Technologies for Safe & Efficient Transportation

THE NATIONAL USDOT UNIVERSITY TRANSPORTATION CENTER FOR SAFETY

Carnegie Mellon University | UNIVERSITY of PENNSYLVANIA

User-Centric Interdependent Urban Systems: Using Energy Use Data and Social Media Data to Improve Mobility

FINAL RESEARCH REPORT

PI: Zhen (Sean) Qian, Ph.D.

Graduate Student Researchers: Weiran Yao, Pinchao Zhang

Department of Civil and Environmental Engineering

Carnegie Mellon University

Contract No. DTRT12GUTG11

DISCLAIMER

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the U.S. Department of Transportation's University Transportation Centers Program, in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof.

1. Project overview

1.1. Background

Central to smart cities is the complex nature of interrelationships among various urban systems. Linking all urban systems is the system users. The individual daily activities engage using those urban systems at certain time of day and locations. There may exist clear spatial and temporal correlations among usage patterns across all urban systems. A general idea is to fuse and analyze user demand and usage data from transportation, energy, water, building systems and social media platforms, as shown in Fig. 1, to discover the spatiotemporal usage patterns among those systems. This enables cross-system demand prediction and management. For some users, the usage of one urban system is likely to be used minutes or hours ahead of their usage of other urban system(s) as a result of daily activity chains. Therefore, the spatiotemporal usage of an urban system can be accurately predicted a few minutes or hours ahead by real-time sensing user patterns of other urban system(s). This is otherwise hard to accomplish by solely monitoring one "siloed" system. Ultimately, real-time control strategies for demand management of one urban system can be developed with efficient real-time demand prediction upon other urban system(s).

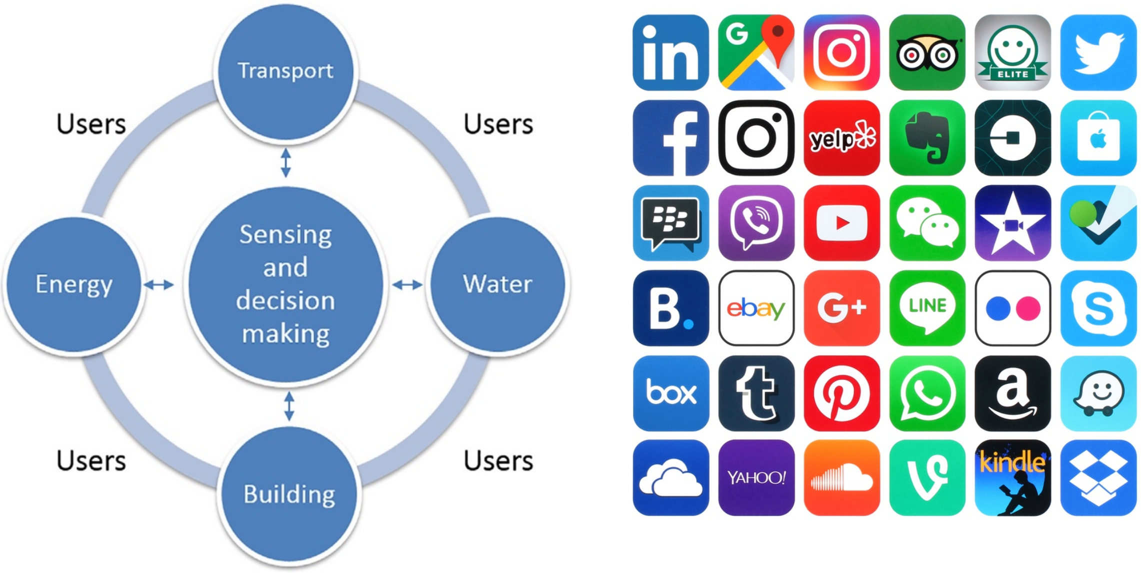
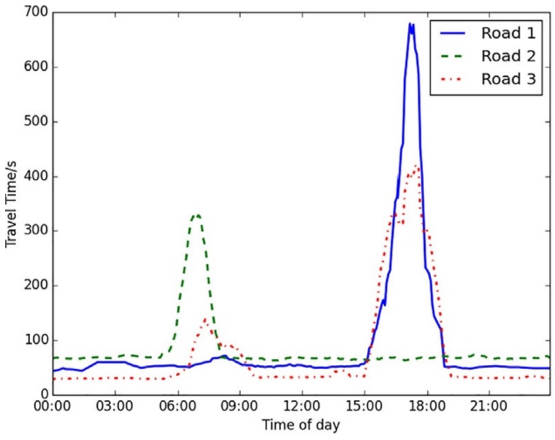


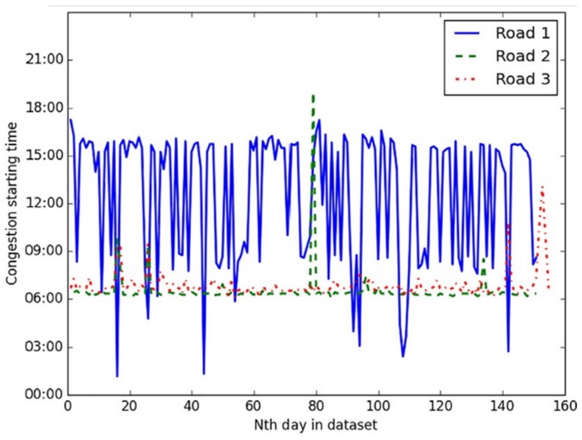
Fig. 1. Interdependency of some urban systems: their system user patterns are inter-related both temporally and spatially (social media platforms graph source: www.prestigesocialmedia.com.au).

This project is the first step to achieve the ultimate goal by developing a reliable traffic prediction framework which makes use of inter-correlations among urban systems for increasing forecasting accuracy and horizons. In this project, we propose a general crowdsourced data-driven framework to effectively collect, sense and analyze people's daily activity patterns from two interdependent urban systems, which includes energy system and social media system, for improving congestion predictions in transportation systems. We further focus our applications on predicting congestions during morning periods before commuter departures in the early morning, such as before 5AM. This problem setting is meaningful in practice because knowing in advance the accurate traffic conditions of a road before leaving homes enables travelers to better plan their trips. Traffic control strategies can also be effectively adjusted in time before peak hours thanks to the long forecasting horizon.



(a)

Fig. 2(a). Morning congestion patterns.



(b)

Fig. 2(b). Variance of congestion starting time.

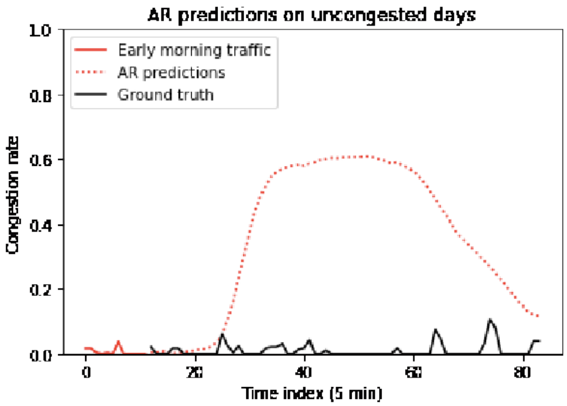


Fig. 2(c). Autoregressive predictions on an uncongested day.

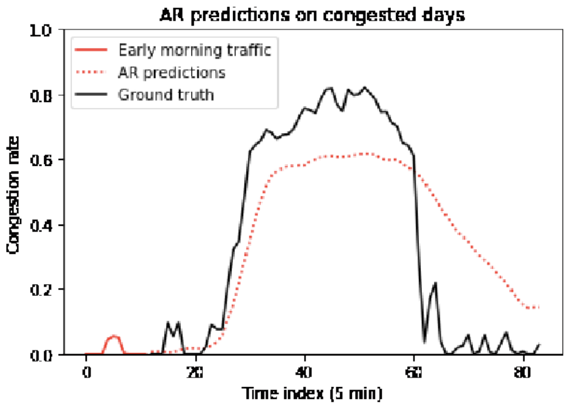


Fig. 2(d). Autoregressive predictions on a congested day.

However, most existing methods which solely use information from traffic systems, such as autoregressive time-series models, could actually fail under this problem setting. Fig. 2 conceptually illustrates why using real-time traffic data is usually not sufficient for morning congestion predictions. Because autoregressive models rely primarily on correlations between future and past traffic states, they could fail if past traffic dynamics contain little information, or in other words, additional factors (e.g. weather, incidents, etc.) have more impacts on future traffic. Morning congestion prediction falls exactly into this situation. Roads to be congested in morning peak hours, such as Road 2 in Fig. 2(a), have been in free-flow conditions for hours before the model makes predictions at 5 AM. Real-time monitoring the speed or travel time does not necessarily help predict the exact time of traffic break-down, nor would historical data help as much due to day-to-day variation as shown in Fig. 2(b). The performance of autoregressive models for morning congestion predictions are shown in Fig. 2(c) and Fig. 2(d). Because early morning traffic contains very little information to explain peak congestion variances, the autoregressive model can at best predict the road's historical average congestion rates. Finally, the morning traffic predictions for congested days, as shown in Fig. 2(c), and uncongested days, as shown in Fig. 2(d) on a same road segment can look very similar.

Things are different if taking energy and social media systems into consideration. There may exist spatiotemporal relations between the morning travel demand and crowd activities patterns (e.g. sleep/wake up time, etc.) extracted from electricity usage data and geocoded social media posting activity and content. Those daily characteristics may be partially attributed to those commuters' activities at midnight or early in the morning (such as 3–5 am). Therefore, real-time traffic prediction can be complemented and enhanced by mining additional real-time electricity usage and social media data in addition to traffic data. We hope to provide more accurate prediction for real-time traffic. In principle, we would expect electricity usage and social media data to add additional insights for a better traffic prediction, such as to explain what kinds of spatiotemporal electricity and social media patterns are precursors for morning congestions.

For commuters, the developed framework is helpful for planning trips because morning congestion with fine-grained spatial and temporal resolutions are predicted long before peak hours. For traffic management agencies, this method is capable of evaluating past traffic network congestion and helping adjust traffic control policies in time because of its longer forecasting horizons. For researchers, this study provides new approaches to sense crowd activity patterns from user interactions with multiple urban systems and illustrates the relationships between people's activities and traffic congestion through energy usage and social media perspectives.

1.2. Problem statement

The main objective of this project is to predict the congestion on a particular road segment in the traffic network some time during morning periods, which are defined as from 5:00/6:00 AM to 10:59 AM, using electricity or social data available until early morning, which is defined as before 5:00 AM on that day. The reference (free-flow) speed v^{ref} of road segment i is calculated as the 85 percentiles of observed speed on that segment for all time periods (Eq. 1), which is a commonly-used way to determine reference speed from probe-based speed data. Congestion is described by congestion rates (r_{it}^{d}). The congestion rate on a road segment is defined in Eq. 2 as the percentage decrease from the free-flow it (reference) traffic speed of the road v^{ref} to the observed speed v_{it}^{d}. A road segment is defined as congested at time t (S_{it}^{d}= 1) if observed speeds drop below rthres of reference speed, i.e., r_{it}^{d} ≥r^{thres}, for at least t min minutes, as defined in Eq. 3. For each segment i on day d, we measure its congestion starting time CST_{i}^{d} and duration CD_{i}^{d}. If multiple congested periods occur, congested starting time is defined as the starting point of the first congestion period. Congestion duration is defined as the interval between the first congestion starting point and the last congestion ending point. If any congestion period exists on segment i during morning periods of day d, we say congestion occurs, i.e., O_{i}^{d}= 1. Three types of congestion measurements are predicted: (1) congestion occurrences O_{i}^{d}, where the output is 100 a binary prediction indicating if any period in morning on day d is congested; (2) congestion starting time (CST_{i}^{d}), where the continuous output is the starting point of congested periods on day d, and (3) congestion duration (CD_i^d), where the continuous output is the duration between the starting and ending point of congested periods.

$$v_i^{ref} = P_{0.85}(v_{it}^d) \tag{1}$$

$$r_{it}^d = 1 - v_{it}^d / v_i^{ref} \tag{2}$$

$$S_{it}^d = \begin{cases} 1, \text{if } r_{it}^d, r_{it+1}^d, \ldots, r_{it+l_{min}}^d \\ 0, \text{otherwise} \end{cases} \tag{3}$$

2. Using time-of-day electricity usage data to predict morning roadway congestion

In this section, we consider predicting traffic congestion at a morning time t, using household-level electricity usage data during the time interval $[t', t'']$ that can be up to a few hours earlier than t on the same day. $[t', t'']$ is set as [12am, 6am] in this section. As we will show later, this predictor can outperform a predictor using only the traffic data, even at the time much closer to t than t″. This is because morning traffic breakdown is largely attributed to random demand characteristics, such as demand level, driving behavior, and departure time from home. Those demand characteristics can hardly be predicted using traffic data only in the real time, and oftentimes the traffic break-down may occur in the morning peak without sufficient prior signs from traffic data. Our hope is that electricity use at midnight or in the very early morning can partially reveal those demand characteristics to some extent, and therefore can help better predict traffic congestion for some locations.

2.1. Data sources

2.1.1. Electricity usage data

The electricity usage data are acquired from an Advanced Metering Infrastructure (AMI) program run by Pecan Street Inc. There were around 400 households participating in the program in the year of 2014. There were some households entering and leaving this program during the year. We choose those households that stayed the entire year of 2014, in all 322 households. For each of those households, the electricity use (in kW h) were recorded in 5 min time intervals throughout the year. We consider all 251 weekdays in 2014. The daily use profiles from midnight to 6 am are normalized in a way that its sum of squares is one. Though each household is referred to with a unique reference ID number, those households/users are completely anonymous. Their locations and any other private information are filtered out and unknown to this research.

2.1.2. Traffic data

Historical travel time data are acquired from National Performance Management Research Data Set (NPMRDS). The travel time data were provided in 5-min time intervals, and cover highways and major roads around the Austin Metropolitan Area. Its spatial resolution is defined by Traffic Management Channels (TMC), and acquired from NPMRDS as well. As shown in Fig. 3, we select 15 TMC road segments in this study.

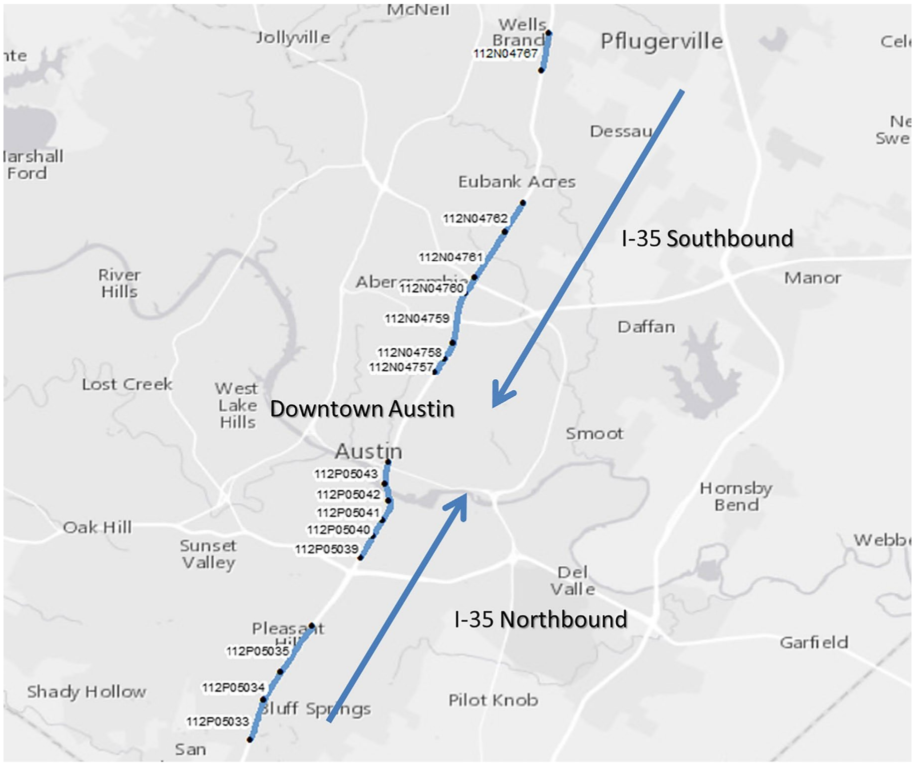


Fig. 3. The 15 road segments used in this section.

2.2. Feature extraction

2.2.1. Cluster analysis

We first use a simple K-means clustering with K = 2 to separate all workdays into two seasons. A larger portion of electricity is consumed during the night in summer than winter, whereas more electricity is used during the morning in winter than summer. In this study, we only use weekdays in summer. By clustering, the summer starts in April and ends in October. To completely filter out the seasonal effect, only weekdays from May to October are used. Furthermore, we also found that the daily patterns on Monday and Friday could be very different from those on other weekdays. Hence, we focus on all Tuesdays, Wednesdays and Thursdays in summer, in all 79 weekdays. Next, we conduct a clustering analysis for all 79 weekdays × 322 households = 25,438 daily profiles. Each daily profile is a vector of 72 elements representing the electricity use of all 5-min time intervals from 12 AM to 6 AM. K = 10 is selected by GAP statistics for the K-means algorithm. Fig. 4 plots the daily profile of cluster average for each of the 10 clusters, representing 10 most representative patterns. We denote the ten clusters as patterns $A, B, C, \dots, J$. The time-of-day electricity use varies quite substantially among those patterns.

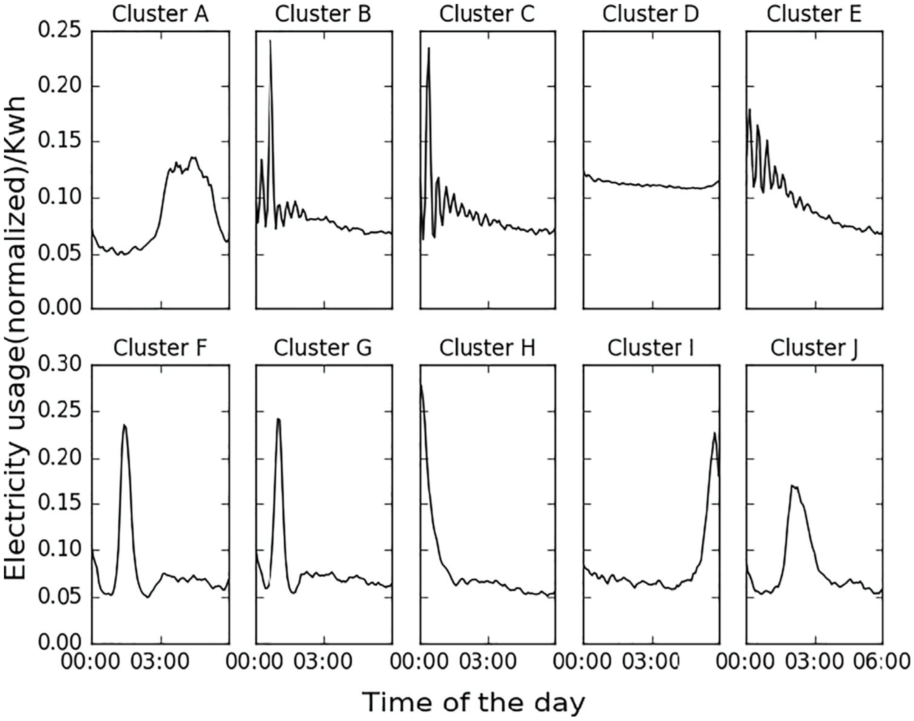


Fig. 4. The 10 most representative electricity usage patterns.

2.2.2. Feature encoding

We construct features on each day that are derived from the clustering results of daily profiles. Thus, each household or user on each day will be assigned to a typical pattern (cluster). Features that are related to traffic congestion can be both aggregate or disaggregate. Aggregate features are highly compressed. They are a vector of $K - 1$ elements for each daily profile. Each element is the ratio of households/users that are under a pattern on that day. The element of one last pattern can be dropped from the vector as a result of redundancy. Those aggregate features would offer effective prediction if the assumption holds that all households/users are homogeneous under the same electricity use pattern.

If this is not reasonable, then disaggregate features allow us to examine household/user-level behavior in full details. On each day, the electricity use pattern of a household is represented by a vector consisting of K−1 binary elements where an element is 1 if a pattern is followed and zero otherwise. The feature vector length is $(K - 1) \times H$, possibly exceeding the sample size.

2.3. Model construction

We assume that congestion starting time on the dth day is predicted by its expected value that is a linear combination of many features:

$$E(y^d) = \beta^T x^d \tag{4}$$

where x^d is a vector of p features observed on the dth day. Those features are derived from daily electricity profiles e^d. The commonly used ordinary least square (OLS) linear predictor is to learn the parameters β such that:

$$\min_{\beta} ||y - x^T\beta||_2 + \alpha||\beta||_1 \tag{5}$$

where $||\beta||_1 = \sum_{i=1}^{d} \beta_i$ denotes the L1-norm of β. The LASSO regression helps select the most critical features that are linearly related to the response.

2.4. Results and discussion

2.4.1. Examination of the relationship between electricity usage patterns and morning congestion

This subsection highlights the correlations between aggregate features and the congestion starting time, which uses all 79 weekdays in regressing the predictor. Fig. 5 shows the coefficients and their respective p-values for all 15 linear predictors. Patterns B, C, E, F and G have positive effects on the congestion starting time. Patterns B, C, and E are households who steadily use electricity after midnight with an oscillating and declining usage over time. It is no surprise that more households in those patterns on a particular day lead to a later congestion starting time next time. We speculate that those households with those types of after-midnight activities are likely commuters. They are likely to commute later if under those patterns than if under patterns A, I and J. Of those positively correlated patterns, B, C, and E have higher positive effects than F and G, and the effects on F and G are not statistically significant. F and G are households who use electricity intensively at around 1:00–1:30 am, but the usage is very low at all other times. This type of midnight activities may not imply commuters, or not necessarily related to their travel activities.

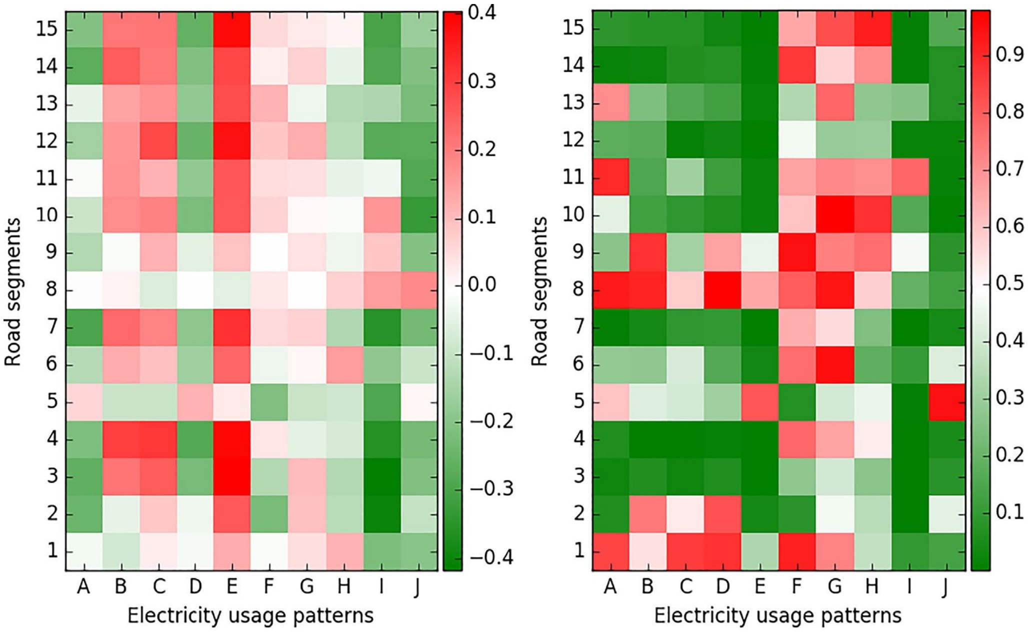


Fig. 5. Regression results using aggregated features for predicting congestion starting time.

Patterns A, D, I and J have negative effects on the congestion starting time, which are statistically significant for most TMCs. The effect by pattern D is slightly milder. Patterns A, I and J each represents a group of households whose electricity use increases from 2 am and then declines before 6 am with possibly a particular work schedule in the early morning, consistent with the speculation that users get up in early morning and leave by 6 am for work. Pattern D shows the electricity use is relatively high and generally stable from midnight to 6 am. However, it still implies a slight decline after midnight and a slight increase after 5 am. It is intuitive that more households under those four patterns imply an earlier departure time from home, thus possibly leading an earlier congestion starting time.

2.4.2. Predictor performances with aggregate and disaggregate feature encodings

When assessing and comparing the overall performance of a predictor denoted by Root Mean Square Error (RMSE) and Mean Average Error (MAE), we use a two-level cross validation. All 79 weekdays are first divided into 3 folds (namely, 3-fold cross validation at the upper level). At each time, two of the three folds are picked out as the training data set, leaving the other fold as the testing data set. We apply 4-fold cross validation (namely the lower level cross validation) to the training data set to learn all parameters for the predictor (such as coefficients of features, and α of the LASSO model). Then the calibrated predictor is used to compute the RMSE or MAE on the testing data. The final RMSE or MAE are averaged from the upper level 3-fold cross validation. Fig. 6 shows the cross-validation performances of our LASSO predictors with electricity usage features (aggregate/disaggregate encodings), compared with two benchmark models: autoregressive–moving-average (ARMA) models using the same length of real-time traffic data and historical mean models that only use historical traffic data.

For predicting morning congestion starting time (CST), as shown in Fig. 6(a) and (b), using disaggregate features of electricity-use data offers a more accurate prediction than using aggregate features for 9 out of 15 TMCs. This is no surprise since it carries more detailed information regarding usage patterns. In most TMCs, using electricity data is far more advantageous than using traffic data only. This result implies that electricity usage pattern is spatially and temporally correlated with highway usage, and it is possible to predict morning congestion from electricity use during midnight and early morning. Clearly the electricity use of those households does not bring in useful information to predict traffic in those two TMCs, comparing to use historical average. For predicting morning congestion durations, as shown in Fig. 6(c) and (d), the predictors using aggregate features and disaggregate features, and historical means are used again. In addition, we create a new predictor that uses the predicted CST (from aggregated features) in addition to aggregate features. Generally, the predictor with disaggregated features has reasonable results and considerably outperforms other predictors in all TMCs except TMC 3. While in most TMCs using aggregate features with and without CST information are very close to the predictor using the historical mean. It seems that neither the aggregate features nor the CST carries useful information to explain the morning congestion duration.

Comparing to predicting CST, predicting the duration has higher RMSE. Clearly the congestion duration is more challenging to predict, given the households information from midnight to 6 am. This is not surprising. The electricity use data from midnight to 6 am can better explain the spatial and temporal distribution of travelers in the early morning than in the later morning. Congestion duration is likely to be affected by many factors other than the starting time of using highways, such as how long travel demand peaks and higher probability of incidents during morning congestion. Those add more complications to the prediction of duration than CST.

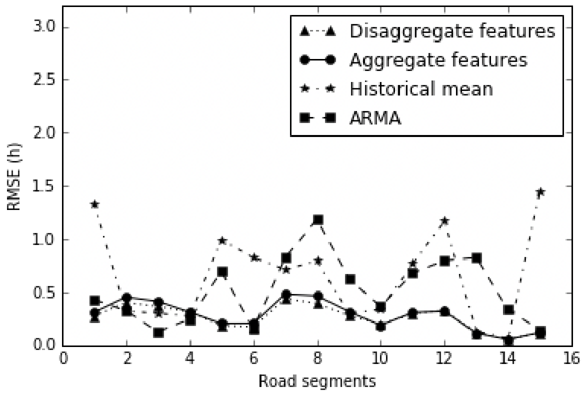


Fig. 6(a) RMSE for predicting congestion starting time.

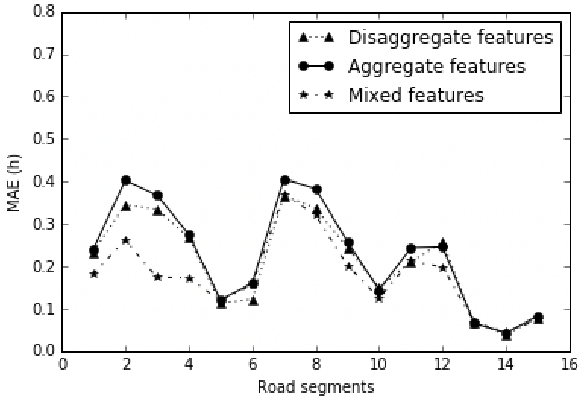


Fig. 6(b) MAE for predicting congestion starting time.

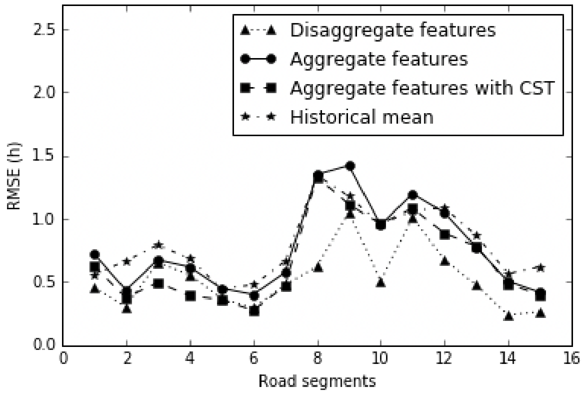


Fig. 6(c) RMSE for predicting congestion duration.

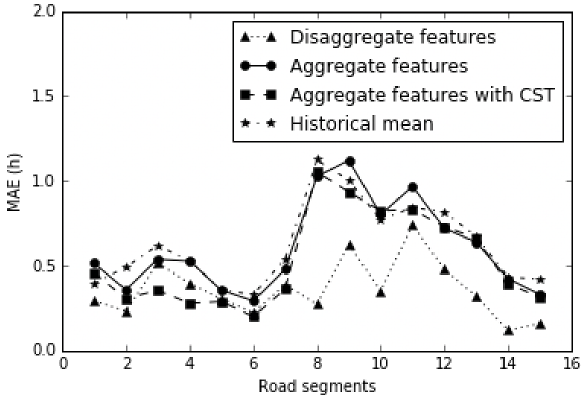


Fig. 6(d) MAE for predicting congestion duration.

3. Using geocoded social media data to improve morning congestion prediction

In this section, we predict the traffic congestion, including congestion starting time (CST_i^d), congestion duration (CD_i^d) and congestion occurrence (O_i^d) on a particular road segment in the network some time during morning periods, which are defined as from 5:00 AM to 10:59 AM, using traffic, social media and auxiliary features (e.g. weather, day-of-week, holiday) available until early morning, which is defined as before 5:00 AM on that day.

3.1. Data sources

This section uses probe-sourced traffic speed data from INRIX Traffic, tweet messages collected from the free Twitter Streaming and User Timeline API services and weather information scraped from Weather Underground for one year from January to December in 2014. The INRIX traffic data were reported every 5 minutes for 1,908 road segments georeferenced by Traffic Message Channel (TMC) code in Allegheny County, Pennsylvania, United States, as shown in Fig. 7(a). Several major US highways including I-376, I-279, I-597, PA-28, etc. are covered in the dataset. Each data record includes TMC code, time stamp, observed speed (mph), average speed (mph), reference speed (mph) and two parameters for the confidence of the speed, namely confidence score and confidence value. In this study, we only use TMC code, timestamps and observed speed fields of INRIX dataset.

We construct our Twitter streaming dataset by collecting all geo-coded tweets posted within the bounding box (-80.20, 40.29; -79.80, 40.62). 1,782,636 tweets from January 1, 2014 to December 31, 2014 in Allegheny County were collected, of which 672,527 (37.72%) tweets posted by 43,670 users are tagged with accurate locations. The Twitter data include date/time, text, user ID, language, latitude, longitude, user profile location, etc. In addition, 2014 US Census Tract Cartographic Boundary Shapefiles and Pittsburgh Zoning Map are used to spatially join geocoded tweets with the neighborhood and land-use information. The coverage of collected tweets is shown in Fig. 7(b). Weather Underground data were reported every one hour during 2014, with each report including time stamps, temperature, dew point, humidity, wind speed, precipitation, visibility, etc. Weekday/weekend and holiday information are from US Federal Holiday Calendar.

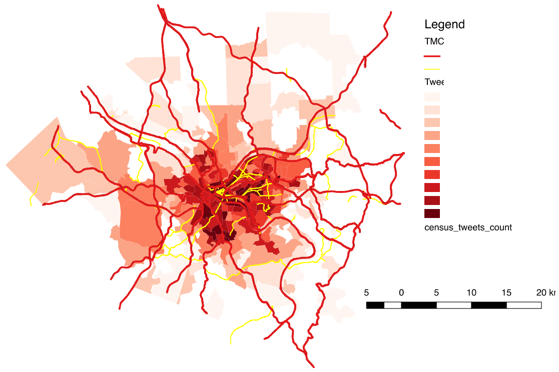


Fig. 7(a) TMC road segments used in this study.

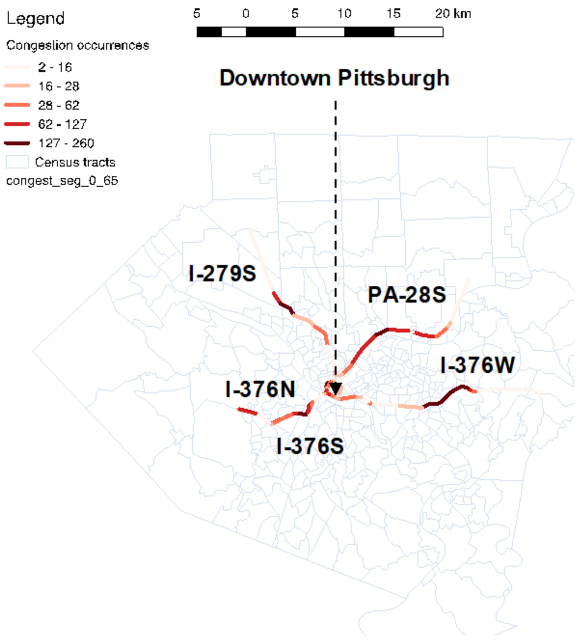


Fig. 7(b) Traffic and social media data coverage in Allegheny County.

3.2. Feature extraction

The workflow of our method consists of four steps: (1) the first identifies typical morning congestion patterns in urban transportation networks; (2) the second processes social media data and extracts spatiotemporal social media features; (3) the third examines the relationship between morning congestion patterns and social media features through regression analysis, and (4) the last makes use of such relationships to construct a predictive model for forecasting morning congestion in the network.

3.2.1. Characterization of morning road congestion

As illustrated in Fig. 8, this step first transforms raw INRIX traffic speed data into congestion rates, and computes congestion status, congestion starting time and duration for road segments during morning periods. Then, congested road segments with at least two congestion occurrences are selected. We ignore segments with only one congestion as they are most likely to be caused by unexpected traffic incidents.

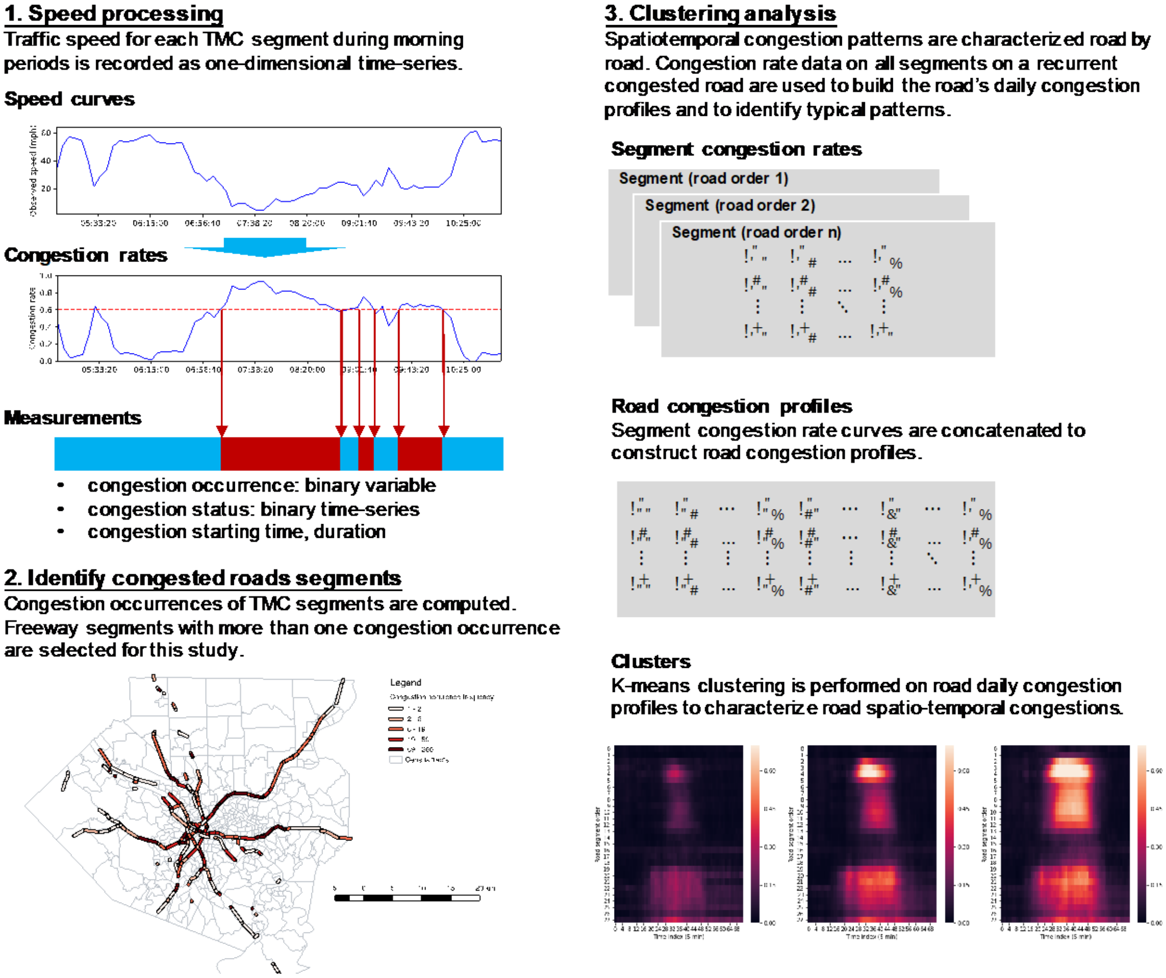

1. Speed processing
Traffic speed for each TMC segment during morning periods is recorded as one-dimensional time-series.
Speed curves
Observed speed (mph)
05:33:20
06:15:00
06:56:40
07:38:20
08:20:00
09:01:40
09:43:20
10:25:00
Congestion rates
Congestion rate
Measurements
congestion occurrence: binary variable
congestion status: binary time-series
congestion starting time, duration
2. Identify congested roads segments
Congestion occurrences of TMC segments are computed. Freeway segments with more than one congestion occurrence are selected for this study.
Legend
3. Clustering analysis
Spatiotemporal congestion patterns are characterized road by road. Congestion rate data on all segments on a recurrent congested road are used to build the road's daily congestion profiles and to identify typical patterns.
Segment congestion rates
Segment (road order 1)
Segment (road order 2)
Segment (road order n)
Road congestion profiles
Segment congestion rate curves are concatenated to construct road congestion profiles.
Clusters
K-means clustering is performed on road daily congestion profiles to characterize road spatio-temporal congestions.
Road segment order
Time index (5 min)

Fig. 8. Characterization of morning road congestion.

Clustering analysis using K-means is then conducted to identify typical daily morning congestion patterns for each road in the network. A daily road congestion profile for road R is a vector of $N \times T$ dimensions, where N denotes the number of congested segments on that road and T is the sampling points during morning periods. Optimal cluster size K is selected by GAP statistics.

3.2.2. Social media processing

This section presents a social media processing workflow to reduce data noise and augment the dataset by user timeline tweets. The flowchart is shown in Fig. 9. User timeline tweets are all tweet messages posted by a user in the past, which can be downloaded through Twitter User Timeline API. Timeline tweets do not contain accurate posting coordinates but the amount is 12.8 times larger on average for users in our dataset. By assuming that local residents stay at their homes every night, non-geocoded tweets can be used to track how people starting trips from 160 that area are active on the previous day and early morning. The difficulty lies in filtering local residents from a large number of noisy users (e.g. visitors, etc.), and to infer Twitter users' home locations.

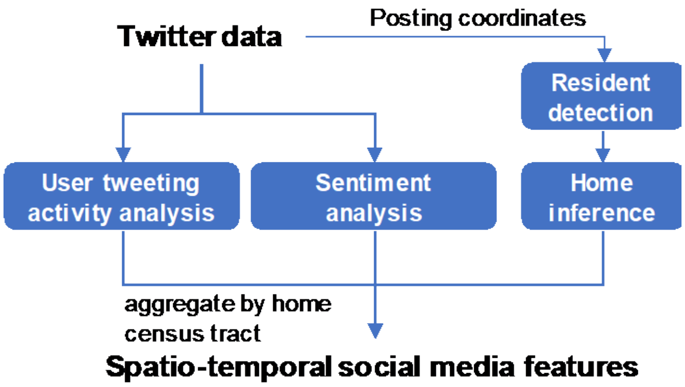

Twitter data
Posting coordinates
Resident detection
User tweeting activity analysis
Sentiment analysis
Home inference
aggregate by home census tract
Spatio-temporal social media features

Fig. 9. Social media processing flowchart.

3.2.2.1. Resident detection and home location inference

We make use of the self-reported account profile locations of Twitter users to identify local residents. For users who posted geocoded tweets within the bounding box, we select those who clearly declare their residence with place names in Allegheny County. A resident classifier is built using regular expressions that match local city names and nicknames (e.g. pittsburgh, pgh, da burgh, steel city, etc.), sports teams (steeler, etc.), zip codes (e.g. 15213, etc.), area code (e.g. 412), universities (e.g. cmu, chatham, etc.), townships, neighborhoods (e.g. shadyside, oakland, etc.) and coordinates within bounding box (e.g. 40.429, -79.932, etc.). Manual inspections are later conducted to check if local place names are matched.

A density-based algorithm called DBSCAN (Density-Based Spatial Clustering of Applications with Noise is applied to first find users' frequently-visited places using tweet check-in coordinates. A rule-based method is later applied to infer user home locations using place land-use variables, check-in ratios, night-time activities and home-related tweet features. As illustrated in Fig. 10(a).

The last step computes the exact coordinates for user home locations. As shown in Fig. 10(b), check-in points of an identified posting coordinate cluster may spread across multiple land use areas. Four-level weights are used to approximate the probability of a point in that area being a residential place, which include: Residence: 1.0; Mixed-use: 0.5; Education, Downtown: 0.2; Industry, Amenity: 0.0. Coordinates in an identified home cluster are then averaged by land-use weights to compute exact home locations. At last, 4,306 local residents in Allegheny County with home locations are identified. Home locations by census tract are visualized in Fig. 10(c).

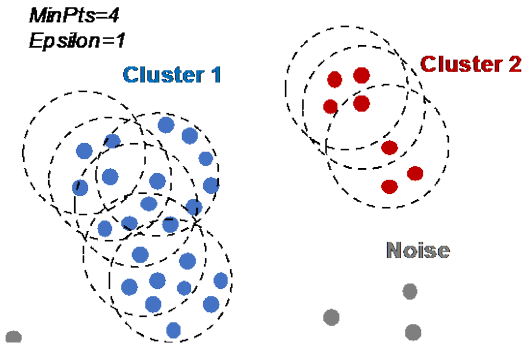


Fig. 10 (a). DBSCAN clustering process.

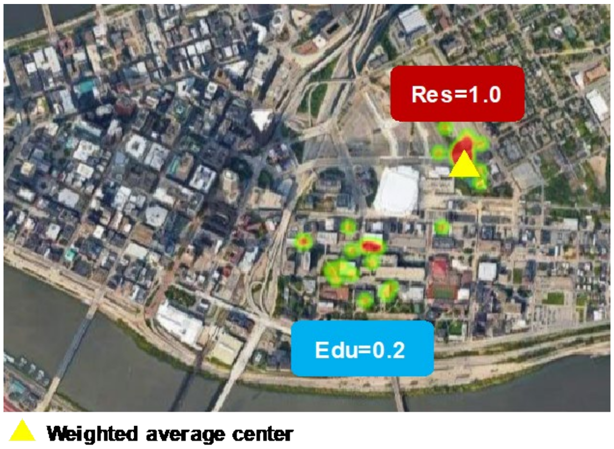


Fig. 10 (b). Home clusters and weighted average inference.

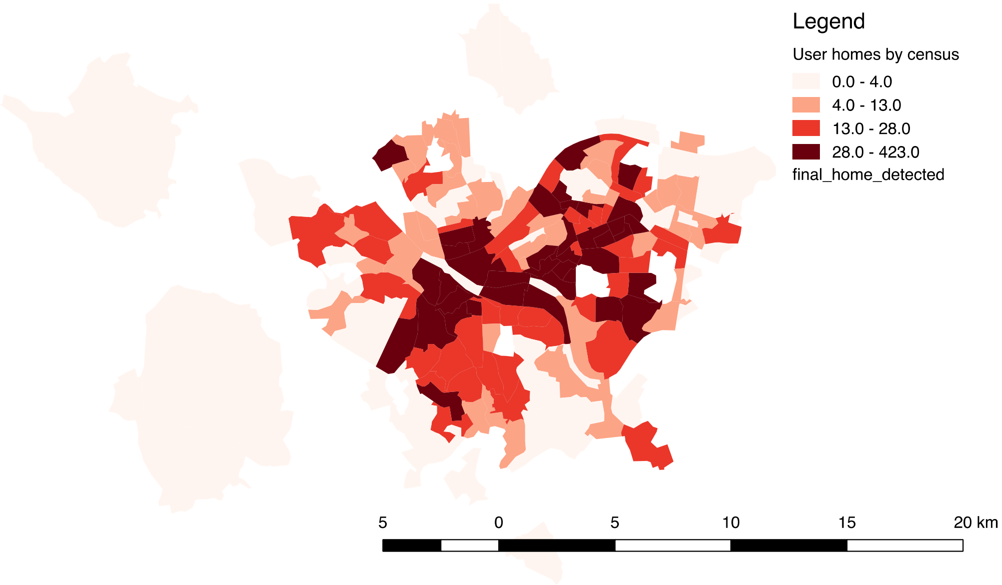


Fig. 10 (c). Visualization of identified home locations by census tract in Allegheny County.

3.2.2.2. User tweeting activity and sentiment feature encoding

We hypothesize that people's daily activities (e.g. asleep/awake time, night activities, etc.) in early morning and on last day have correlations with road congestion in morning peak hours. We use user tweeting activities to probe people's activities. A user i's activity profile on day d is characterized by a N-dimension vector $\mathrm{A_i^d} = [\mathrm{a_{i0}^d}, \mathrm{a_{i1}^d}, \dots, \mathrm{a_{iN}^d}]$, where $\mathrm{a_{it}^d}$ is the user i tweeting counts during time interval $\mathrm{t} - 1$ and t of day d. Note that a day starts from the morning of day d (e.g. 5 AM) to the early morning of day d + 1. we aggregate user activities by home census tract to extract normalized spatiotemporal feature vectors. $\mathrm{PAct_c^d}$ describes how people starting trips from census c on day d are active on previous day $\mathrm{d} - 1$ and on the early morning of d.

Abnormal tweeting trends by day are used to reflect urban activity trend. We encode the total tweet counts on day d as a feature Act^d. The tweet sentiment analysis a bidirectional-LSTM neural network with soft attention. We fit the model with Stanford's Sentiment140 dataset, which has 1,600,000 labeled tweets for positive or negative sentiment. With 220 a hold-out validation set of 80,000 tweets, the model is trained with the remaining samples using the early-stopping criterion, which stops the training iterations when accuracy observed on validation set starts to decrease. The final model achieved an accuracy of 86.7% for classifying tweet sentiment on the validation set. We use the last sigmoid layer output, i.e, the probability of the tweet being positive $\mathrm{p^{pos}}$, to compute the sentiment score for our study. The sentiment score of a tweet is defined as $2\mathrm{p^{pos}} - 1$. The score ranges from -1 to +1, where -1 indicates very negative 225 while +1 indicates very positive sentiments. The daily urban sentiment $Sent^d$ is the average of all tweet sentiment on the day d.

3.2.2.2. Auxiliary feature encoding

Besides social media features, weather, weekday/weekend, month-of-year, and holiday information are used as control variables. Weather variables are apparent temperature T_{AT}^d and precipitation status P^d observed on the early morning of the day before congestion prediction time. Apparent temperature is computed by combinations of Heat Index (HI), which measures "how hot it really feels when relative humidity is factored in with the actual air temperature", and Wind Chill Temperature (WC), which measures "the lowering of body temperature due to the passing-flow of lower-temperature air". HI and WC are calculated using Meteorological Calculator provided on National Weather Service. We apply these two measures as apparent temperatures if their conditions can be met. Otherwise, we use air temperature directly. Precipitation status is a binary variable indicating if the pavement condition is wet. Weekend/weekday variable W^d.indicates if day d is during weekends and holiday variable H^dindicates if day d is an official national holiday. Finally, explanatory variable vector x^d can be defined as the concatenation vector $[\mathrm{PAct_c^d}, Act^d, Sent^d, W^d, H^d]$. We normalize each variable to have "zero mean and unit variance."

3.3. Model construction

Two kinds of models are constructed for this study. The first is an examination model to explain how social media precursors affect the morning road congestion patterns. We use an ordered logit regression model to correlate congestion cluster index with social media features, while controlling the effects of weather, weekday/weekend, month-of-year, and holiday information. The second is a segment-based stacked linear predictors (logistic regression + LASSO) for predicting the congestion occurrence, congestion starting time and congestion duration on each day.

3.3.1. Examination model

Because of traffic propagation effects, the identified road congestion clusters generally show ordered spatiotemporal patterns. Ordered logit model is performed to evaluate the impacts of social media on morning road congestion scales, while controlling for the effects of weather, weekday/weekend, month and holiday effects. Y^d is an ordered categorical variable indicating the congestion cluster observed on day d. x^d is a vector of explanatory variables observed on previous day $d-1$ and early morning of day d before congestion prediction time. θ_c is the learned thresholds for classifying cluster level c and β^T are the variable weights.

$$P(Y^d \leq c) = \sigma(\theta_c - \beta^T x^d + \epsilon^d) \tag{6}$$

Due to high dimensions and co-linearity of the extracted features, recursive feature elimination (RFE) and L2-norm regularization are applied on model coefficients to remove irrelevant spatiotemporal variables and to learn stable relationships. If a road has non-ordinal congestion clusters, we split its road segments until each sub-road only has ordered congestion patterns.

3.3.2. Prediction model

We then build l1-regularized linear congestion predictors for each segment on the road, with the learned β^T in examination model added to each predictor feature set to push it to select features that explain within-cluster variances. As shown in Fig. 11, predictors for each road segment form a pipeline that consists of three linear models: binary logistic regression classifier is trained to predict if congestion O_i^d will occur on the segment (Eq. 7); Lasso1 (Eq. 8) and Lasso2 (Eq. 9) are both linear regressors trained on days when congestion occurs, to respectively predict congestion starting time and duration. During the prediction phase, logistic regression is first used to classify congested days. Only if the segment is predicted to be congested, Lasso1 and Lasso2 will be performed to predict congestion starting time CST_i^d and duration CD_i^d. Otherwise, the congestion starting time is regarded as being indefinitely late, thus predicted as the ending point of defined morning periods and congestion duration is 0.

$$\min_{\beta_i} -\sum_{d} \log P(C_i^d | x^d; \beta_i, \beta) + \alpha_i ||\beta_i||_1 \tag{7}$$

$$\min_{w_i} ||CST_i - w_i^T [x^d, \beta^T x^d]||_2 + \alpha_i ||w_i||_1 \tag{8}$$

$$\min_{\gamma_i} ||CST_i - \gamma_i^T [x^d, \beta^T x^d]||_2 + \alpha_i ||\gamma_i||_1 \tag{9}$$

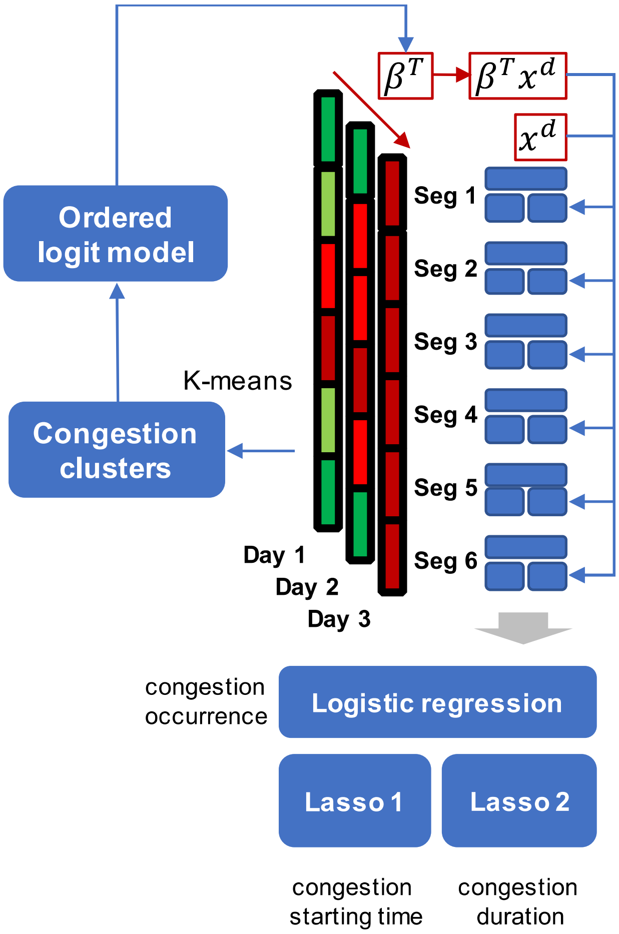


Fig. 11. The proposed predictive model in this research.

3.4. Results and discussion

3.4.1. Examination of relationship between social media and morning congestion patterns

3.4.1.1. Cluster analysis

K-means clustering is performed for all road congestion profiles to find spatiotemporal congestion patterns of that road. Most roads have ordered congestion clusters that reflect the overall congestion scales when K reaches the optimal GAP statistics, except for PA-28S, which has two groups of clusters representing different congestion locations. As shown in Fig. 12, we split the road into two sub-roads by the transition segment and perform clustering on two sub-roads separately to find ordinal clusters. We manually order the clusters according to congestion coverage and duration. The ordered clusters are summarized in Fig. 12. The congestion propagation dynamics have been observed in the results, where the early congestion of downstream segments can cause congestion of upstream segments with time lags.

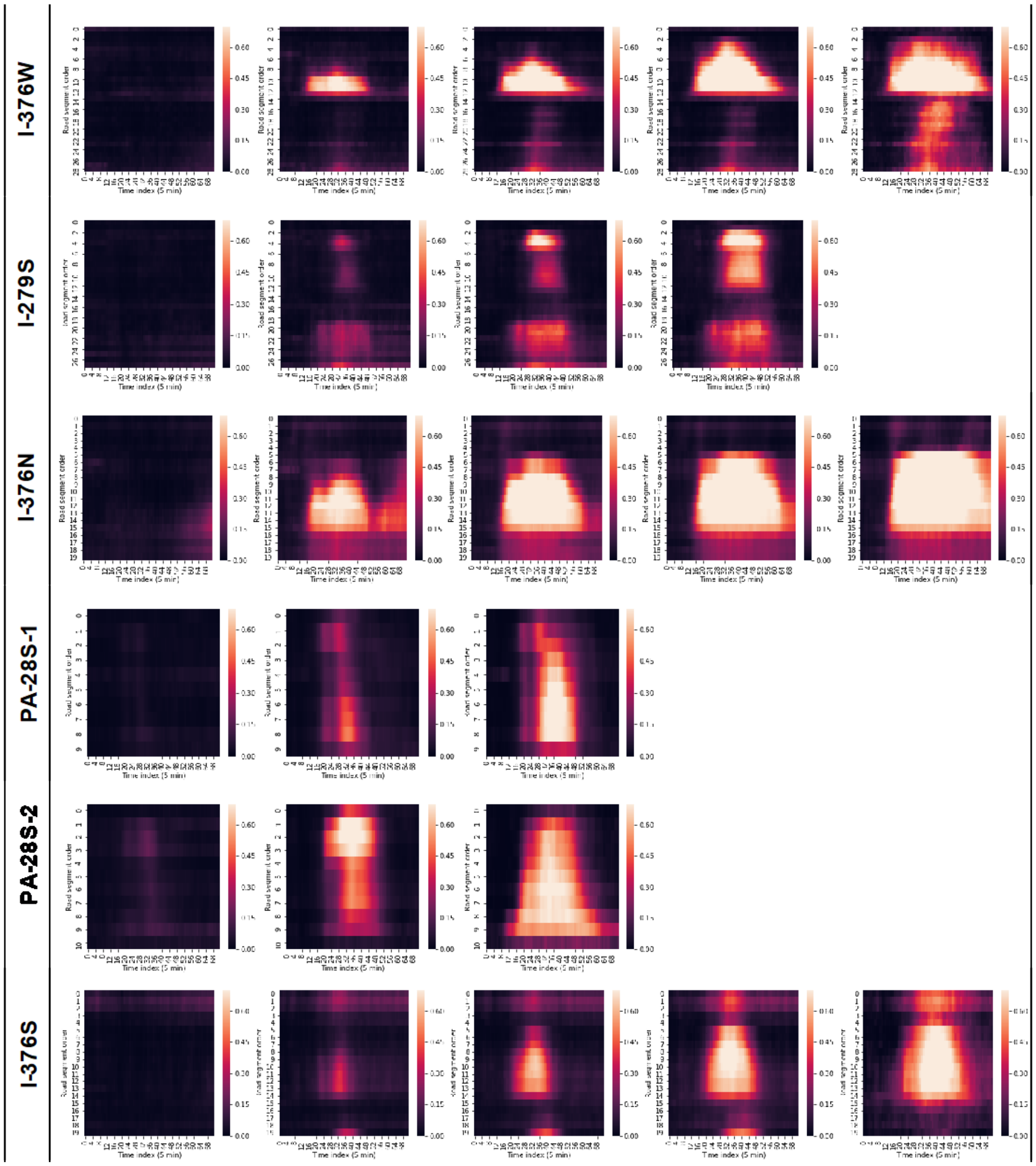


Fig. 12. Morning congestion clustering results.

3.4.2.1. Regression results

Ordered logistic regression with l2 regularity $\alpha = 100$ is performed to associate ordered congestion cluster index of the road with social media features and control variables. To deal with the issue of imbalanced class, the data sets were over-sampled from the 2014 data. Over-sampling over-selects the minority class in order to achieve balanced training. We define five periods for visualization: last morning (LM: 06:00AM - 09:59AM), last daytime (LA: 10:00AM -05:59 PM), last evening (LE: 06:00PM - 11:59PM), night (NT: 0:00AM - 03:59AM) and early morning (EM: 04:00AM - 05:59AM). User tweeting activities in four of the five periods, including LM, LE, NT, and EM, are used as features to account for congestion scales. The model coefficients are shown in Fig. 13 with variables ordered from left to right by the time of day. The resulting relationships are surprisingly simple and powerful. Generally, we discover that the earlier people go to sleep, the more congested the roads will be in the next morning. The early-sleeping patterns are represented by high tweeting activities in last evening (LE) together with low tweeting activities at night (NT), which results in the early activities drop in selected spatial areas. In addition, people's tweeting activities in the early morning (EM) is positively associated with morning congestion scales, which is intuitive as the earlier people get up, the high chances they will commute and contribute to morning congestion. People's activities in last morning (LM) periods are also positively-correlated but the relationship is not consistently observed on PA-28S-2 and I-376S. It is expected that weekends, holidays variables are negatively correlated with morning congestion scales as the morning commuting demands are decreased on these days. Weather variables, including precipitation status and apparent temperature, are positively correlated with congestion scales. However, the effects of overall tweeting activities of the last day and sentiment are not significant.

3.4.2. Prediction performances

With forecasting horizons set as six hours, we use tweet or traffic data from the last day until 5 AM to predict congestion between 5:00 AM and 10:59 AM in the morning. The prediction errors by segment are shown in Fig. 14. The results suggest that our proposed method (Logit+Lasso+Olm) outperforms other benchmarks in terms of predictions for all four congestion measurements. Note that the Root Mean Square Error (RMSE) of congestion start time and durations are a lot higher than Mean Absolute Error (MAE) because the prediction errors on days when congestion occurrences are incorrectly predicted are much higher than normal errors and RMSE tends to enlarge the effects of big errors. Therefore, MAE is considered as a more reasonable metric to reflect congestion starting time or duration errors in reality. Our method is capable of forecasting morning congestion occurrences with high precision and recall (F1=0.83), with congestion starting time predicted with an average error of 21 min and duration predicted with an average error of 10 min across all segments. The predictions accuracy of binary congestion status time-series is relatively low (F1=0.69) because our method is unable to predict very short congestion (<10min) and multi-period congestion. However, our method still outperforms other benchmarks in terms of this congestion measurement.

When further inspecting prediction errors by segment in Fig. 14. we observe that the prediction performances of different methods are correlated, which is reasonable because if congestion on some road segments has low variances, all prediction methods are supposed to have high prediction performances.

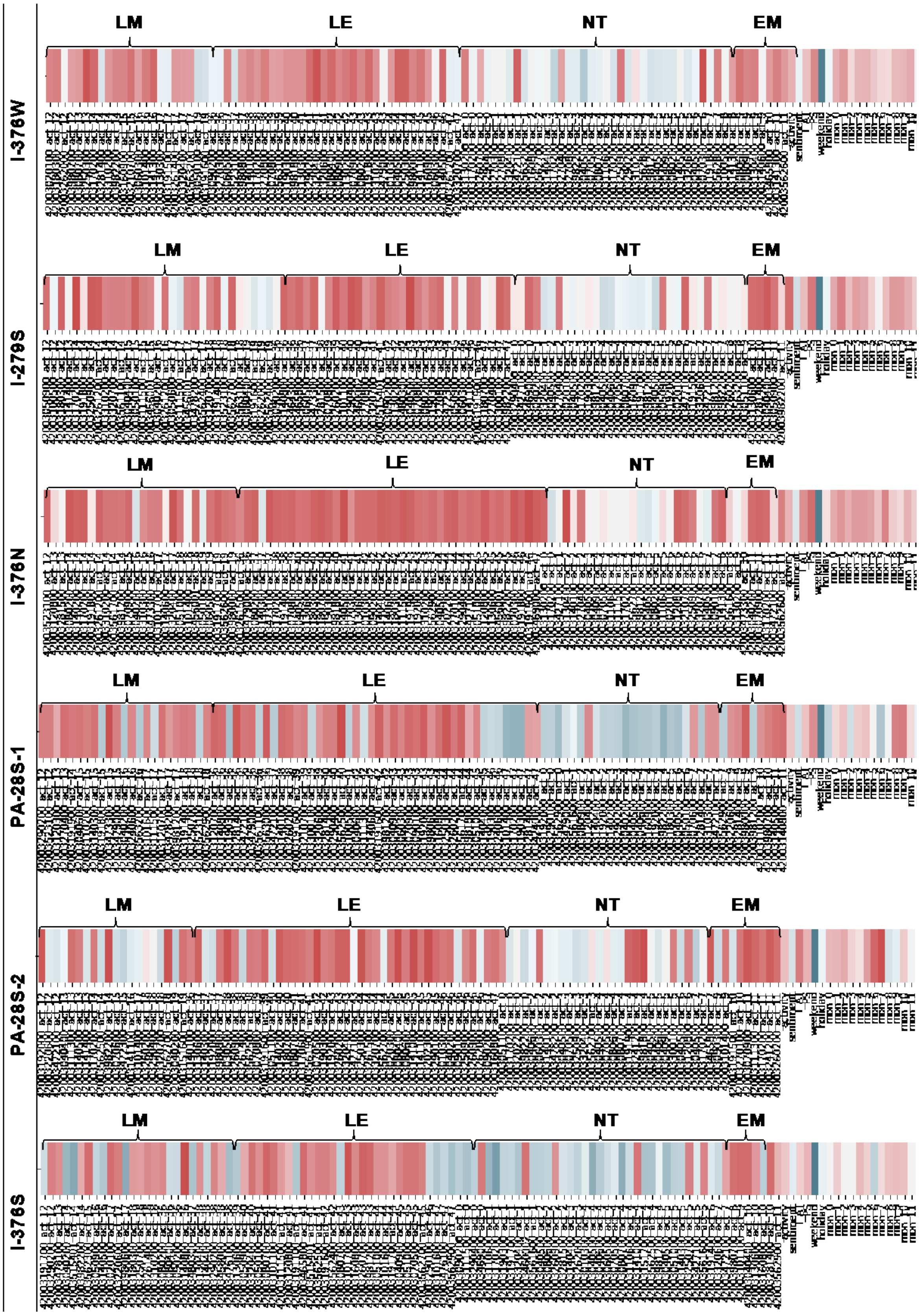


Fig. 13. Regression coefficients

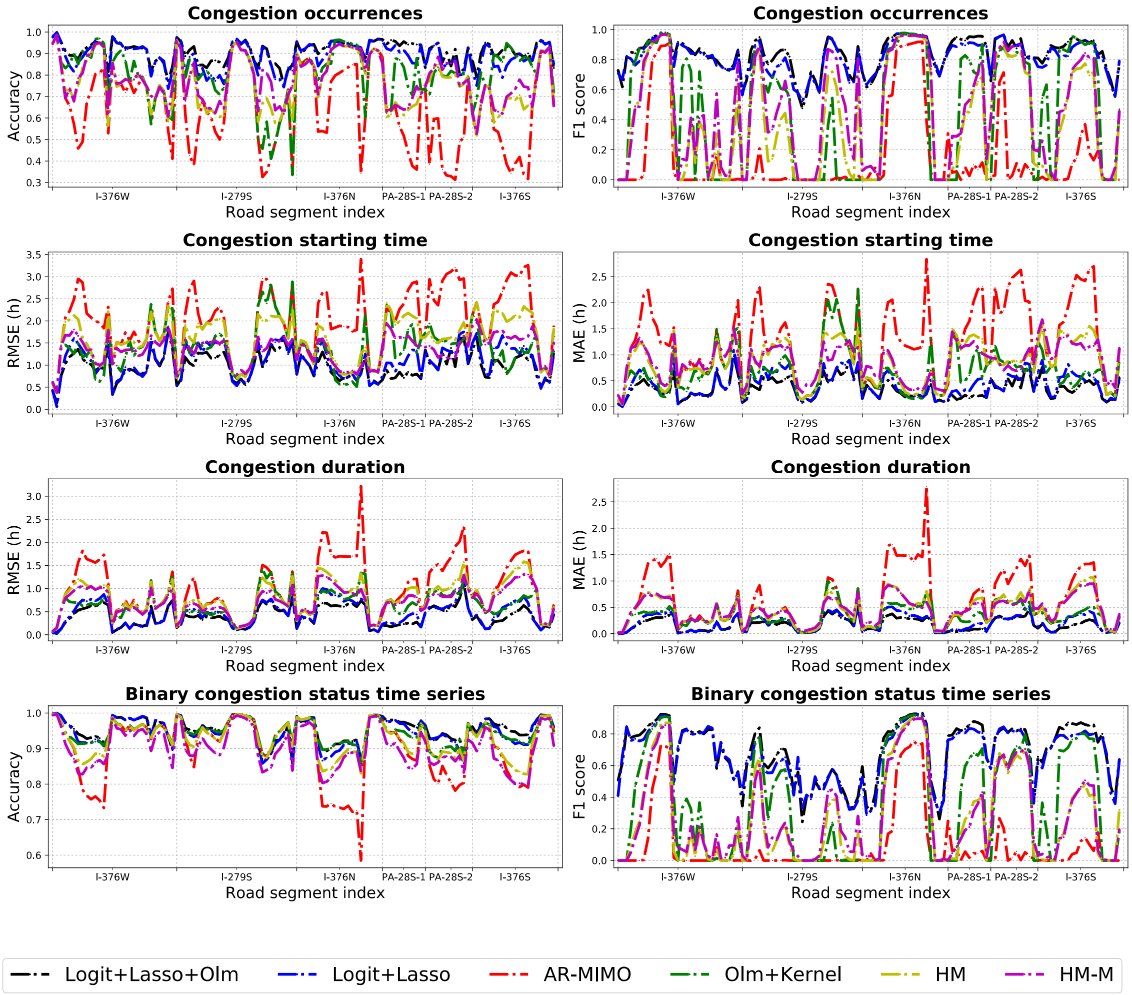


Fig. 14. Predictor performances

The performances of all six methods can be categorized into 3 groups. The first group includes "Logit+Lasso+Olm" and "Logit+Lasso". These two methods both use social media features and build models for each segment. Their performances are generally the best among all methods. "Logit+Lasso+Olm" outperforms "Logit+Lasso" on some segments in the middle of roads. The second group is "Olm+Kernel" method. This method performs very unstable among the segments, with the prediction errors on some segments in the middle of the road even lower than "Logit+Lasso+Olm", but also with errors of segments on both ends of the road higher than the worse method "AR". In general, the model performance is in the middle of the three groups and has a better ability to predict congestion occurrences and binary congestion state time-series, compared with its other predictions. The third group contains historical average methods ("HM", "HM-M") and autoregressive methods "AR-MIMO". Not surprisingly, these three methods perform badly in terms of all congestion measures because historical methods can't capture day-to-day variances and autoregressive methods can't extract useful traffic dynamics from early morning traffic. Autoregressive method "AR-MIMO" is the worst among all methods, with low chances of correctly predicting congestion occurrence (F1=0.148) and high errors of predicting congestion starting time (MAE=1.282 h) and duration (MAE=0.636

h). The results make it an inappropriate method for predicting morning congestion with long forecasting horizons, such as using data before 5 AM.

In summary, the methods that make use of social media data perform better than methods that make use of traffic data for morning congestion predictions. The methods that build models for each segment outperforms methods that build models for the whole road.

4. Conclusions

This project proposes a general framework to explore the spatial and temporal correlation among usage patterns of energy systems and social media platform with roadway systems, and make use of such relationships to increase the forecasting accuracy and horizons of morning congestion predictions, which has huge practical values for helping travels' plan travel choices and supporting active traffic control.

For using energy usage data for morning congestion prediction, we propose a methodology along with data analytics for the City of Austin to predict morning congestion starting time and duration using the time-of-day electricity use data from 322 anonymous households with no spatial location information. The results are very compelling and encouraging. We show that using sampled household-level electricity data from midnight to early morning, even from midnight to 2 am, can reliably predict congestion starting time (CST) of many highway segments that are otherwise hard to predict using only real-time travel time data (through time series or the historical mean).

For using social media data for morning congestion prediction, the resulting relationships are surprisingly simple and powerful. Generally, we discover that the earlier people go to sleep, which can be sensed by social media platforms, the more congested roads will be in the next morning. The early-sleeping patterns are represented by high tweeting activities last evening together with low tweeting activities at night, which results in the early activities drop in selected spatial areas. In addition, people's tweeting activities in early morning are positively associated with morning congestion scales, which is intuitive as the earlier people get up, the high chances they will commute and contribute to morning congestion. These relationships are powerful because most tweeting activity features (last evening, night, etc.) are readily available many hours ahead of the next morning and therefore can be used to improve long-term morning congestion predictions. A predictive pipeline consisting of an ordered logit model, a logistic regression model, and two Lasso models have been presented to predict multiple congestion measurements including morning congestion occurrences, congestion starting time and duration, and binary congestion status time series. The results show that our method is capable of forecasting morning congestion occurrences with high precision and recall (F1=0.83), with congestion starting time predicted with an average error of 21 min and duration predicted with an average error of 10 min across all segments.

This project demonstrates the potential of cross-system prediction and control as a proof of concept, which hopefully will open the door for future research along this arena. As a first attempt, this project explores the relationship between energy system and social media platform usage and highway system usage, and proposes a general data analytics and prediction framework that can potentially work for any pair of cyber or physical systems. Transportation system use may be revealed partially ahead of time by monitoring systems other than the energy system and social media, such as water/sewer system. We plan to future extend our framework to include those systems in the near future.

Note: this project is also partially supported by an NSF grant. The NSF grant focuses on developing a general framework for mining multi-source data to predict traffic in advance, while this project focuses on real-world technology deployment test in both Austin and Pittsburgh.