Modeling Traffic Congestion and an Examination of the Factors Influencing Chicago Public Transit Ridership in the City of Chicago

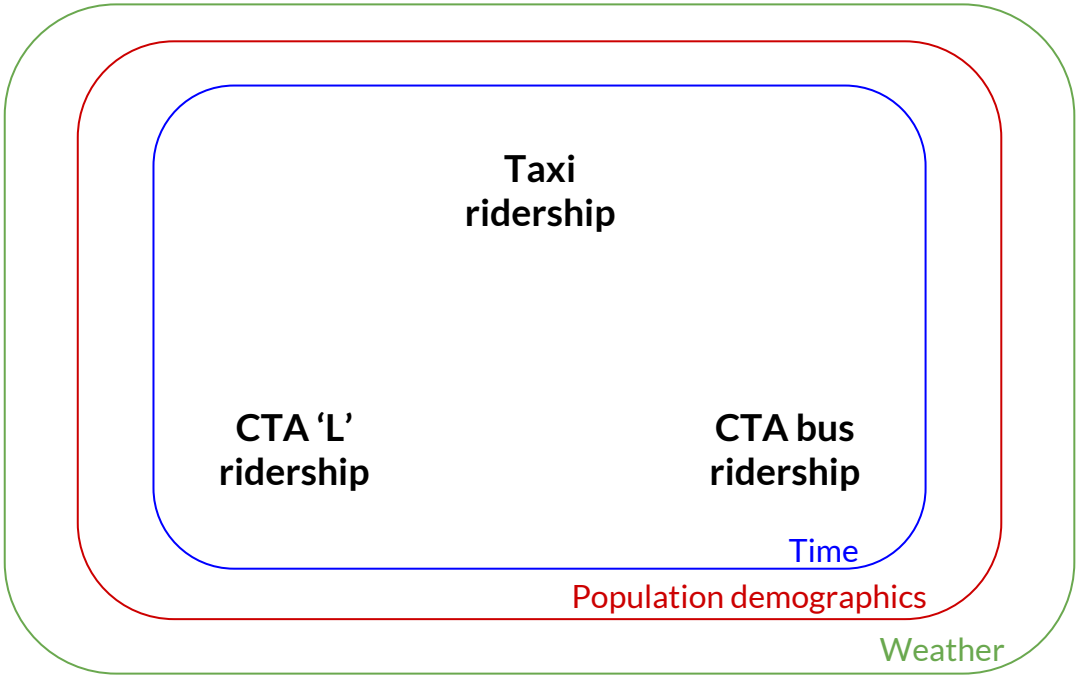
CSP 571, Spring 2018

Elliot Chibe
Andrew Hile
Linxi Li

Max Oellien
Calin Segarceanu
Christopher Ver Hoef

ILLINOIS INSTITUTE OF TECHNOLOGY

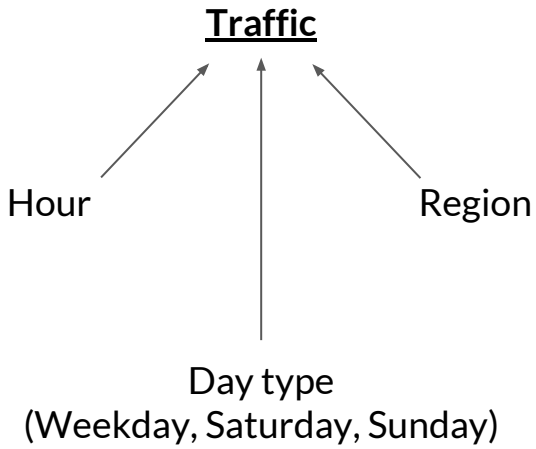
Objective: Explore public transportation behavior



Deliverable: Explore the factors influencing the ridership of each mode of Chicago public transportation

Objective: Model Chicago traffic congestion

Deliverable: Predict traffic congestion at a given time and location



Models chosen for testing:

- Simple linear regression
- Linear regression including second and third-order interaction terms
- kNN regression (resampled using repeated 10-fold cross-validation)

Validation: Final model applied to 12 days' worth of out-of-sample data.

Project management

Task delegation

Inferential	Taxi	Weather	Census	'L'	Traffic	Bus
Predictive	LR: training	LR: test	LR+: training	LR+: test	kNN: training	kNN: test

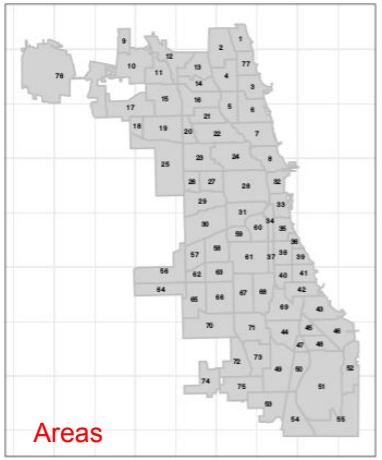
Timeline

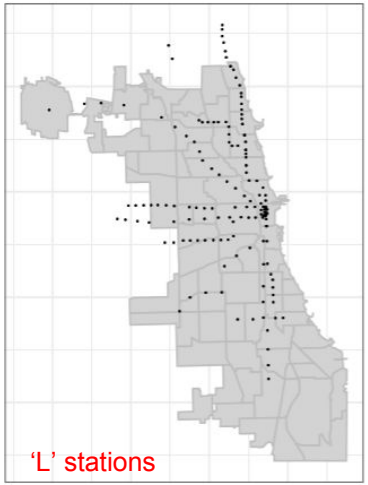
Study obstacles

Issue	Implications
Limitation: data for potentially relevant factors is unavailable • Ridesharing and Metra ridership • Geographical factors (e.g. number of intersections, total road length) • Public transit vehicle volume and timetable	Unable to account for these factors
Limitation: Traffic data encompassed the shortest timeframe (01/2013 - 01/2015) due to limited data collection	Investigation scope narrowed to the traffic data timeframe
Limitation: Valid timeframe for modeling further limited due to traffic data observation recording issues (01/2015)	• Model scope narrowed to the valid timeframe • Separate validation set scraped 4/10-4/22 • Unable to include weather in the model
Assumption: 2010 census results are representative of the study timeframe	Normalized responses by population counts that were not current

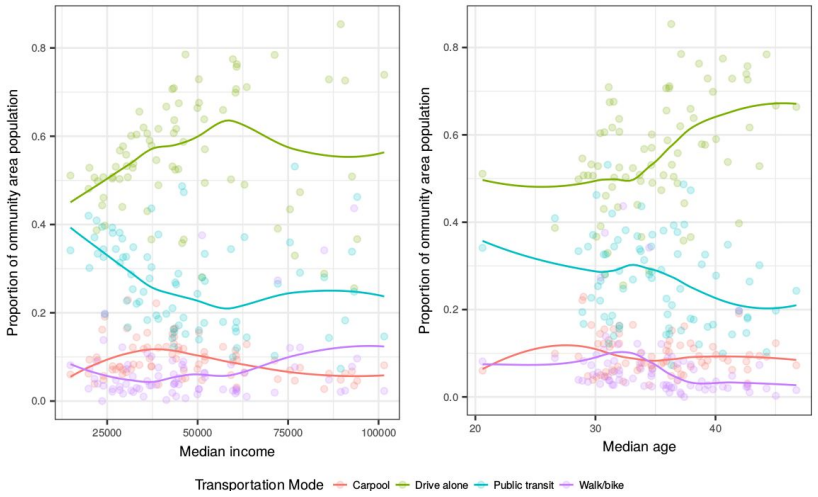
Data pipeline challenges

Issue	Solution
Location convention	Utilized *rgeos*, *rgdal*, *sp* R packages to align all datasets to community area
Taxi dataset size (~5GB/113M rows + 23 columns)	Aggregated taxi rides by day via Java
Time convention • Traffic: timestamp • Others: day	Aggregated traffic congestion by day via R





Public transportation factors: demographic

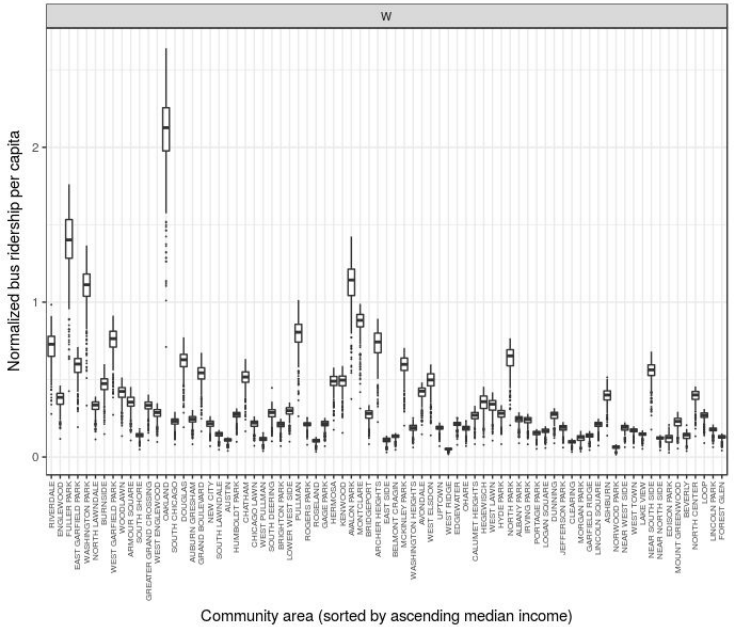


	Drive alone	Carpool	Public transit	Walk/bike
Median income ↑	↑	~	↓	~
Median age ↑	↑	~	↓	↓

Public transportation factors: demographic

Validating census observations on public transit ridership data:

- Areas with low median age did not exhibit lower ‘L’ or bus ridership per capita compared to areas with high median age
- Areas with low median income only showed the expected behavior when using bus ridership per capita as a response

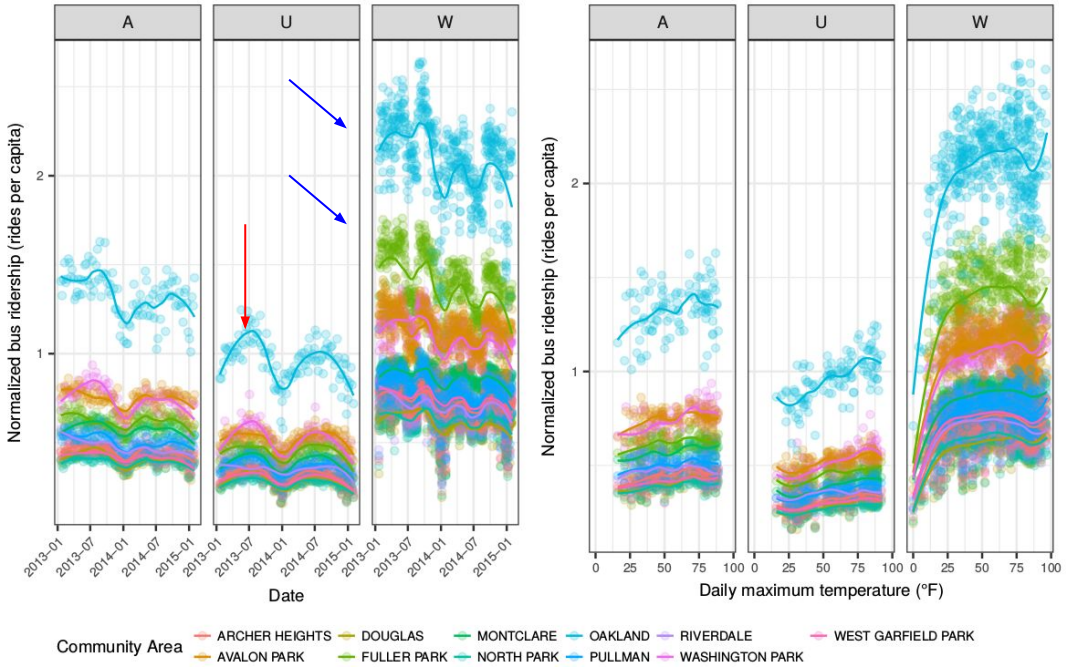


Public transportation factors: time/weather

Bus and 'L' ridership appears cyclic over the course of a year

- Case 1: sinusoidal
  - More typical/ of weekend ridership
  - Aligns with ridership trend with respect to temperature
- Case 2:
  - Jan - Jun: Relatively flat
    Jun - Aug: Step function down
    Aug - Nov: Step function up, potentially beyond initial baseline
    Nov - Dec: Steep decline
  - Likely driven by time-related factors (e.g. academic breaks, seasonal vacations)

Taxi ridership did not exhibit similar behavior



Predictive model results

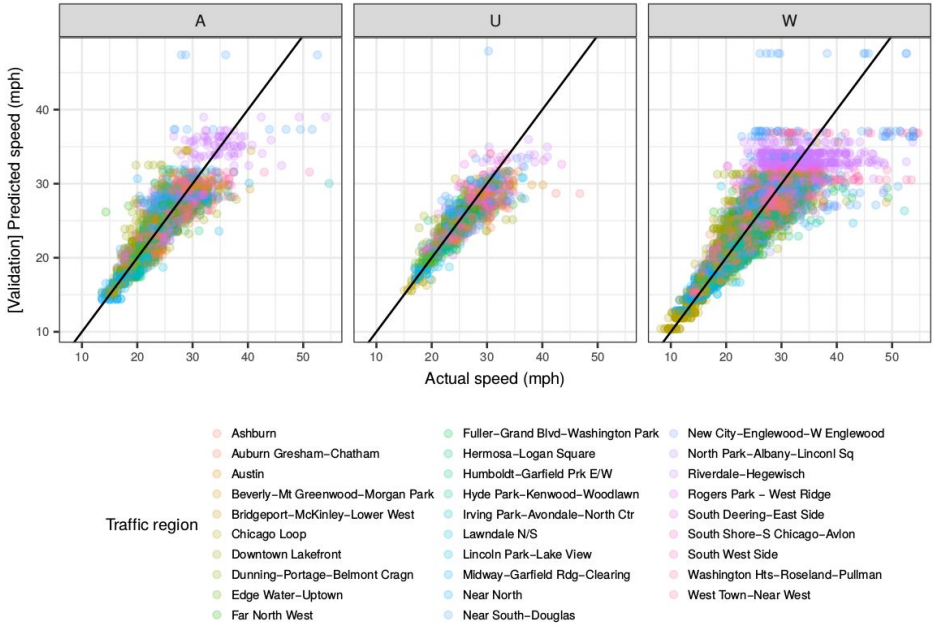
Model quality

- LR+ offers marginal fit improvement over LR at the cost of increasing DoF from 31 to 173.
- kNN improves training R^2 by **46%** and RMSE by 30% over LR! (...but is **SLOW**)
- Optimal $k = 9$

kNN residual analysis

- Model prediction accuracy and precision improves during periods of heavy congestion.
- Weekend observations exhibit this effects to a much lower degree.

Model	Training R^2	Training RMSE	Test RMSE	Validation RMSE
Linear regression	0.5245	3.348	3.2830	
Linear regression with interaction effects	0.542	3.288	3.2248	
kNN regression	0.7663	2.3465	2.2525	2.2129



Executive summary

Deliverable: Explore the factors influencing each mode of Chicago public transportation (taxi, CTA bus, and 'L')

Method: Qualitative analysis of responses with respect to meteorological, demographic and temporal factors

Findings:

- Distribution modality of all responses was defined by day of week and community area.
- Income and age connected to area-to-area distribution bias.
- Area distribution variation governed by time-related factors, with some influence by temperature.

Opportunities:

- Examine similar locations (e.g. NYC) that have published ridesharing data to use as an analog to Chicago
- Add bike-share (Divvy) data to investigation scope

Deliverable: Predict traffic congestion at a given time and location

Method: Prototype parametric (linear regression variants) and non-parametric (kNN regression) models

Results:

- kNN regression yielded the best fit quality, which was sustained in out-of-sample validation.
- All models tended to overestimate congestion when low.
- Model predictions were more precise during periods of heavy congestion.

Opportunities:

- Scrape current data for at least one year in order to enable weather to be added as a factor (current scope: one month)
- Change location resolution to the traffic segment level