide the accuracy of each model individually. Also, top- and bottom-coding can be used to replace the tails of distributions with a threshold value, such as $150,000 for income or 10 kilowatts-hours for household energy usage). Top- and bottom-coding limit attribute disclosure risk (*i.e.*, preventing knowledge of specific values within a time series), but may not be effective at limiting identification disclosure risk (*i.e.*, preventing the identification of an entire time series). Top- and bottom-coding could have an effect similar to adjusting for outliers which improves forecast accuracy when the outliers are close to the forecast origin (Chen & Liu, 1993).

* 1. *Adjusted Forecasts*

Privacy methods adjust forecasts by altering the original time series data. Similar to judgmental adjustments, this presents the forecaster with multiple forecasts to choose from. We reference the long history on judgmental forecasting (Petropoulos et al., 2022, see sections 2.11.2 and 3.7.3) investigating how the direction and magnitude of adjustments, and the volatility of forecasts affect forecast accuracy.

There are two critical differences between privacy adjustments and judgmental adjustments.

First, judgmental adjustments alter a forecast after it is output from a forecasting model. The underlying time series and their features are not changed. For the direction of the adjustment, Davydenko & Fildes (2013) found that both positive and negative adjustments can improve accuracy, but positive adjustments tend to give only a marginal improvement. Khosrowabadi et al. (2022) similarly found that beneficial positive adjustments tended to be small, and beneficial negative adjustments tended to be large. Fildes et al. (2009) showed that negative adjustments reduce forecast bias, whereas positive adjustments maintain bias or exacerbate it. The magnitude of judgmental adjustments is also positively associated with the size of accuracy improvements when adjustments are based on reliable information. For volatility, accuracy improvements are greater for time series that have forecast errors with low volatility, presumably because adjusters struggle to assess the effect of future events accurately when a time series is more volatile (Fildes et al., 2009).

Second, the motivation for judgmental adjustments is different. Motivations include gaining control of the forecasting process, incorporating practitioner expectations, and compensating for judgmental biases (Petropoulos et al., 2022, sec. 3.7.3). The goal is to incorporate the intuition and experience of the adjuster, knowledge of special events, or insider or confidential information to improve forecast accuracy (Fildes et al., 2019). Despite varying motivations, judgmental adjustments have been shown to improve forecast accuracy by 5-10% on average (Davydenko & Fildes, 2013; Khosrowabadi et al., 2022). For privacy adjustments, the goal is to improve privacy by blurring the data. The assumption is that forecast accuracy will not improve – instead, utility (forecast accuracy) will tradeoff with privacy (Duncan & Stokes, 2004).

1. **The k-nearest Time Series (k-nTS) Swapping Method**

We solve the data protection problem for the data owner using a matrix-based k-nTS (k-nearest time series) swapping method, where the data owner releases a set of protected time series where is based on , the original values of all series through time . To create a protected series , the *k*-nTS swapping method finds the k most similar time series to where similarity is based on the time series features. For each period *t*, it randomly chooses one of the k similar series to and replaces with the original value at time *t* from the randomly chosen series.

Depending on the quantity of available data, *k*-nTS swapping can use rolling windows of data that adjust for dynamic changes in time series features. For example, if we choose a rolling window of size *n*, then where . Protection in subsequent periods from to rolls forward by one time period. We label the time series features for the current window as which we refer to as the feature vector for time series *j* in time period *t* based on the *n* values in For simplicity, we omit the *t* subscript for the feature vectors and write .

For each time series , the data owner computes the feature vector . This vector can contain any single-valued feature calculated from the values in , such as the strength of the trend and seasonality, the spectral entropy, or the mean value of the current window. Let be the set of *m*-vectors containing the features from each of the *J* time series windows. For each , the data owner computes a set of squared distances of the elements of . We define as the distance between and , *i.e.*, the feature vectors corresponding to two distinct time series from . Without loss of generality, we use the Euclidean norm, or ℓ2-norm, as a distance metric[[4]](#footnote-4). Since our case is multivariate and partially ordered, we can get a totally ordered set based on the Euclidean distance.

We define as the *k*th nearest neighbor of , with the corresponding feature vector . Then, for a time series , we have such that for any integers where . Note that and the superscript means the *i*th order statistic of the related Euclidean distances of all from . Thus, for a given time series vector , its *k*-nearest time series can be represented as the set based on an ordered set .

f

A screenshot of a computer

Description automatically generated with medium confidence

(1)

**Algorithm 1: The k-nTS Swapping Method**

**Require [Initialization]:**

: the matrix of original time series.

: the matrix of time series features.

: the feature distance matrix.

**for do**

1. Find the set for by sorting the th column of from smallest to largest and finding the th smallest component.
2. Replace the last component of with the last component of for a randomly chosen

**end for**

We calculate the matrix using the fact that , which can be written as the following:

, (2)

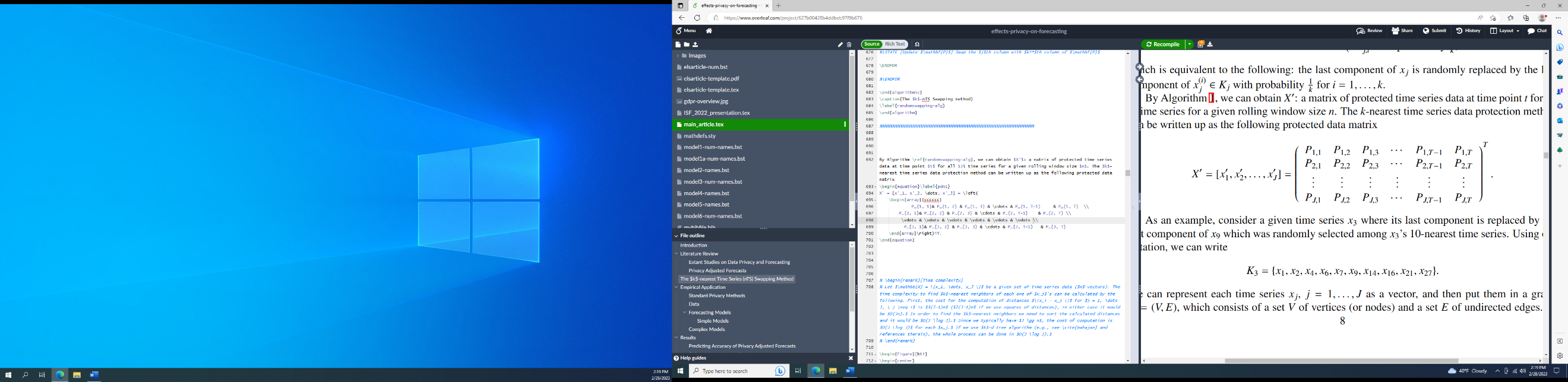
where **1** denotes a column vector of ones[[5]](#footnote-5). It is easy to see that the column vector Let denote the *j*th column of . Then we can write the distance matrix where In the general case where , for each time series we sort and take the *k* smallest components so that we have

. (3)

That is, the data owner selects a value of *k* from 1 to a maximum of and selects the *k*-nearest time series to based on the *m* features. Let the *i*th most similar time series to be . Swapping the last component of with the last component of one of its *k*-nearest time series , is:

with probability (4)

By Algorithm 1, we can obtain , a matrix of protected time series data through time *t* for all *J* time series. All of the first values of each time series are swapped based on the *k*-nearest time series to the first rolling window. For each successive time period, the window is rolled forward, the *k*-nearest neighbors are re-calculated, andswapping is performed on a rolling basis. The output of the *k*-nTS swapping method can be written as the following protected data matrix,



(5)

* 1. **The *k*-nearest Time Series + (k-nTS+) Swapping Method**

The k-nTS+ swapping method adds a feature selection process to k-nTS swapping which selects features that are good predictors of forecast accuracy. The data owner can provide a large selection of time series features to the method. However,using all of the features for the basis of swapping would significantly increase the dimensionality and reduce the efficiency of a swapping process. To address this problem, a random forest-based recursive feature elimination (RFE) algorithm is applied to features initially selected by RReliefF (Robnik-Sikonja & Kononenko, 2003). Prior work has shown that random forest-based RFE is efficient when applied to sets of highly correlated features (Gregorutti et al. 2017), ensuring that the method selects a small set of features that predict forecast accuracy well.

**Figure 2: Flowchart of the k-nTS+ data protection process.**

A diagram of a company

Description automatically generated

Figure 2 illustrates the k-nTS+ swapping process, which can be used collaboratively between a data owner and the forecaster. The forecaster specifies their preferred forecasting model The data owner generates protected versions of the original data set up through time period using baseline privacy methods with varying levels of privacy strength. The data owner computes the forecast errors from using the model to forecast time periods through for each version of the data. These errors are the target variable in the feature selection process, which uses RReliefF and random forest-based RFE to select an efficient set of features for predicting the forecast errors of the model The selected features are then used as the basis for swapping to create the protected data set from through original data set up through time period . The data owner can perform the swapping process over successive time periods to release protected data for times and beyond, and may choose to perform the feature selection process again for any time period. Algorithm 2 specifies the k-nTS+ swapping method below.

**Algorithm 2: The k-nTS+ Swapping Method**

**Require [Initialization]**:

: the matrix of original time series.

: the baseline privacy methods.

: the desired forecasting model.

: set of functions for calculating time series features.

: the number of nearest neighbor time series to consider for swapping.

: the number of nearest neighbor time series to consider for RReliefF.

the number of recursive feature elimination iterations

~~: recursive feature elimination prediction error threshold~~

*h:* forecast horizon

*Step 1: Create Baseline Protected Datasets*

1. Store the data values from all series from time(s) through in as a test set:
2. Create protected data set for each baseline privacy method

*Step 2: Generate Baseline Forecasts*

1. Generate forecasts using for time(s) through based on the original data and each protected data set

*Step 3: Measure Forecast Accuracy*

1. Compute average forecast error at the series level for the original data forecasts and for the protected data forecasts across times through

*Step 4: Extract and Select Time Series Features*

1. Extract cross-sectional time series features matrices from and , from
2. Create cross-sectional feature matrix and forecast error vector by concatenating the feature matices and error vectors from the baseline protected and original data sets
3. Treat forecast errors as the target and time series features as the predictors. Generate weight for each of the N featuresRReliefF(Robnik-Sikonja & Kononenko, 2003) with nearest neighbor parameter .
4. Select features which contains the names of the features with RReliefF weights greater than zero.
5. Create cross-sectional feature matrix from such that contains the features
6. **for** :
   1. Train a random forest to predict using .
   2. Calculate , the mean-abs1olute error of the random forest out-of-bag predictions for iteration and number of features .
   3. Calculate importance of each feature as the change in mean-absolute error of the out-of-bag predictions after permuting the feature in
   4. **for** subset size :
      1. Drop feature with the lowest importance from such that features remain.
      2. Assign rank to for iteration .
      3. **if** Repeat steps (b.) and (c.)

**end for**

**end for**

1. Compute the average mean-absolute error of the out-of-bag predictions using features for
2. Compute , the average rank of each feature
3. Identify the number of features with the minimum
4. Calculate for the percentage increase in the average out-of-bag mean absolute error from using features in the random forest model
5. Set to the smallest value of with
6. Select the features with the best (lowest) average ranks , denoted

*Step 5: Create Protected Data Set using k-nTS+ Swapping*

1. Use input *k* and the feature functions selected in *Step 4* to perform swapping through time *T* using **Algorithm 1: The k-nTS Swapping Method.**
2. **Illustrative Example Using Simulated Time Series**
   1. *Data*

We simulate two sets of ten time series each with “desirable” and “undesirable” features, respectively. To simulate a time series with desirable features, we first sample a target spectral entropy value from a uniform distribution over the range [0.3, 0.4]. We then sample a target value for the first autocorrelation coefficient from a uniform distribution over the range [0.7, 0.9]. The target for the second autocorrelation coefficient is sampled from a uniform distribution over the range . The target feature values for the series with undesirable features are sampled in a similar fashion, where the distributions for and are defined over the ranges and , respectively. Our goal is to create sets of time series with relatively high signal-to-noise ratios and strong autocorrelations (desirable features) compared to relatively low signal-to-noise ratios and weak autocorrelations (undesirable features).

We sample target and values for each of the desired series and use GRATIS to simulate a time series with the corresponding target features by sampling a time series from a mixture autoregressive (MAR) model optimized by a genetic algorithm to simulate time series with features that are as close as possible to the target features (Kang et al., 2020). Each of the simulated time series is scaled using a mean and standard deviation sampled from uniform distributions over the ranges and , respectively. If necessary, we shift entire time series by a constant value to ensure that the minimum value in any of the time series is one. Figure X plots the four sets of simulated time series.

**Figure X:** Plots of the sets of ten time series. The series with desirable features (high autocorrelation, low spectral entropy) are shown on the left, and sets with undesirable features (low autocorrelation, high spectral entropy) are shown on the right.

A graph of different colored lines

Description automatically generated

The rest of the example proceeds as follows. Subsection 4.2 defines the time series features used for the k-nTS+ swapping method, subsection 4.3 describes the privacy methods and defines identification disclosure risk, subsection 4.4 presents the privacy and forecast accuracy results, and subsection 4.5 analyzes how the time series features change after privacy protection.

* 1. *Time Series Features for Forecast Accuracy*

For k-nTS+, we include all features which could help improve accuracy in the swapping process and which are easily computable using the *tsfeatures* package in R. Examples include *Spike*, *Max Variance Shift*, and *Max Level Shift*. We allow our proposed feature selection method to select the efficient set of features which will balance the trade-off between forecast accuracy and data privacy well. We refer the reader to Hyndman et al. (2022) for a detailed explanation of time series features and provide further mathematical detail on several time series features in the Appendix.

* 1. *Privacy Protection*
     1. *Differential Privacy*

We follow the interpretation and implementation of differential privacy from Gonçalves et al. (2021a). A mechanism satisfies -differential privacy by guaranteeing that, for every output of and every pair of series and which differ on at most one observation,

. (6)

A differentially private time series can be created using a randomized mechanism that adds a vector of random noise values, each of which is drawn from a Laplace distribution with scale parameter , to an original time series . The sensitivity is determined as the maximum absolute difference between two time series and which differ in at most one observation, where . We assess thirty-nine values of ranging from 20 to 1, which ranges from least private to most private.

* + 1. *Additive Noise*

Additive noise adds a normally distributed random number with mean zero and standard deviation to each value in a time series . Protected values can be written , where and . The protection parameter denotes the number of standard deviations of and we assess thirty-nine values of ranging from approximately 0.08 to 3.00, which ranges from least private to most private.

* + 1. *k-nTS and k-nTS+*

The k-nearest series for k-nTS and k-nTS+ swapping are determined using features computed by the data owner from the original data set. To initialize the swapping process, the features and neighbors are calculated for the first window of *n* from each original series. Swapping is performed for each of the *n* values using this same set of neighbors. The window is then rolled forward by one period, the features and neighbors are re-calculated, and a new value is swapped into the last period of the window. This process continues until all the values from time 1 to *T* in each of the original series have been replaced. For the purposes of this example, we select and use the features described in Section 4.2.

To perform feature selection for *k*-nTS+, the data owner first creates protected versions of the original data using differential privacy from time period 1 to . Then, using this protected data from time period 1 to , the data owner generates forecasts for each of the protected data sets at time and computes the absolute error of each forecast for each series. The k-nTS+ swapping method is applied to each forecasting model separately in order to detect the variation in forecast accuracy due to changes in time series features (and not the forecasting model), but our final protected data sets are created using the features which have the highest average importance across forecasting models. For this example, we use differentially private data sets created using the 39 epsilon values described in Section 4.3.1.

We select so that the prediction error of a random forest model using the selected features is within 5% of the minimum average prediction error of the best random forest model. For each k-nTS+ protected data set, we use the features with the highest average rank across the RFE iterations for all forecasting models Next, the data owner uses these features swap all series for time periods 1 to T. The data owner shares this protected data with forecasters who forecast time period . We test k-nTS+ for .

* + 1. *Identification disclosure risk*

For our privacy metric, we assess the ability of each privacy method to protect against *identification disclosure*, which occurs when an adversary correctly predicts the identity of a protected time series. Each protected data set consists of the protected series along with a pseudo identifier, *i.e.*, . Identification disclosure occurs if an adversary (or forecaster) correctly predicts the identity of one or more of the protected time series based on the protected time series and some outside information the adversary possesses. For example, identification disclosure occurs when an adversary correctly states, "Series X comes from the monthly sales of the Roseville, Minnesota Target store.”

We perform simulations of a privacy attack in which an adversary uses original time series values to identify the protected time series. In each simulation, we sample ten sequential values from each original time series and treat these as external information available to the adversary. The adversary predicts the identity of each protected series based on which original values are closest to the protected values from the same time periods.

The metric we use is identification disclosure risk for time series (Nin & Torra, 2006, 2009), , the average proportion of the time series which are correctly identified across the simulated privacy attacks,

(7)

where is the adversary’s prediction of the identity of the th protected time series. We use

to denote the indicator function which is equal to one when identification disclosure occurs, *i.e.*, when the predicted identity is equal to the true identity . We refer the reader to the Appendix for added mathematical details.

* 1. *Simulation Results*

We generate fifty protected data sets for each of the parameter values for each of the privacy methods. For each protected data set we perform the privacy simulation discussed in Section 4.3.4 and generate forecasts using SES and DES forecasting models.

The top row of Figure XX shows the average across the fifty simulations for each privacy method and parameter. The time series with undesirable features tend to be less identifiable than series with desirable features. Overall, none of the privacy methods achieve an acceptable tradeoff between privacy and forecast accuracy. Even under the strongest privacy parameters tested (), the noise-based privacy methods do not achieve acceptably low rates of identification risk (over 30% identified on average) and this still results in increases in forecast error of over 300% and 7% for the series with desirable and undesirable features, respectively. k-nTS+ provides good protection against reidentification, where less than 10% of time series are identified on average, which indicates than an attacker is better off randomly guessing the identities of the protected time series. However, this tends to produce extreme increases in forecast error due to large differences in important time series features, such as the mean.

**Figure XX:** Average identification disclosure risk (top row) and forecast MAE (bottom row) averaged across the fifty protected data sets for each privacy parameter and privacy method. Forecast error is also averaged across the forecasting models. The red line denotes the average values observed for the ten time series with desirable features and the blue corresponds to the ten time series with undesirable features.

A graph of data sets

Description automatically generated

Overall, when time series are sufficiently different from each other or have specific patterns, it will not be possible to obtain acceptable privacy protection and forecast accuracy. Either the time series values (and features) will change significantly to achieve acceptable protection and unacceptable forecast accuracy, or forecasts and features are preserved with little to no privacy. To address this problem, we propose transforming the time series to make them less distinguishable. The U.S. Census has …… (Abowd et al., 2013). In essence, we are forcing similarity between the time series features and values. This forced similarity means that less noise/perturbation is required to achieve acceptable levels of privacy, therefore we are more likely to achieve acceptable forecast accuracy at the same time.

We convert each simulated time series into a series of rates using the following formula

where . We set to begin the conversion. We repeat the data protection and forecasting simulation described above on the transformed series, and the results are depicted in Figure XXX. Overall, we see a much more favorable trade-off between privacy and forecast accuracy for all privacy methods. Interestingly, the series with desirable features are now less identifiable than the series with undesirable features after the rate transformation.

**Figure XXX:** Average identification disclosure risk (top row) and forecast MAE (bottom row) averaged across the fifty protected *rate* data sets for each privacy parameter and privacy method. Forecast error is also averaged across the forecasting models. The red line denotes the average values observed for the ten time series with desirable features and the blue corresponds to the ten time series with undesirable features.

A graph of data sets

Description automatically generated

Table XXX summarizes the simulation results for each privacy method and data set based on the privacy parameter that provided the strongest privacy protection while increasing forecast error by no more than 15%. For method-data set combinations where this criteria was not met for any privacy parameter, we include results for the privacy parameter that resulted in the best forecast accuracy. Converting to rates allows the noise-based privacy methods to offer significantly stronger protection against re-identification at similar percent increases in MAE of the forecasted rates. knTS+ also offers improved privacy protection, with vastly improved forecast accuracy on the series with desirable features. The right-most column of Table XXX contains the percentage change in MAE after transforming the rate forecast back into the original time series space. This is applicable for data owners who retain ownership of the original confidential data, and want to compute forecasts on the original time series scale. The data owner can obtain a forecast for the original confidential series using the rate forecast and known confidential data point (*i.e.*, the forecast origin) as follows:

.

We find that on average the converted rate forecasts from the protected series with desirable features are over 20% more accurate than the corresponding original forecasts. Furthermore, the converted rate forecasts for the series with undesirable features are about 5% less accurate on average than the corresponding original forecasts and more accurate than the forecasts from the protected time series on the original scale while offering lower privacy risk.

**Table XXX:** Average percentage of identified time series, percent change in average MAE, and percent change in average MAE after converting the forecasted rate back to the original time series space for the strongest privacy parameter that resulted in no more than 15% increase in forecast error. If this criteria was not met, we include the results for the privacy parameter that had the smallest increase in forecast error.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Method | Data Set | Rate | Parameter | Average Percent Identified | Percent Change MAE | Percent Change MAE (converted rate) |
| Additive Noise | Desirable |  | s = 0.154 | 100% | 12% | - |
|  | Desirable | X | s = 2.462 | 13% | 14% | -24% |
|  | Undesirable |  | s = 3.000 | 47% | 8% | - |
|  | Undesirable | X | s = 3.000 | 14% | 8% | 5% |
| Differential Privacy | Desirable |  | 20\* | 100% | 22% | - |
|  | Desirable | X | 3.5 | 16% | 10% | -28% |
|  | Undesirable |  | 2.5 | 48% | 6% | - |
|  | Undesirable | X | 2.5 | 16% | 10% | 5% |
| k-nTS+ | Desirable |  | k = 2\* | 18% | 1011% | - |
|  | Desirable | X | k = 2\* | 8% | 23% | -21% |
|  | Undesirable |  | k = 4 | 11% | -8% | - |
|  | Undesirable | X | k = 9 | 5% | 6% | 3% |

In the Appendix Section XX.XX, we include the results from the same simulation applied to sets of one hundred time series. In summary, we find that…. More detailed results can be found in the Appendix.

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

The organizers of the early M competitions did not disclose the true identity of the time series used in their competitions (Makridakis & Hibon, 2000). For our application, this provides a natural connection to privacy because we can compute the identification disclosure risk of each protected time series. We define identification disclosure risk as the probability of matching a protected time series to its original time series in the original data set. Good privacy implies the identification disclosure risk is low or similar to random guessing. To be conservative, we assume that an adversary (possibly a forecaster) has external data on at least one original time series and attempts to match it to the protected time series. The data owner seeks to alter the time series with privacy methods to reduce the identification disclosure risk while maintaining as much forecast accuracy as possible.

Recent work by Spiliotis et al. (2020) shows that the M3 competition data contain time series features representative of real world data which makes it suitable for illustrating our feature-based k-nTS+ swapping method. The data owner protects every single time series value from time period 1 to T. The protected time series are given to forecasters to produce one-step ahead forecasts for time T + 1. The data owner then measures forecast accuracy using the protected and original data against the actual values from T + 1.

We assume the forecaster may be an adversary attempting to identify an original time series by using the protected time series. For calculating identification disclosure risk (see subsection 4.3.4 for further details), we take the most conservative approach and assume that the adversary knows the length of each original series which makes identification easier. Thus, we separate the privacy analysis and protection into the groups of time series with equal lengths within each of the M3 competition data sets. Random guessing averages out to approximately 1.27%. We note that in order to test our proposed method with the desired values of (see Section 4.3.3) we excluded any series that did not have the same length as at least fifteen other series. This left us with 2363 of the original 3003 M3 competition series.

The rest of the empirical application is outlined as follows. Subsection 4.2 defines the time series features used for the k-nTS+ swapping method, subsection 4.3 describes the privacy methods and defines the identification disclosure risk, subsection 4.4 presents the privacy and forecast accuracy results, and subsection 4.5 analyzes how the time series features change after privacy protection. Subsection 4.5 also analyzes whether the volatility of the original time series and magnitude or direction of privacy adjustments maintained forecast accuracy.

* 1. *Time Series Features for Forecast Accuracy*

For comparison, and to demonstrate the uplift in forecast accuracy from using our machine learning based feature selection method, we also perform a manual feature selection and generate k-nTS protected data sets using the time series features that had a relationship with forecast accuracy based on the literature review in Section 2.3, namely, spectral entropy, hurst, skewness, kurtosis, error ACF, trend, seasonality, mean, and variance. We omit stability and non-linearity since these features had little to no effect on accuracy, and frequency because none of the privacy methods we consider change the frequency of the original data.

*Privacy Protection*

* + 1. *Differential Privacy*

We follow the interpretation and implementation of differential privacy from Gonçalves et al. (2021a). A mechanism satisfies -differential privacy by guaranteeing that, for every output of and every pair of series and which differ on at most one observation,

. (6)

A differentially private time series can be created using a randomized mechanism that adds a vector of random noise values, each of which is drawn from a Laplace distribution with scale parameter , to an original time series . The sensitivity is determined as the maximum absolute difference between two time series and which differ in at most one observation, where . We use the values of which ranges from least private to most private.

* + 1. *Additive Noise*

Additive noise adds a normally distributed random number with mean zero and standard deviation to each value in an original time series . Protected values can be written , where and . The protection parameter denotes the number of standard deviations of and we set which ranges from least private to most private.

* + 1. *k-nTS and k-nTS+*

The k-nearest series for k-nTS and k-nTS+ swapping are determined using features computed by the data owner from the original data set. To initialize the swapping process, the features and neighbors are calculated for the first window of *n* from each original series. Swapping is performed for each of the *n* values using this same set of neighbors. The window is then rolled forward by one period, the features and neighbors are re-calculated, and a new values is swapped into the last period of the window. This process continues until all the values from time 1 to *T* in each of the original series have been replaced. We select for the monthly series and… We use the features described in Section 4.2.

To perform feature selection for *k*-nTS+, the data owner first creates protected versions of the original data using additive noise and differential privacy from time period 1 to . Then, using this protected data from time period 1 to , the data owner generates forecasts for each of the protected data sets at time and computes the absolute error of each forecast for each series. The k-nTS+ swapping method is applied to each forecasting model separately in order to detect the variation in forecast accuracy due to changes in time series features (and not the forecasting model), but our final protected data sets are created using the features which have the highest average importance across forecasting models

We select so the features are within 5% of the minimum average prediction error from the best random forest model. For each k-nTS+ protected data set, we use the features with the highest average rank across the RFE iterations for all forecasting models. Next, the data owner uses these features swap all series for time periods 1 to T. The data owner shares this protected data with forecasters who forecast time period . We test the k-nTS and k-nTS+ methods for .

* + 1. *Identification disclosure risk*

As previously mentioned, the forecasters of the M3 competition did not know the identities of the original time series. For our privacy metric, we assess the ability of each privacy method to protect against *identification disclosure*, which occurs when an adversary correctly predicts the identity of a protected time series. Each protected data set consists of the protected series along with a pseudo identifier, *i.e.*, . The pseudo identifier in our application is the `Series` column from the original M3 data, which contains a PID for each time series, *e.g.*, `N1402`. Identification disclosure occurs if an adversary (or forecaster) correctly predicts the identity of one or more of the time series in the M3 data set based on the protected time series and some outside information the adversary possesses. For example, identification disclosure occurs when an adversary correctly states, "Series N1402 comes from the monthly sales of the Roseville, Minnesota Target store.”

We perform simulations of a privacy attack in which an adversary uses original time series values to identify the protected time series. In each simulation, we sample ten sequential values from each original time series and treat these as external information available to the adversary. The adversary predicts the identity of each protected series based on which original values are closest to the protected values from the same time periods.

The metric we use is identification disclosure risk for time series (Nin & Torra, 2006, 2009), , the average proportion of the time series which are correctly identified across the simulated privacy attacks,

(7)

where is the adversary’s prediction of the identity of the th protected time series. We use

to denote the indicator function which is equal to one when identification disclosure occurs, *i.e.*, when the predicted identity is equal to the true identity . We refer the reader to the Appendix for added mathematical details.

* 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. Similar to the M3 Competition, 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. The LGBM, RNN, and VAR models are trained separately on the subsets of time series with equal lengths. We perform minimal data pre-processing and use the standard settings in the off-the-shelf packages. [[6]](#footnote-6)

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

|  |  |
| --- | --- |
| **Model Name** | **Variant** |
| SES | - |
| DES | Additive trend |
| TES | Additive trend/seasonality |
| 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%.[[7]](#footnote-7)

**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 ­. 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*

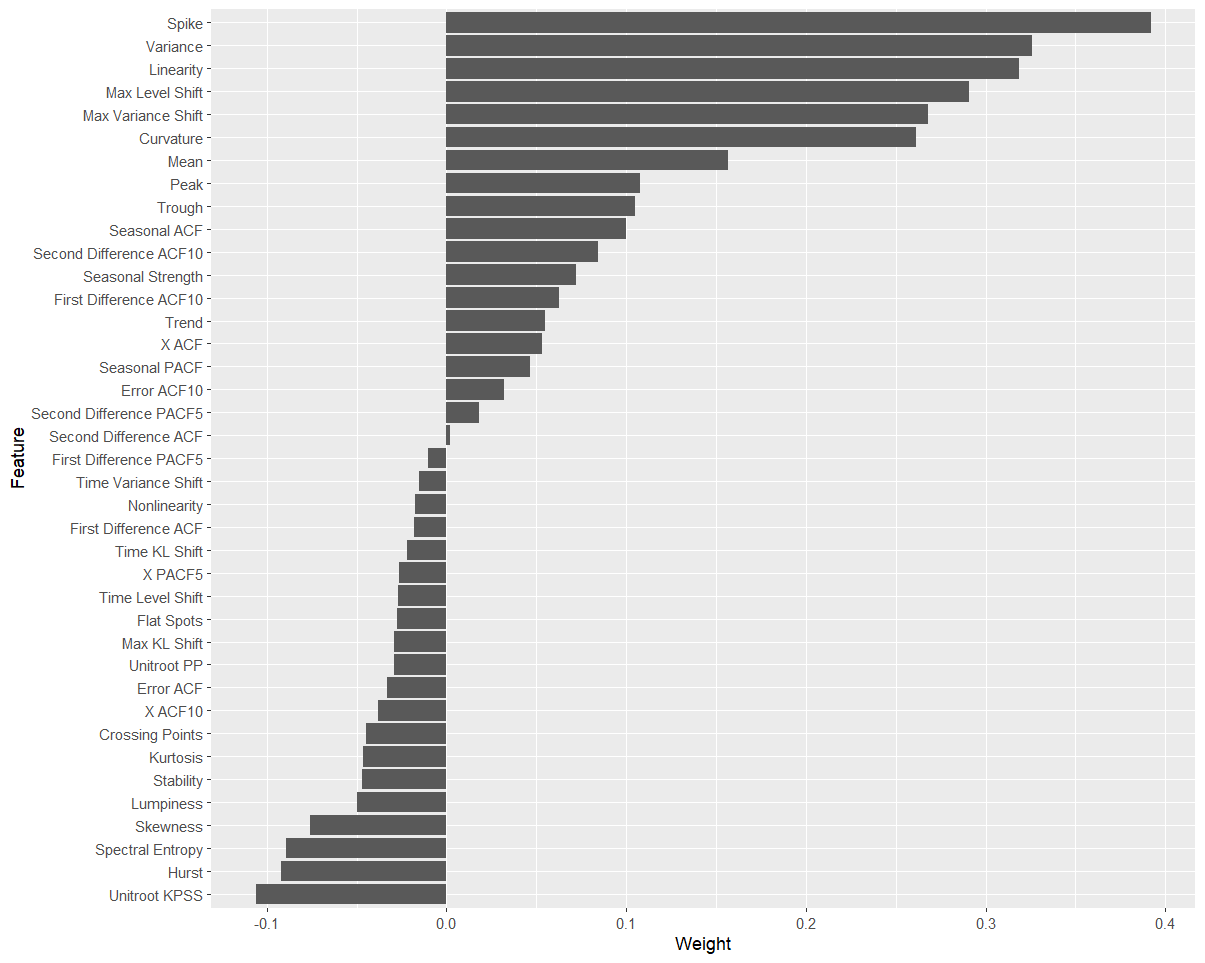
Let denote one of the nearest neighbor feature vectors to , where is the number of nearest neighbors considered by RReliefF, and let and denote the forecast errors for the corresponding time series. Let and denote the events that series and have different forecast errors and different values for feature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

. (8)

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.**

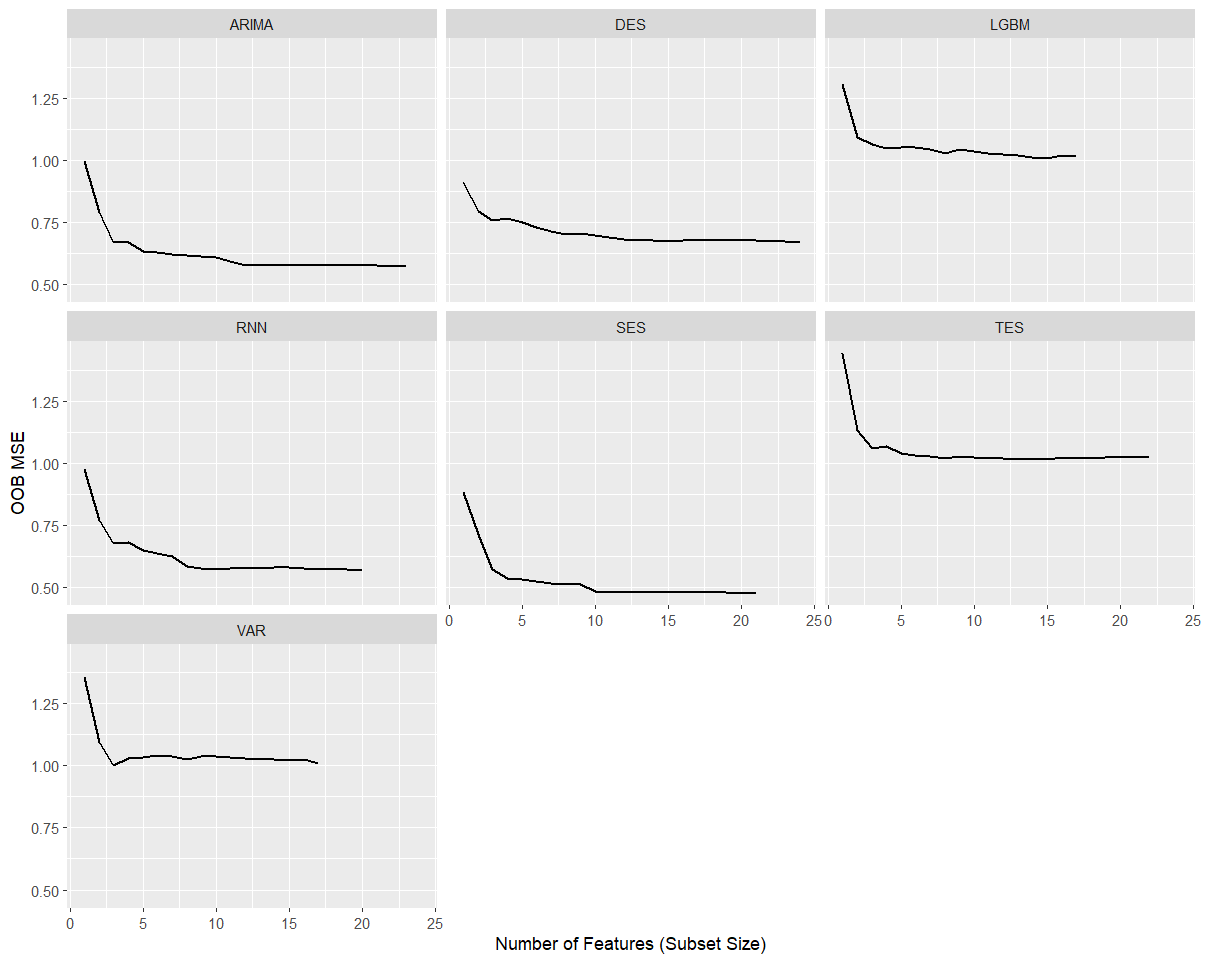


Figure 5 summarizes the results of RFE in Figure 4 and displays the permutation-based importance values for each forecasting model's six most highly ranked features. Some features, such as *spike,* *max variance shift, max level shift, mean,* and *variance* are highly ranked across most or all forecasting 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.**

Chart

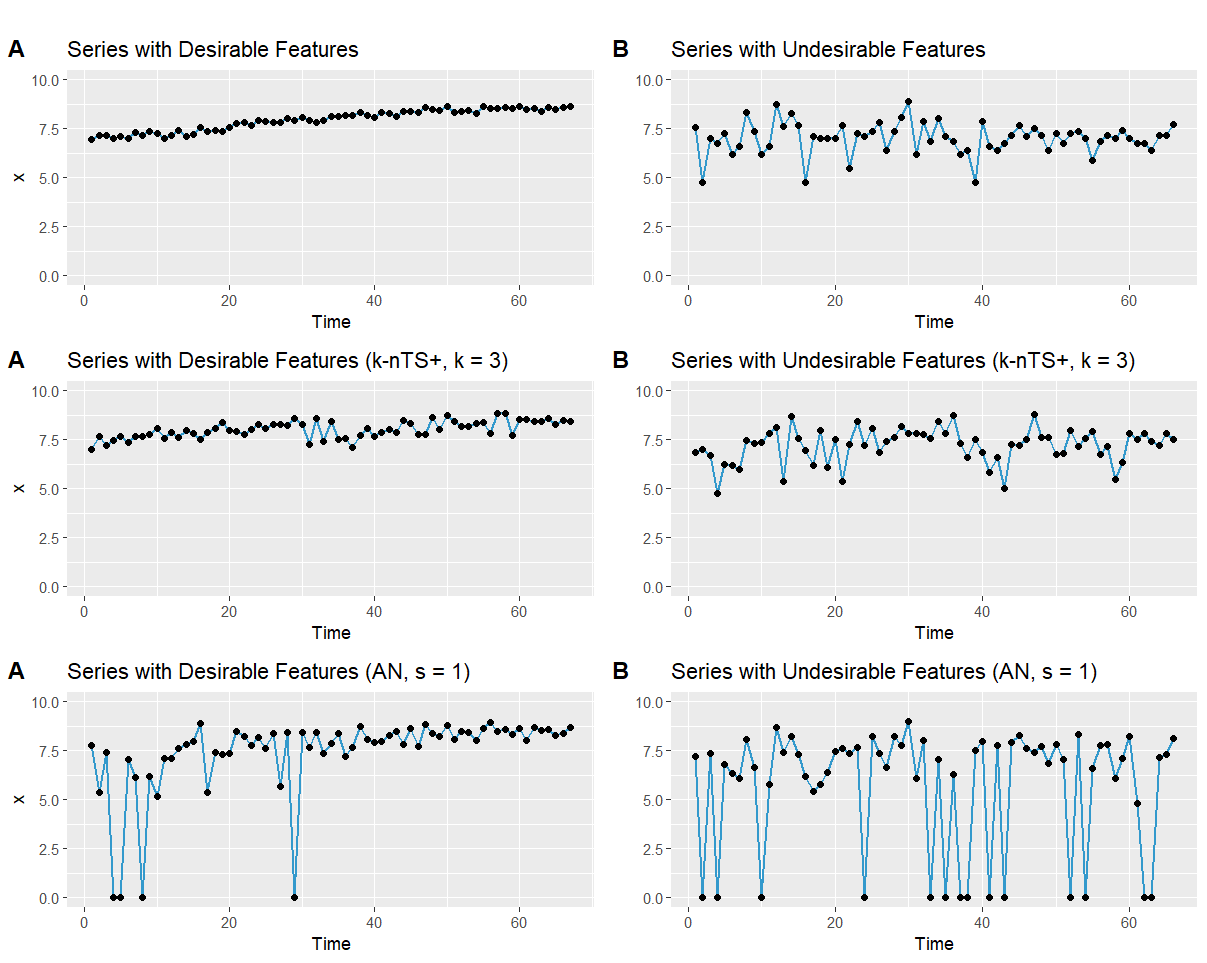
Description automatically generated

* + 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.

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

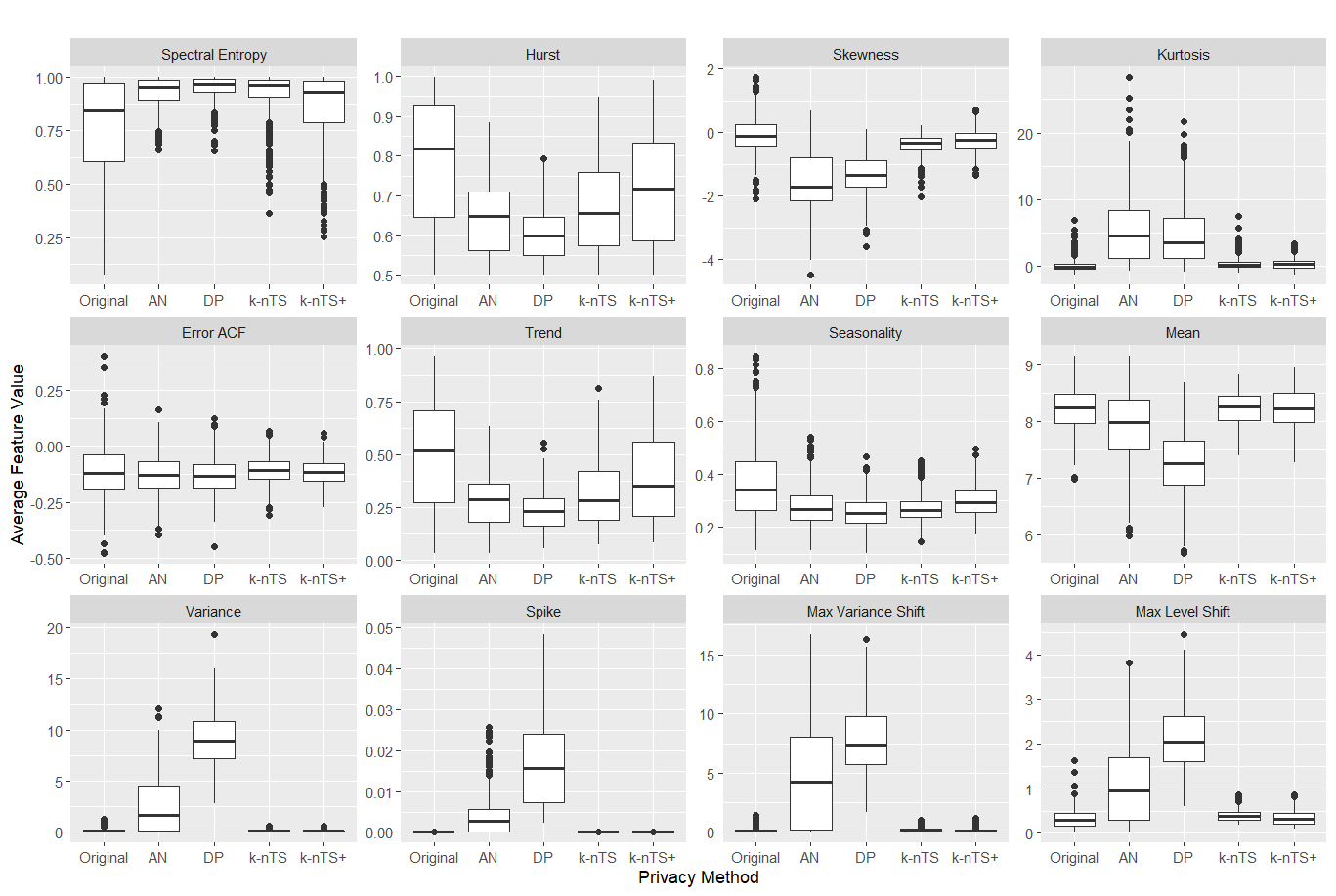


**Table 5: Time series feature values from undesirable and desirable time series.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **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.**



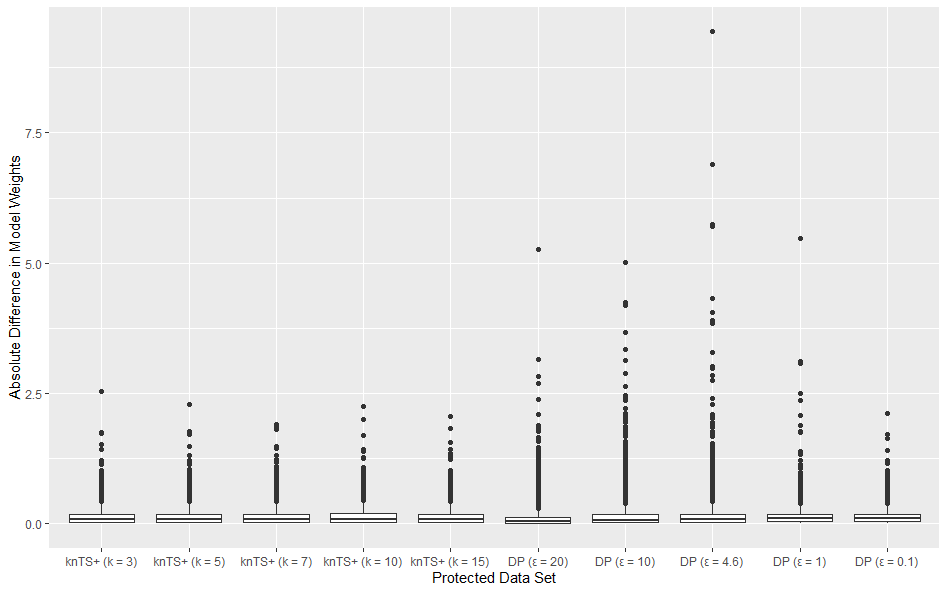
kWe note that while *k*-nTS performed swapping based on the values of *Spectral Entropy, Hurst*, and *Seasonality*, it does not preserve these feature distributions as well as *k*-nTS+. One reason could be that all three of these features are based on autocorrelation and we found that k-nTS swapping 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. *Forecasting Model Parameters*

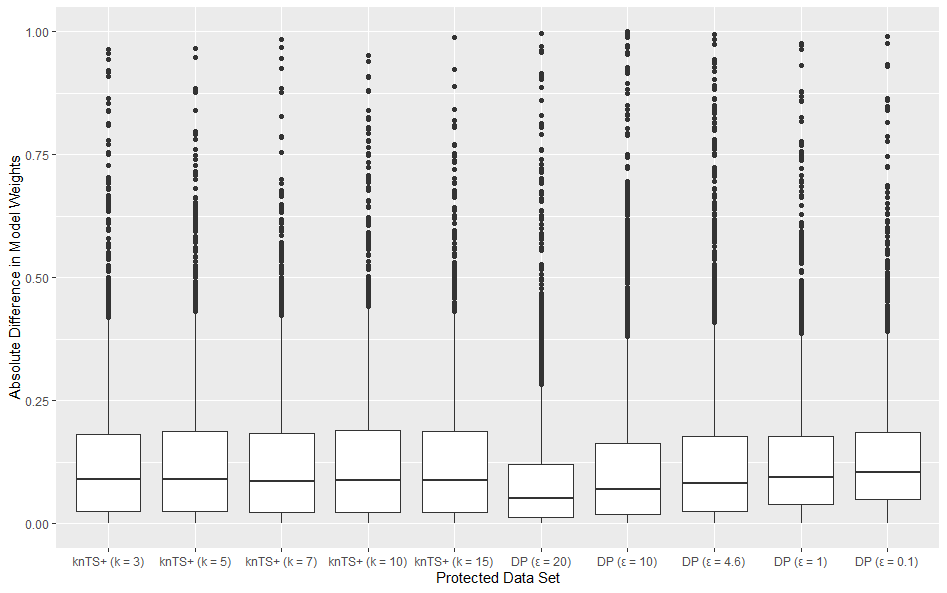
We investigate a real-world setting where a forecaster seeks to utilize a VAR model with a one period lag to illustrate how swapping affects the parameters and forecasts from the trained model. Equations XX show the parameters of a VAR model trained on five of the M3 monthly micro time series.

Add equations from protected data here.

**FIGURE 8**: Absolute differences in VAR model weights for k-nTS+ and DP protected data sets.



**FIGURE 9**: Absolute differences in VAR model weights (trimmed at 1) for k-nTS+ and DP protected data sets.

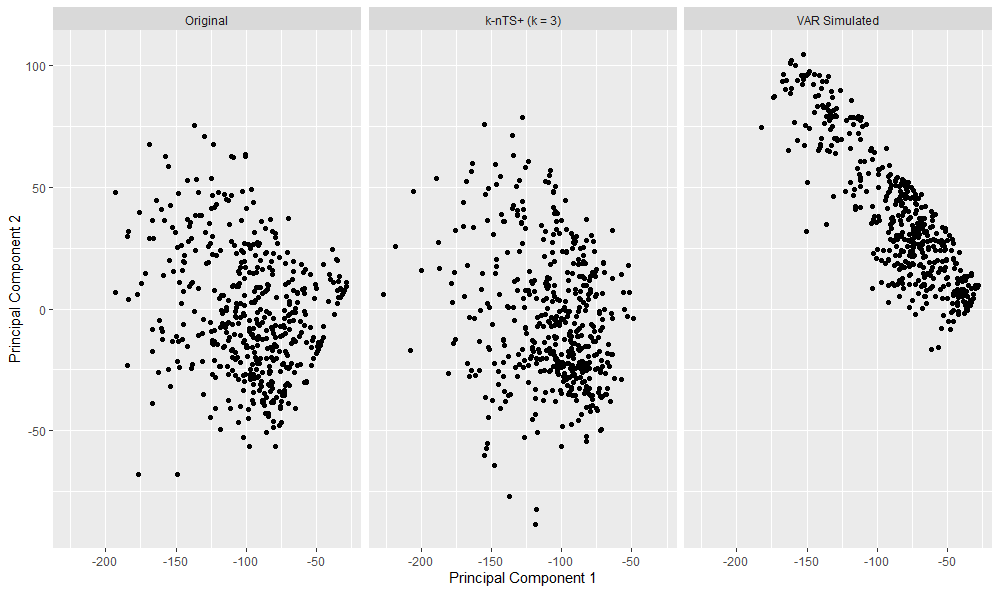


Figures 8 and 9 show…

An alternative solution to sharing protected time series data is sharing models (or model weights) trained on the original data set. This approach is common in scenarios with multiple data owners (*e.g.*, see Gonçalves et al., 2021a). In the centralized scenario, this approach has the downsides of requiring the data owner to train and adjust models on behalf of the forecaster and cannot be used for specific use cases that require a time series data set, such as a forecasting competition.

We consider the case of the data owner sharing the weights of a VAR model with the forecaster, who can then generate their own forecasts and/or simulate time series. We simulate one time series corresponding to each of the training time series in the M3 monthly micro data set. Figure XX shows the projection of the features of the simulated time series from the VAR models compared to the projection of the k-nTS+ (k = 3) protected data onto the first two principal components defined by the time series features of the original M3 monthly micro data.

**FIGURE 10:** first two principal components of the feature vectors from the original M3 monthly micro, k-nTS+ (k = 3) protected monthly micro, and simulated series from VAR weights trained on the original M3 monthly micro.

****

Visually, the features from the k-nTS+ (k = 3) protected data preserve a much more realistic representation of the original feature space. Many of the simulated series contain combinations of features that do not occur in the original data and overall do not provide nearly as much coverage over the total feature space as the k-nTS+ (k = 3) protected data. We performed the privacy attack described in Section 4.3.4 on the simulated series and found that, on average, only 5.8% of time series were identified using their simulated counterparts.

* + - 1. *Computational Details*

Table XXX contains the time (in seconds) to perform each step of the data protection and forecasting process for different numbers of time series. We see that…

* + 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,

(9)

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 Davydenko & Fildes, (2013)) to compare the relative accuracy of the adjusted and original forecasts. The AvgRelAE of the adjusted forecasts is computed as

, (10)

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.[[8]](#footnote-8) We remove the forecasts with the 5% smallest and 5% largest ratios ( to prevent extreme outliers from affecting AvgRelAE (Davydenko & Fildes, 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.

1. **Conclusions**

This paper examined the impact of data privacy on forecast accuracy in a centralized scenario where a data owner shares a protected data set with forecasters. Our proposed *k*-nTS+ swapping method used time series features to swap the values between time series to maintain forecast accuracy. We demonstrated the effectiveness of our privacy method using data from a well-known forecasting competition where the identities of the time series needed to be kept confidential. The proposed method limited the average reduction in forecast accuracy to +13.9% of the original forecast accuracy. Nearly all other privacy methods we studied degraded forecast accuracy to unusable levels (over 100%) at similar levels of privacy.

To the best of our knowledge, this paper is the first to create a protected time series data set tailored to maintain forecast accuracy. We did this by carefully investigating the similarity between time series features rather than time series values. The protected data also preserved important features for forecasting such as spike, max variance shift, max level shift, mean, variance, strength of trend, and strength of seasonality. Furthermore, we showed that our k-nTS+ protected data was more representative of the original time series, potentially leading to increased trust and adoptability among organizations.

A substantial portion of the privacy literature is focused on theoretical privacy guarantees such as differential privacy. Our findings agree with past research (Goncalves et al. 2021a) and show that differential privacy (and additive noise) generates unusable forecasts at reasonable levels of privacy. This undesirable privacy-utility tradeoff has also been demonstrated in contexts other than forecasting. For example, a recent paper by Blanco-Justicia et al. (2022) found that much of the work on differential privacy and deep learning utilized relaxed versions of differential privacy with large values of that theoretically do not provide meaningful levels of privacy protection. Their experiments found that model regularization (e.g., L2-regularization) provided comparable privacy protection with better accuracy and lower model learning cost than differential privacy. In our application, we found that our k-nTS+ swapping method had better forecast accuracy at comparable levels of identification disclosure risk with differential privacy.

Our proposed privacy method can support organizations making efforts to use privacy to improve consumer experiences or perceptions. For example, automating the data protection process using k-nTS+ can limit human intrusions on personal data (Goldfarb & Tucker, 2013). Prioritizing privacy by implementing privacy methods could create a competitive advantage through increased consumer loyalty, trust, and positive performance (Martin & Murphy, 2017). These effects could mitigate damaging outcomes, such as poor stock returns and consumers falsifying information, that arise when consumers feel that their data is vulnerable. Furthermore, replacing sensitive data with protected data could dampen the negative effects of a data breach if the data were to be compromised (Martin et al., 2017). There are also potential benefits from complying with privacy law by using protected time series, such as avoiding the need for consumer consent to re-use data, removing data retention limits, and enabling cross-border data transfers (GDPR Recital 26; Arbuckle & El Emam, 2020). Using protected data could also help organizations avoid fines such as the recent $1.3 billion fined against Meta for transferring EU user data to the US[[9]](#footnote-9).

It is possible for an adversary to use other techniques to compromise the privacy of the protected data, such as reconstruction attacks (Goncalves et al., 2021a). Such an attack is unlikely to succeed since all past data is protected by releasing a single protected data point for each time period, and even if some of these points are similar to the original points, the adversary does not necessarily know which points those are. Further, performing swapping using a rolling window allows the set of *k* nearest time series to change for each successive time period which will incorporate data from a set of time series potentially much larger than *k*. While differentially private time series would be immune to such attacks in theory, we found that forecasts from differentially private data at acceptable levels of are unusable. Increasing improves accuracy but also results in significantly higher identification disclosure risk (~X% for ). Overall, differential privacy could not provide an acceptable trade-off between privacy and forecast accuracy in our application since the differentially private time series were either not useful for forecasting or not protective against identification disclosure.

When applying our proposed privacy method, we only considered the scenario where all time series are stored in a centralized scenario. Our method could also be extended to decentralized scenarios where data is spread across multiple owners. In this case, each data owner could apply the method to their own time series before sharing/pooling data together. In this case, the time series from each data owner could only be swapped with values from the same data owner. Whether this results in adequate privacy and forecast accuracy will depend on how many series the data owner has available for swapping and whether specific values are considered sensitive. In cases of limited data, data owners could consider swapping with time series that are simulated to have the desired features (CITE GRATIS). This situation also presents additional privacy attacks, such as identifying which data owner supplied which time series, rather than individually identifying the series themselves. Rather than data owners performing the data protection, our proposed method could be applied by a central trusted party after data from multiple owners has been pooled. This could increase willingness to share data by alleviating privacy concerns with storing sensitive data in a central location (CITE Goncalves paper).

Although we showed that *k*-nTS+ swapping balanced the tradeoff between forecast accuracy and privacy well, future work could also examine the utility of use cases beyond forecasting. Since many of the time series features were preserved and the entire protected data set was shared, forecasters could use this time series data for other applications. One limitation of our study was that we did not consider privacy metrics other than identification disclosure risk, such as attribute disclosure risk. Further research could pursue this area or address whether combinations of forecasts using multiple protected data sets improve the privacy-utility tradeoff.

*This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.*

**References**

Abowd, J. M., Schneider, M. J., & Vilhuber, L. (2013). Differential privacy applications to Bayesian and linear mixed model estimation. Journal of Privacy and Confidentiality, 5(1).

Arbuckle, L., & El Emam, K. (2020). Building an anonymization pipeline: creating safe data. O'Reilly Media.

Bandara, K., Bergmeir, C., & Smyl, S. (2018). Forecasting Across Time Series Databases using Recurrent Neural Networks on Groups of Similar Series: A Clustering Approach. ArXiv:1710.03222 [Cs, Econ, Stat]. http://arxiv.org/abs/1710.03222

Chen, C., & Liu, L.-M. (1993). Forecasting time series with outliers. Journal of Forecasting, 12(1), 13–35. https://doi.org/10.1002/for.3980120103

Davydenko, A., & Fildes, R. (2013). Measuring forecasting accuracy: The case of judgmental adjustments to SKU-level demand forecasts. International Journal of Forecasting, 29(3), 510–522. https://doi.org/10.1016/j.ijforecast.2012.09.002

Duncan, G. T., & Stokes, S. L. (2004). Disclosure risk vs. data utility: The RU confidentiality map as applied to topcoding. Chance, 17(3), 16-20.

Dwork, C. (2011). A firm foundation for private data analysis. Communications of the ACM, 54(1), 86-95.

European Parliament and Council of European Union. Regulation (EU) 2016/679 of the  
European Parliament and of the Council. 2016. URL: https://eur-lex.europa.eu/  
eli/reg/2016/679/oj

Fildes, R., Goodwin, P., Lawrence, M., & Nikolopoulos, K. (2009). Effective forecasting and judgmental adjustments: An empirical evaluation and strategies for improvement in supply-chain planning. International Journal of Forecasting, 25(1), 3–23. https://doi.org/10.1016/j.ijforecast.2008.11.010

Fildes, R., Goodwin, P., & Önkal, D. (2019). Use and misuse of information in supply chain forecasting of promotion effects. International Journal of Forecasting, 35(1), 144–156. https://doi.org/10.1016/j.ijforecast.2017.12.006

Fulcher, B. D., & Jones, N. S. (2014). Highly Comparative Feature-Based Time-Series Classification. IEEE Transactions on Knowledge and Data Engineering, 26(12), 3026–3037. https://doi.org/10.1109/TKDE.2014.2316504

Goerg, G. (2013). Forecastable component analysis. In International conference on machine

learning (pp. 64-72). PMLR.

Goldfarb, A., & Tucker, C. (2013). Why managing consumer privacy can be an opportunity. MIT Sloan Management Review, 54(3), 10.

Gonçalves, C., Bessa, R. J., & Pinson, P. (2021a). A critical overview of privacy-preserving approaches for collaborative forecasting. International Journal of Forecasting, 37(1), 322–342. https://doi.org/10.1016/j.ijforecast.2020.06.003

Gonçalves, C., Bessa, R. J., & Pinson, P. (2021b). Privacy-Preserving Distributed Learning for Renewable Energy Forecasting. IEEE Transactions on Sustainable Energy, 12(3), 1777–1787. https://doi.org/10.1109/TSTE.2021.3065117

Gonçalves, C., Pinson, P., & Bessa, R. J. (2021c). Towards Data Markets in Renewable Energy Forecasting. IEEE Transactions on Sustainable Energy, 12(1), 533–542. https://doi.org/10.1109/TSTE.2020.3009615

Herzen, J. et al. (2022). Darts: User-friendly modern machine learning for time series. Journal of

Machine Learning Research, vol. 23, num. 124, pg. 1-6.

URL:http://jmlr.org/papers/v23/21-1177.html.

Hewamalage, H., Bergmeir, C., & Bandara, K. (2022). Global models for time series forecasting: A Simulation study. Pattern Recognition, 124, 108441. https://doi.org/10.1016/j.patcog.2021.108441

Hyndman, R.J., & Athanasopoulos, G. (2021) Forecasting: principles and practice, 3rd edition, OTexts: Melbourne, Australia. OTexts.com/fpp3. Accessed on Feb. 23rd, 2023.

Imtiaz, S., Horchidan, S.-F., Abbas, Z., Arsalan, M., Chaudhry, H. N., & Vlassov, V. (2020). Privacy Preserving Time-Series Forecasting of User Health Data Streams. 2020 IEEE International Conference on Big Data (Big Data), 3428–3437. https://doi.org/10.1109/BigData50022.2020.9378186

Kang, Y., Hyndman, R. J., & Smith-Miles, K. (2017). Visualising forecasting algorithm performance using time series instance spaces. International Journal of Forecasting, 33(2), 345–358. https://doi.org/10.1016/j.ijforecast.2016.09.004

Kang, Y., Hyndman, R. J., & Li, F. (2020). GRATIS: GeneRAting TIme Series with diverse and controllable characteristics. Statistical Analysis and Data Mining: The ASA Data Science Journal, 13(4), 354-376.

Kang Y, Cao W, Petropoulos F, et al. Forecast with forecasts: Diversity matters[J]. European Journal of Operational Research, 2022, 301(1): 180-190.

Khosrowabadi, N., Hoberg, K., & Imdahl, C. (2022). Evaluating human behaviour in response to AI recommendations for judgemental forecasting. European Journal of Operational Research, 303(3), 1151–1167. https://doi.org/10.1016/j.ejor.2022.03.017

Koning, A. J., Franses, P. H., Hibon, M., & Stekler, H. O. (2005). The M3 competition: Statistical tests of the results. International Journal of Forecasting, 21(3), 397–409. https://doi.org/10.1016/j.ijforecast.2004.10.003

Li, L., Kang, Y., & Li, F. (2022). Bayesian forecast combination using time-varying features. International Journal of Forecasting, S0169207022000930. https://doi.org/10.1016/j.ijforecast.2022.06.002

Li L, Kang Y, Petropoulos F, et al. Feature-based intermittent demand forecast combinations: accuracy and inventory implications[J]. International Journal of Production Research, 2022: 1-16.

Liyue Fan & Li Xiong. (2014). An Adaptive Approach to Real-Time Aggregate Monitoring With Differential Privacy. IEEE Transactions on Knowledge and Data Engineering, 26(9), 2094–2106. https://doi.org/10.1109/TKDE.2013.96

Loning, M., et al. (2022). sktime/sktime: v0.13.4. doi: 10.5281/zenodo.7117735.

Luo, J., Hong, T., & Fang, S.-C. (2018). Benchmarking robustness of load forecasting models under data integrity attacks. International Journal of Forecasting, 34(1), 89–104. https://doi.org/10.1016/j.ijforecast.2017.08.004

V[0akis, S., & Hibon, M. (2000). The M3-Competition: Results, conclusions and implications. International Journal of Forecasting, 16(4), 451–476. https://doi.org/10.1016/S0169-2070(00)00057-1

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). The M4 Competition: Results, findings, conclusion and way forward. International Journal of Forecasting, 34(4), 802–808. https://doi.org/10.1016/j.ijforecast.2018.06.001

Martin, K. D., & Murphy, P. E. (2017). The role of data privacy in marketing. Journal of the Academy of Marketing Science, 45, 135-155.

Montero-Manso P, Athanasopoulos G, Hyndman R J, et al. FFORMA: Feature-based forecast model averaging[J]. International Journal of Forecasting, 2020, 36(1): 86-92

Nin, J., & Torra, V. (2009). Towards the evaluation of time series protection methods. Information Sciences, 179(11), 1663–1677. https://doi.org/10.1016/j.ins.2009.01.024

Nogueira, Fernando (2014). Bayesian Optimization: Open source constrained global optimizatio

tool for Python. URL: https://github.com/fmfn/BayesianOptimization.

Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M. Z., Barrow, D. K., Taieb, S. B., Bergmeir, C., Bessa, R. J., Bijak, J., Boylan, J. E., Browell, J., Carnevale, C., Castle, J. L., Cirillo, P., Clements, M. P., Cordeiro, C., Oliveira, F. L. C., Baets, S. D., Dokumentov, A., … Ziel, F. (2022). Forecasting: Theory and practice. 167.

Petropoulos, F., & Siemsen, E. (2022). Forecast Selection and Representativeness. Management Science, mnsc.2022.4485. https://doi.org/10.1287/mnsc.2022.4485

Qi, L., Li, X., Wang, Q., & Jia, S. (2022). fETSmcs: Feature-based ETS model component selection. International Journal of Forecasting, S0169207022000954. https://doi.org/10.1016/j.ijforecast.2022.06.004

Rose, O. (1996). Estimation of the Hurst parameter of long-range dependent time series (Vol. 137).

Report.

Schneider, M. J., Jagpal, S., Gupta, S., Li, S., & Yu, Y. (2018). A flexible method for protecting marketing data: An application to point-of-sale data. Marketing Science, 37(1), 153-171.

Seabold, S. & Perktold, J. (2010). statsmodels: Econometric and statistical modeling with python.

9th Python in Science Conference.

Sobolev, D. (2017). The effect of price volatility on judgmental forecasts: The correlated response model. International Journal of Forecasting, 33(3), 605–617. https://doi.org/10.1016/j.ijforecast.2017.01.009

Sommer, B., Pinson, P., Messner, J. W., & Obst, D. (2021). Online distributed learning in wind power forecasting. International Journal of Forecasting, 37(1), 205–223. https://doi.org/10.1016/j.ijforecast.2020.04.004

Spiliotis, E., Kouloumos, A., Assimakopoulos, V., & Makridakis, S. (2020). Are forecasting competitions data representative of the reality? International Journal of Forecasting, 36(1), 37–53. https://doi.org/10.1016/j.ijforecast.2018.12.007

Talagala, T. S., Li, F., & Kang, Y. (2022). FFORMPP: Feature-based forecast model performance prediction. International Journal of Forecasting, 38(3), 920–943. https://doi.org/10.1016/j.ijforecast.2021.07.002

Véliz, C., & Grunewald, P. (2018). Protecting data privacy is key to a smart energy future. Nature Energy, 3(9), 702-704.

Wang, X., Smith, K., & Hyndman, R. (2006). Characteristic-Based Clustering for Time Series Data. Data Mining and Knowledge Discovery, 13(3), 335–364. https://doi.org/10.1007/s10618-005-0039-x

Willinger, W., Paxson, V., & Taqqu, M. S. (1998). Self-similarity and heavy tails: Structural modeling of network Traffic. R. Adler, R. Feldman, and MS Taqqu, editors, A Practical Guide to Heavy Tails: Statistical Techniques for Analyzing Heavy Tailed Distributions, Birkhauser Verlag, Boston.

1. **Appendix**
   1. **Implementation Details**

*Preprocessing*

Time series were pre-processed by taking the log of the original series. For the VAR model, we took the first difference of the log series. For the global models (RNN and LGBM), we divided each log series by its respective mean.

*Model Implementations*

All local models (SES, DES, TES, and AutoARIMA) were implemented using the *sktime* forecasting module (Loning et al., 2022). We used additive trend and seasonality components for DES and TES. We set AutoARIMA to apply a seasonal model (SARIMA) with a maximum of twenty-five iterations.

The VAR model from the *statsmodels* (Seabold & Perktold, 2010) module was applied to subsets of approximately five time series at a time, ensuring that each subset consisted only of time series with the same length.

The LGBM and RNN models were implemented using the *darts* module (Herzen et al., 2022). For both models, we reserved a validation time period immediately prior to the desired forecast horizon. We used Bayesian optimization (Nogueira, 2014) to optimize the hyperparameters of each model to minimize the absolute forecast error (L1 loss) in the validation time period. We retrained each model using the optimized hyperparameters and the full training data (including the validation period) prior to generating forecasts for the desired forecast horizon.

The Bayesian optimizer was initialized for the LGBM models using ten starting points and run for 50 iterations. For the RNN the optimizer was initialized using five starting points and run for 25 iterations.

The hyperparameters for the LGBM and RNN models and their corresponding ranges are shown in Tables 9 and 10. We limited the RNN to the last ten input-output window samples from each time series for computational efficiency and trained ten RNN models taking the median of the forecasts as the final forecast (Hewamalage et al., 2022).

**Table 9: LGBM hyperparameter ranges.**

|  |  |
| --- | --- |
| **Hyperparameter** | **Range** |
| Input Window Length | 25 |
| Learning Rate | [0.01, 0.1] |
| Number of Boost Rounds | [50, 1000] |
| Number of Leaves | [10, 100] |
| Bagging Frequency | [1, 5] |
| Bagging Fraction | [0.01, 1] |
| L2 Regularization Parameter | [0, 0.5] |
| Minimum Observations in Leaf | [10, 100] |

**Table 10: RNN hyperparameter ranges.**

|  |  |
| --- | --- |
| **Hyperparameter** | **Range** |
| Input Window Length | 25 |
| Training Length | 30 |
| Learning Rate | [0.001, 0.1] |
| Weight Decay | [0.0001, 0.0008] |
| Number of Layers | [1, 2] |
| Hidden Dimension | [20, 50] |
| Batch Size | [200, 700] |
| Number of Epochs | [3, 30] |
| Dropout Rate | [0.1, 0.5] |

**6.2 Mathematical Details of Identification and Attribute Disclosure**

To perform identification disclosure, we assume a third party possesses some original data on a unit of interest in the protected dataset. Denote this original data which contains a direct identifier (e.g., the identity of time series ) and original data which contains a sequence of values that are components of the original time series .

We let denote the random variable (from the perspective of the third party) that indicates the corresponding for , i.e., when the values in are components of the original version of the protected series . Since the true value is unknown, the third party predicts the value of to be the series with the highest match probability, conditional on the known values, as follows,

, (1)

where identification disclosure occurs when The probability is calculated as follows. Let denote the protected values of each time series that occur in the same time periods as . The third party computes the similarity between and the protected values using the Euclidean distance,

.

Using these similarities, the third party builds a probability mass function for over all protected series in as

,

and predicts as in (1).

To estimate the risk of identification disclosure, we perform simulations in which we sample sequential values from each original time series , and we measure the average proportion of series which are identified. The sampled values are denoted . Each of the vectors corresponds to one of the original time series, and we compute conditional on the sampled from series . We repeat this simulation times to obtain , and computethe average proportion of correctly identified time series across all external data samples and original time series,

These simulations assume that the third party in possession of predicts the match for each vector independently of the predicted matches for other vectors. The risk estimate from a given simulation is equivalent to the identification risk when independent third parties are each in possession of one of the vectors and each attempts identification risk as described above. Overall, multiple vectors may be matched to the same protected time series.

**6.3 Equations for Important Time Series Features from the Literature**

*Spectral Entropy*

Suppose is a univariate stationary time series with a finite mean and constant variance. The spectral density of is estimated as the scaled Fourier transform of the autocovariance function of . The spectral density can be thought of as the probability density function of a random variable on the unit circle (Goerg, 2013), where for a non-zero integer , when , the spectral density will have a peak at the corresponding frequency . The forecastability, or spectral entropy, of is measured using the Shannon entropy of , given by

,

where the maximum entropy occurs when . In practice, estimates of , where high values represent a low signal-to-noise ratio, indicating that is difficult to forecast (Kang et al., 2017).

*Hurst*

Next, we consider a self-similarity feature quantified using the Hurst parameter (Wang et al., 2006), which measures the long-range dependence of a time series. Spiliotis et al. (2020) found this feature had the largest effect on forecast accuracy. We use the definition of self-similarity of a time series described by (Willinger et al., 1998). Suppose that is the increment process of , *i.e.*, . An aggregate sequence, denoted , is created by averaging over non-overlapping blocks of size , where

and indexes the block. If is a self-similar time series, then

for all integers . We use the definition of second-order self-similarity, where s exactly second-order self-similar if has the same variance and autocorrelation as for all values of , or is asymptotically second-order self-similar if this holds as (Rose, 1996). The parameter is the Hurst exponent, which is estimated using the differencing term from a fractional ARIMA model, i.e., FARIMA(0, , 0) (Wang et al., 2006; Hyndman et al., 2022), where

5.

Estimates of range from 0 to 1, where corresponds to a random walk (Sobolev, 2017), corresponds to anti-persistent or mean-reverting series, and corresponds to persistent time series that are more likely to maintain their current trend.

*Skewness*

*Skewness* measures the lack of symmetry in the distribution of the values of (Wang et al., 2006), where positive (or negative) values are associated with a right- (or left-) skewed data distribution,

*Kurtosis*

We measure *Kurtosis* relative to the standard normal distribution (Wang et al., 2006). Positive *Kurtosis* corresponds to distributions that tend to have a distinct peak near the mean with heavy tails, whereas negative *Kurtosis* corresponds to distributions that are relatively flat near the mean,

where 3 is the *Kurtosis* of the standard normal distribution.

*Error Autocorrelation Function (Error ACF)*

Next, we perform STL decomposition (Cleveland et al. 1990) to obtain the trend, seasonal, and remainder components of . We use the approach of Hyndman et al. (2019) to obtain

,

where , , and are the trend, seasonal, and remainder components, respectively.

We extract the first-order autocorrelation coefficient of the detrended and deseasonalized series, referred to as 'linearity' by Spiliotis et al. (2018):

*Error ACF* .

*Error ACF* is a measure of the predictability of a time series after the trend and seasonality have been accounted for (Kang et al. 2017).

*Trend and Seasonality*

We also compute the strength of trend (*Trend*) and strength of the seasonal component () as follows,

,

and

.

In practice, the values of and range from 0 to 1 (Hyndman 2022).

*Mean and Variance*

The next two features are the *Mean* and *Variance*, also used by Bandara et al. (2018) to cluster similar time series for forecasting, which are written as follows,

.

We also included many other features from the *tsfeatures* package in R. We refer the reader to (Hyndman et al., 2022) for explanation of these features.

**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. [↑](#footnote-ref-1)
2. cite Verizon DBIR numbers 2019 to 2022. [↑](#footnote-ref-2)
3. For legislation examples in the United States, see [this](https://iapp.org/resources/article/us-state-privacy-legislation-tracker/) map. See articles 6, 45, and 46 of the GDPR. [↑](#footnote-ref-3)
4. All norms on are equivalent to the Euclidean norm. [↑](#footnote-ref-4)
5. Note that we could also define a distance matrix based on the actual time series values , where would become a function of rather than . [↑](#footnote-ref-5)
6. Implementation details can be found in the appendix. [↑](#footnote-ref-6)
7. 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-7)
8. AvgRelAE can be generalized to accommodate multiple forecasts for each series. See Davydenko & Fildes, (2013) for the AvgRelMAE. [↑](#footnote-ref-8)
9. [E.U. slaps Meta with record $1.3 billion fine for data privacy violations](https://www.msn.com/en-us/news/other/e-u-slaps-meta-with-record-1-3-billion-fine-for-data-privacy-violations/ar-AA1bvYOW) [↑](#footnote-ref-9)