JOURNAL OF
EMERGING INVESTIGATORS

Article

Differential privacy in machine learning for traffic forecasting

Sunny Vinay[1], Rachel Redburg[2]

[1] Leland High School, San Jose, California

[2] Computer Science, UC Santa Barbara, Santa Barnara, California

SUMMARY

Time series data has many applications to understand real-world phenomena. A common application is traffic congestion forecasting. Preserving privacy with traffic data is also essential, and an emerging solution is differential privacy, which causes a tradeoff in utility. In this paper, we measured the privacy budgets and utilities of different differentially private mechanisms combined with different machine learning models that forecast traffic congestion at future timestamps. We expected the ANNs combined with the Staircase mechanism to perform the best with every value in the privacy budget range, especially with the medium-high values of the privacy budget. In this study, we used the Autoregressive Integrated Moving Average (ARIMA) and neural network models to forecast and then added differentially private Laplacian, Gaussian, and Staircase noise to our datasets. We tested two real traffic congestion datasets, experimented with the different models, and examined their utility for different privacy budgets. We found that a favorable combination for this application was neural networks with the Staircase mechanism. Our findings identify the optimal models when dealing with tricky time series forecasting and can be used in non-traffic applications like disease tracking and population growth.

INTRODUCTION

Today, the increase in knowledge, technology, and innovation in the physical and virtual worlds has created a demand for the collection and analysis of real-world data. This demand is fueled by humanity's desire to know the patterns of phenomena around us. By understanding patterns, we can make useful predictions for businesses or for ensuring safety. For example, traffic monitoring can be used to see the public's interest at specific locations and times. This information is precious to owners who seek to optimize the number of customers in their business. Furthermore, disease and natural disaster tracking can save lives. The speed and ease of tracking have only been improving with the advancements in machine and deep learning. All of this new technology is extremely important and exciting, but it comes with the major concern of privacy. These models must use real preexisting data to make predictions about the future. The problem is that this data stores information about individuals. In the case of traffic monitoring, GPS service providers and sensors are used to collect the locations and speeds of individuals at specific times. These advancements amplify the danger of hackers, who exploit these databases to gain information about specific individuals even though these people trust that their sensitive data is being kept private.

Many of the older, simpler approaches to privacy, such as data anonymization and k-anonymity, don't work as well anymore because hackers have more sophisticated methods for pinpointing individuals. Data anonymization doesn't account for linkage attacks since hackers use multiple databases, and k-anonymity only works for large databases with very simple fields (1). Fortunately, there is a modern approach to preserving the utility and privacy of this data.

This approach is known as differential privacy, which is the study of quantifying and limiting how much privacy is lost when extracting useful information from a dataset (1). There are many differentially private algorithms, but all provide the same guarantee. This guarantee is that no individual can be targeted from the output of a query on a dataset. This is mostly done by perturbing the input or output using a randomized algorithm. The output of this algorithm remains similar even if a single data point in the dataset is modified (2). Traffic congestion forecasting is a key application of differential privacy in a few ways. First, many of the existing traffic datasets, including the one used in this research, use map data from user devices to track travel times and traffic. Second, data that includes pictures of license plates or passengers are frequently recorded by law enforcement and points to specific drivers on the road. Third, the time series nature of our traffic data matches the form of other time series data like disease tracking and sales, which could reveal the personal information of individuals. All the data in each of these cases is very specific to individuals and can be used to track their location and movements, making differential privacy a key technique in protecting user privacy.

There is existing research in the field of machine learning with differential privacy, specifically differential privacy applied to time series data and forecasting. Various studies we found are examples of research focusing on similar approaches (2-6). However, there are a few key differences among the approaches. Most of them use time series data in a different way, as opposed to direct forecasting. Many are related to sharing real-time aggregate statistics of private data, but our research aims to fill the gaps and limitations in time series differential privacy. We also expand upon all the existing research by comparing and trying different forecasting models beyond simple regression models. For example, Fan and Xiong compared various differentially private mechanisms but with a constant forecasting model (2). Our research aims to fill the gap of the exact effects of different models and mechanisms on the utility and privacy when making predictions by experimenting with the machine learning and differential privacy parts to draw a conclusion on the most favorable privacy-preserving combination for traffic monitoring and other time series data. This will give more insight into the effect of machine learning on differential privacy by experimenting with a variety in terms of data, models, and popular mechanisms that simulate the relevant options in the field of machine learning and differential privacy.

First, we define some key terminologies. Machine learning (ML) is a sub-field of artificial intelligence (AI) and the study of where computer systems use data and algorithms to improve their knowledge for completing a given task. In supervised learning, the data consists of predetermined mappings between the input (x) and the output (y) (7). We used time series data for this paper, a sequence of measurements of some variable (traffic congestion) at successive points in time with an equal time interval between every point (7). Auto-regressive integrated moving average (ARIMA) is a model that uses past time series values and linearly maps them into an output. It is a widespread model when working with time series data (8).

Inspired by the biological neural networks that make up our brains and emerging technology within AI, artificial neural networks (ANNs) are made up of many neurons, which receive and output information. The two types of ANNs that we studied are feed-forward neural networks (FFNN) and convolutional neural networks (CNN). FFNNs are the simplest type of ANNs. The connections between neurons do not form any cycles, so the flow of information is straightforward through the network. An improved version of FFNNs, CNNs excel at mapping inputs to outputs. Models like long short-term memory (LSTM) and recurrent neural networks (RNN) are commonly used in time series problems, but FFNNs and CNNs are simpler and less used, so we chose to explore them in our time series context.

Finally, differential privacy (DP) is a cybersecurity technique that prevents learning information about a specific individual in a dataset. Differentially private mechanisms add noise to functions to guarantee that an adversary won't learn anything new about an individual, meaning that individuals can be taken out or added without affecting the overall results (10). The privacy budget (ε) determines the level of privacy and utility. Generally, lower values of the privacy budget provide more perturbation, which leads to higher levels of privacy and lower levels of utility. In contrast, higher values provide less privacy and more utility. We study three differentially-private mechanisms: the Laplace mechanism, the Gaussian mechanism, and the Staircase mechanism. Dwork *et al*. have proven that the Laplace mechanism preserves differential privacy if noise is added (11). It is good for low-sensitivity queries, where sensitivity is defined as the change in the dataset if one element is changed. The Gaussian mechanism is a common, more flexible alternative to the Laplace mechanism (10). Both the Laplace and Gaussian mechanisms are very popular in differential privacy. Geng and Viswanath proposed the Staircase mechanism to optimize the Laplace mechanism (12). Similar to how we chose the ML models, we chose two common (Laplace and Gaussian) and one less common (Staircase) DP mechanism to balance our research.

Our research's main contribution and objective are to figure out the best combination of a forecasting model and a differentially private mechanism for making predictions from time series data. For our paper, we specifically used two traffic congestion datasets. We defined this combination as the best one that guarantees a high level of privacy and utility. These are the two biggest concerns, as an increase in the level of privacy decreases the level of accuracy. In our case, the accuracy of a model depends on how close its predictions were compared to the actual values. We tried three different common forecasting models (ARIMA and two neural network models) and three different differentially private mechanisms (Laplace, Gaussian, and Staircase). We expected the ANNs combined with the Staircase mechanism to perform the best with every value in the privacy budget range, especially with the medium-high values of the privacy budget. The Staircase and Gaussian mechanisms should both outperform the Laplace mechanism. Non-parametric models, such as neural networks, work better with non-linear data, such as traffic congestion. Within neural networks, we expect CNNs to outperform FFNNs because CNNs are optimized FFNNs.

RESULTS

In trying to figure out the best combination of a forecasting model and a differentially private mechanism, we implemented each of the models and mechanisms discussed in the Introduction (FFNN and CNN with Laplace, Gaussian, and Staircase) with their best hyperparameters (explained in **Figure 6**). We plotted a summary of the raw values when the privacy budget is equal to 0.1 (**Table 1**). The range of the privacy budget goes from 0.001 to 2, and the corresponding error using mean absolute error (MAE) is plotted (**Figures 1-5**). MAE is a common accuracy metric that measures the average magnitude of the errors in a set of predictions without considering their direction (9). We chose this over other metrics like mean square error (MSE) and root-mean-square deviation (RSME) for two reasons. First, MAE does not give a very high weight to outliers in a dataset, which works well since we are only comparing privacy and ML models. Second, other metrics tend to be increasingly larger than MAE as the

Noise Mechanism	ARIMA	FFNN	CNN
No Noise	2.7464	0.0046	0.0049
Laplace	2.7689	0.0195	0.0133
Gaussian	3.1339	0.2397	0.5981
Staircase	**2.7522**	**0.0053**	**0.0055**

Table 1: MAE values for each mechanism and model combination. MAE was calculated after each noise mechanism and forecasting model was applied to the dataset (ε = 0.1). Bolded values are the best-performing ones in the group. The privacy budget was held constant at ε = 0.1 because it is the middle value in our range. Out of the noise mechanisms, the Staircase mechanism consistently performed the best, while the Gaussian mechanism consistently performed the worst with every forecasting model.

test sample size increases. Since we have two differently sized datasets, MAE was better when comparing accuracy between datasets. We explain the following datasets in more detail in the Methods section.

Dataset #1

From TomTom traffic indexing, this dataset marks urban congestion. We only used ARIMA, since it contains a smaller number of data points than Dataset 2. Neural network models were unsuccessful since they require much more data to make predictions accurately. The algorithm predicted the traffic congestion for 100 time steps in the future and compared those predictions to the actual values (**Figure 1**). These predictions matched the trends and extrema very closely, but the slight error came from the model and the small amount of training data in this set. This error of 2.74 is only about 0.2 less than the error when we added noise (**Table 1**).

We then added three types of differentially private noise (Laplace, Gaussian, Staircase) to Dataset #1 before training it with ARIMA (**Figure 2**). In general, all mechanisms trended downward. As the privacy budget increased, the error decreased, which was expected since a lower budget guarantees more privacy by adding more noise. All mechanisms seem to have at least one major drop between an interval. For the Laplace mechanism, the blue line shows a decrease in error that reaches zero towards the end. The biggest drop in error occurred between privacies 0.01 and 0.1. For the Gaussian mechanism, the behavior of this mechanism was the most unexpected. There was a slight increase in error between 0.01 and 0.1, followed by a massive drop. However, the error was the highest of the three mechanisms, which is unexpected since the noise distribution is less steep than the Laplace distribution. The Staircase mechanism steadily decreased like the Laplace mechanism. The biggest drop in error occurred between privacies 0.001 and 0.01. This mechanism gave the least error.

Dataset #2

We used FFNN and CNN on this dataset since these were more accurate and successful than when run on the first dataset (**Figures 3-5**). The graph of the forecasted versus actual values is not shown because there are many data points, making the graph cluttered and less readable. We plotted FFNN with all mechanisms (**Figure 3**), CNN with the Laplace and Staircase mechanisms (**Figure 4**), and CNN with the Gaussian mechanism. The overall results were very similar for both neural network models, which is expected since these models work similarly. For the Laplace mechanism, the blue line shows a decrease in error that reaches zero towards the end. The error associated with FFNN decreases in almost equal intervals (**Figure 3**). CNN behaves strangely, as there is a slight spike in error between 0.01 and 0.1 (**Figure 4**). For the Gaussian mechanism, the red line shows an overall decrease in error. It starts off higher and meets with the Laplace and Staircase mechanisms as the privacy budget increases (**Figure 3**). This unexpected behavior of the highest error is like what we observed with Database #1. Unlike Database #1, the error never increased in any privacy interval. For the Staircase mechanism, the yellow line shows a decrease that eventually plateaus. The error values were very similar to the Laplace mechanism with FFNN (Figure 3). With CNN, there is a bigger drop in error between 0.001 and 0.01 (**Figure 4**).

Overall, the most accurate differentially private mechanism

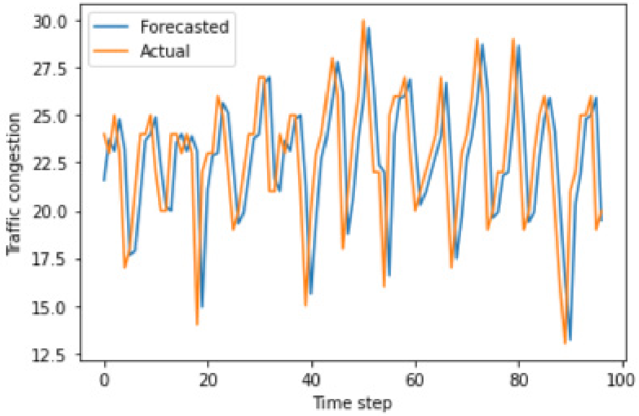

Figure 1: Forecasted vs Actual values from ARIMA on Dataset 1 without noise. The ARIMA forecasting model was used to predict traffic congestion in future time steps. The blue line shows the predictions, while the orange line shows the actual values. The model was fairly accurate as the lines are very close to each other.

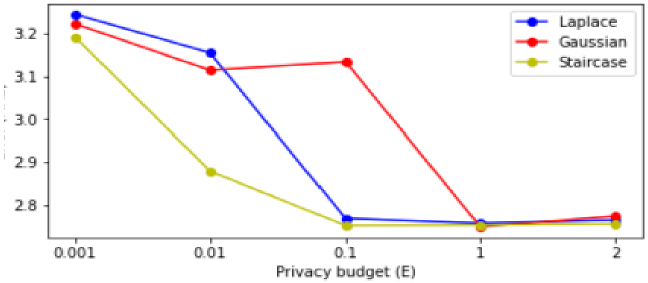

Figure 2: Utility vs Privacy from ARIMA on Dataset 1 with noise. Graph shows privacy budget vs error (MAE) for every differential privacy mechanism. The ARIMA forecasting model was used on Dataset 1 with added noise from each mechanism. Different privacy budgets were tested, and the results consistently show that a greater privacy budget (less protection) means less error (greater accuracy).

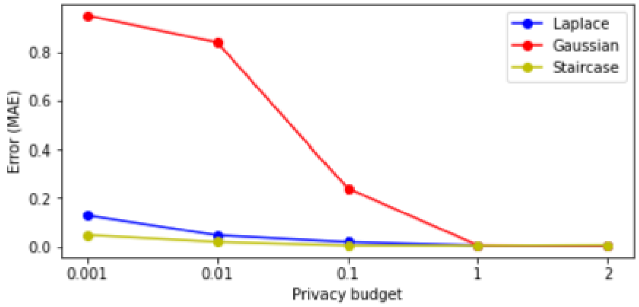

Figure 3: Utility vs Privacy from FFNN on Dataset 2 with noise. Graph shows privacy budget vs MAE for every differential privacy mechanism. The FFNN forecasting model was used on Dataset 2 with added noise from each mechanism. Different privacy budgets were tested, and the results consistently show that a greater privacy budget (less protection) means less error (greater accuracy).

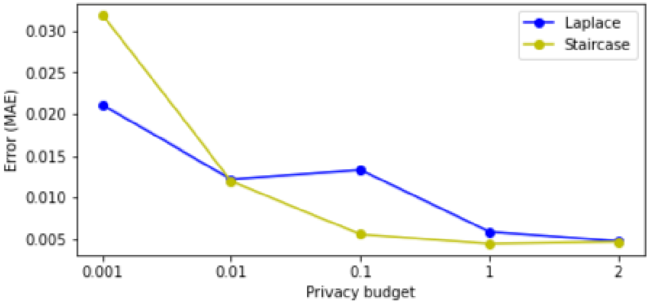

Figure 4: Utility vs Privacy from CNN on Dataset 2 with noise. The CNN forecasting model was used on Dataset 2 with added noise from the Laplace and Staircase mechanisms. Different privacy budgets were tested, and the results show that a greater privacy budget (less protection) means less error (greater accuracy).

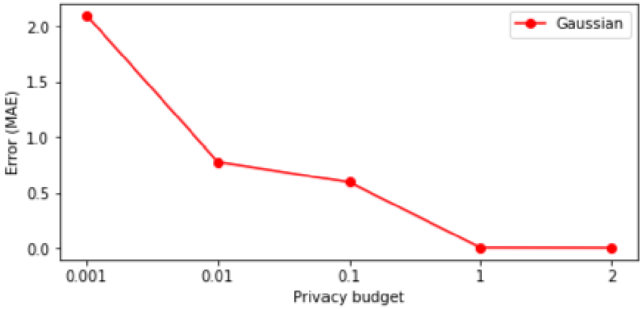

Figure 5: Utility vs Privacy from CNN on Dataset 2 with noise. The CNN forecasting model was used on Dataset 2 with added noise from the Gaussian mechanism. Different privacy budgets were tested, and the results show that a greater privacy budget (less protection) means less error (greater accuracy).

was the Staircase mechanism. It outperformed the Gaussian mechanism and performed slightly better than the Laplace mechanism. The privacy budget depends on the application and how much utility can be given up for an increase in privacy.

DISCUSSION

Our results are consistent with differential privacy. As the privacy budget increased, the mechanisms added less noise to the data, yielding a smaller error. This small error seemed consistent for the Laplace and Staircase mechanisms, with the slight exception of the strange behavior of the Gaussian mechanism. With the Gaussian mechanism, error increased during one interval of ARIMA and exhibited greater error than the Laplace mechanism for most of the privacy budgets. This large error is most likely due to the difference in sensitivity, which was a general challenge due to the nature of the time series datasets, where every data point is in a specific order, and the values and intervals are extremely dependent on each other. Another potential reason is the difference in using the Laplace and Gaussian mechanisms. The Gaussian mechanism is more favorable in certain situations, for example, when using larger datasets, which could explain why it still had a continuous decrease with Dataset #2 (13). Even with the lowest privacy budget (0.001), the Gaussian mechanism still yielded errors higher than the other mechanisms.

After experimentation with commonly used models, we found a favorable combination of machine learning and differential privacy for the specific application of traffic congestion forecasting. Adversaries exploit time series data by bypassing older, simpler privacy methods, thus increasing privacy concerns. Differential privacy can solve this with many different mechanisms that protect information about individuals and allow us to see important patterns. Our experimentation involved two different time series datasets, three machine learning models, and three differentially private mechanisms. After formatting data and choosing the hyperparameters, we added calibrated noise to the input data and trained our forecasting models. We evaluated the utility, or accuracy, compared to the range of private budgets. We were able to draw a conclusion on which models worked the best. Furthermore, although our experimentation was specific to traffic congestion, other similar applications with time series data, such as disease or weather forecasting, would have very similar results. Still, our research was limited in the number of models and mechanisms we were able to try, and future research should include the utility versus privacy study of more machine learning models that work well with time series, such as exponential smoothing or deep learning models like Recurrent Neural Networks (8, 14). Since adding noise to each time step creates a background, we could even consider representing the data differently by doing a Fourier transform and adding noise in the frequency domain (15).

We conclude that, for time series datasets with appropriate amounts of data points, a neural network approach, such as Feed-forward Neural Networks, and adding noise to the data through the Staircase mechanism ensures the best utility no matter the privacy budget. The Laplace mechanism is an excellent second option since the Gaussian mechanism displayed some unexpected results. These results are significant in a world where privacy for individuals is a huge concern.

MATERIALS AND METHODS

Figure 6 shows the flow of the methodology.

Datasets

We chose two datasets to test the models and mechanisms. With time series problems, the time between any pair of recorded data points must be the same for all pairs in the dataset.

Dataset #1 is from TomTom traffic indexing (Ref). Using free-flow travel times of vehicles on the road, TomTom compiled this traffic data for COVID traffic tracking. There are 482 data points (385 for training, 97 for testing) representing the traffic congestion in a San Francisco Street on 1-day intervals, meaning that we aimed to predict daily traffic.

Dataset #2 is from the UC Irvine Machine Learning Repository (13). It contains 15 months' worth of daily data that describes the occupancy rate of different car lanes of the San Francisco Bay Area freeways. We chose data from one specific sensor, so the total data is 440 days x 144 10-minute intervals = 63,360 data points (38,448 for training, 24912 for testing). We aimed to predict traffic 10 minutes in advance.

Data Preparation

While there are differences in preparing the data between the different datasets and models, many parts are the same. Our end goal was to separate the train and test data, with

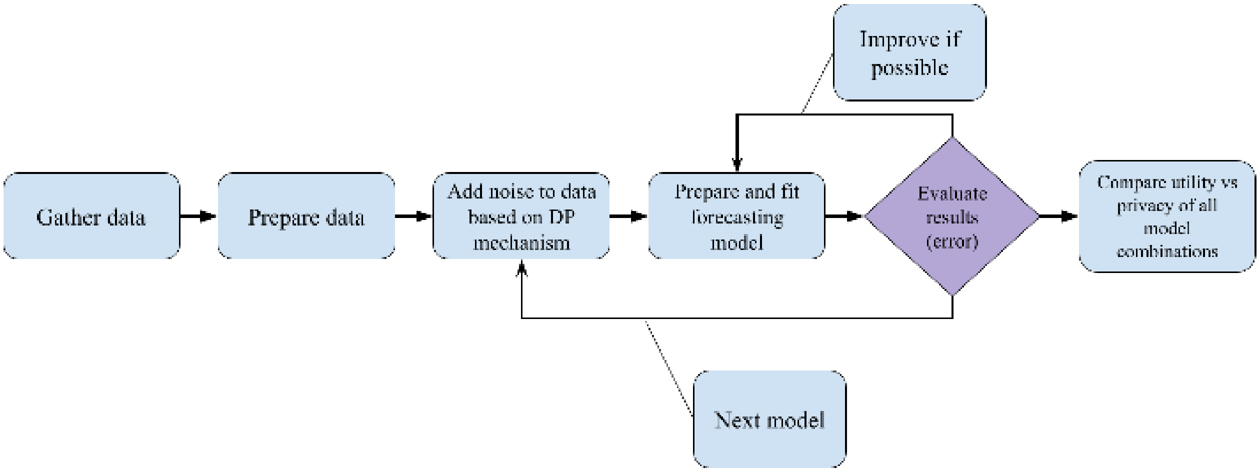
Improve if possible
Gather data
Prepare data
Add noise to data based on DP mechanism
Prepare and fit forecasting model
Evaluate results (error)
Compare utility vs privacy of all model combinations
Next model

Figure 6: Experiment Methodology. Shows the steps for data preparation, modeling, and comparing.

only the time step and traffic congestion value. In Dataset #1, we removed the unnecessary columns of "country", "city", and "diffRatio". We also removed the first couple of months of this dataset, as this part has been affected by the COVID-19 pandemic. In Dataset #2, we also used data from just one of the 900 sensors provided in the dataset.

For the train and test split, we use a standard 80% train and 20% test. Dataset #2 is pre-split into training and testing data. For Dataset #1, we used Python 3.8 and the latest version of NumPy to do the split.

The next step was to scale the data. We scaled the data to a range of values between 0 and 1. Dataset #1 had to be manually scaled, but Dataset #2 was already scaled.

NumPy methods were used to transform the input and output. We created input and output (X and Y) representing the traffic congestion for each of training and testing.

Machine Learning

Hyperparameters are constant parameters that machine learning models depend on. We chose our hyperparameters from previous experimentation and by conducting our own tests (8).

ARIMA

We only used ARIMA on Dataset #1 due to its smaller size. To find the degree of differencing (d), we use the Augmented Dickey-Fuller test. To find the lag order (p) and moving average (q), we use the ACF and PACF plots of the data and examine the trends and spikes. We use p = 1, d = 1, q = 0: (1,1,0).

FFNNs and CNNs

For both neural network models, we chose 30 hidden layers, 60 epochs, a learning rate of 0.01, and a batch size of 30. We chose 50 neurons for each FFNN layer and 200 neurons for each CNN layer.

DP Mechanisms

These differentially private mechanisms (proved in Section I) add noise (ξ), based on their corresponding distributions, to each of the T time steps of the input data.

$$\text{Input training data + Noise } (\xi) = \text{Secure dataset}$$

Epsilon (ε), also known as the privacy budget, contributes to how much noise is added and was varied through a range (0.001…2)

We used the random.laplace() and random.noral() methods from NumPy to draw samples from the Laplace and Gaussian mechanisms, respectively. For the Staircase mechanism, we use Diffprivlib to generate the noise.

Received: June 9, 2022
Accepted: October 30, 2022
Published: December 21, 2022

REFERENCES

1. Brubaker, Marcus, and Simon Prince. "Tutorial #12: Differential Privacy I: Introduction." Borealis AI, 10 Feb. 2021, www.borealisai.com/research-blogs/tutorial-12-differential-privacy-i-introduction/.
2. Fan, Liyue and Xiong, Li. "An Adaptive Approach to Real-Time Aggregate Monitoring With Differential Privacy." *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 9, pp. 2094-2106, Sept. 2014, doi: 10.1109/TKDE.2013.96.
3. Rastogi, Vibhor and Nath, Suman. "Differentially private aggregation of distributed time-series with transformation and encryption." *Association for Computing Machinery*, pp. 735-746, Jun. 2010, doi: 10.1145/1807167.1807247.
4. Couchot, Jean-François, *et al.* "Anonymously forecasting the number and nature of firefighting operations." *Association for Computing Machinery*, no. 30, Jun. 2019, doi: 10.1145/3331076.3331085.
5. Eibl, Gunther, *et al.* "The influence of differential privacy on short term electric load forecasting." *Energy Informatics*, no. 48, Oct. 2018, doi: 10.1186/s42162-018-0025-3.
6. Smith, David, *et al.* "Privacy-Preserved Optimal Energy Trading, Statistics, and Forecasting for a Neighborhood Area Network." *Computer*, vol. 53, no. 5, pp. 25-34, May 2020, doi: 10.1109/MC.2020.2972505.
7. Athanasopoulos, George and Hyndman, Rob J. "Forecasting: Principles and Practice." *Otexts*, 2nd ed., 2018.
8. Aronsson, Linus and Bengtsson, Aron. "Machine

learning applied to traffic forecasting." *Chalmers University of Technology, Department of Computer Science and Engineering*, 2019, from odr.chalmers.se/bitstream/20.500.12380/300031/1/CSE%2019-05%20CPL%20Aronsson%20Bengtsson.pdf.

9. Wesner, Janet. "Mae and RMSE - Which Metric Is Better?" Medium, Human in a Machine World, 23 Mar. 2016, medium.com/human-in-a-machine-world/mae-and-rmse-which-metric-is-better-e60ac3bde13d.
10. Fathima, Shaistha. "What Is Differential Privacy?" Medium, Becoming Human: Artificial Intelligence Magazine, 1 Oct. 2020, becominghuman.ai/what-is-differential-privacy-1fd7bf507049.
11. Dwork, Cynthia and Roth, Aaron. "The Algorithmic Foundations of Differential Privacy." *Foundations And Trends In Theoretical Computer Science*, vol. 9, no. 3-4, pp. 211-407, Aug. 2014, doi: 10.1561/0400000042
12. Geng, Quan and Viswanath, Pramod. "The optimal mechanism in differential privacy." *IEEE international symposium on information theory*, pp. 2371-2375, Jun. 2014, doi: 10.1109/ISIT.2014.6875258.
13. Cuteri, Marco. (2011). "PEMS-SF Data Set." California Department of Transportation, UCI Machine Learning Repository [archive.ics.uci.edu/ml/datasets/PEMS-SF#].
14. Kairouz, Peter, *et al*. "The Composition Theorem for Differential Privacy." *IEEE Transactions on Information Theory*, vol. 63, no. 6, pp. 4037-4049, Jun. 2017, doi: 10.1109/TIT.2017.2685505
15. Rastogi, Vibhor and Nath, Suman, "Differentially private aggregation of distributed time-series with transformation and encryption." *2010 ACM SIGMOD International Conference on Mangement of Data,* pp. 735-746, Jun. 2010, doi: 10.1145/1807167.1807247.