NYC Trip Fare Prediction Using Temporal Convolutional Networks (TCN)

1st Md Nafiur Rahman Konok
dept. of CSE
Ahsanullah University of Science and Technology
Dhaka, Bangladesh
email: nafiur.cse.200104030@aust.edu

2nd Arafat Islam Shopnil
dept. of CSE
Ahsanullah University of Science and Technology
Dhaka, Bangladesh
email: arafat.cse.200104031@aust.edu

3rd Yasir Arafah Prince
dept. of CSE
Ahsanullah University of Science and Technology
Dhaka, Bangladesh
email: yasir.cse.200104042@aust.edu

4th Sumaiya Shejin
dept. of CSE
Ahsanullah University of Science and Technology
Dhaka, Bangladesh
email: sumaiya.cse.200104043@aust.edu

***Abstract*—This report examines taxi utilization patterns in New York City, focusing on how drop-off locations affect the time it takes for taxis to find their next fare. Using a 2013 dataset obtained through a Freedom of Information Law (FOIL) request, the analysis explores how factors such as borough and time of day influence taxi availability. Taxi utilization, defined as the percentage of time a taxi is occupied, is shown to vary based on geographic location, with taxis in busy areas like Manhattan finding new passengers quickly, while those in remote locations experience longer waits. To capture these patterns, a Temporal Convolutional Network (TCN) is applied to model the time series data, revealing temporal dependencies in taxi utilization. By leveraging advanced temporal and geospatial analysis techniques in Spark, this report provides key insights into the dynamics of taxi transportation across New York City.**

***Index Terms*—Temporal CNN, Heatmap, Time-series, Geospatial**

I. Introduction

Accurately estimating taxi costs in a busy city like New York City is essential to enhancing user experience and improving urban mobility. Conventional fare prediction models often fall short in capturing the dynamic and intricate nature of urban transportation systems, as they frequently rely on elementary machine learning techniques or simple linear regressions. This study offers a novel method for predicting NYC taxi fares using Temporal Convolutional Networks (TCN) combined with time series analysis. By incorporating temporal variables such as day of the week, time of day, and month, we aim to more accurately model the dynamic fluctuations in taxi fares.

Rather than relying solely on static features, we employ temporal CNNs to capture complex patterns in fare changes over time. This allows us to model the dependencies between trip features and time, reflecting real-world conditions like traffic patterns and demand surges. By integrating structured data such as trip distances and temporal variables with time series analysis, our approach seeks to create a more reliable and accurate fare prediction system.

This research not only improves the accuracy of fare forecasts but also demonstrates the broader utility of CNNs in processing complex, multimodal datasets. The work holds significant implications for urban transportation providers, offering the potential for better resource allocation, optimized routes, and increased passenger satisfaction through improved fare prediction.

In summary, this study introduces a pioneering methodology that leverages temporal CNNs and time series analysis for taxi fare prediction. By capturing the intricate temporal patterns and integrating them with traditional data, the proposed model promises to outperform existing methods, paving the way for smarter and more efficient urban mobility solutions.

II. Related Work

This section of study looked at several new publications about Geospatial and Temporal Data Analysis on New York City Taxi Trip Data.

Antoniades et al. [1] aimed to predict taxi fare and ride duration in New York City using data available at the ride's start. They utilized Linear Regression, Lasso, and Random Forest models based on features like pickup and dropoff coordinates, trip distance, and time. The study found that Random Forest models provided the most accurate predictions. For fare amount, the Random Forest model achieved an accuracy with a mean prediction error of 14.0% and a Root Mean Square Error (RMSE) of $2.28 on the validation dataset. For ride duration, mean prediction error was 24.3% and RMSE was 5.24 minutes on the validation dataset.

Chou et al. [2] focused on developing predictive models for taxi demand and fare using advanced machine learning techniques. The study incorporates a hybrid model combining Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) with Mixture Density Network (MDN) for taxi demand

Identify applicable funding agency here. If none, delete this.

forecasting, and an ensemble learning model blending Linear Regression (LR), Ridge Regression (RR), and Multilayer Perceptron (MLP) for fare prediction. The LSTM-RNN with MDN model achieved a Root Mean Squared Error (RMSE) of 3.31 for taxi demand prediction. The ensemble learning model used for fare prediction achieved the best RMSE of 3.24.

Umang Patel and Anil Chandan [3], explored the utilization of Big Data techniques to analyze taxi trip and fare data in New York City. This study leverages the data collected by the NYC Taxi & Limousine Commission, which includes 180 million taxi rides in 2014 alone. The findings from the Big Data analysis offer insights into optimizing taxi dispatch to high-demand areas and adjusting fare rates dynamically based on demand.

Christie Natashia Archie and Shubashini Rathina Velu [4] presented a comprehensive analysis of taxi trip data in New York City, aiming to derive insights into urban mobility patterns, economic activities, and optimal fleet management. Using datasets from 2009 to 2015, the authors applied various data preprocessing techniques, feature engineering, and exploratory data analysis (EDA) to understand taxi demand and supply dynamics.

Poongodi et al. [5] explores the application of machine learning techniques to predict taxi trip durations in New York City. The study focuses on the use of Multi-Layer Perceptron (MLP) and Extreme Gradient Boosting (XGBoost) models to predict how long a taxi trip will take. The results demonstrated that XGBoost outperformed MLP in terms of prediction accuracy. XGBoost had an average Root Mean Square Error (RMSE) of 0.39 for the training dataset and 0.44 for the testing dataset. MLP, on the other hand, showed a training accuracy of about 0.2740 and a testing accuracy of approximately 0.41, indicating that XGBoost was slightly better in performance.

The research contributes to the field of urban mobility and fare prediction by introducing a novel methodology that integrates geospatial images with traditional trip and fare data. Unlike existing models that rely solely on numerical and categorical data, this approach leverages the spatial context captured in satellite images of pickup and dropoff locations. By using Convolutional Neural Networks (CNNs), the model extracts complex spatial patterns and combines them with temporal and trip-specific features, providing a more accurate and robust prediction of taxi fares. This method enhances the precision of fare prediction, offering a significant improvement over traditional models.

Furthermore, the use of CNNs for regression tasks in this context is relatively unexplored, showcasing the versatility of CNNs beyond their conventional applications in image classification. By demonstrating the effectiveness of this approach, the research paves the way for innovative applications of deep learning in transportation analytics and urban planning, ultimately contributing to more efficient and user-friendly taxi services.

III. Dataset

The dataset utilized in this study comes from the NYC Taxi and Limousine Commission (TLC), specifically the 2013 NYC taxi trip and fare dataset. This dataset provides a comprehensive overview of taxi operations in New York City, capturing millions of individual trips and corresponding fare information. It is a rich source of data for transportation research, offering detailed insights into trip patterns, fare structures, and the factors influencing urban mobility. By leveraging this extensive dataset, our study aims to develop more accurate taxi fare prediction models, harnessing the large volume of data to explore temporal patterns and dependencies within NYC's dynamic transportation network.

A. Dataset Collection

The dataset utilized in this research is the NYC taxi trip and fare data from the year 2013. It is divided into two main categories: Trip and Fare. The trip data consists of 12 csv files, each with dimensions of 14,776,615 rows by 14 columns and on the other hand the fare data also comprises 12 csv files, each with dimensions of 14,776,615 rows by 11 columns. Both the categories have some common columns including unique identifiers for the taxi, taxi driver and vendor. Timestamp of when the trip started. In Trip category there has information about the code indicating the rate type, flag indicating if the trip record was stored and forwarded due to connection issues, timestamp of when the trip ended, number of passengers, duration of the trip in seconds, distance of the trip in miles, longitude and latitude of pickup and dropoff locations. On the other hand, the Fare category has information about the method of payment, base fare for the trip, additional surcharges, Metropolitan Transportation Authority tax, driver tip, amount paid for tolls and finally total fare amount for the trip, including all charges. Table 1 shows the columns of each category.

Table 1 shows the columns of each category,

TABLE I
DATASET COLUMNS OF EACH CATEGORY

Trip	**Fare**
medallion	medallion
hack_license 1	hack_license
vendor_id	vendor_id
rate_code	pickup_datetime
store_and_fwd_flag	payment_type
pickup_datetime	fare_amount
dropoff_datetime	surcharge
passenger_count	mta_tax
trip_time_in_secs	tip_amount
trip_distance	tolls_amount
pickup_longitude	total_amount
pickup_latitude	
dropoff_longitude	
dropoff_latitude	

B. Data Pre-processing

By addressing missing values and eliminating duplicates, we cleansed the data. The current features were transformed into new features through the use of feature engineering approaches. To be more precise, we took the day of the week and the hour, day, and month out of the pickup datetime in order to record different time-related elements of every journey. Furthermore, we calculated the difference between the pickup and drop-off times to get the journey duration in minutes. More in-depth research is made possible by these additional features, which offer deeper insights into trip characteristics and temporal trends.

C. Train-Test splits

To evaluate the performance of our models, we applied 80% of the data used to train the models, allowing them to learn patterns and relationships, while the remaining 20% was set aside for testing. The test data, which the models had not seen during training, was used to assess their performance and generalization ability on unseen data, ensuring that the models can make accurate predictions in real-world scenarios.

IV. Methodology

A. Geospatial Analysis

In our analysis using geospatial techniques, we explored several key metrics related to trips, payments, and locations. First, we identified common pickup and drop-off locations, which allowed us to visualize and understand the most frequent starting and ending points of trips in specific geographic areas. This information helps in optimizing resources and understanding high-demand zones. We then analyzed the trip distances, where we computed both the longest and shortest trips based on the distance traveled, giving insights into the range of trips typically taken.

Next, we calculated the average fare amounts for each pickup location, helping to determine fare trends in specific regions. We also examined the most common pickup and drop-off location pairs, identifying frequently traveled routes, which could indicate regular commuting patterns or popular areas for travel. To further delve into the fare data, we explored the top highest and lowest fares, revealing the extremes in pricing and their possible geographical patterns.

We also assessed payment behaviors by computing the average fare amounts for different payment types and categorized payment types used (e.g., cash, credit card). This breakdown provided a clearer understanding of how passengers prefer to pay for their rides. Additionally, we calculated the average trip distance for different payment types, offering insights into how payment preference might vary with trip length.

Lastly, we calculated the average total amounts (including additional charges) for each payment type, providing a comprehensive view of the payment landscape and its association with trip features.

we calculated the average trip distance for different passenger counts. This helped identify how trip lengths vary with the number of passengers, revealing whether trips with more passengers tend to be longer or shorter. This insight is valuable for optimizing fleet management and pricing strategies based on group travel patterns.

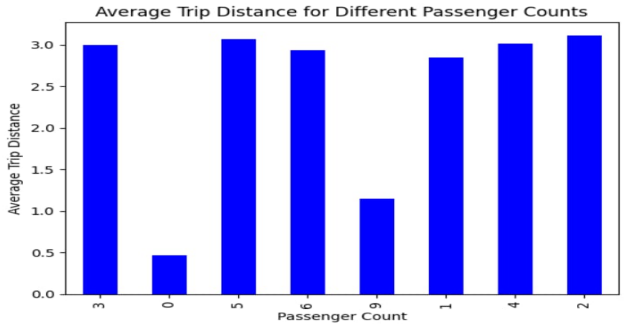

Fig. 1. Average trip distance for different passenger counts

Fig. 2. Average fare amount for different passenger counts

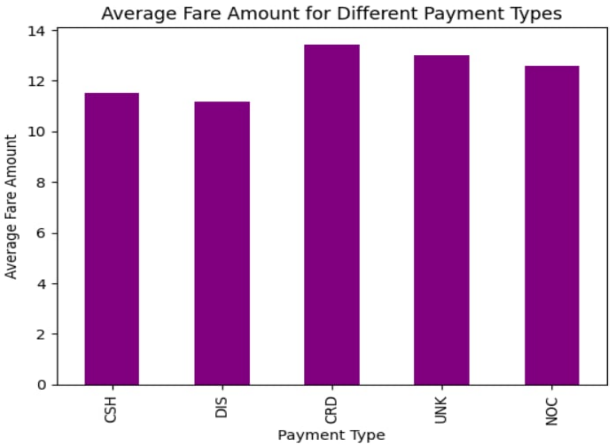

Fig. 3. Average fare amount for different payment types

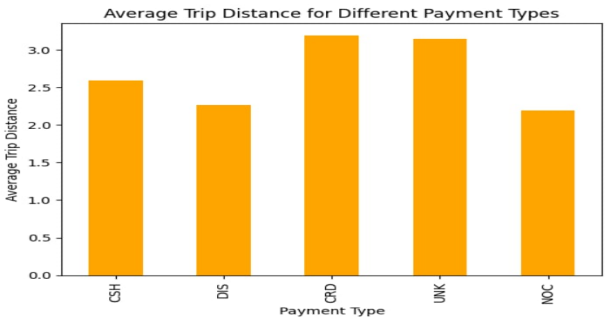

Fig. 4. Average trip distance for different payment types

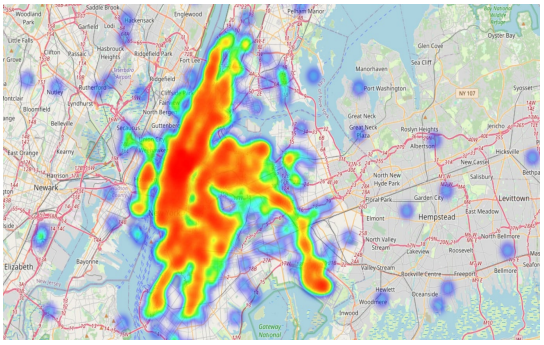

Fig. 5. Heatmap based on pickup location

B. Time-series Analysis

In our time series analysis, we explored various patterns related to taxi pickups, drop-offs, fare amounts, and other trip characteristics over time. Starting with peak hours for taxi pickups and drop-offs, we identified the times of day when the highest volume of rides occurred, providing insights into demand trends. We then analyzed how the average fare amount varies by hour, revealing fluctuations in pricing across different times of the day, which can be driven by factors like traffic, demand, and trip length.

Next, we examined the average trip distance by hour of the day, helping us understand how trip lengths change over time, possibly indicating longer trips during off-peak hours or shorter trips during high-demand times. We also analyzed passenger count across different hours, which provided insights into how many people typically ride during certain times and ride-sharing strategies.

We looked into the hourly total fare amounts, identifying specific hours that generated the most revenue, and analyzed how payment types vary across different hours, shedding light on how passengers prefer to pay (e.g., cash or card) at different times. We also explored whether drop-off times vary by hour, providing insights into trip durations and potential traffic patterns throughout the day.

Further, we investigated how the tip amount varies by hour, which could reveal when passengers are more likely to tip generously. We extended our analysis to understand pickup and drop-off variations by day of the week, highlighting daily travel patterns and identifying peak days for taxi services.

We explored how the average fare amount and total fare amount fluctuate by the day of the week, helping us pinpoint which days are the most profitable. We also compared peak pickup hours on weekdays vs. weekends, showing how demand shifts depending on whether it's a workday or a leisure day.

In terms of fare trends, we assessed whether the average fare amount varies by hour on weekdays vs. weekends, which may indicate differences in fare structures due to different demand patterns. Additionally, we identified the most popular pickup locations by hour, allowing us to map out hot spots throughout the day.

Finally, we explored how trip distances vary by pickup location and time of day, offering insights into location-specific travel behavior.

By leveraging time series analysis, we were able to uncover temporal trends and patterns that are crucial for optimizing taxi operations, pricing strategies, and resource allocation.

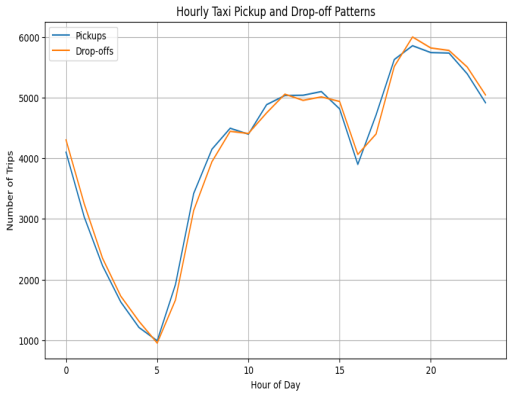

Fig. 6. Hourly taxi pickup and dropoff pattern

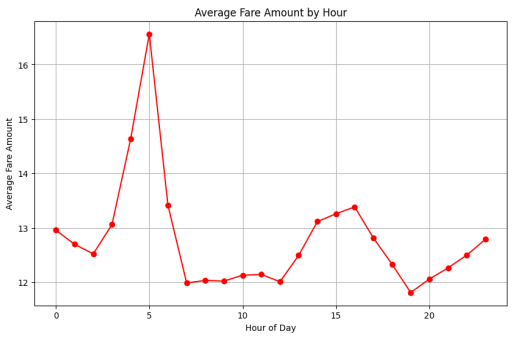

Fig. 7. Average fare amount by hour

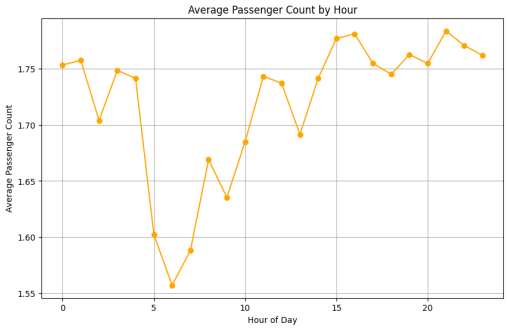

Fig. 8. Average passenger count by hour

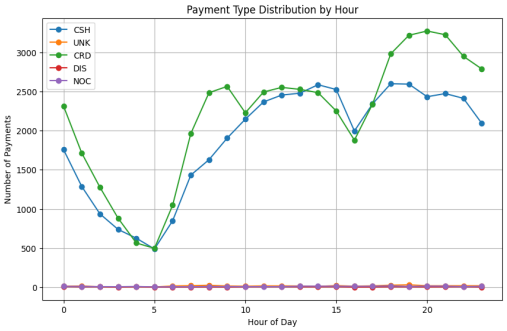

Fig. 9. Payment type distribution by hour

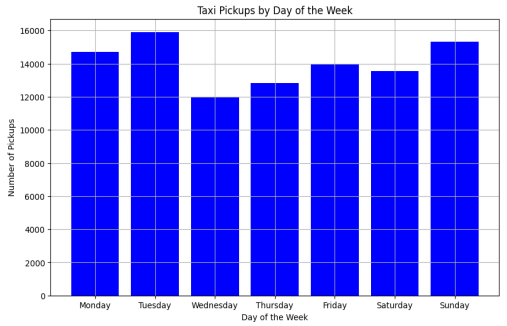

Fig. 10. Taxi pickup by day of the week

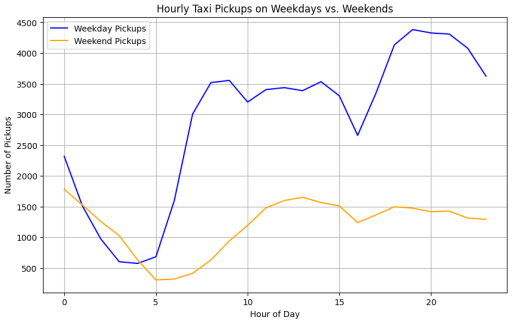

Fig. 11. Hourly taxi pickups by weekdays and weekends

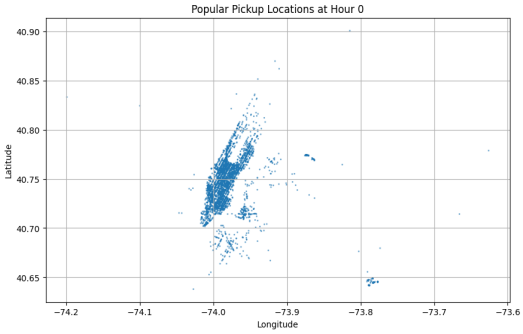

Fig. 12. Popular pickup location by hour 0

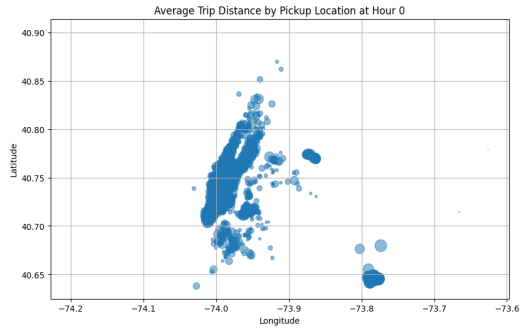

Fig. 13. Average trip distance by pickup location by hour 0

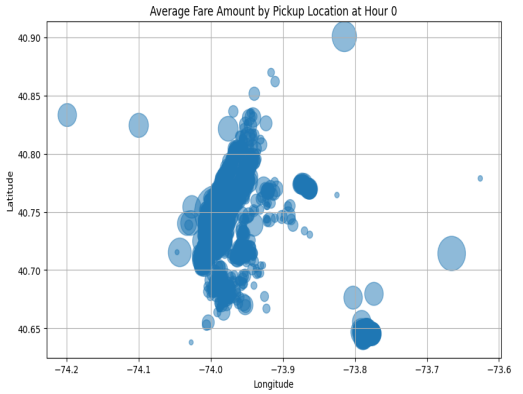

Fig. 14. Average fare amount by pickup location by hour 0

V. Result Analysis

In this study, we utilized a Temporal Convolutional Network (TCN) to predict NYC taxi fares by learning from time series data. The results indicate that the Temporal CNN model achieved strong generalization and convergence. The training and validation losses stabilized at approximately 15 Mean Squared Error (MSE), with the validation loss remaining consistently low throughout the training process. This convergence suggests that the model did not overfit and was able to generalize effectively to unseen data. The minimal gap between training and validation losses demonstrates that the Temporal CNN was able to learn effectively from the temporal features in the dataset without significant overfitting.

s

In terms of fare prediction accuracy, the model's predicted fare amounts closely followed the actual fare trends. It was able to capture temporal patterns effectively, as shown by the close alignment between actual and predicted fare amounts. However, the model experienced some difficulty predicting outliers or extreme fare values, where deviations were more noticeable. This suggests that further tuning or the inclusion of additional features may improve the model's performance, particularly in predicting extreme fares.

Overall, the Temporal CNN model shows strong potential for predicting taxi fares by learning from temporal features. It successfully captured the underlying patterns in the dataset, as evidenced by the stable loss behavior across epochs and the close correspondence between actual and predicted fares. This highlights the model's effectiveness in improving fare prediction accuracy, which could contribute to more efficient urban mobility solutions.

Fig. 15. Training and validation loss

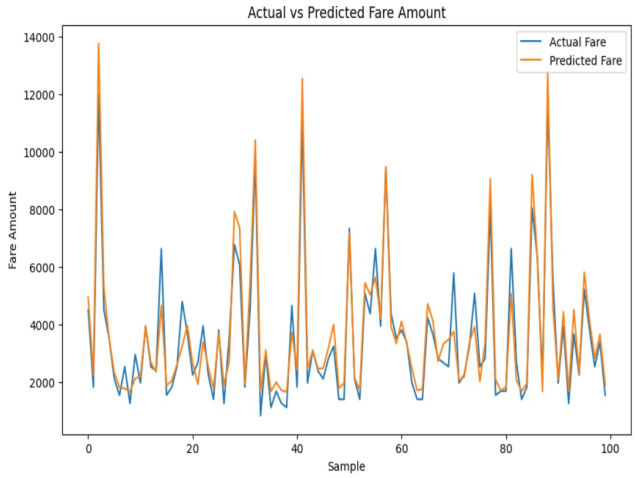

Fig. 16. Actual and predicted fare amount

VI. Conclusion & future work

In conclusion, this study demonstrates that the application of Temporal Convolutional Networks (TCN) for predicting New York City taxi fares is a promising and effective approach. The model exhibited strong generalization capabilities, with stable training and validation losses, indicating its capacity to capture temporal patterns and dependencies within the data without overfitting. While the model performed well in predicting typical fare ranges, there were occasional deviations in the prediction of extreme or outlier fares, suggesting opportunities for further enhancement.

Future work could focus on improving the model's performance by incorporating additional contextual features such as weather conditions, real-time traffic data, or socioeconomic factors that influence fare variability. Further exploration of hyperparameter tuning and the integration of other advanced machine learning techniques may also improve the model's ability to handle outliers. Additionally, extending the methodology to other cities or across longer time periods would provide valuable insights into the model's generalizability and robustness in diverse urban environments.

References

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.
[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
[3] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
[4] K. Elissa, "Title of paper if known," unpublished.
[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.
[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.