MIT Supply Chain
MANAGEMENT

Using smartphone data to predict Beijing bike-sharing demand

by Terri Fu, Leonhard Fricke

Supply Chain Management Capstone Project
Advisor: Dr. Inma Borrella

MIT Supply Chain
MANAGEMENT

Leonhard Fricke

Terry Liu

Dr. Inma Borrella

Agenda

MIT Supply Chain MANAGEMENT

1 **Research Motivation and Methodology**

2 Drivers of Bike-sharing Demand

3 Forecasting Models

4 Taking insights into practice

The bicycle is a simple solution to some of the world's most complicated problems.

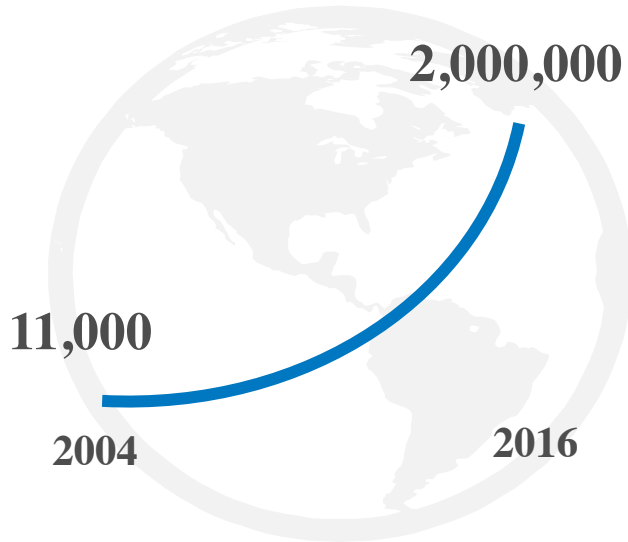
Bike-sharing as new transportation mode

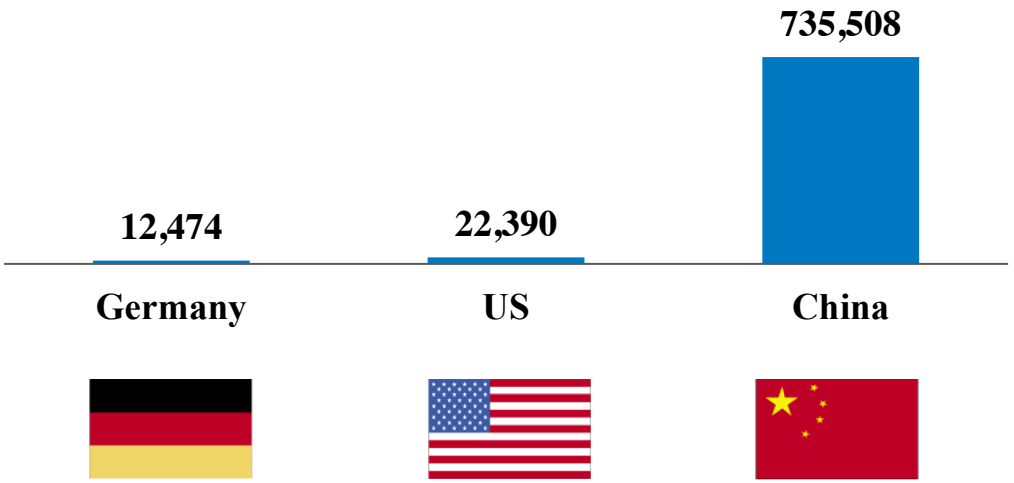
Free-floating

Sign Up **Unlock** **Ride!** **Lock**

Bike-sharing Research Motivation

of shared bicycles





https://www.ft.com/content/5efe95f6-0aeb-11e7-97d1-5e720a26771b

Bike-sharing Congestion

https://www.theatlantic.com/photo/2018/03/bike-share-oversupply-in-china-huge-piles-of-abandoned-and-broken-bicycles/556268/

Bike-sharing is driven by different factors

MIT Supply Chain MANAGEMENT

Factors related to bicycle usage

Individual characteristics

- Gender, Age, Income, Employment, Education, etc.

Societal

- Social norms, safety, law restrictions, etc.

Infrastructure

- Urban form, Hilliness, cycling lanes, parking, etc.

Environmental

- Temperature, pollution, wind speed, etc.

DeMaio 2009: 'Bike-sharing : History , Impacts , Models of Provision , and Future'
Shaheen, Guzman, and Zhang 2010: 'Bikesharing in Europe, the Americas, and Asia: Past, Present, and Future'
Fishman, Washington, and Haworth 2013: 'Bike share: a synthesis of the literature'

We tested certain drivers of Bike-sharing Demand

MIT Supply Chain MANAGEMENT

Time and Day

- Time of the day
- Day of the week
- Weekend
- Public Holiday

Environmental

- Temperature
- Humidity
- Pressure
- Wind speed
- Air Quality

The Case of Smartphone Data

MIT Supply Chain MANAGEMENT

TalkingData

7+ billion mobile devices

Online Resources

WEATHER UNDERGROUND

Google Calendar

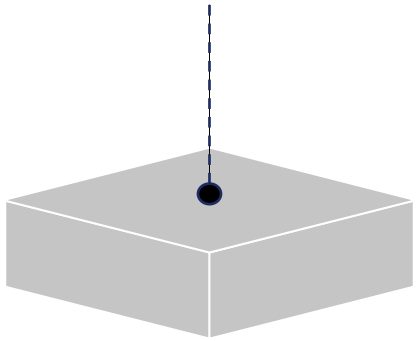
U.S. Department of State

1 Month Data Sample for Analysis

From Exploration to Forecasting Methods

1

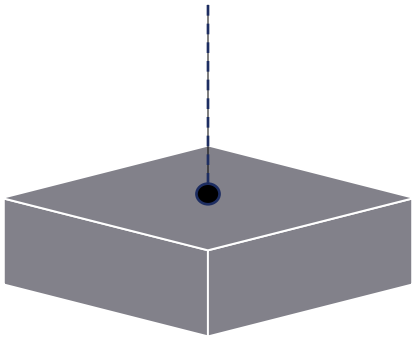
Exploratory Analysis



- Visualization
- Descriptive Statistics
- Stakeholder Interviews

2

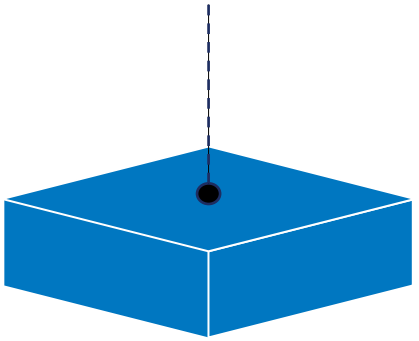
Identify demand drivers



- Polynomial Regression

3

Forecasting Models



- Linear Regression
- Neural Network
- Random Forest

Agenda

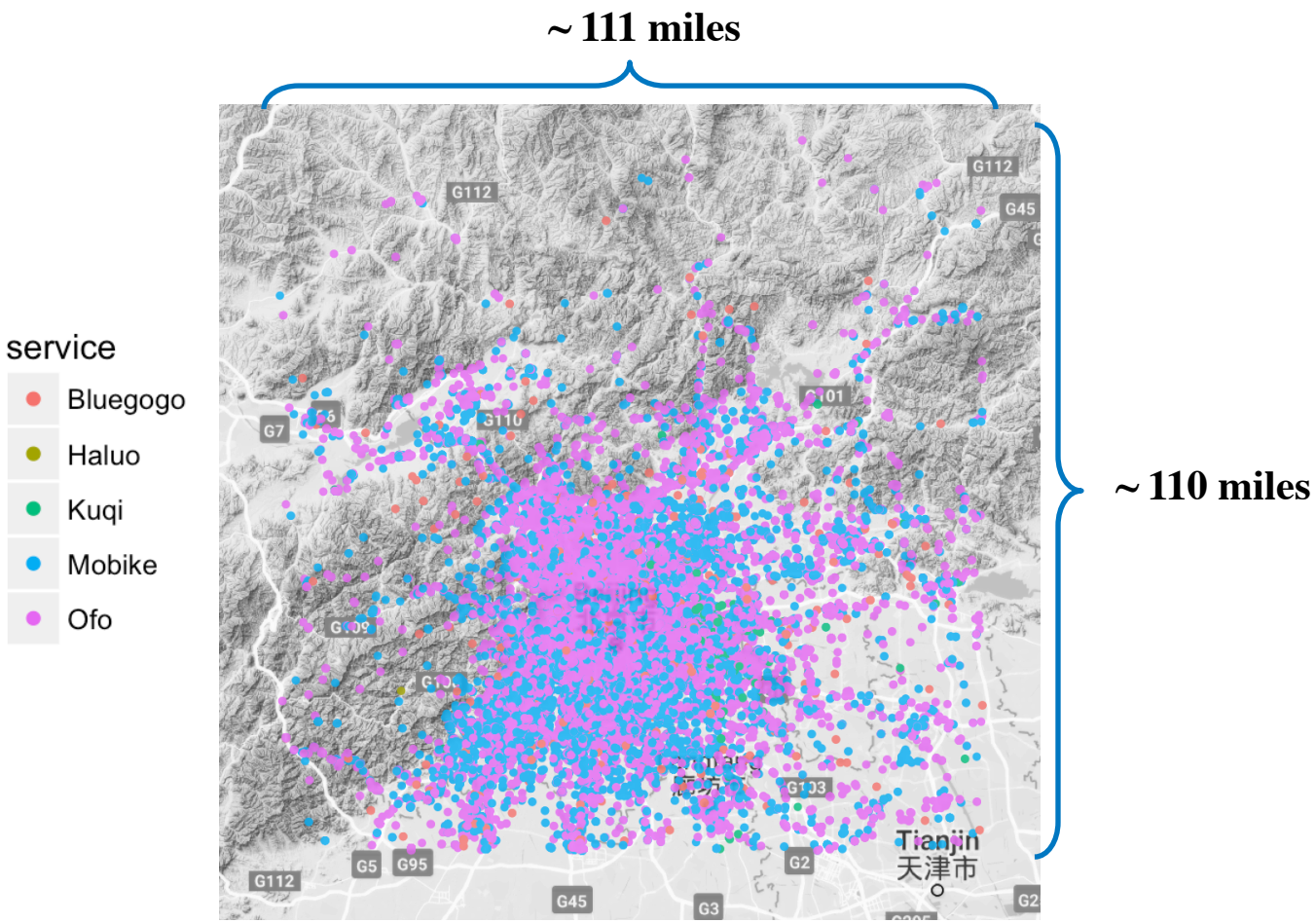
1 Research Motivation and Methodology

2 **Drivers of Bike-sharing Demand**

3 Forecasting Models

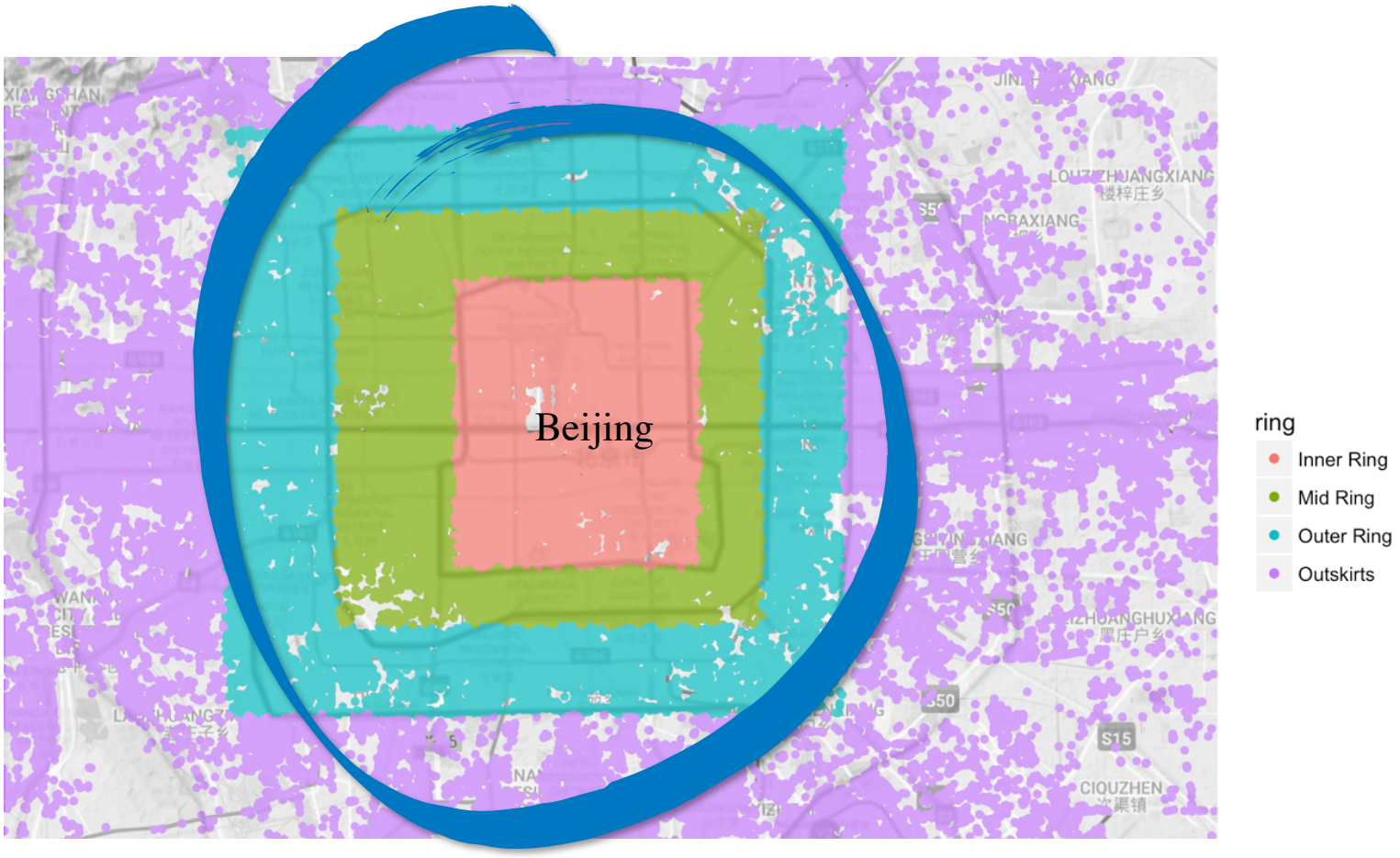
4 Taking insights into practice

Our Dataset



1,215,894 single observations over a period of one month (June 2017)

Our Focus Area



We tested certain drivers of Bike-sharing Demand

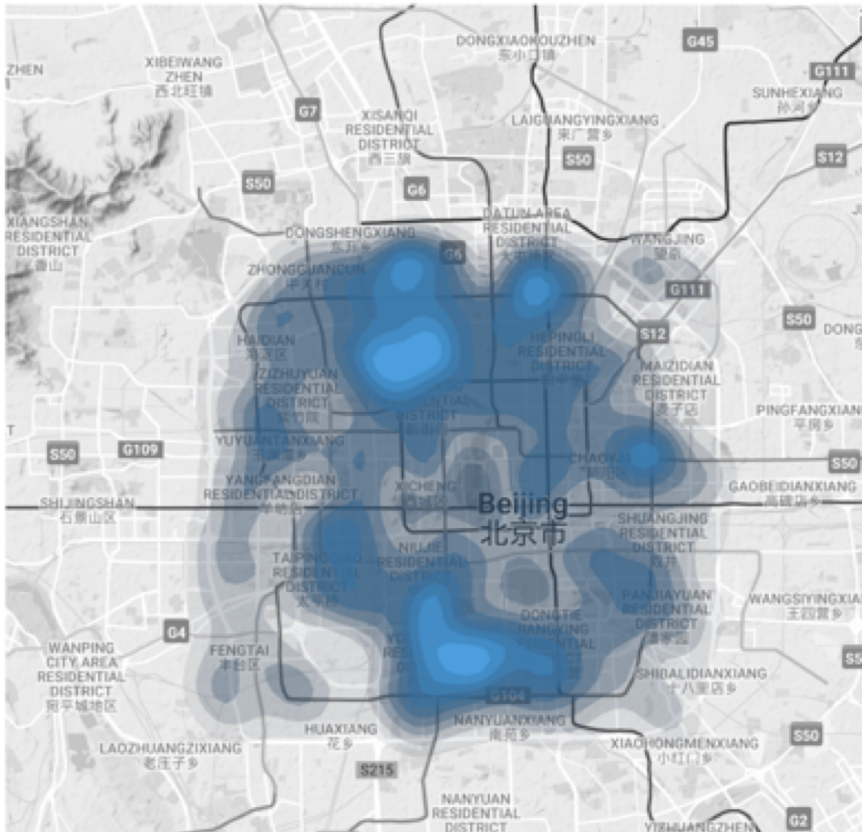
Time and Day

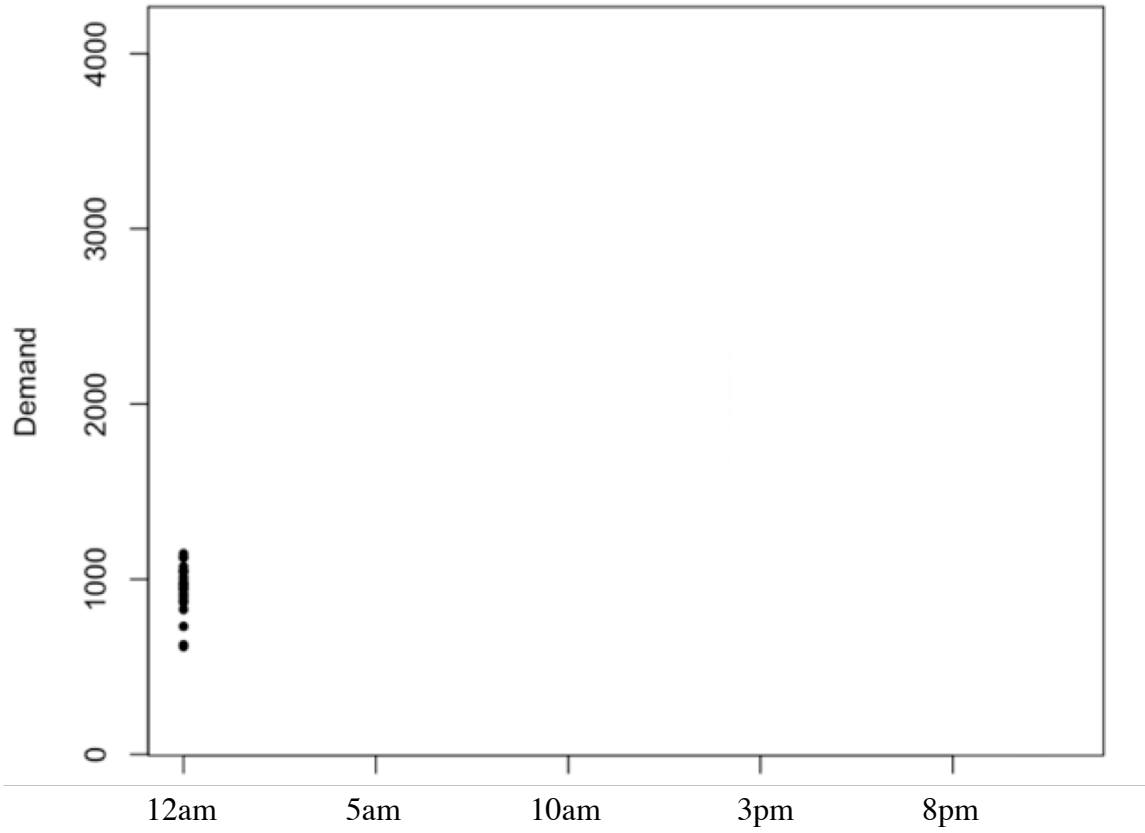
- Time of the day
- Day of the week
- Weekend
- Public Holiday

Bike-sharing Demand throughout the Day

MIT Supply Chain MANAGEMENT

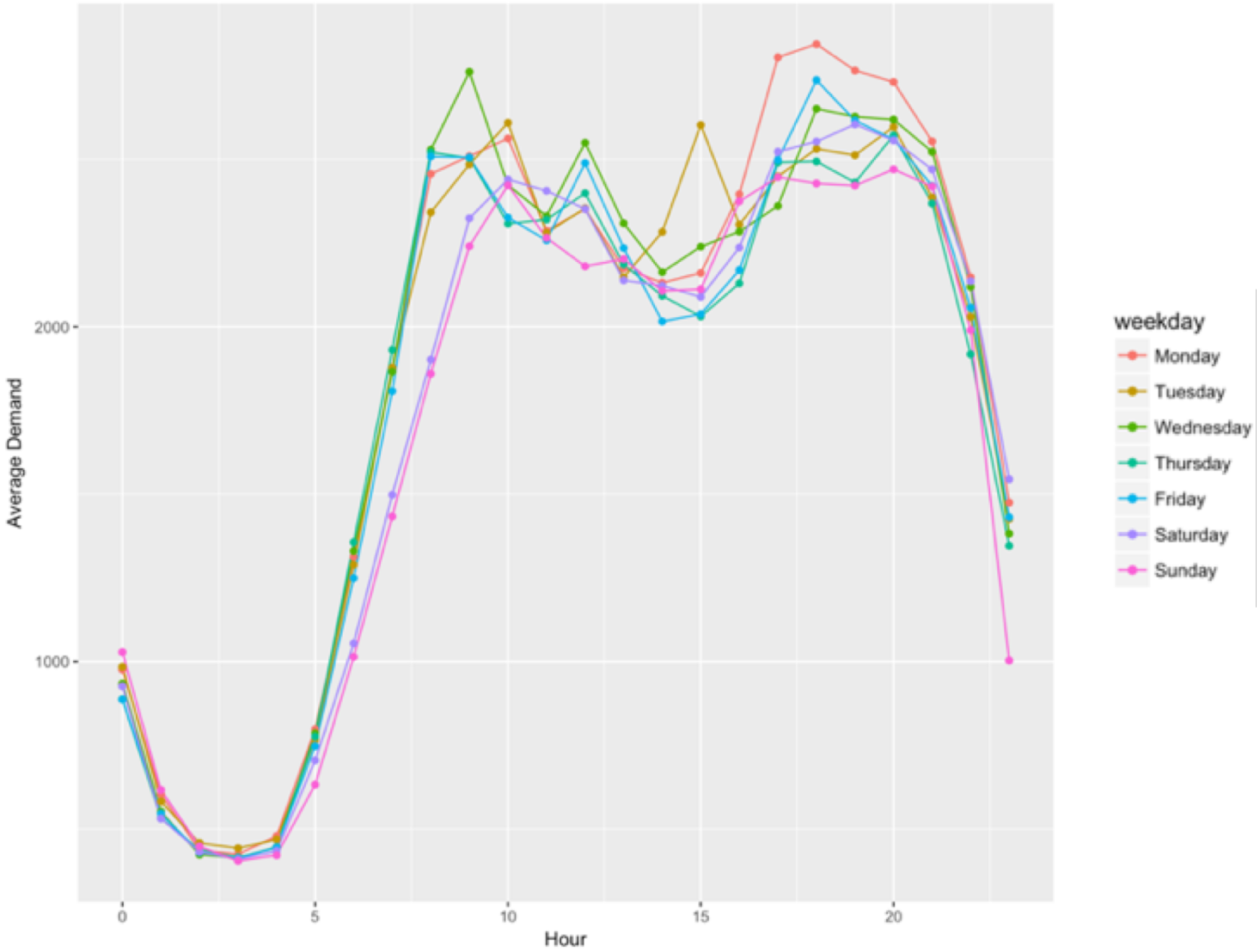
11 pm





Bike-sharing demand differs per location and time of the day

Average Bike-sharing Demand per Day



We tested certain drivers of Bike-sharing Demand

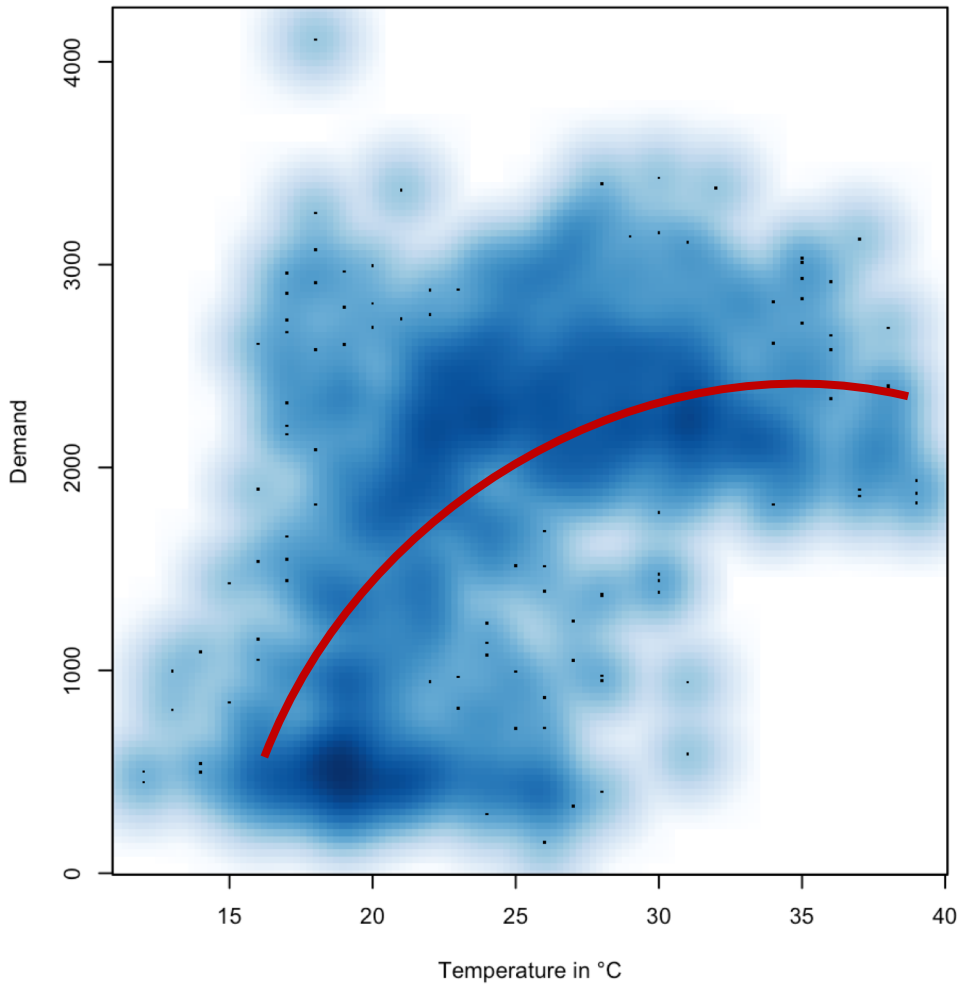
Environmental

- **Temperature**
- **Humidity**
- **Pressure**
- **Wind speed**
- **Air Quality**

Temperature

MIT Supply Chain MANAGEMENT

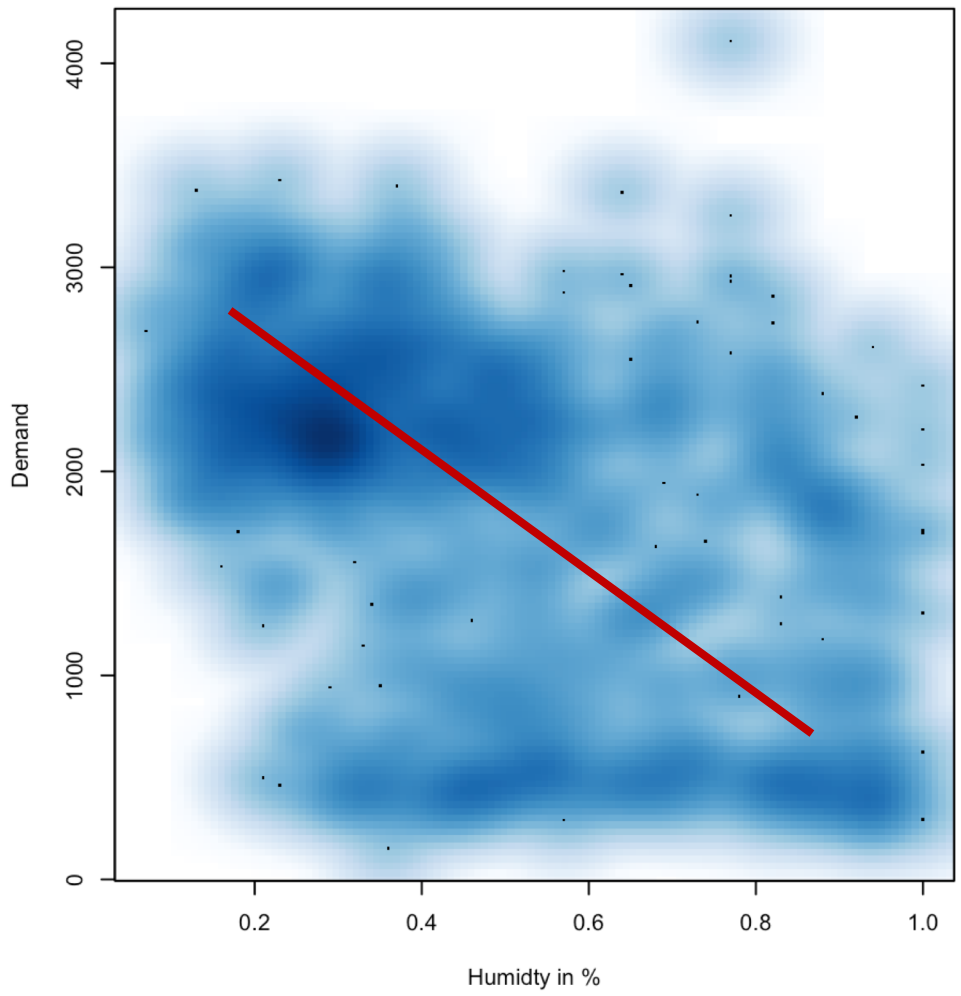
Observations



Humidity

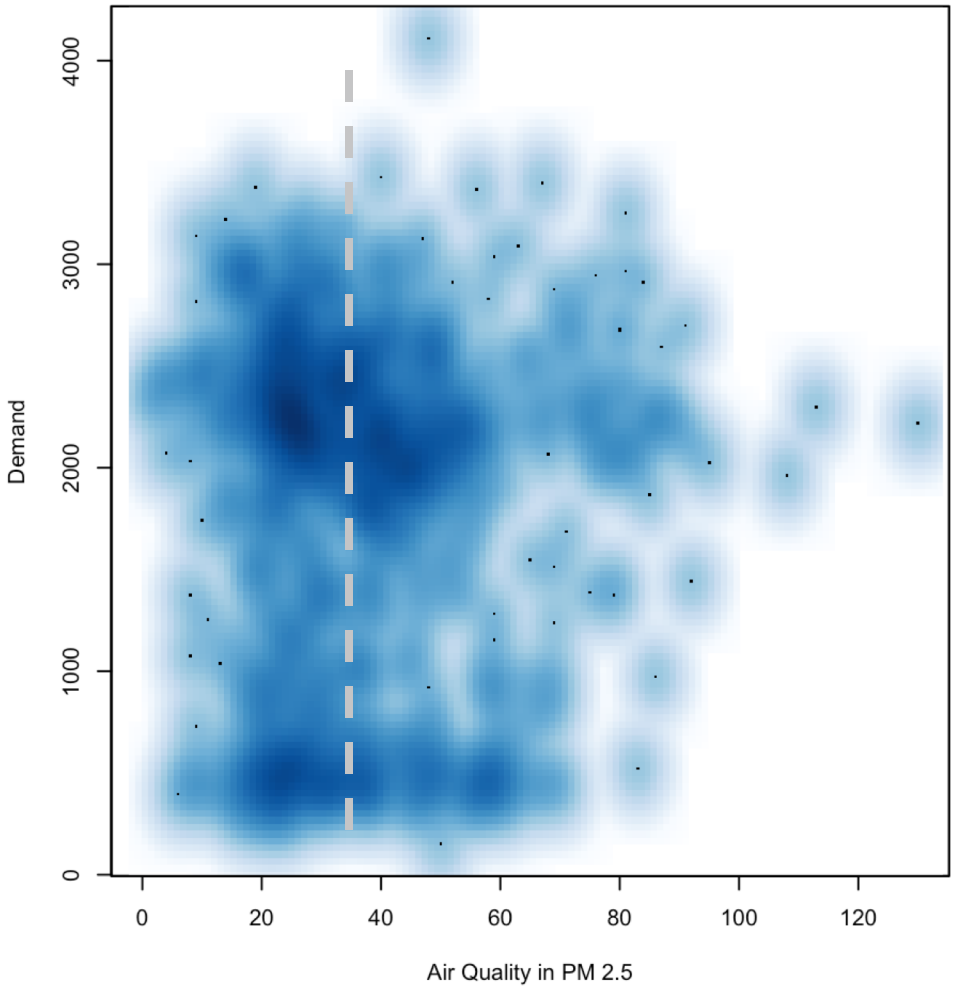
MIT Supply Chain MANAGEMENT

Observations



Air Quality

Observations



https://blissair.com/what-is-pm-2-5.htm

We find significant Drivers of Bike-sharing Demand

Time and Day

- Time of the day
- Day of the week
- Weekend
- Public Holiday

Environmental

- Temperature *
- Humidity
- Pressure *
- Wind speed *
- Air Quality

*** polynomial Relationship**

Agenda

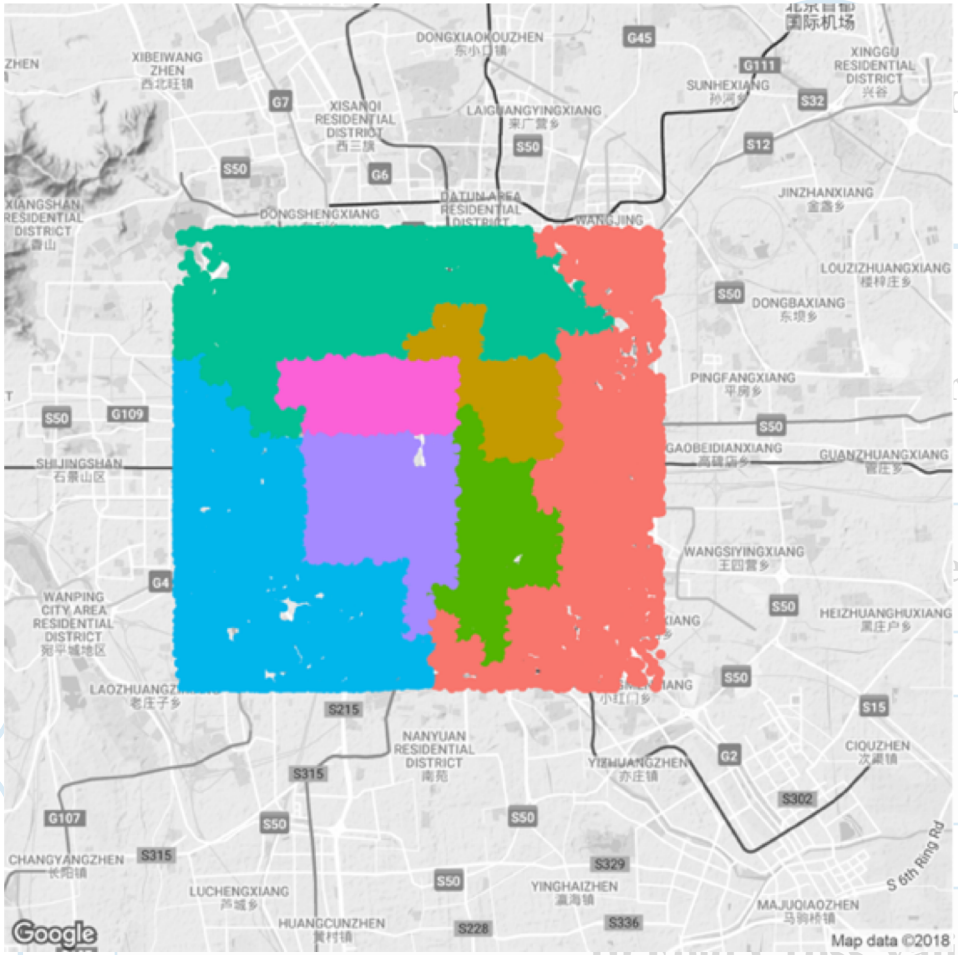
1 Research Motivation and Methodology

2 Drivers of Bike-sharing Demand

3 **Forecasting Models**

4 Taking insights into practice

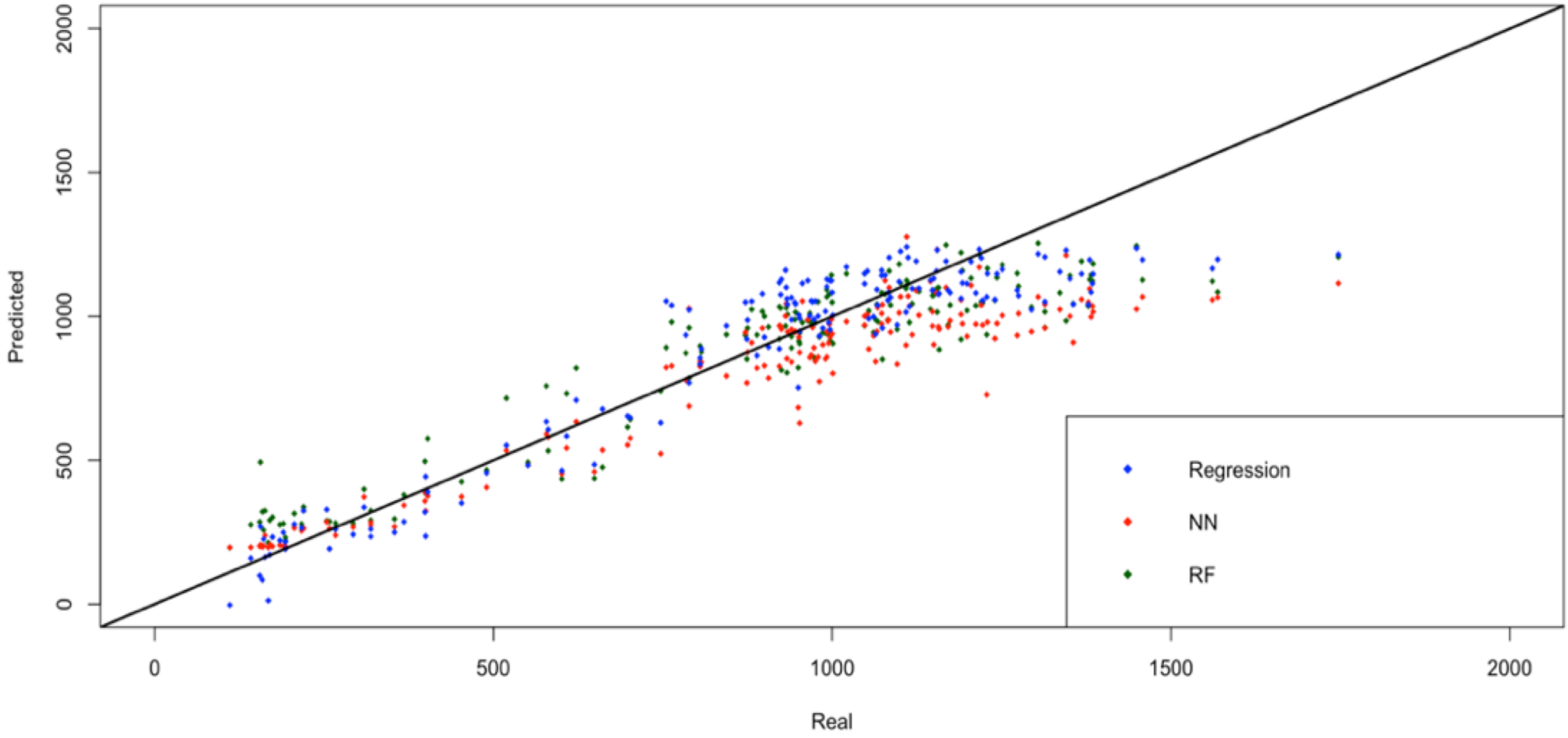
Forecasting Models Fitting and Validation



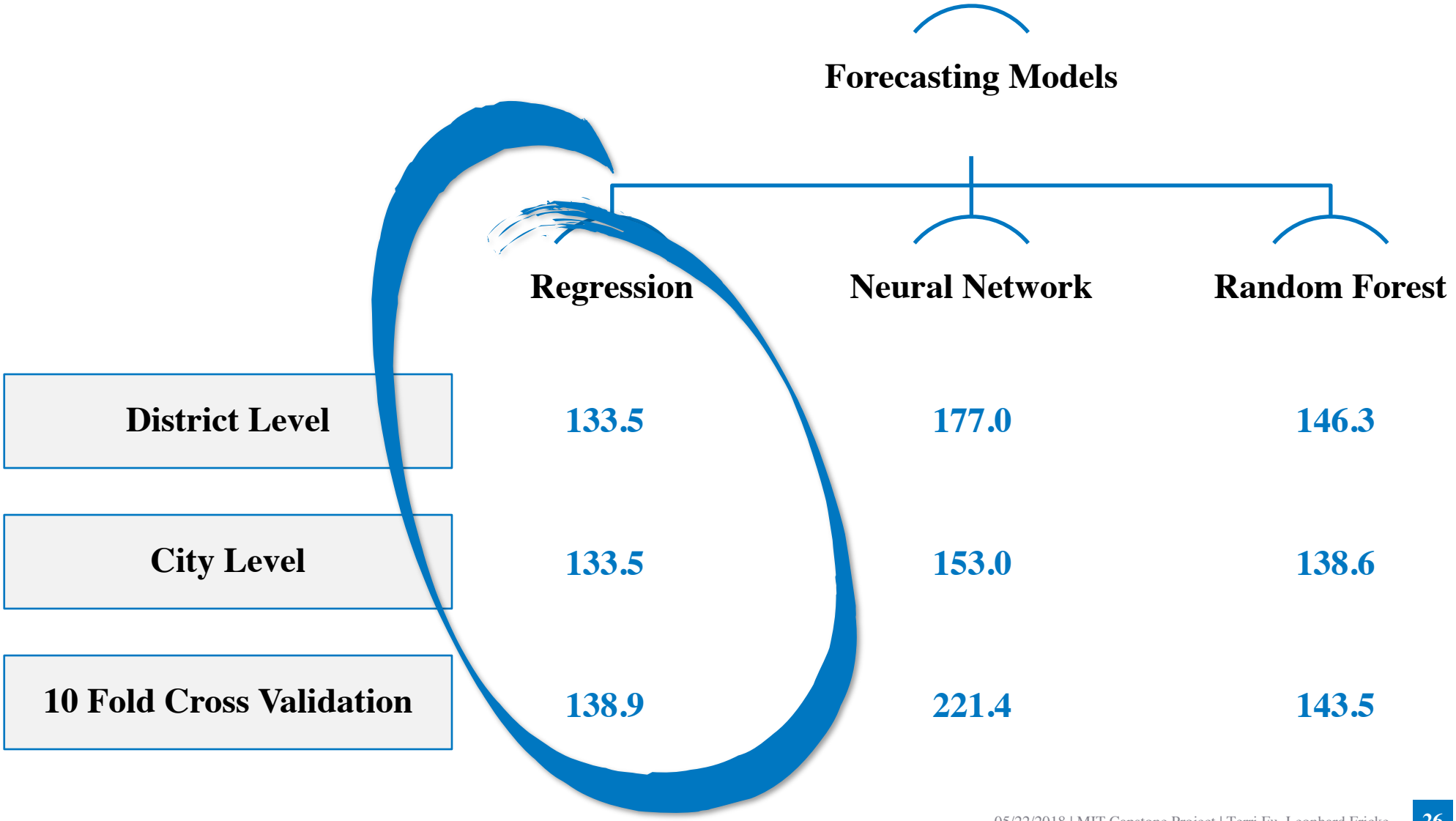
Real vs. Predicted Value

MIT Supply Chain MANAGEMENT

City Level



RMSE as Forecasting Models Result



Immediate Benefits

MIT Supply Chain MANAGEMENT

Predict The Future

Drive Price & Promotion

TalkingData

Customer Satisfaction

Remain Competitive

Agenda

1 Research Motivation and Methodology

2 Drivers of Bike-sharing Demand

3 Forecasting Models

4 Taking insights into practice

Contribution

Private Sector

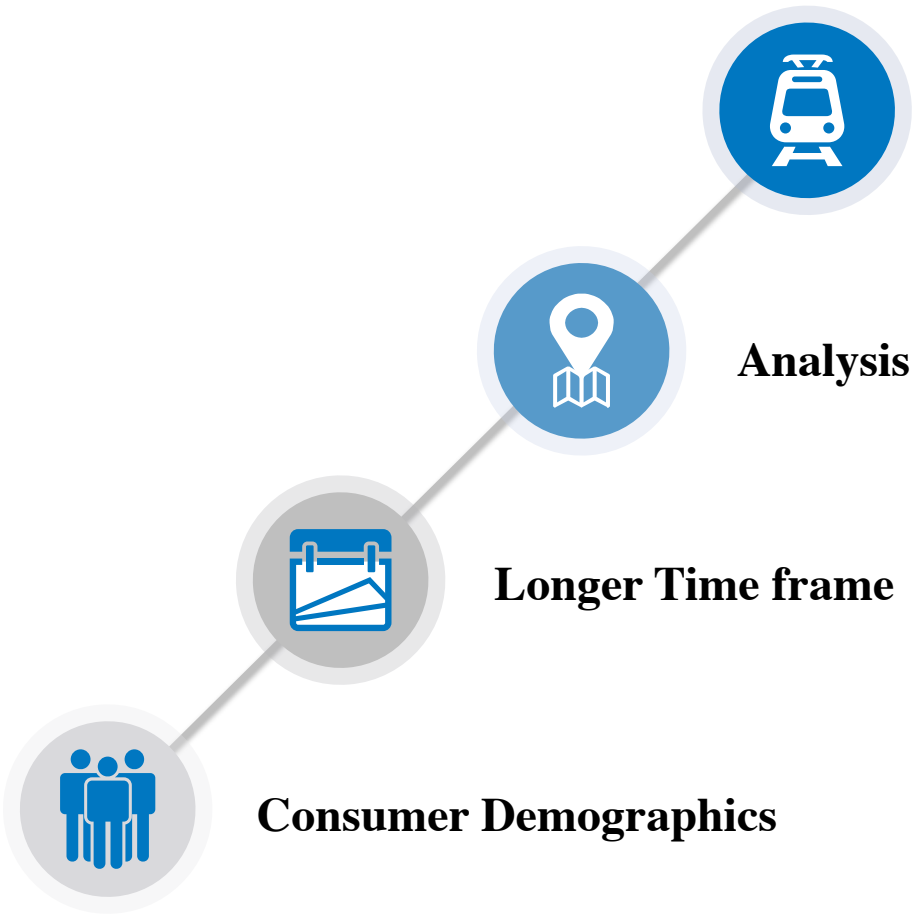
- Sharing bike operators, demand planning, replenishment, ground operations
- Transferrable to
  - other commute types such as car-sharing
  - other cities and countries

Public Sector

- Local regulators, department of transportation, researchers
  - public service planning
  - Infrastructure investment
  - Policy-making on sharing transportation

Future opportunities for TalkingData

MIT Supply Chain MANAGEMENT



Analysis by Points of Interest

Analysis by Admin. Districts

Longer Time frame

Consumer Demographics

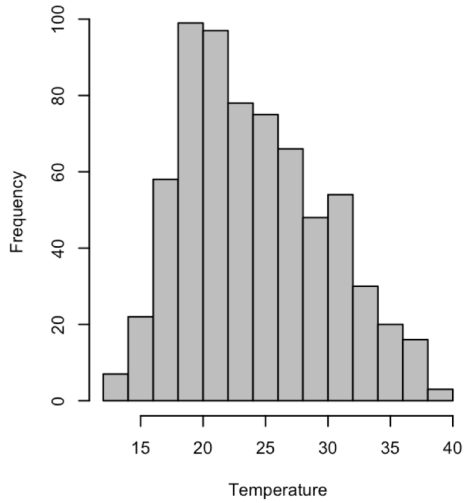
Any Questions?

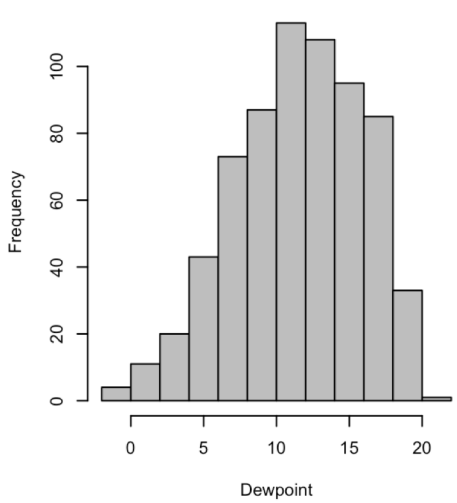
Backup

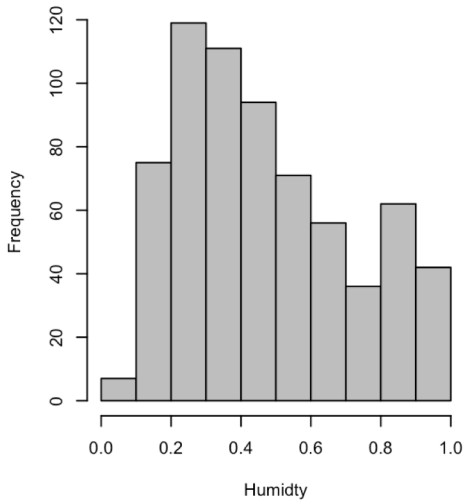
Title	Modeling Methodologies	Valuation	Conclusion
Forecasting Utilization in City Bike-Share Program (a)	(1) Neural Network (2) Poisson Regression (3) Markov Model (4) Mean Value Benchmark	RMSLE	Neural Network is found to perform the best which is 0.49.
Predicting Capital Bikeshare Demand in R (b)	(1) Regression (2) Generalized Boosted Models	RMSLE	The GBM model clearly performed better than our linear regression model
Bike Share Demand Prediction using RandomForests (c)	(1) Random Forest (2) Enhancing RM model using TuneRF (3) Conditional Inference Trees (4) Generalized Boosted Models	RMSLE	Random Forest with TuneRF performance is 0.49
Research on Antecedents and Consequences of Factors Affecting the Bike Sharing System (d)	(1) Multiple linear regression (2) Poisson and Topology method	P-Values	Temperature that people feel is most important determinant
Forecasting Bike Rental Demand (e)	(1) Basic Linear Regression (2) Generalized Linear Models with Elastic Net Regularization (3) Generalized Boosted Models (4) Principal Component Regression (5) Support Vector Regression (6) Random Forest (7) Conditional Inference Trees	RMSLE	Two of the Tree based models are found to perform the best: • CTree: 0.46 • Random Forest: 0.50
Demand Prediction of Bicycle Sharing Systems (f)	(1) Ridge Regression (2) Support Vector Regression (3) Random Forest (4) Gradient Boosted Tree	RMSLE	Random forest is found to perform best
Bike Sharing Demand: Forecast use of a city bikeshare system (g)	(1) Gradient Boosted Decision Trees (GBDT)	RMSLE	RMSLE value of GBDT Loop is 0.5683
Forecasting Bike Rental Demand Using New York Citi Bike Data (h)	(1) Linear Regression (2) Neural Network (3) Decision Tree (4) Random Forest	RMSLE	Random forest model has shown relative best performance
Prediction of Bike Sharing Demand (i)	(1) SVM (2) Neural Network (3) Poisson Regression (4) Random Forest (5) Extra Trees Regressor (6) GBM (7) Linear Combination Model (8) Discriminating Linear Combination Model	RMSE	Linear Combination model and Discriminating Linear Combination model are good models for predicting bike sharing demand with RMSe being close to 0.36

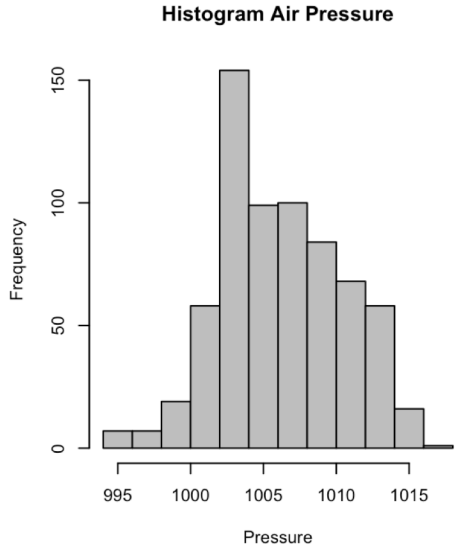
Legend: (a) Lee, Wang, and Wong 2014; (b) Liu 2015; (c) Patil, Musale, and Rao 2015; (d) Jing and Zhao 2015; (e) Du, He, and Zhechev 2014; (f) Yin, Lee, and Wong 2012; (g) Tian, Li, and Zhang, n.d. ; (h) Wang 2016; (i) Sachdeva and Sarvanan 2017

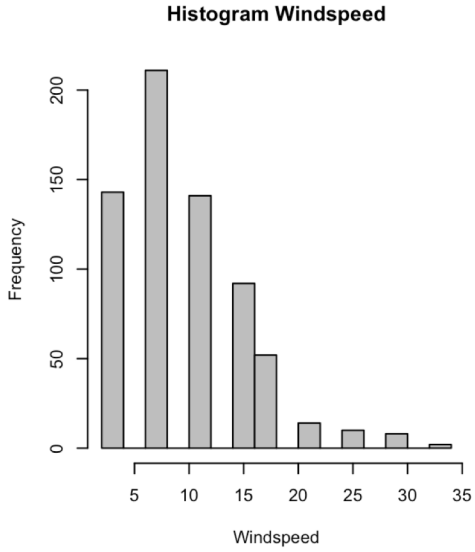
Variable	Description
Demand *	Number of unique devices with active bike-sharing applications on a smartphone
Year	Year of observation (i.e. 2017)
Month	Month of observation (i.e. 1 for January)
Day	Day of observation (i.e. 1 for 1st for first day of the month
Day of week	String naming the respective weekday to a date (i.e. Monday, Tuesday, etc.)
Public holiday	Indicating whether respective date is a public holiday (1) or not (0)
Working weekend	Indicating whether respetive weekend day is a working day (1) or not (0)
Hour	Hour of the day (0-24 hours)
Time	Time of the day (0-12 am/pm)
Temp. (in °C)	Temperature is a measure of the warmth or coldness of an object or substance with reference to a standard value (i.e. 1 for 1°C)
Dew Point (in °C)	The temperature at which the air temperature must be cooled for water vapor to condense, forming water droplets, fog, or cloud (i.e. 1 for 1°C)
Humidity (in %)	The amount of water vapor in the air, expressed as a percentage. Relative humidity is the ratio of the amount of moisture in the air to the amount that is needed to saturate the air. Thus, it is a function of both moisture content and temperature, as its name states, humidity is "relative" to temperature.
Pressure (in hPa)	The force exerted by the weight of the atmosphere and gravity (i.e. 1085 for 1085 hPa)
Win dir	The direction that the wind is blowing from expressed in cardinal directions (i.e. S for South)
Wind speed (in km/h)	The rate at which air is moving horizontally past a point. It may be a 2-minute average speed, or an instantaneous speed (i.e. 10 for 10 km/h)
Conditions	String naming the Weather condition (i.e. clear, rainy, etc.)
Air quality (in PM 2.5)	PM2.5 readings are often included in air quality reports from environmental authorities and companies. They refers to atmospheric particulate matter (PM) and to fine particles.

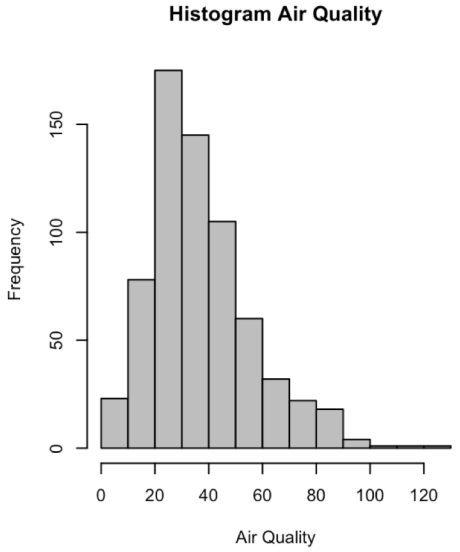

Histogram Temperature
Frequency
0
20
40
60
80
100
15
20
25
30
35
40
Temperature

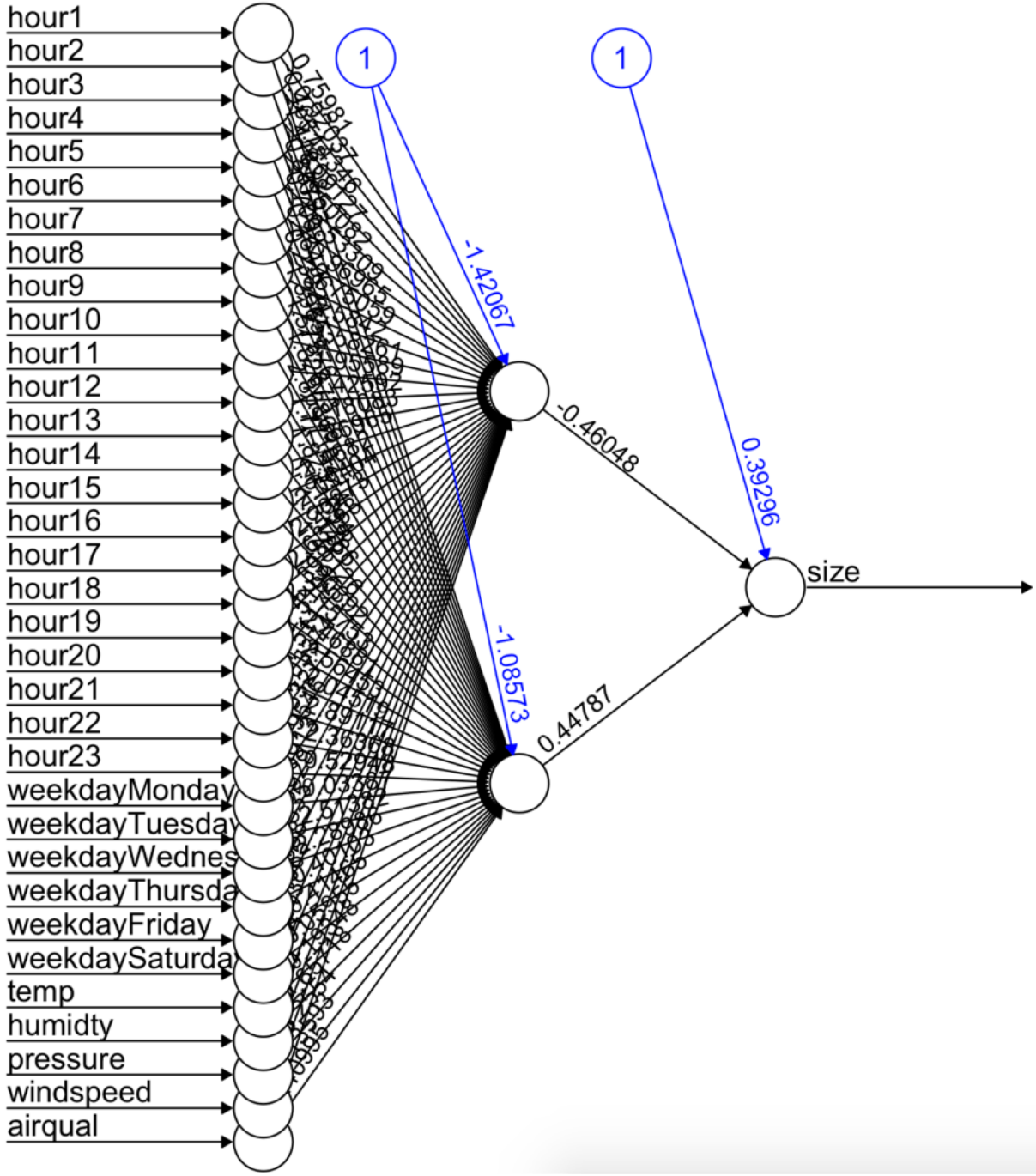

Histogram Dewpoint
Frequency
0
20
40
60
80
100
0
5
10
15
20
Dewpoint


Histogram Humidty
Frequency
0
20
40
60
80
100
120
0.0
0.2
0.4
0.6
0.8
1.0
Humidty

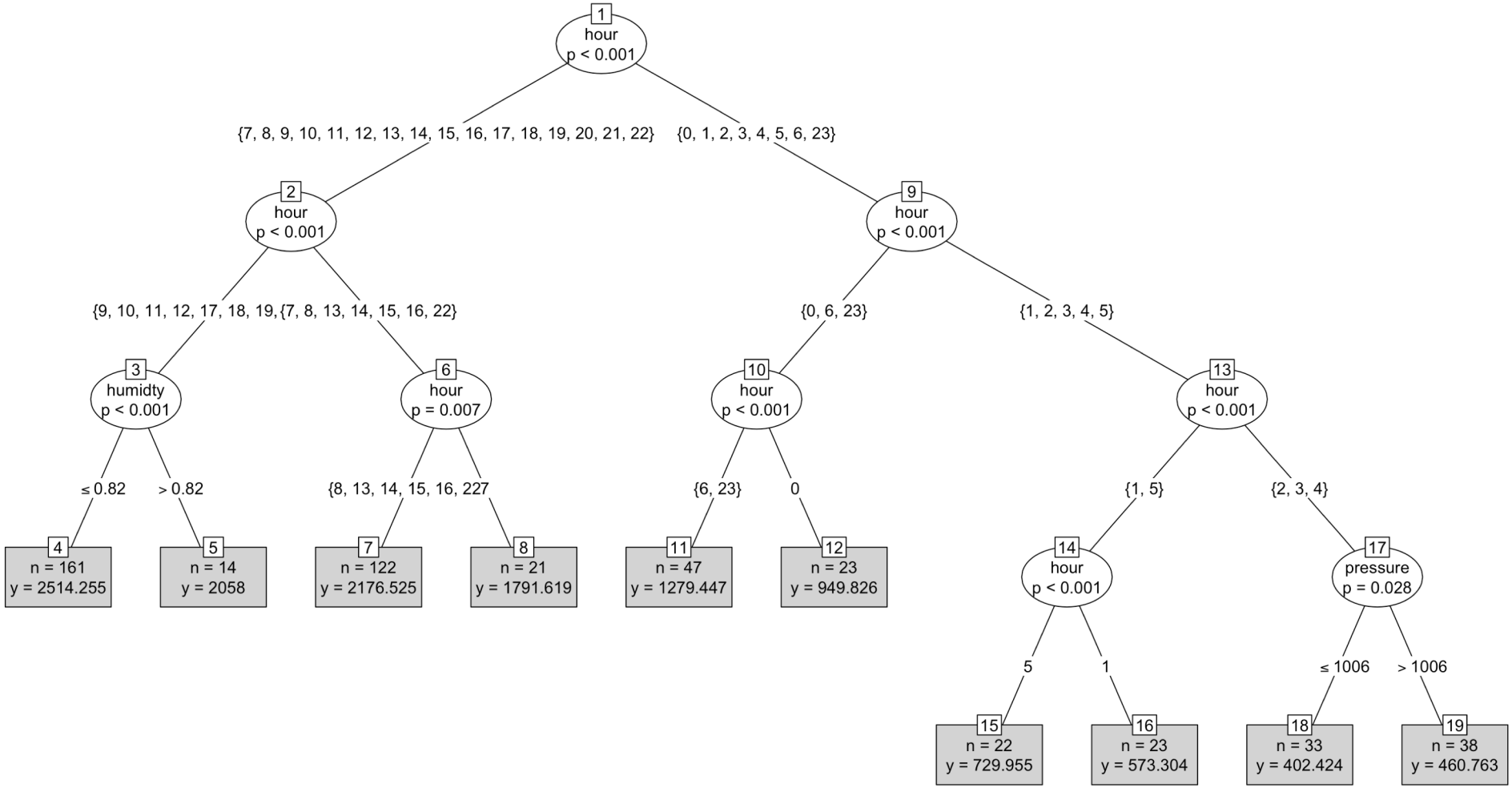

Histogram Air Pressure
Frequency
0
50
100
150
995
1000
1005
1010
1015
Pressure


Histogram Windspeed
Frequency
0
50
100
150
200
5
10
15
20
25
30
35
Windspeed

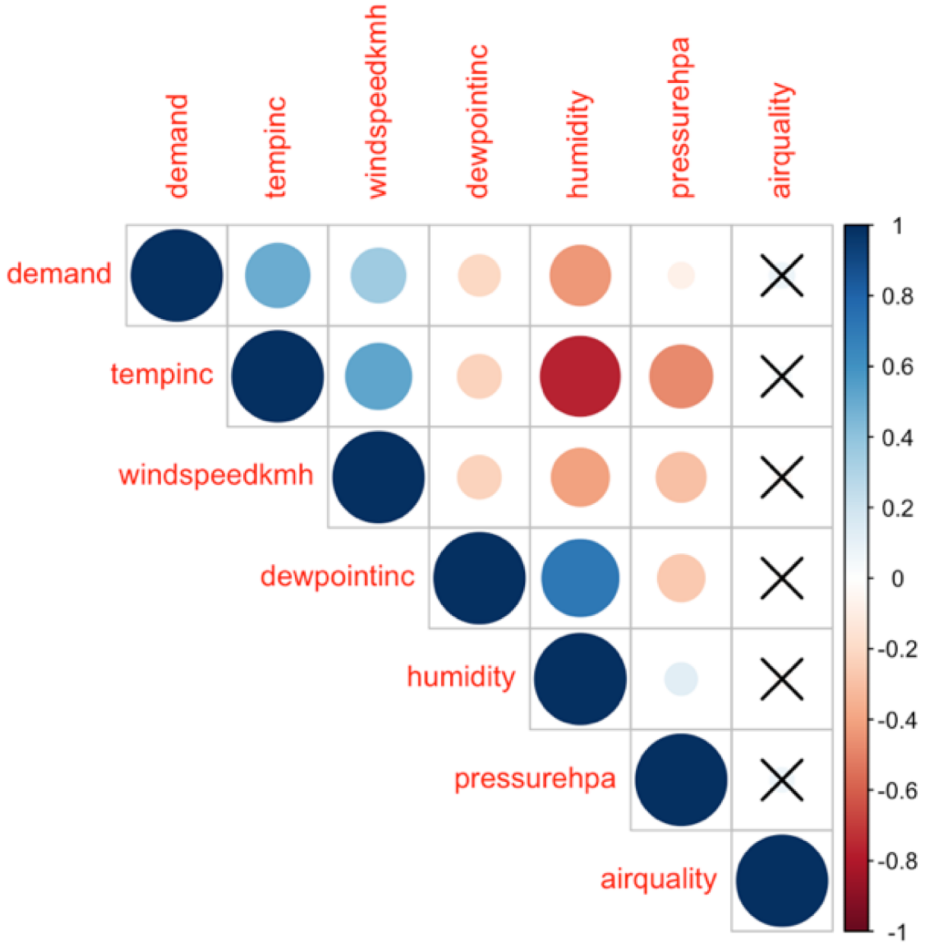

Histogram Air Quality
Frequency
0
50
100
150
0
20
40
60
80
100
120
Air Quality

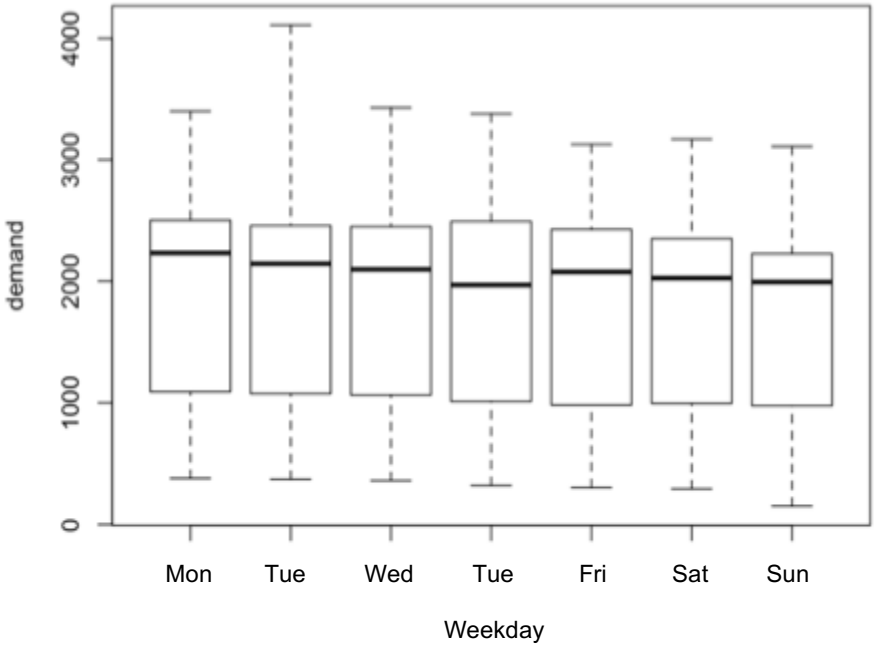


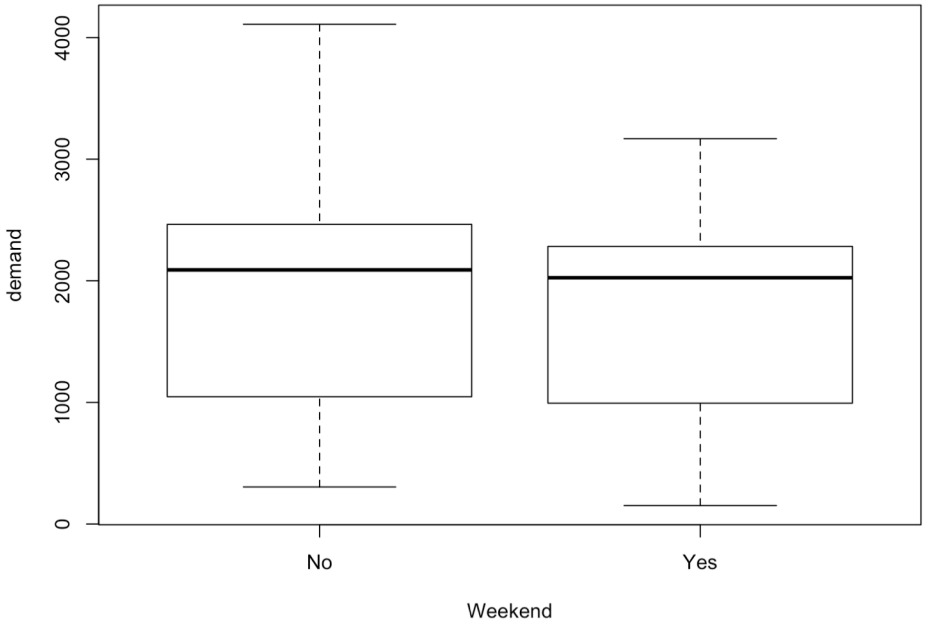
MIT Supply Chain MANAGEMENT


1
hour
p < 0.001
{7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22}
{0, 1, 2, 3, 4, 5, 6, 23}
2
hour
p < 0.001
9
hour
p < 0.001
{9, 10, 11, 12, 17, 18, 19,
{7, 8, 13, 14, 15, 16, 22}
{0, 6, 23}
{1, 2, 3, 4, 5}
3
humidty
p < 0.001
6
hour
p = 0.007
10
hour
p < 0.001
13
hour
p < 0.001
≤ 0.82
> 0.82
{8, 13, 14, 15, 16, 227
{6, 23}
0
{1, 5}
{2, 3, 4}
4
n = 161
y = 2514.255
5
n = 14
y = 2058
7
n = 122
y = 2176.525
8
n = 21
y = 1791.619
11
n = 47
y = 1279.447
12
n = 23
y = 949.826
14
hour
p < 0.001
17
pressure
p = 0.028
5
1
≤ 1006
> 1006
15
n = 22
y = 729.955
16
n = 23
y = 573.304
18
n = 33
y = 402.424
19
n = 38
y = 460.763

Individual characteristics	Societal context	Physical Infrastructure	Others
- Gender - Age - Income - Employment status - Household structure - Bicycle & car ownership - Physical activity - Level of education - Attitudes - Environmental beliefs - Habits	- Social norms - Costs of cycling - Costs of alternatives - Safety - Education (training cyclists, informing car-users, etc.)	- Urban form - Hilliness - Cycling infrastructure - System infrastructure - Bicycle parking	- Weather / Climate - Seasons - Day of the week - Daytime - Working day / holiday - Events


demand
tempinc
windspeedkmh
dewpointinc
humidity
pressurehpa
airquality
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1


demand
0
1000
2000
3000
4000
Mon
Tue
Wed
Tue
Fri
Sat
Sun
Weekday

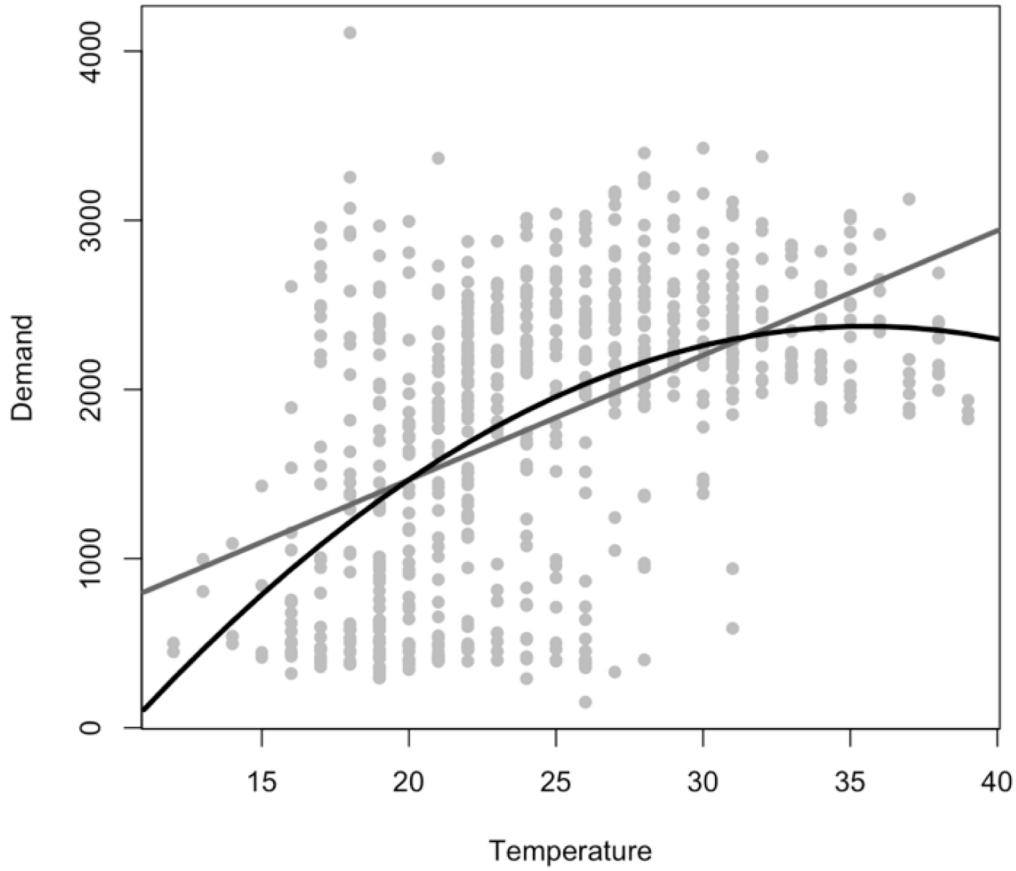

demand
0
1000
2000
3000
4000
No
Yes
Weekend

	Dependent variable:				
	Demand				
	(1)	(2)	(3)	(4)	(5)
dayperiodmorning	**1,253.000*****	**1,254.000*****	**1,253.000*****	**1,253.000*****	**1,253.000*****
	(48.120)	(48.260)	(48.160)	(48.100)	(48.290)
dayperiodafternoon	**1,384.000*****	**1,385.000*****	**1,384.000*****	**1,384.000*****	**1,384.000*****
	(53.010)	(53.170)	(53.060)	(52.990)	(53.200)
dayperiodevening	**1,677.000*****	**1,678.000*****	**1,678.000*****	**1,678.000*****	**1,678.000*****
	(56.950)	(57.120)	(57.010)	(56.920)	(57.150)
weekdayMonday	**206.500*****		**177.400****		
	(73.510)		(70.970)		
weekdayTuesday	**162.800****		**133.700***		
	(73.510)		(70.970)		
weekdayWednesday	**149.700****		**150.900****		
	(70.910)		(70.970)		
weekdayThursday	**77.550**		**78.800**		
	(70.910)		(70.970)		
weekdayFriday	**98.400**		**99.640**		
	(70.910)		(70.970)		
weekdaySaturday	**56.650**		**57.900**		
	(70.910)		(70.970)		
weekend1				**-99.290****	
				(41.960)	
pubhol	**-121.400**				**-41.940**
	(80.980)				(73.320)
Constant	**776.600*****	**874.300*****	**775.000*****	**903.000*****	**877.400*****
	(56.630)	(32.930)	(56.670)	(34.980)	(33.410)
Observations	673	673	673	673	673
R^2	0.658	0.653	0.657	0.656	0.653
Adjusted R^2	0.653	0.651	0.653	0.654	0.651
Residual Std. Error	492.500 (df = 662)	493.900 (df = 669)	492.900 (df = 663)	492.200 (df = 668)	494.200 (df = 668)
F Statistic	127.600*** (df = 10; 662)	419.100*** (df = 3; 669)	141.200*** (df = 9; 663)	317.900*** (df = 4; 668)	314.100*** (df = 4; 668)
Note:					*p<0.1; **p<0.05; ***p<0.01

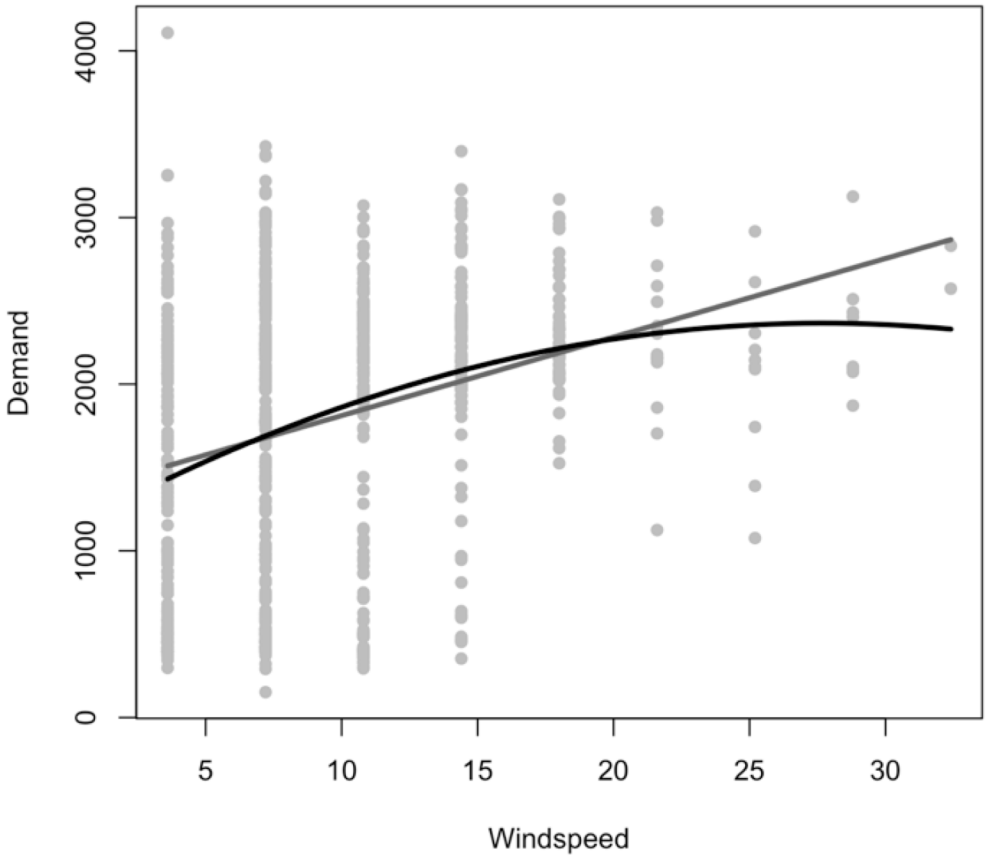
	Dependent variable:					
	Demand					
	(6)	(7)	(8)	(9)	(10)	(11)
temp	**76.3*****	**73.7*****	**54.1*****	**88.2*****	**68.0*****	**75.0*****
	(10.0)	(4.9)	(7.6)	(5.5)	(5.8)	(5.0)
humidty	**-193.0**		**-596.0*****			
	(194.0)		(177.0)			
pressure	**37.8*****			**39.9*****		
	(8.2)			(7.4)		
windspeed	**11.3***				**11.2***	
	(5.9)				(5.9)	
airqual	**-2.0**					**-2.4**
	(1.5)					(1.5)
Constant	**-38,040.0*****	**-7.2**	**759.0*****	**-40,553.0*****	**22.4**	**53.4**
	(8,429.0)	(125.0)	(259.0)	(7,559.0)	(126.0)	(130.0)
Observations	671	673	673	671	673	673
R^2	0.3	0.2	0.3	0.3	0.3	0.3
Adjusted R^2	0.3	0.2	0.3	0.3	0.3	0.2
Residual Std. Error	709.0 (df = 665)	726.0 (df = 671)	720.0 (df = 670)	712.0 (df = 668)	724.0 (df = 670)	725.0 (df = 670)
F Statistic	54.0*** (df = 5; 665)	222.0*** (df = 1; 671)	118.0*** (df = 2; 670)	130.0*** (df = 2; 668)	113.0*** (df = 2; 670)	112.0*** (df = 2; 670)

Note: *p<0.1; **p<0.05; ***p<0.01

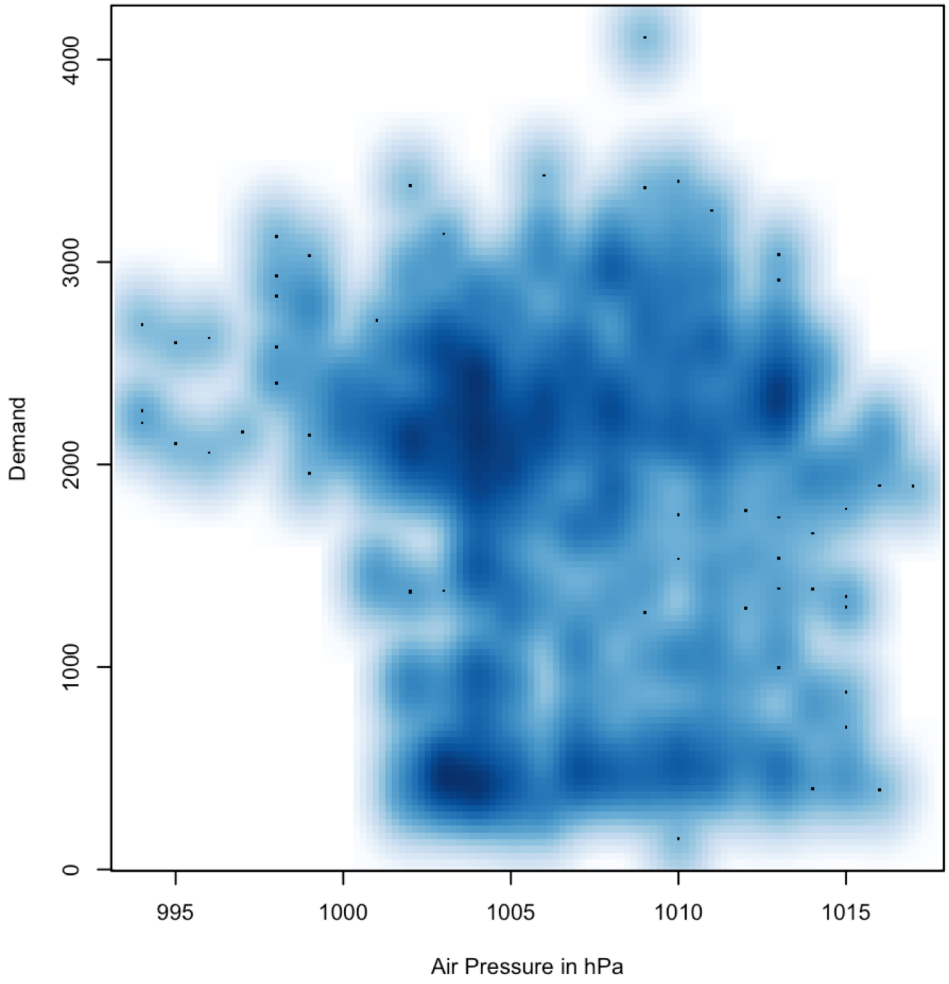
	Dependent variable:		
	Demand		
	(12)	(13)	(14)
temp	**73.70*****	**268.00*****	**-212.00**
	(4.95)	(39.50)	(213.00)
temptemp		**-3.77*****	**15.40***
		(0.76)	(8.39)
temptemptemp			**-0.25****
			(0.11)
Constant	**-7.23**	**-2,383.00*****	**1,470.00**
	(125.00)	(495.00)	(1,752.00)
Observations	673	673	673
R^2	0.25	0.27	0.28
Adjusted R^2	0.25	0.27	0.28
Residual Std. Error	726.00 (df = 671)	713.00 (df = 670)	711.00 (df = 669)
F Statistic	222.00*** (df = 1; 671)	127.00*** (df = 2; 670)	87.00*** (df = 3; 669)
Note:			*p<0.1; **p<0.05; ***p<0.01



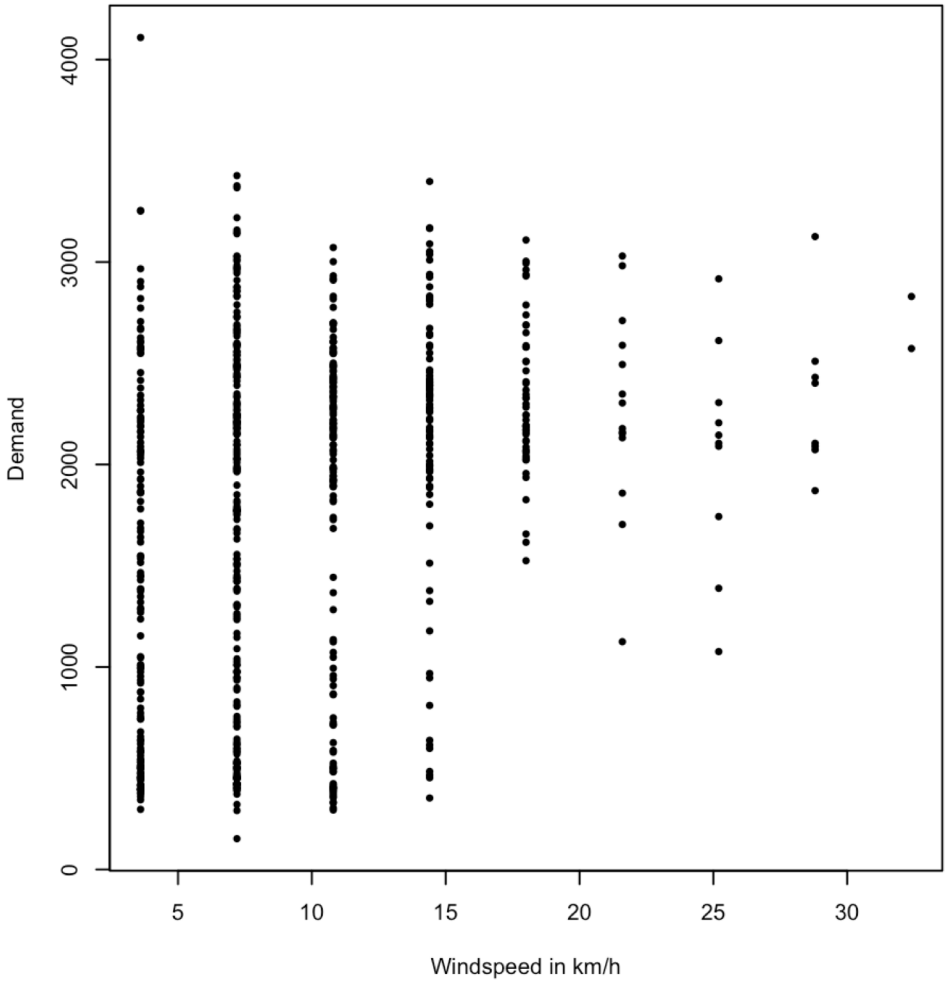
	Dependent variable:		
	Demand		
	(15)	(16)	(17)
windspeed	**47.10*****	**89.30*****	**78.60**
	(5.52)	(18.50)	(51.50)
windwind		**-1.61****	**-0.75**
		(0.68)	(3.96)
windwindwind			**-0.02**
			(0.09)
Constant	1,340.00***	1,130.00***	1,164.00***
	(62.70)	(108.00)	(188.00)
Observations	673	673	673
R^2	0.10	0.11	0.11
Adjusted R^2	0.10	0.10	0.10
Residual Std. Error	795.00 (df = 671)	792.00 (df = 670)	793.00 (df = 669)
F Statistic	72.90*** (df = 1; 671)	39.60*** (df = 2; 670)	26.40*** (df = 3; 669)
Note:			*p<0.1; **p<0.05; ***p<0.01

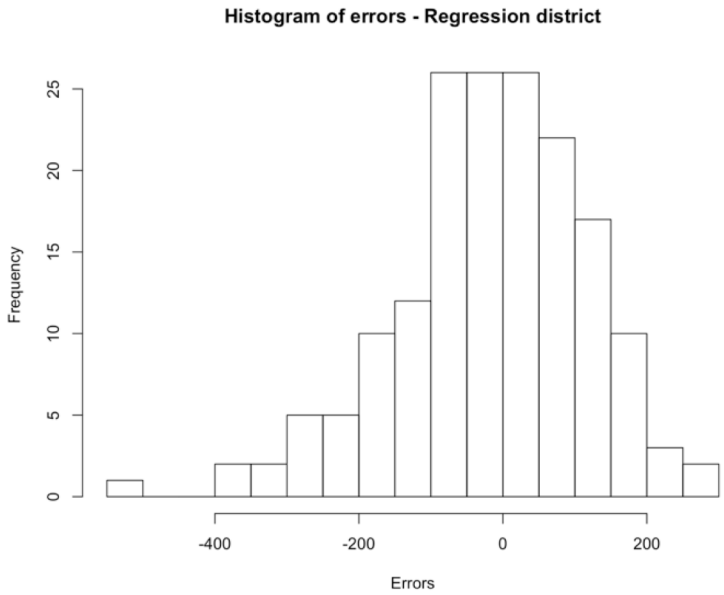


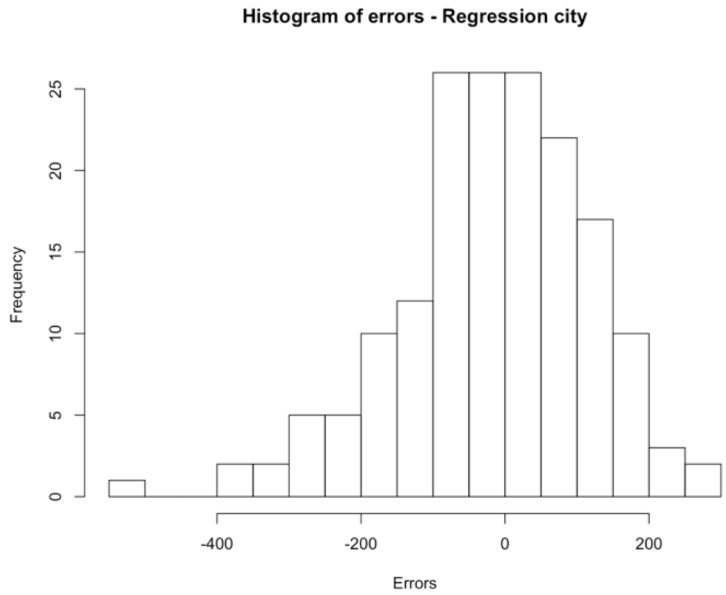
Air pressure

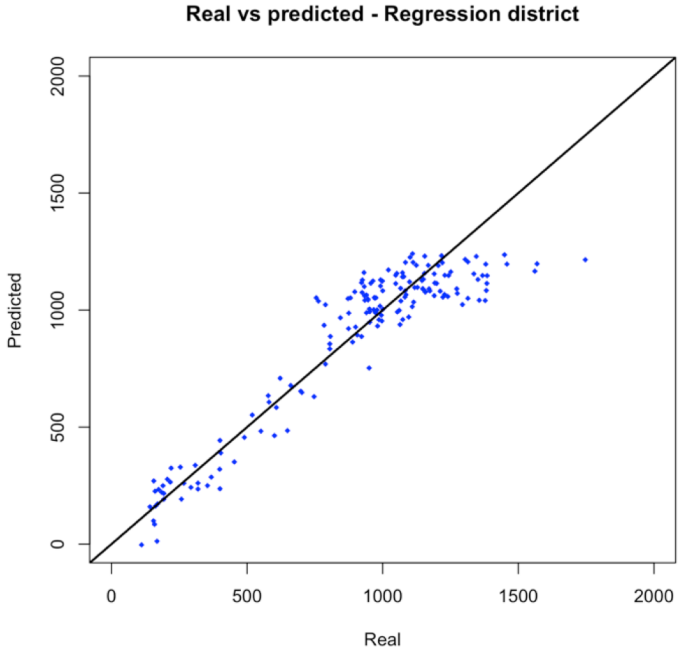


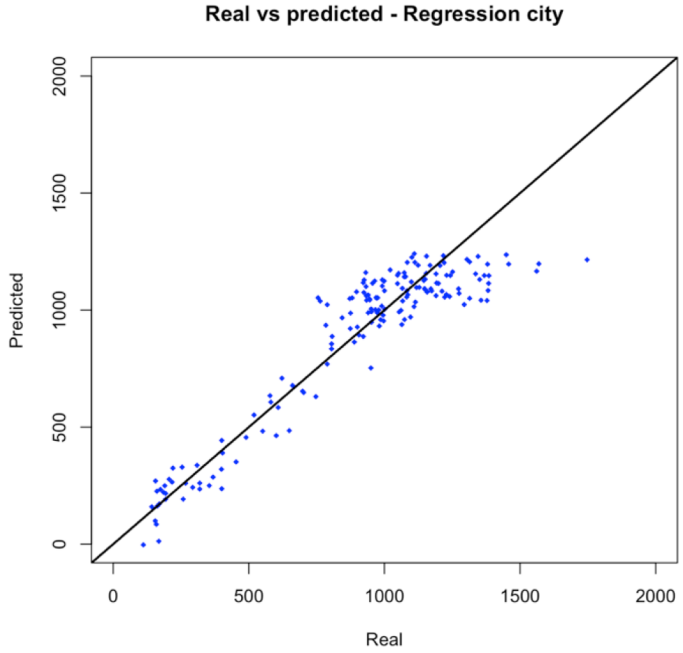
Windspeed

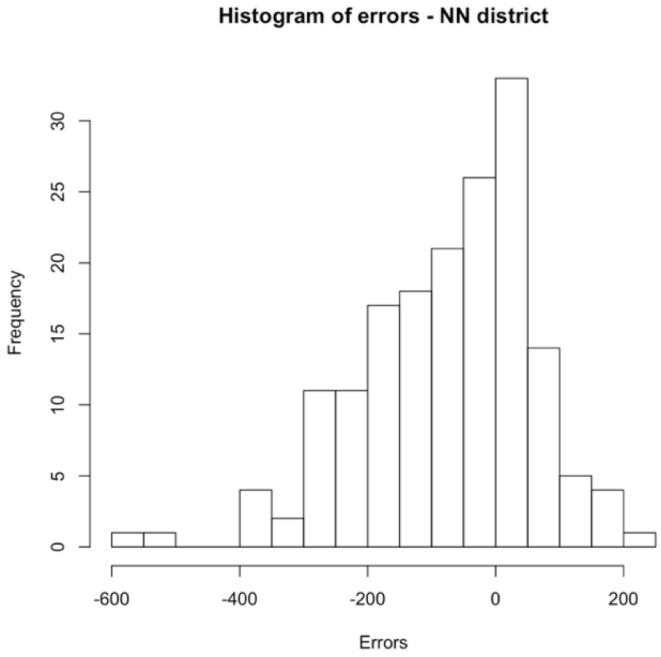


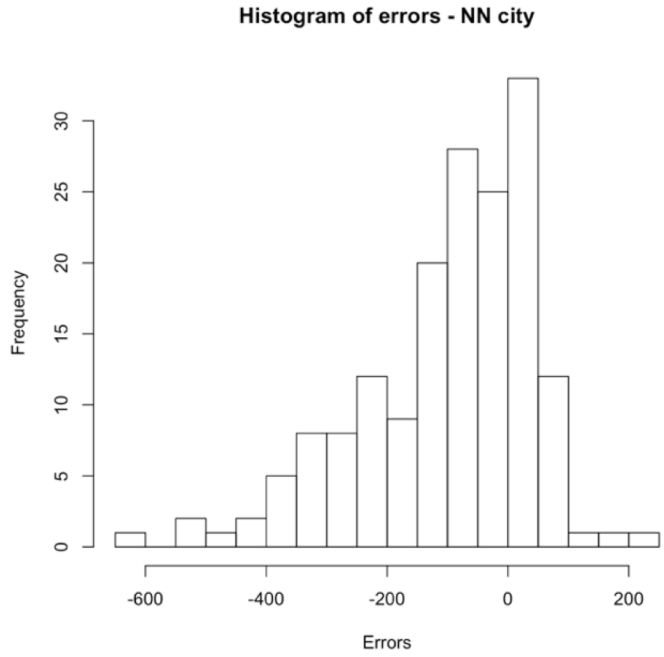

Histogram of errors - Regression district
Frequency
0
5
10
15
20
25
-400
-200
0
200
Errors

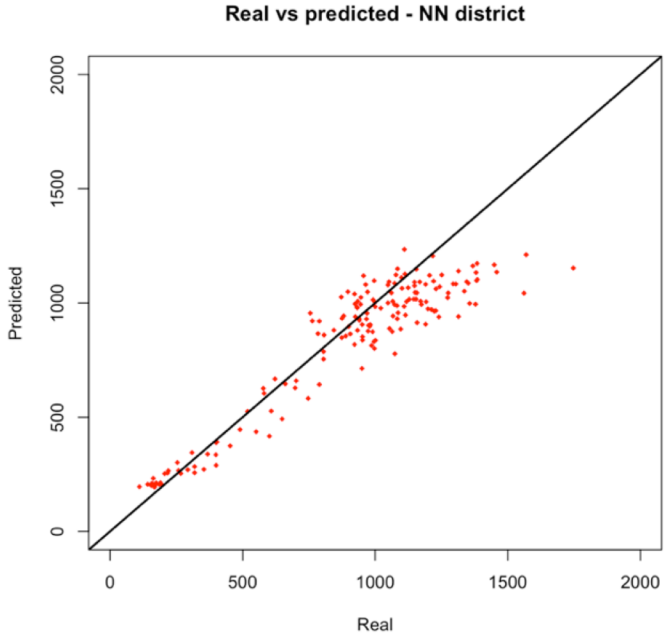

Histogram of errors - Regression city
Frequency
0
5
10
15
20
25
-400
-200
0
200
Errors

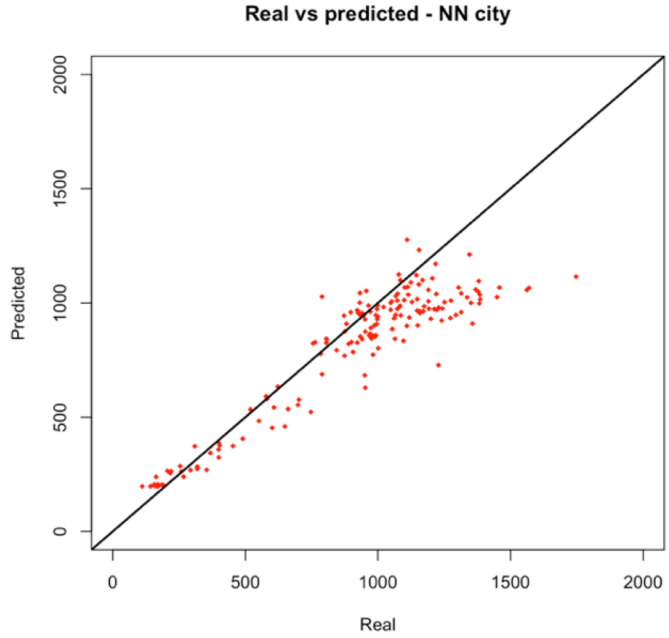

Real vs predicted - Regression district
Predicted
0
500
1000
1500
2000
Real

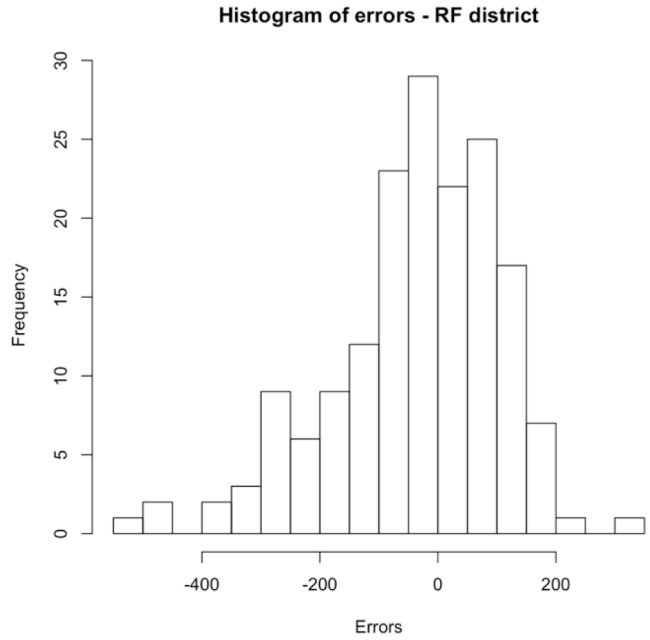

Real vs predicted - Regression city
Predicted
0
500
1000
1500
2000
Real

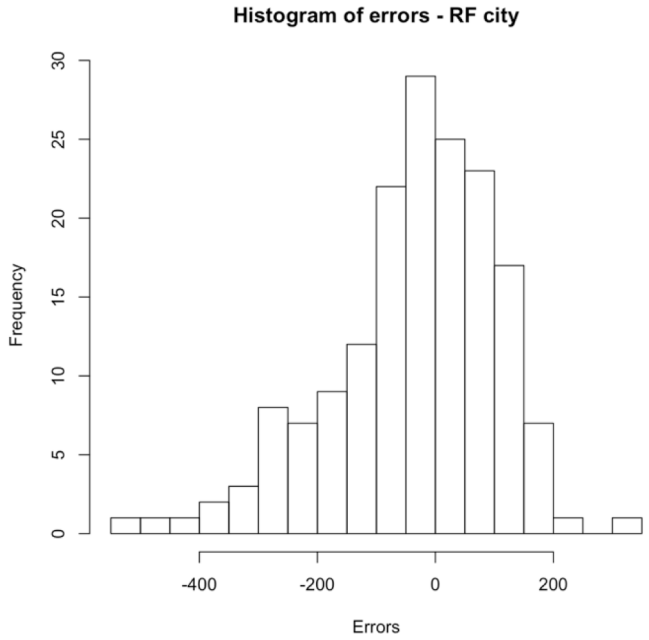

Histogram of errors - NN district
Frequency
0
5
10
15
20
25
30
-600
-400
-200
0
200
Errors

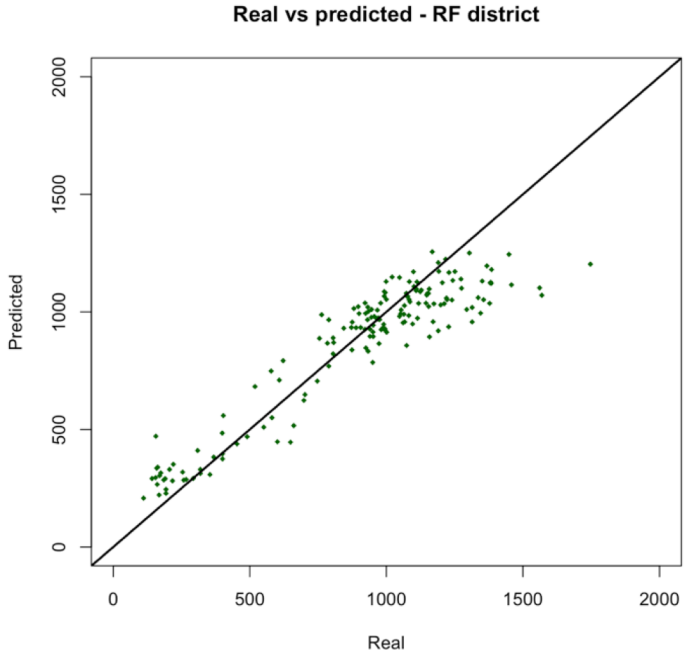

Histogram of errors - NN city
Frequency
0
5
10
15
20
25
30
-600
-400
-200
0
200
Errors

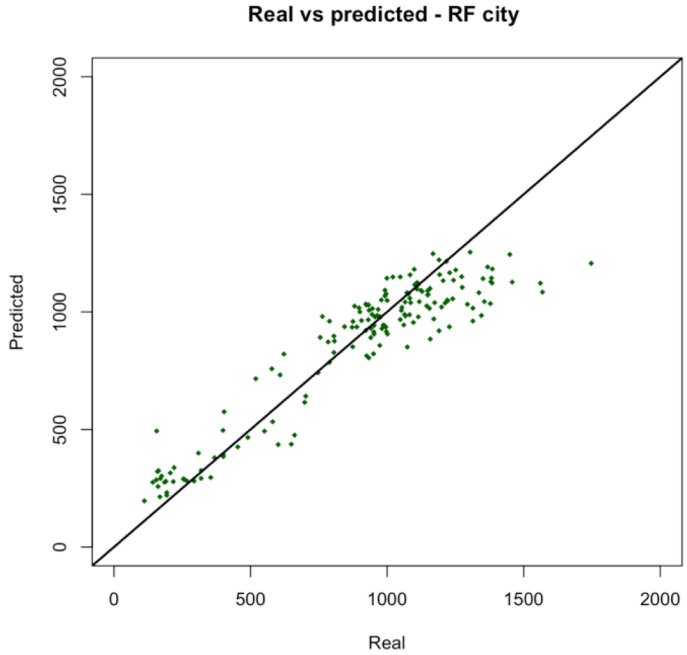

Real vs predicted - NN district
Predicted
0
500
1000
1500
2000
Real

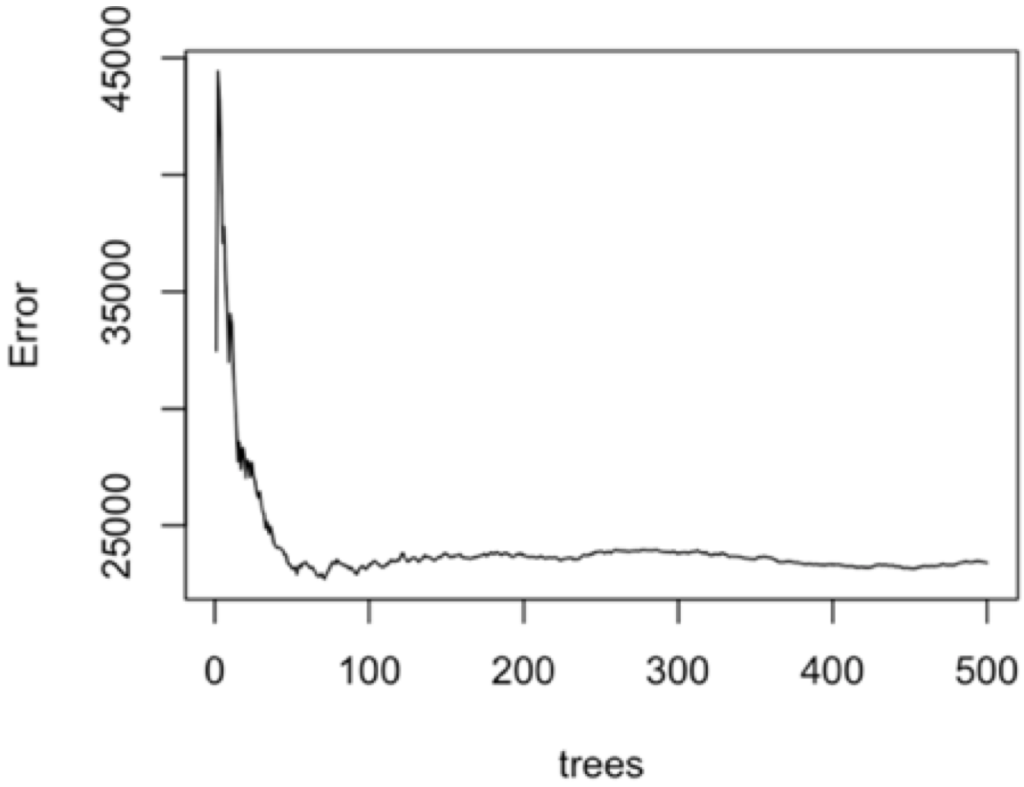

Real vs predicted - NN city
Predicted
0
500
1000
1500
2000
Real


Histogram of errors - RF district
Frequency
0
5
10
15
20
25
30
-400
-200
0
200
Errors


Histogram of errors - RF city
Frequency
0
5
10
15
20
25
30
-400
-200
0
200
Errors


Real vs predicted - RF district
Predicted
0
500
1000
1500
2000
0
500
1000
1500
2000
Real


Real vs predicted - RF city
Predicted
0
500
1000
1500
2000
0
500
1000
1500
2000
Real


Error
45000
35000
25000
0
100
200
300
400
500
trees