The Effects of Privacy Protection on Forecast Accuracy

1. **Introduction**
   1. Private time series could be utilized in many domains, such as tourism demand forecasting (Guizzardi et al., 2021; J. Li et al., 2018), or forecasting sales based on consumer behavior (Boone et al., 2019).
   2. Forecasting using differentially private data has been studied in domains such as collaborative solar power (Gonçalves et al., 2021), consumer health data (Imtiaz et al., 2020), and stock price prediction (X. Li et al., 2019), but no work examining the effects of other data protection methods on forecast accuracy.
   3. We compare the accuracies of off-the-shelf and open-source forecasting models under different data protection methods.
   4. We guide forecasters on selecting an optimal forecasting model given data characteristics and protection method.
   5. Based on the findings of (Gonçalves et al., 2021), we believe that random noise protection (additive noise or differential privacy) will generally reduce forecast accuracy even for small amounts of noise (low privacy). On the other hand, top and bottom coding may improve forecast accuracy by attenuating the effect of outliers on forecasts (Chen & Liu, 1993). We examine whether improvements in accuracy post-data protection are related to upward or downward adjustments to the forecasted points and relate these findings to the judgmental forecasting literature (e.g., (Fildes et al., 2009)). Furthermore, we characterize time series by applying PCA to time series features including spectral entropy, the strength of trend and seasonality, first order autocorrelation, and the optimal Box-Cox transformation parameter (Hewamalage et al., 2022). We examine how principle component scores change after data protection and use these changes to explain the effects of data protection on forecast accuracy.
2. **Literature Review**
   1. Data protection for time series
      1. Sharing sensitive data in collaborative forecasting environments
         1. (Sommer et al., 2021) Propose methods for privacy-preserving distributed learning of linear, separable forecasting models. Not useful if forecasting using a centralized dataset, or if more complicated non-linear models are of interest.
         2. (Gonçalves et al., 2021) Review methods for privacy-preserving collaborative forecasting with VAR models. Relevant to our work is that protecting series with differential privacy leads to reduced forecast accuracy, which we believe will hold true for other models and datasets.
      2. Circumventing data sharing
         1. Transfer learning solutions involve transferring pre-trained models instead of sensitive data, which can overcome privacy barriers or regulations (Wellens et al., 2021). Works for complicated global models, e.g., what was used in M5, but not for univariate statistical models.
      3. Top/bottom coding, additive noise, differential privacy, etc.
   2. Judgmental forecasting
3. **Experimental Design**
   1. Selection of forecasting models
      1. Popular univariate methods found in R, Python and/or serving as benchmarks in forecasting competitions
         1. SES, DES, TES, ARIMA, Prophet (Facebook), Greykite (LinkedIn)
         2. The pre-processing performed for the exponential smoothing models (and for all univariate models in the future) is minimal – we allow the models to capture the important aspects of the time series.
      2. Popular and/or cutting-edge multivariate models that have won forecasting competitions
         1. VAR, LGBM, NeuralProphet, Lasso-Var(?), RNN (LSTM cells)
         2. We perform pre-processing for the LGBM models (and for all multivariate models in the future) that is consistent with best practice.
      3. Parameters to vary
         1. Forecast horizon
   2. Selection of data protection methods
      1. Common methods
         1. Top/bottom coding, differential privacy, additive noise
      2. Proposed method
         1. *K*-nTS swapping
      3. Parameters to vary
         1. Level of protection (e.g., % top/bottom coded, , etc.)
   3. Data
      1. M3 monthly micro, M4 weekly finance
         1. The m4 weekly finance data was chosen to obtain initial results since it contains a relatively low number of series (164) and would enable quick code implementation.
         2. The m3 monthly micro data was chosen for comparison because machine learning methods perform well on this data, and it provides a reasonably high number of time series (474).
   4. Accuracy measures
      1. MAPE, MdAPE, RMSE
   5. Time series feature extraction
      1. characterize time series by their spectral entropy, strength of trend and seasonality, first order autocorrelation, and optimal box-cox transformation parameter (Hewamalage et al., 2022) and examine how these traits interact with data protection to affect forecast accuracy.
   6. Complexity vs. Error Analysis
      1. Regression of model complexity against…
   7. Experimental Framework
      1. Obtain original series
      2. Create protected series
      3. Perform pre-processing on and :
         1. For SES, DES, TES:
            1. As needed for , Truncate values < 1 to 1 (enables taking the log)
            2. Take the log (removes multiplicative effects and stabilizes variance) – helps with model convergence.
         2. Pre-processing for other univariate methods…
         3. For LGBM (Hewamalage et al., 2022; Makridakis et al., 2018)
            1. As needed for , Truncate values < 1 to 1 (enables taking the log)
            2. Mean normalize to put data on same scale across series
            3. Take the log (removes multiplicative effects and stabilizes variance)
            4. Perform conditional deseasonalization (conditional on 90% confidence auto-correlation test): - *I believe this is the equation used, which is described in the M4 competition benchmark documentation. I need to verify with python package developers that this is the case.*
            5. Convert time series into windows
            6. Normalize each window by removing the trend of each window
         4. Pre-processing for other multivariate methods…
      4. Train each forecasting model:
         1. SES, DES(additive trend), TES(additive trend and seasonality), LGBM(window length= 20), other models…
      5. Generate forecasts for and
      6. Post process the forecasts
         1. Reverse the pre-processing, including bias correction for exponential transformation (Hyndman & Athanasopoulos, 2021)
      7. Compare forecast accuracies
4. **Results**
   1. Global accuracy measures for all models
      1. 1-Step Forecast horizon
         1. M3 Monthly Micro

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Error Metric | Forecasting Model | Original | DP () | Add. Noise (1 SD) | Top Coding (10%) | Bottom Coding (10%) |
| MAPE | SES | 0.262 | **0.336** | **0.298** | 0.257 | 0.281 |
| DES | 0.252 | 0.405 | 0.354 | 0.247 | 0.268 |
| TES | 0.230 | 0.514 | 0.492 | **0.223** | **0.249** |
| LGBM | **0.227** | 0.505 | 0.380 | **0.223** | 0.261 |

* + - * 1. *Compare values within columns first (most accurate are highlighted)*

LGBM is best on original data

SES is best under DP and additive noise (by a lot)

LGBM/TES compete under top and bottom coding

* + - * 1. *Comparing values across columns*

All methods have worse accuracy under DP, additive noise, and bottom coding

All methods have improved accuracy under top coding relative to the original data (we will see if we can relate this to downward adjusting forecasts in the judgmental forecasting literature)

* + - 1. M4 Weekly Finance

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Error Metric | Forecasting Model | Original | DP () | Add. Noise (1 SD) | Top Coding (10%) | Bottom Coding (10%) |
| MAPE | SES | **0.023** | 0.251 | **0.184** | 0.120 | 0.072 |
| DES | **0.023** | **0.248** | 0.204 | **0.119** | **0.071** |
| TES | 0.026 | 0.344 | 0.395 | 0.120 | 0.074 |
| LGBM | 0.032 | 0.382 | 0.477 | 0.125 | 0.078 |

* + - * 1. *Compare values within columns first*

SES/DES are best on original data and under data protection

* + - * 1. *Comparing values across columns*

All methods have worse accuracy under any data protection

* + 1. Accuracies for upward vs. downward adjusted forecasts – make table like e.g., (Fildes et al., 2009):

A screenshot of a computer

Description automatically generated with medium confidence

1. **Conclusion**
2. **References**

Boone, T., Ganeshan, R., Jain, A., & Sanders, N. R. (2019). Forecasting sales in the supply chain: Consumer analytics in the big data era. *International Journal of Forecasting*, *35*(1), 170–180. https://doi.org/10.1016/j.ijforecast.2018.09.003

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

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

Gonçalves, C., Bessa, R. J., & Pinson, P. (2021). 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

Guizzardi, A., Pons, F. M. E., Angelini, G., & Ranieri, E. (2021). Big data from dynamic pricing: A smart approach to tourism demand forecasting. *International Journal of Forecasting*, *37*(3), 1049–1060. https://doi.org/10.1016/j.ijforecast.2020.11.006

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., & Athanasopoulos, G. (2021). *5.6 Forecasting using transformations | Forecasting: Principles and Practice (3rd ed)* (3rd ed.). https://otexts.com/fpp3/ftransformations.html

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

Li, J., Xu, L., Tang, L., Wang, S., & Li, L. (2018). Big data in tourism research: A literature review. *Tourism Management*, *68*, 301–323. https://doi.org/10.1016/j.tourman.2018.03.009

Li, X., Li, Y., Yang, H., Yang, L., & Liu, X.-Y. (2019). DP-LSTM: Differential Privacy-inspired LSTM for Stock Prediction Using Financial News. *ArXiv:1912.10806 [Cs, q-Fin]*. http://arxiv.org/abs/1912.10806

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and Machine Learning forecasting methods: Concerns and ways forward. *PLOS ONE*, *13*(3), e0194889. https://doi.org/10.1371/journal.pone.0194889

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

Wellens, A. P., Udenio, M., & Boute, R. N. (2021). Transfer learning for hierarchical forecasting: Reducing computational efforts of M5 winning methods. *International Journal of Forecasting*, S0169207021001606. https://doi.org/10.1016/j.ijforecast.2021.09.011