Forecasting Traffic Flow: Short Term, Long Term, and When It Rains

Hao Peng, Santosh U. Bobade, Michael E. Cotterell and John A. Miller

Department of Computer Science

University of Georgia

System Motivations

- Performance parallel/distributed
- Support Advanced Time Series Analysis with uniform interface
- Provide Database storage and flexible queries
- Provide support for automation through ontology
- Exploit Theoretical and Simulation Models

Application Motivations

- Availability of Big Data through sensors
- Benefits of Research
  - Intelligent Transportation System
  - Traffic Apps
  - Advanced Trip Planning
- Many existing studies
  - Tend to focus on the immediate short term (e.g., 1-step ahead forecasts)
  - Do not consider factors such as weather conditions

Time Series Analysis in the Big Data Era

- Database and Statistical Packages
  - MySQL + R
- JVM based Big Data Frameworks
  - SparkSQL + Spark
  - ScalaTion TSDB + ScalaTion Analytics
- Python based Big Data frameworks
  - pandas + tensorflow + keras

ScalaTion TSDB + Scalation Analytics

- val start = new TimeNum("2017-01-01 -06:00", "yyyy-MM-dd Z")
- val y = traffic.where[TimeNum](("Read Date",(x:TimeNum) => x >= start))
- .select("Volume")
- .toVectorD(0)

- val model = SARIMA (y)
- model.train().eval()
- println(model.fit)
- println(model.forecast(24))

Big Data

- Research is progressing towards larger datasets
- Current Study
  - traffic in GA, ~50MB
    - Volume, Rainfall
- Next Study
  - Austin, TX, ~150MB
    - Volume, Speed, Occupancy, Temperature, Relative Humidity, Wind Speed, Rainfall, Dewpoints
- Future Study
  - PeMS, San Francisco & Los Angeles, CA, ~GBs

Progress in Analyzing Traffic

- Past: univariate time series, small number of datasets, immediate short term forecasts (e.g., minutes)
- Current: deep learning, feed data into deep NNs
- Research Agenda
  - Multivariate Time Series
    - Weather, Spatial Dependencies, Events, Accidents, Road Repair Schedule
  - Time Series Database (TSDB)
    - Easy to use, combine database queries with analytics
  - Theory driven models
    - Theories can help to guide/restrict the model building process

ScalaTion Project

- A Scala-based project for analytics, simulation and optimization
- Open source under an MIT License
- Forecasting models used in this study include
  - Seasonal ARIMA
  - Dynamic Regression
  - Exponential Smoothing
  - Feedforward multi-layer Neural Networks
  - Long Short-Term Memory Neural Networks (under development)
- www.cs.uga.edu/~jam/scalation.html

Seasonal ARIMA

$$Y_t = \sum_{i=1}^{p} \phi_i Y_{t-i} + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} + \sum_{i=1}^{P} \Phi_i Y_{t-is} + \sum_{i=1}^{Q} \Theta_i \epsilon_{t-is} + \epsilon_t$$

- Uses lagged and correlated values of the time series and errors to make forecasts
- Differencing may be necessary to make the time series stationary
- SARIMA(1,0,1)x(0,1,1)120 was the chosen model, as in [Williams and Hoel, 2003], [Shekhar and Williams, 2008] and [Lippi et. al, 2013].
- Automated order search based on AICc as described in [Hyndman and Khandakar, 2007] was also attempted, but the automated models only yielded better results than SARIMA(1,0,1)x(0,1,1)120 for approximately one-third of the traffic sensor data.

Dynamic Regression

$$Y_t = f_{\text{SARIMA}} + \epsilon_t$$
$$\epsilon_t = \beta x_t + z_t$$

- Uses external variables such as rainfall to further explain additional variabilities in the traffic flow time series
- A simple, two-step process for forecasting using both a time series model and a regression model

Exponential Smoothing

$$S_t = \alpha(Y_t - c_{t-L}) + (1-\alpha)(S_{t-1} + b_{t-1}) \ ,$$
$$b_t = \beta(S_t - S_{t-1}) + (1-\beta)b_{t-1} \ ,$$
$$c_t = \gamma(Y_t - S_t) + (1-\gamma)c_{t-L} \ .$$

- Triple Exponential Smoothing with additive seasonality
- 12-step ahead within sample forecast SSE was minimized to find the 3 smoothing parameters
- The default 1-step ahead SSE was attempted, but resulted in very poor forecasting results for higher steps, possibly due to the lack of need to rely on seasonal components make good forecasts for 1-step ahead.

Feedforward Neural Networks

$$a_{\text{out}} = \sigma(\mathbf{w} \cdot \mathbf{a_{in}} + b)$$

- Tanh activation function was used
- Data were normalized to [-0.8, 0.8]
- Back-propagation was used to learn the weights and biases that minimize MSE
- 4 layer structure
  - Input layer of size 50, including the data from previous 24-hr period, the 24-hr period in the previous week, the day of the week and time of the day
  - Two hidden layers of size 40 and 30
  - Output layer of size 24, corresponding to 24-step ahead forecasts
- Other parameters were optimized using grid search

Dataset

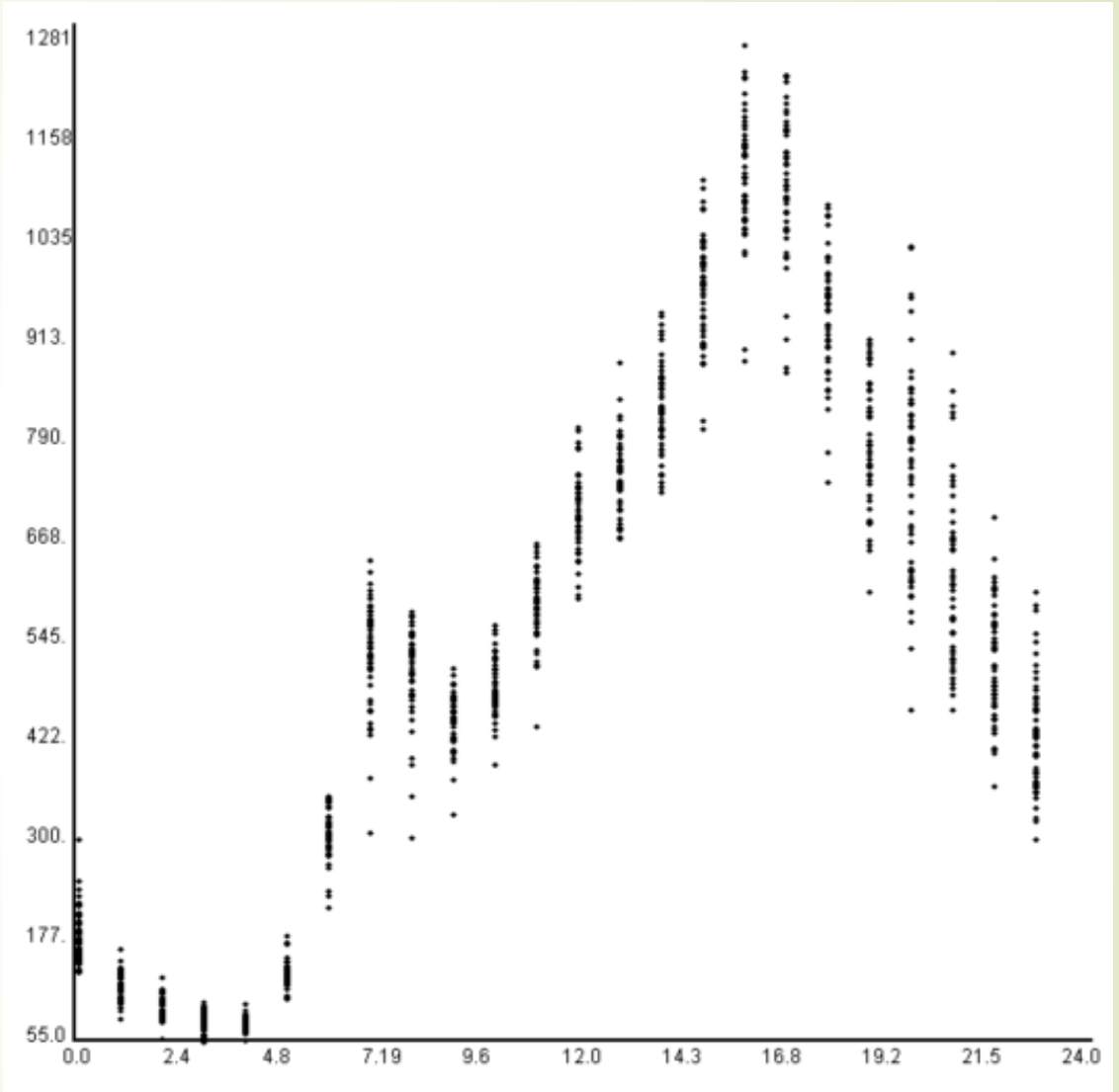
- Hourly traffic flow data
  - Georgia Department of Transportation
  - 74 sensors (both directions), mostly urban areas and major freeways
  - Jan 2013 – June 2017
  - http://www.dot.ga.gov/DS/Data
- Hourly Precipitation Data
  - Automated Surface Observing System (ASOS)
  - 14 sensors that are paired with nearby traffic sensors
  - https://mesonet.agron.iastate.edu/request/download.phtml?network=GA_ASOS

Traffic Sensors

County	Nearby City, Freeway or Point of Interest	Traffic Station ID
Banks	Commerce, I-85, Tanger Outlets	011-0103
Bibb	Macon, I-16, I-75,	021-0116 021-0132 021-0158 021-0258 021-0267 021-0334 021-0349 021-0372 021-0376 021-0541 021-0587
Bryan	Savannah, I-16	029-0103
Camden	I-95	039-0145 039-0218
Chatham	Savannah, I-16, I-516, I-95	051-0107 051-0109 051-0132 051-0137 051-0138 051-0318 051-0334 051-0383 051-0443 051-0509 051-0649
Clarke	Athens, University of Georgia (UGA)	059-0014 059-0087 059-0118 059-0367 059-0611 059-0613
Clayton	Atlanta, I-285	063-0383 063-1023 063-1032 063-1085 063-1172 063-1201
Cobb	Atlanta	067-2334 067-2623
Fulton	Atlanta, I-75, I-85	121-0178 121-0190 121-5110 121-5114 121-5225 121-5374 121-5463 121-5468 121-5486 121-5524 121-5633 121-5969 121-6370
Glynn	Brunswick, St. Simon's Island	127-0105 127-0107 127-0236 127-0289 127-0456
Gwinnett	Atlanta, I-85, Mall of Georgia	135-0298 135-0305 135-0563
Houston	South of Macon	153-0143 153-0189 153-0332 153-0365
Muscogee	Columbus, I-185	215-0165 215-0336
Oconee	Watkinsville, UGA	219-0203
Richmond	Augusta, I-20, I-520	245-0214 245-0218 245-0223 245-0233 245-0303 245-0947

Friday Traffic on US 23 in Atlanta, GA

- Much busier afternoon rush hours since everyone is getting off work and trying to go home
- The same road but for the other direction has much busier morning rush hours



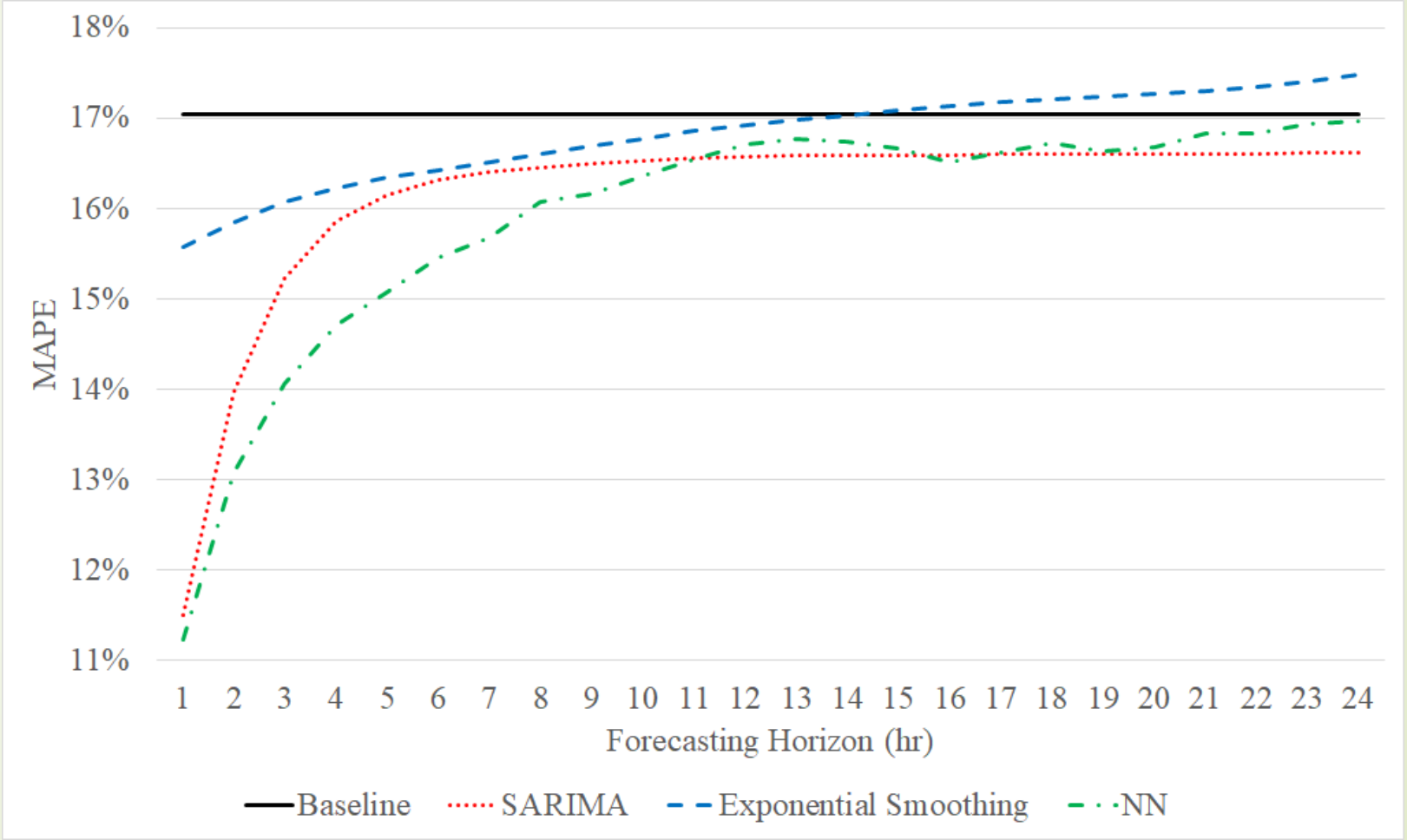
Testing Platform

- Sapelo cluster from Georgia Advanced Computing Resource Center
  - 48-core AMD Opteron Machine
  - Parallel training and testing were conducted per sensor per direction
  - https://gacrc.uga.edu/

Performance Evaluation

- Rolling Forecasts
  - 12 weeks of data as training set (24 weeks if considering rainfall)
  - 8 weeks of data as testing set
  - Sliding window is 8 weeks
- Forecasts were made for 24 hours/steps into the future
- Models that incorporated rainfall data only made 1 hour/step ahead forecasts
  - Difficult to make reliable, long-term weather forecasts
- Forecasts were only produced within the 7:00AM to 7:00PM range on weekdays
- Weekly historical averages by hours were used as baselines
- Mean Absolute Percentage Error (MAPE) was the metric of evaluation

Performance Comparison



Performance Comparison in Rainy Weather

	Baseline	Baseline (12 weeks)	SARIMA	Exponential Smoothing	Neural Network
Traffic Flow Data Only	42.43%	41.62%	20.11%	34.05%	15.12%
With Rain-fall Data	38.03%	37.15%	19.79%	33.41%	14.95%

Conclusions and Future Work

- Short and Long Term Traffic Flow Forecasting
- Weather conditions including rainfall can help to further explain traffic conditions
- Neural Networks performed well
  - Preliminary results show that LSTM can be very promising
- Future Work
  - Larger datasets with higher resolutions
  - Additional models
  - Exogenous variables, Multivariate Time Series

Questions