UC San Diego

Prediction of Highway Traffic Flow Based on Artificial Intelligence Algorithms Using California Traffic Data

Junseong Lee
Yoonju Cho
Yejin Shin
Seoyoon Choi
Jaegwan Cho

Content

1 Background

2 Previous Work

3 Key Feature

4 Work Flow

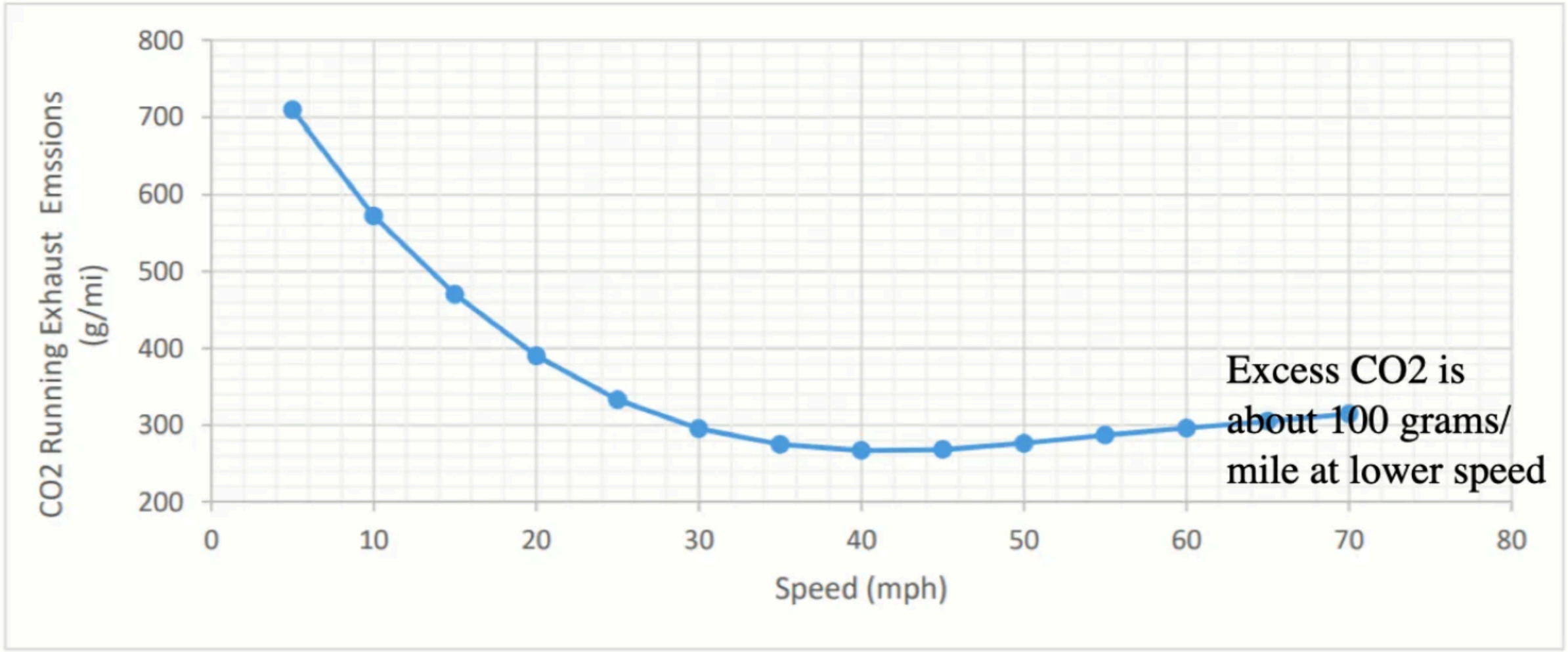
5 Dataset

6 Experiment

7 Result

8 Conclusion

Background



Source: California Air Resource Board

Air pollution and health risks due to vehicle traffic

[Kai Zhang et al., 2013]

Traffic on roads has significantly increased in the U.S. and elsewhere over the past 20 years (Schrank and Lomax, 2007). In many areas, vehicle emissions have become the dominant source of air pollutants, including carbon monoxide (CO), carbon dioxide (CO_2), volatile organic compounds (VOCs) or hydrocarbons (HCs), nitrogen oxides (NO_x), and particulate matter (PM) (Transportation Research Board (TRB), 2002). The increasing severity and duration of traffic congestion have the potential to greatly increase pollutant emissions and to degrade air quality, particularly near large roadways. These emissions contribute to risks of morbidity and mortality for drivers, commuters and individuals living near roadways, as shown by epidemiological studies, evaluations of proposed vehicle emission standards, and environmental impact assessments for specific road projects (World Health Organization (WHO), 2005; Health Effects Institute (HEI), 2010).

America's Ten Busiest States (by interstate vehicle-miles traveled) in 2011

California	84.681 billion
Texas	55.734 billion
Florida	34.689 billion

Source : U.S Department of Transportation

UC San Diego

Time-Gap Based Traffic Model for Vehicular Traffic Flow

[Cho et al., 2014]

The **time** a vehicle needs to maintain to **avoid colliding** with the leading vehicle

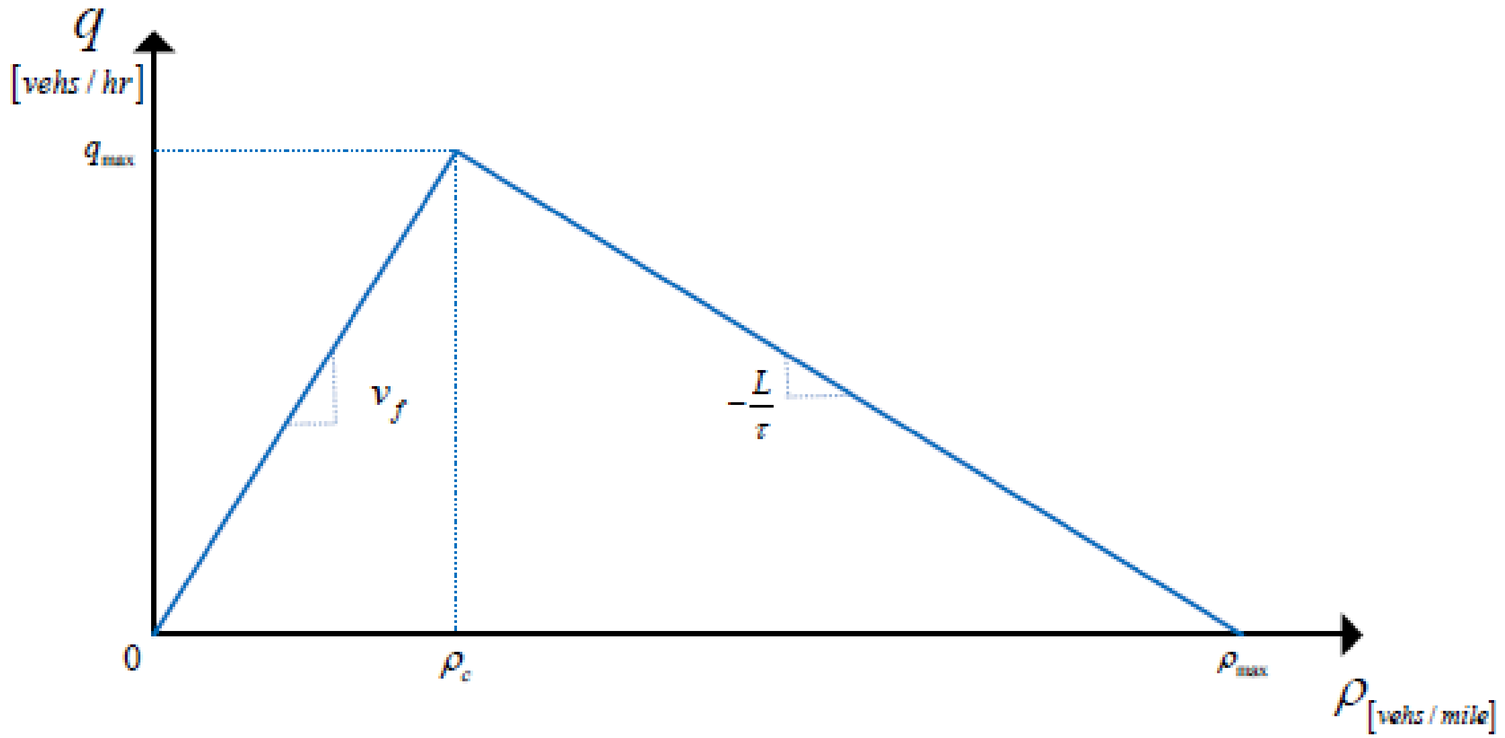
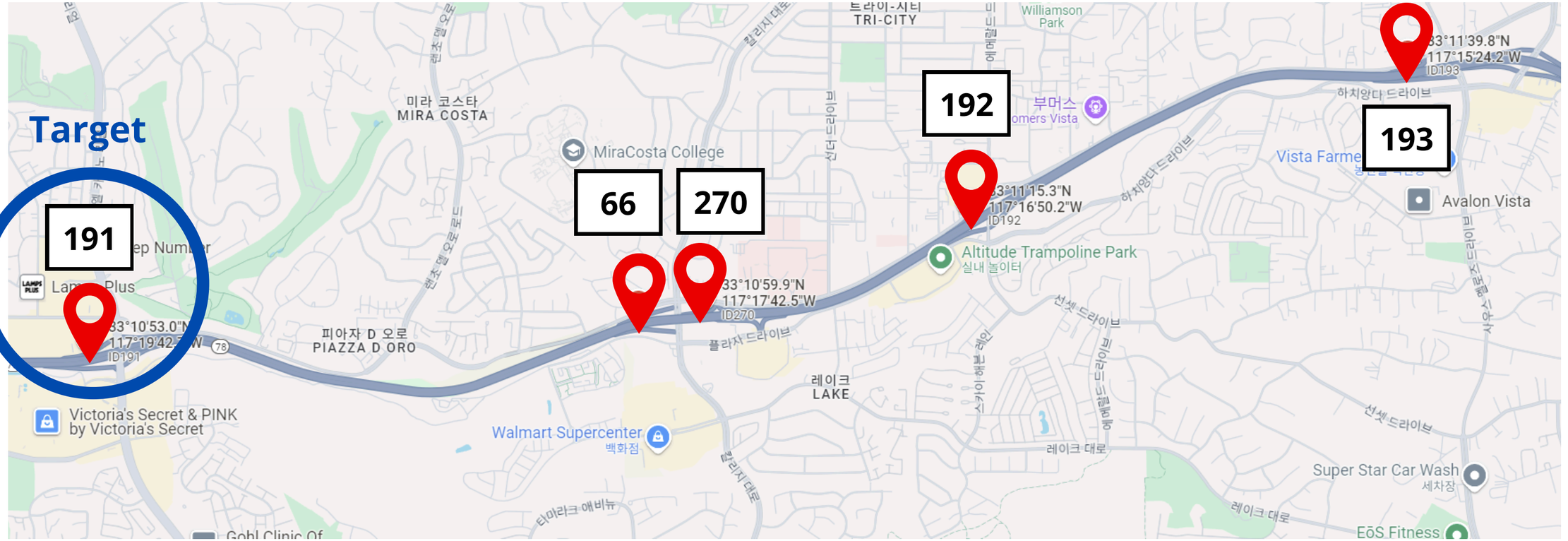


Fig. 1. Traffic Flow on a Road

UC San Diego

Purpose: Predict Traffic Volume of Detector 191, El Camino Real



Westbound

UC San Diego

Previous Research

- Used **only historical data up to 10 minutes** before.
- **Included 470s detector**, which is a **RADAR** detector. However, this may **not be sufficient** for explaining traffic information.

Study on Traffic Flow Prediction on Highways based on Artificial Intelligence Algorithms Using Traffic Data Measured in California, U.S. [Seokjin Choe et al., 2024]

This Study

- Test **different collected time interval**, not just “10 minutes before”.
  - Reconstruct the original 30-second interval data into **1, 2, 5, 10, 15 minute** intervals.
- **Exclude 470s detector** for relialbe prediction.

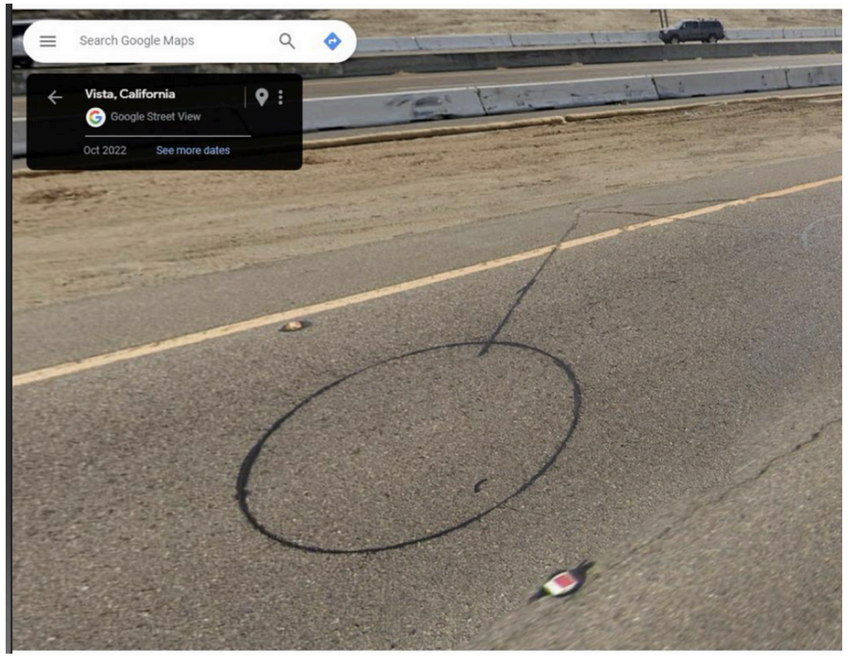
Find the best historical input time period.

Previous Work

Not “Radar” , but “Loop Detector”

Date	Time	Id	Location	Route	Direction	Poll History	MlDetPostr	Fep	Host	Line	Poll Type	Drop	Type
3/1/22	5:00:00	191	El Camino F	78	WB	. 2022-03-0	1.387	5	10.239.88.	135	WAN	8	METER
3/1/22	5:00:00	66	College Blv	78	WB	. 2022-03-0	3.245	5	10.239.88.	135	WAN	7	METER
3/1/22	5:00:00	470	WB @ Colle	78	WB	MM 2022-0	3.2875	5	10.239.88.	135	TCP	8	TMS
3/1/22	5:00:00	270	College Blv	78	WB	. 2022-03-0	3.33	5	10.239.88.	135	WAN	6	METER
3/1/22	5:00:00	471	78 WB E/O	78	WB	MM 2022-0	3.588	5	10.239.88.	135	TCP	2	TMS

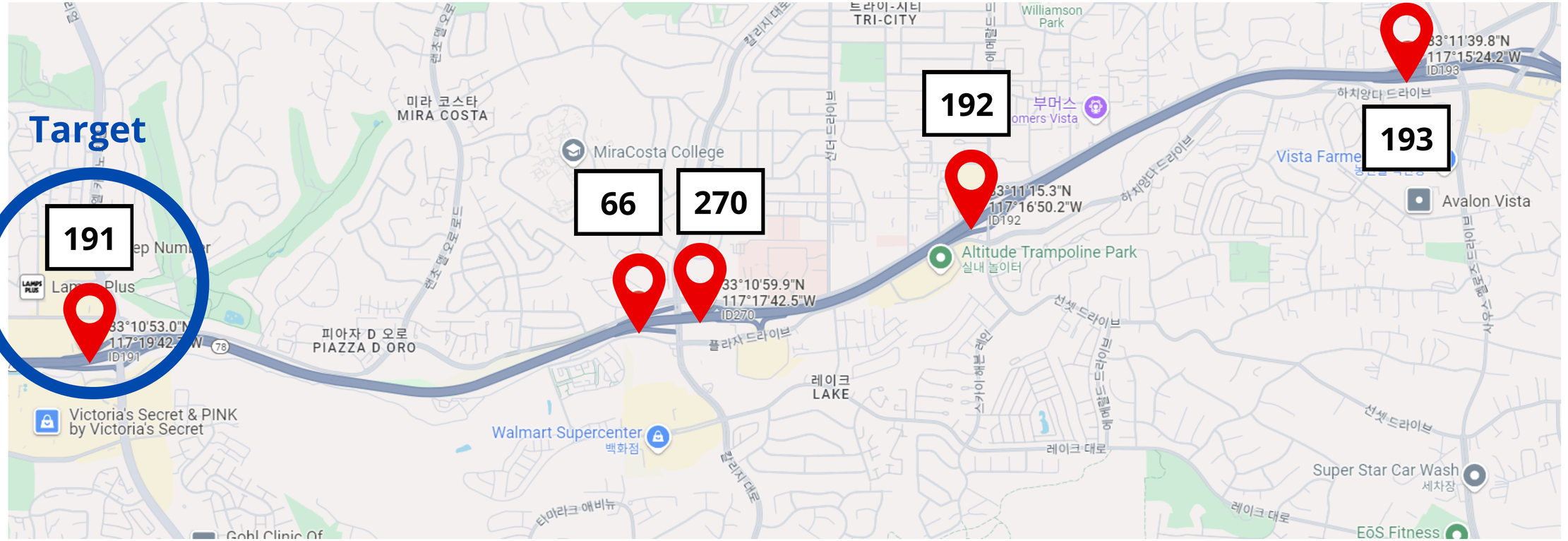
Radar



Loop Detector

UC San Diego

Purpose: Predict Traffic Volume of Detector 191, El Camino Real



Westbound

UC San Diego

Occupancy

The **occupancy count of vehicles** passing through the detector during 30 seconds

Volume

The **number of vehicles** passing through the detector during 30 seconds

30Hz => 30 ticks per second

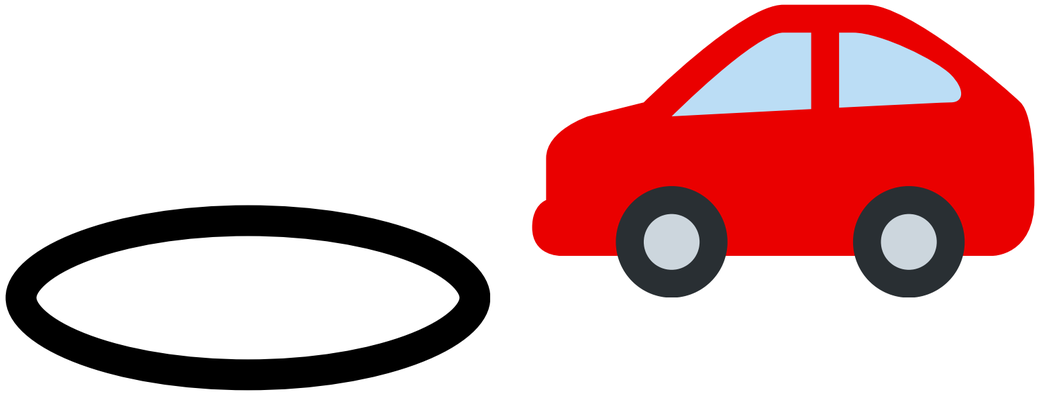
Total **900 ticks** for **30 seconds**

Occupancy = Occupancy Count / 900

Occupancy Count = # of ticks when a car is over the detector

UC San Diego

1st Tick

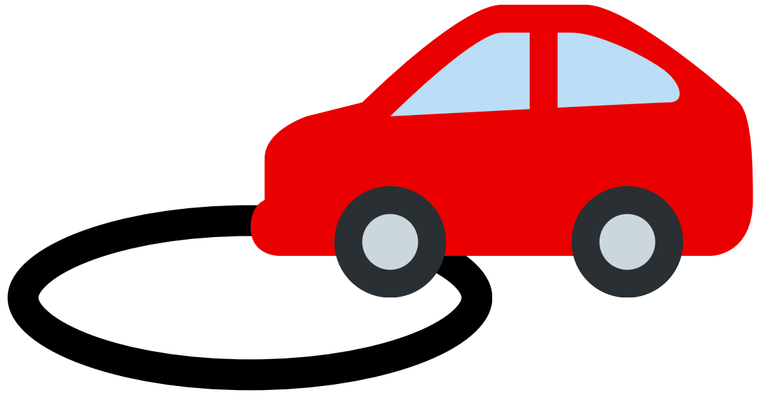


Occ_cnt doesn’t increase!

Occ_cnt = 0

UC San Diego

2nd Tick



Occ_cnt increase by 1

Occ_cnt = 1

UC San Diego

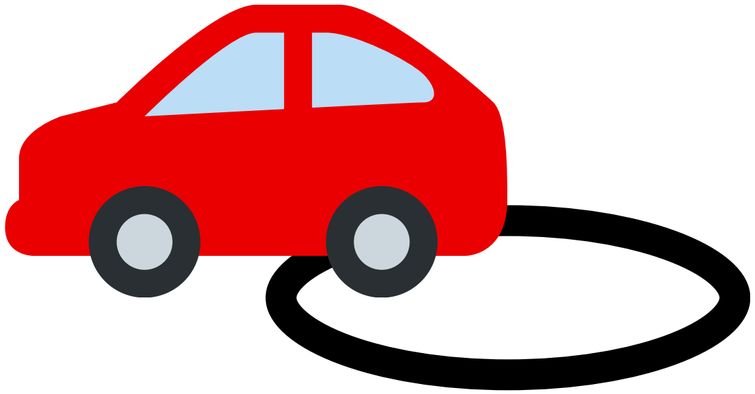
3rd Tick

Occ_cnt increase by 1

Occ_cnt = 2

UC San Diego

4th Tick

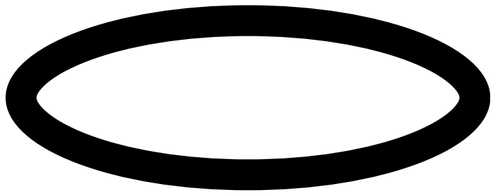


Occ_cnt increase by 1

Occ_cnt = 3

UC San Diego

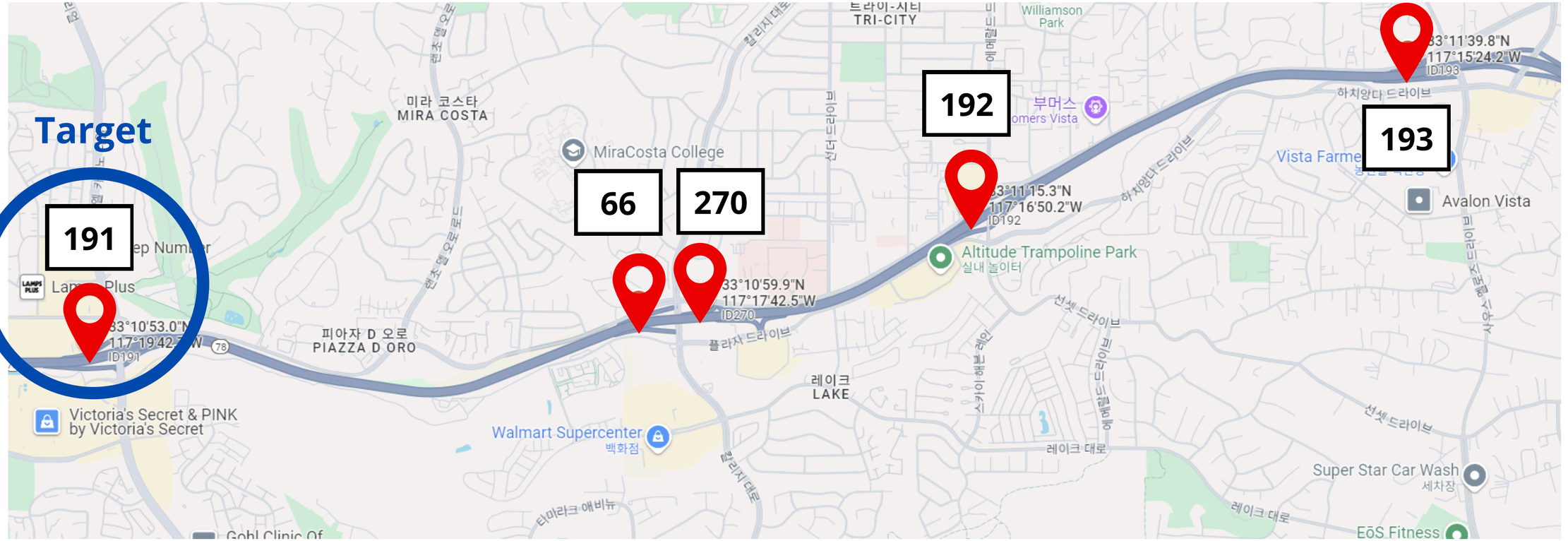
5th Tick



Occ_cnt doesn’t increase!

Occ_cnt = 3

Purpose: Predict Traffic Volume of Detector 191, El Camino Real



Westbound

UC San Diego

1. Original Dataset

- Collected Time Interval 30s

Caltrans: California Department of Transportation

2. New Dataset

3. Model Training & hyperparameter tuning

4. Prediction & Analysis

5. Optimal Collected Time Interval

Work Flow

1. Original Dataset

2. New Dataset

- **Varying Collected Time Interval (30s, 1m, 2m, 5m, 10m, 15m)**

3. Model Training & hyperparameter tuning

4. Prediction & Analysis

5. Optimal Collected Time Interval

Work Flow

UC San Diego

2. New Dataset

1. Original Dataset

3. Model Training & Hyperparameter Tuning

- **Multiple Linear Regression (MLR)**
- **Random Forest (RF)**

4. Prediction & Analysis

5. Optimal Collected Time Interval

Work Flow

UC San Diego

2. New Dataset

1. Original Dataset

4. Prediction & Analysis

- **R^2, RMSE**
- **Scaled RMSE**

3. Model Training & hyperparameter tuning

5. Optimal Collected Time Interval

Work Flow

UC San Diego

2. New Dataset

1. Original Dataset

3. Model Training & hyperparameter tuning

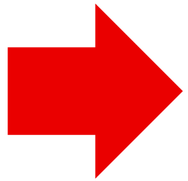
4. Prediction & Analysis

5. Optimal Collected Time Interval

UC San Diego

Original Data from “California Department of Transportation”

30s Data



{30s, 1m, 2m, 5m, 10m, 15m}:
Collected Time Interval

Our Data

- **30s Data**
- **1m Data**
- **2m Data**
- **5m Data**
- **10m Data**
- **15m Data**

UC San Diego

30s of Collected Time Interval - Basic Dataset

Month	Time	191_Vol	191_Occ	66_Vol	66_Occ	270_Vol	270_Occ	192_Vol	192_Occ	193_Vol	193_Occ
7	0:00:00	5	39	7	44	6	45	4	45	4	24
7	0:00:30	1	7	1	7	4	30	4	29	6	38
7	0:01:00	2	16	2	51	4	67	7	44	4	25
7	0:01:30	5	32	6	39	5	39	3	18	6	46
7	0:02:00	3	23	6	34	6	43	1	8	3	17
7	0:02:30	2	16	2	10	2	15	8	57	2	10
7	0:03:00	4	68	3	17	2	14	3	16	1	7
7	0:03:30	5	38	6	33	5	36	2	11	4	23

UC San Diego

“Collected Time Interval (T)”
30s
(Original Dataset)

30s

Time_period_30s				
ID	Month	Time	191_Occ	191_Vol
Row0	7	00:00:00	5	39
Row1	7	00:00:30	1	7
Row2	7	00:01:00	2	16
Row3	7	00:01:30	5	32

Data merge

“Collected Time Interval (T)”
1min

1m

ID	Month	Time	191_Occ	191_Vol
Row0	7	00:00:00	6	46
Row1	7	00:01:00	7	48

UC San Diego

1m of Collected Time Interval

Month	Time	191_Vol	191_Occ	66_Vol	66_Occ	270_Vol	270_Occ	192_Vol	192_Occ	193_Vol	193_Occ
7	0:00:00	6	46	8	51	10	75	8	74	10	62
7	0:01:00	7	48	8	90	9	106	10	62	10	71
7	0:02:00	5	39	8	44	8	58	9	65	5	27
7	0:03:00	9	106	9	50	7	50	5	27	5	30
7	0:04:00	3	18	3	17	4	37	1	5	4	25

UC San Diego

2m of Collected Time Interval

Month	Time	191_Vol	191_Occ	66_Vol	66_Occ	270_Vol	270_Occ	192_Vol	192_Occ	193_Vol	193_Occ
7	0:00:00	13	94	16	141	19	181	18	136	20	133
7	0:02:00	14	145	17	94	15	108	14	92	10	57
7	0:04:00	11	76	5	27	5	43	6	36	6	42
7	0:06:00	9	80	8	50	6	42	6	40	12	116
7	0:08:00	6	44	10	54	11	81	15	132	20	115

UC San Diego

5m of Collected Time Interval

Month	Time	191_Vol	191_Occ	66_Vol	66_Occ	270_Vol	270_Occ	192_Vol	192_Occ	193_Vol	193_Occ
7	00:00:00	30	257	36	252	38	326	33	233	34	215
7	00:05:00	23	182	20	114	18	129	26	203	34	248
7	00:10:00	32	274	32	236	33	284	43	312	46	324
7	00:15:00	27	226	28	202	34	279	38	248	49	343
7	00:20:00	35	253	32	223	42	331	42	294	37	263

UC San Diego

10m of Collected Time Interval

Month	Time	191_Vol	191_Occ	66_Vol	66_Occ	270_Vol	270_Occ	192_Vol	192_Occ	193_Vol	193_Occ
7	00:00:00	53	439	56	366	56	455	59	436	68	23.33452
7	00:10:00	59	500	60	438	67	563	81	560	95	13.43503
7	00:20:00	66	506	56	355	68	522	68	465	66	60.10408
7	00:30:00	44	310	37	266	49	398	52	379	64	41.7193
7	00:40:00	48	467	42	340	47	397	43	317	53	28.99138

UC San Diego

15m of Collected Time Interval

Month	Time	191_Vol	191_Occ	66_Vol	66_Occ	270_Vol	270_Occ	192_Vol	192_Occ	193_Vol	193_Occ
7	00:00:00	85	713	88	602	89	739	102	748	114	55.89574
7	00:15:00	93	732	84	557	102	801	106	713	115	82.51263
7	00:30:00	66	523	59	469	74	622	76	562	94	35.92121
7	00:45:00	64	541	48	315	62	478	60	431	75	13.6504
7	01:00:00	49	374	54	365	54	420	65	460	75	18.50225

Table 1: Percentage Of Missing Value Imputed

Month	Total Values	Missing Values	Percentage (%)
7	89280	401	0.449
8	89280	211	0.236
9	86400	45	0.052
10	86400	42	0.049
11	74880	4191	5.596

Total Percentage of Missing Value = **1.14%**

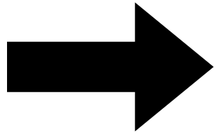
=> Imputed Missing Values with Linear Interpolation
(Test with 30s of collected time interval)

UC San Diego

Input

- **Month**
- **Time**
- **Detector ID 191_Occ_T**

(T = 30s, 1m, 2m, 5m, 10m, 15m)



Output

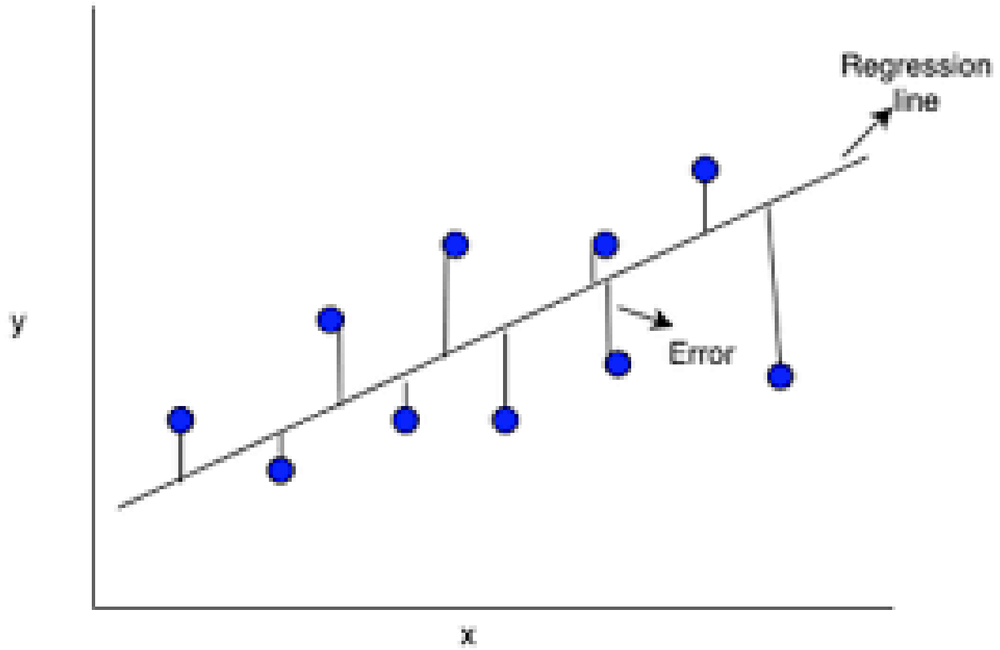
Detector ID 191_Vol_T

(T = 30s, 1m, 2m, 5m, 10m, 15m)

T = Collected Time Interval

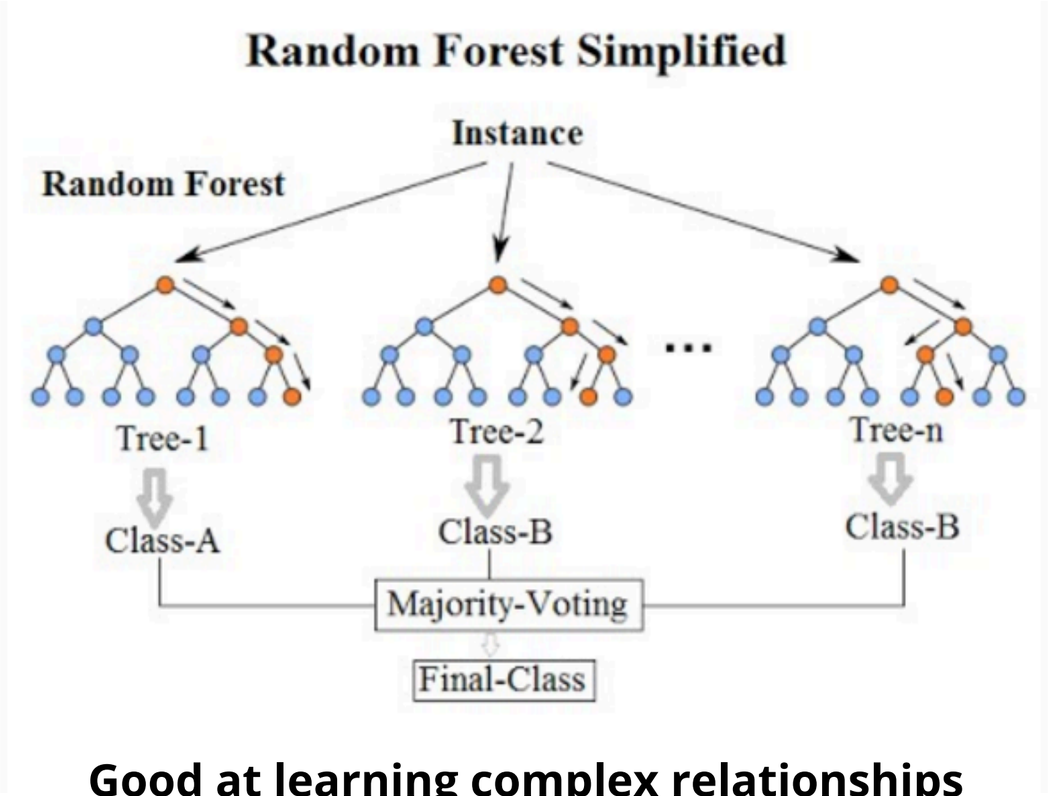
UC San Diego

1. MLR



Intuitively understand the impact of each variable on traffic volume through coefficients

2. RF



Good at learning complex relationships between multiple factors

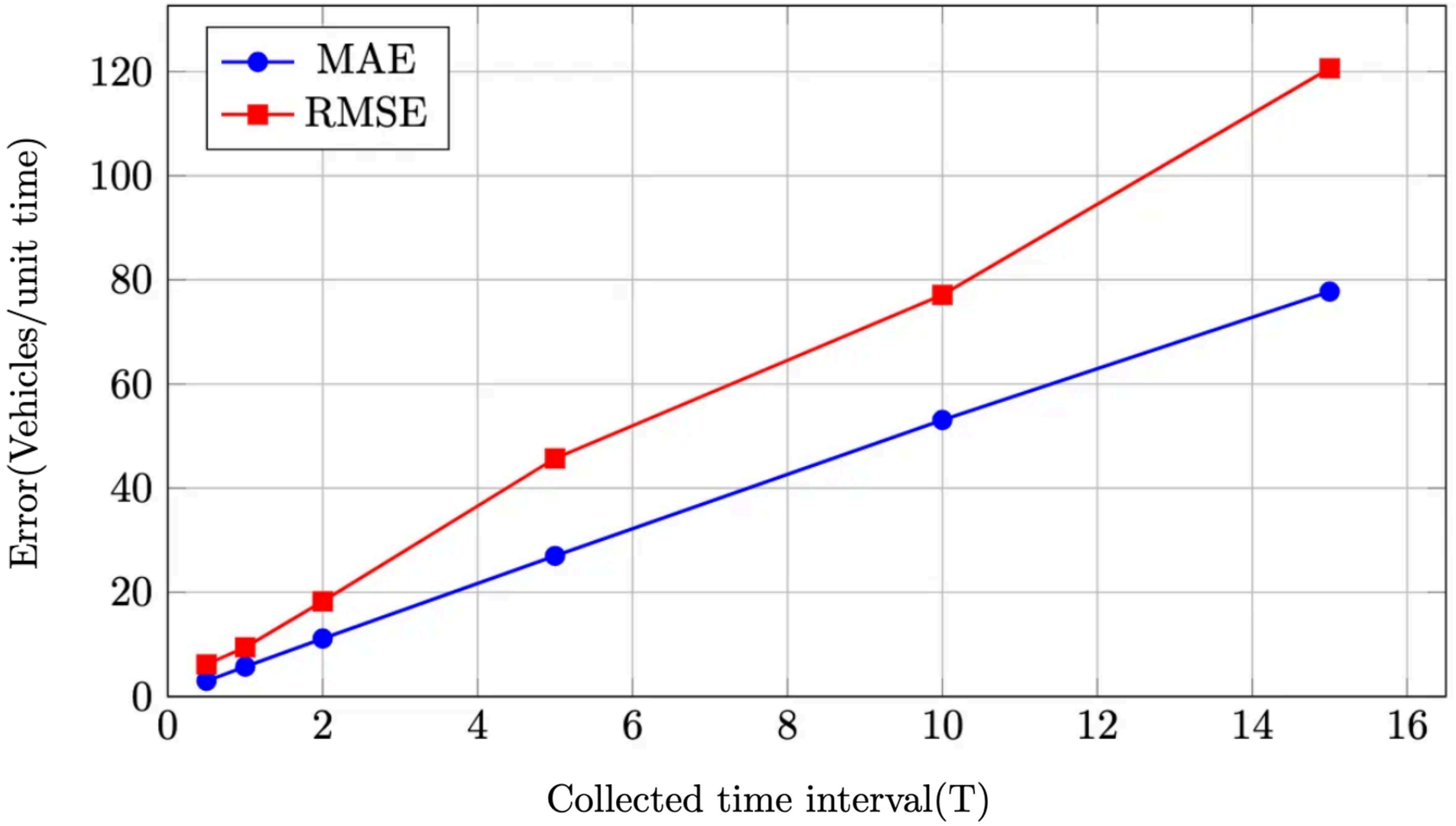
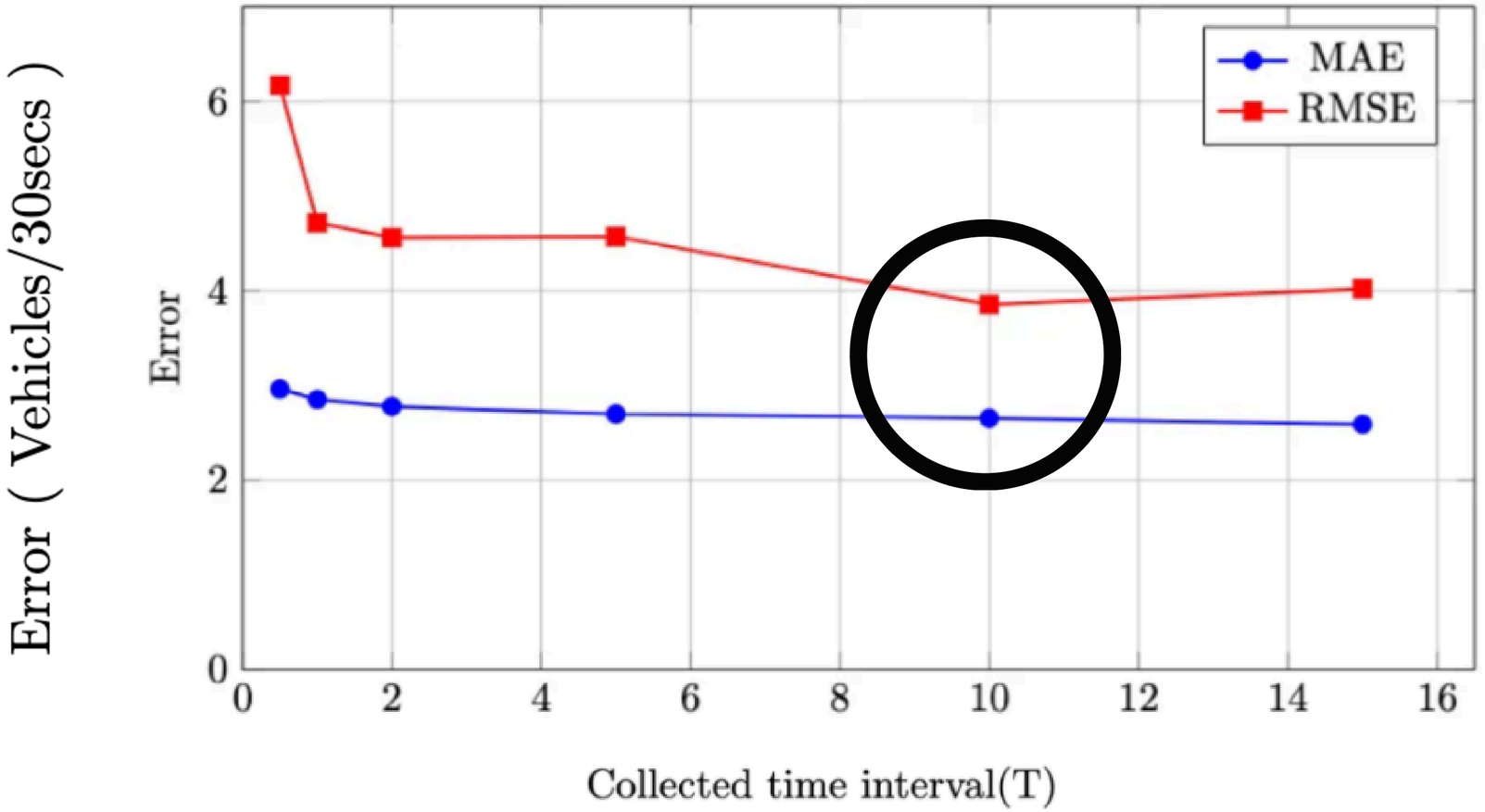


Figure 1: MAE and RMSE over different time intervals

Figure 2: Scaled MAE and RMSE over different time intervals



$$Error_{scaled} = \frac{Error_{origin} * 0.5min}{T} \quad (T = \{0.5,\ 1,\ 2,\ 5,\ 10,\ 15\}\, min)$$

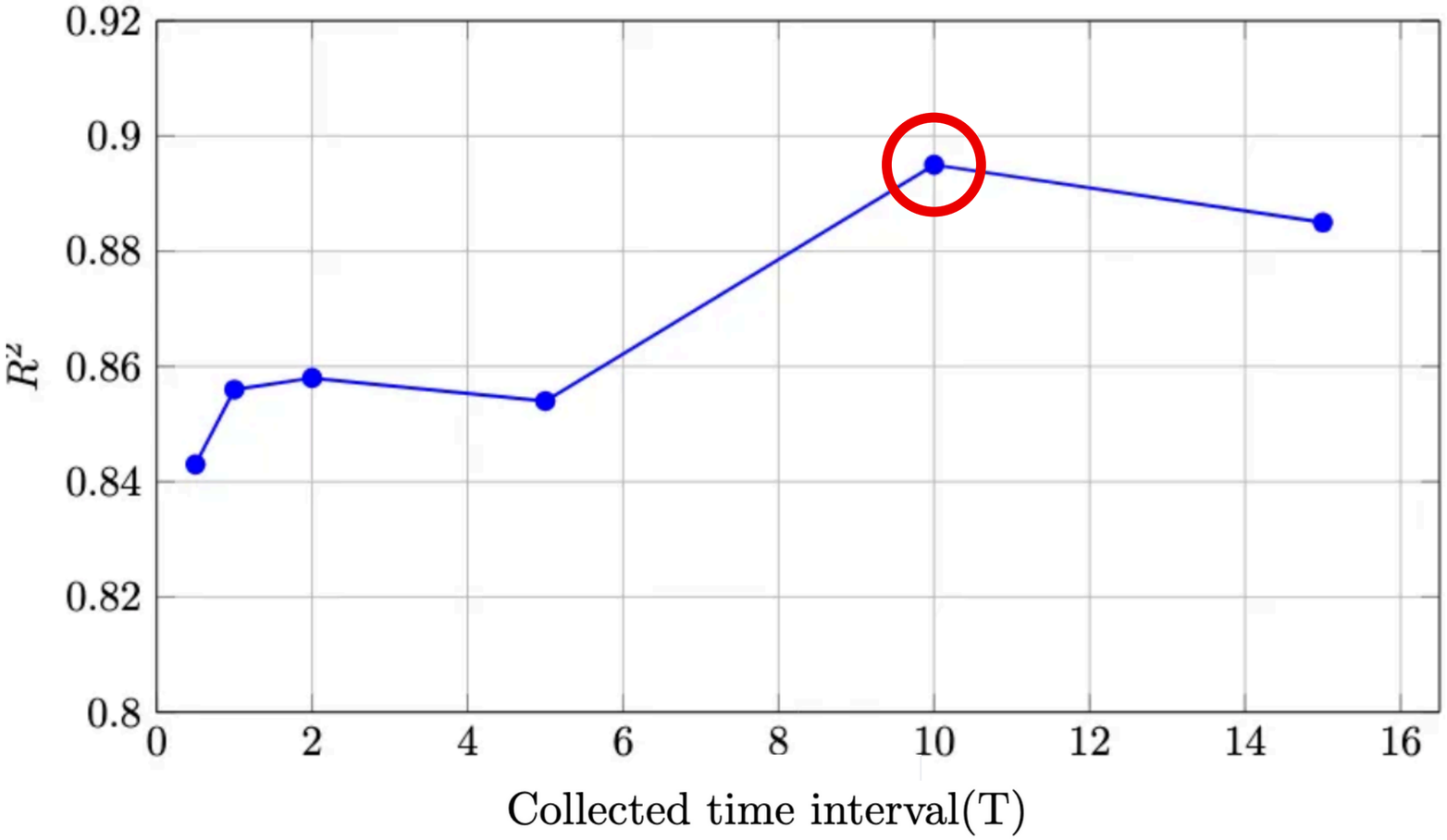


Figure 3: R^2 value over different time period

Optimal Collected Time Interval : 10m

Result - Random Forest

Time (min)	Hyperparameters		Performance Metrics		
	Min_child_node	Tree_depth	R^2	RMSE	Scaled_RMSE
15	7	11	0.991	33.658	1.112
10	7	22	0.989	23.800	1.195
5	9	25	0.987	13.101	1.310
2	15	25	0.984	6.063	1.515
1	4	23	0.978	3.687	1.843
0.5	11	22	0.969	2.228	2.228

Table 2: RF Model Optimal Hyperparameters and Performance Metrics

Min_child_node : 3~20

Tree_depth : 2~25

Brute Force

Result - Random Forest

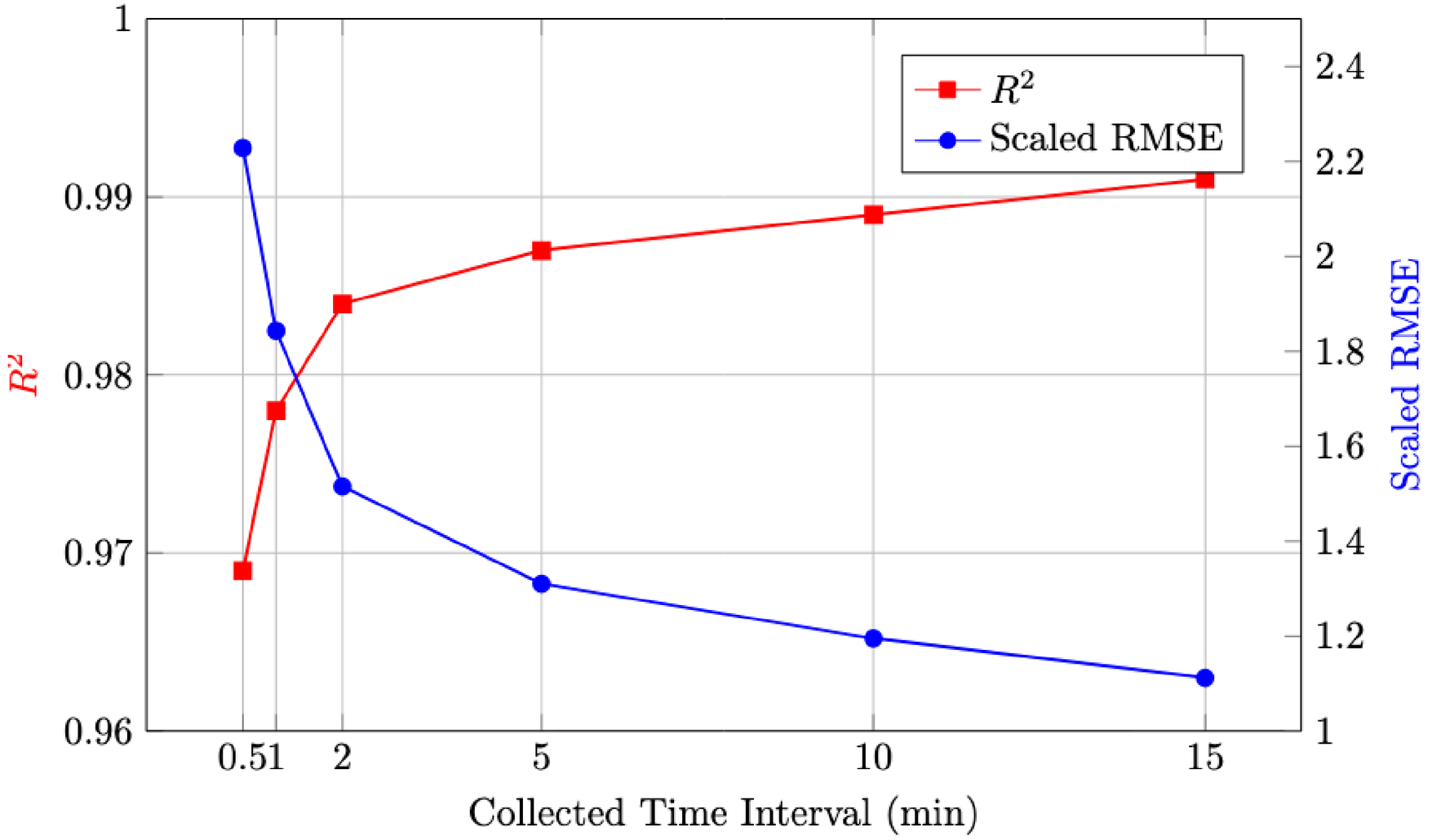


Figure 4: RF Performance Metrics by Different Time Interval

Conclusion

	Detector ID 191	Multiple Detector IDs
Optimal Time Interval(T)	10 min	Not Enough Interval to find optimal time interval

Proposed Alternatives

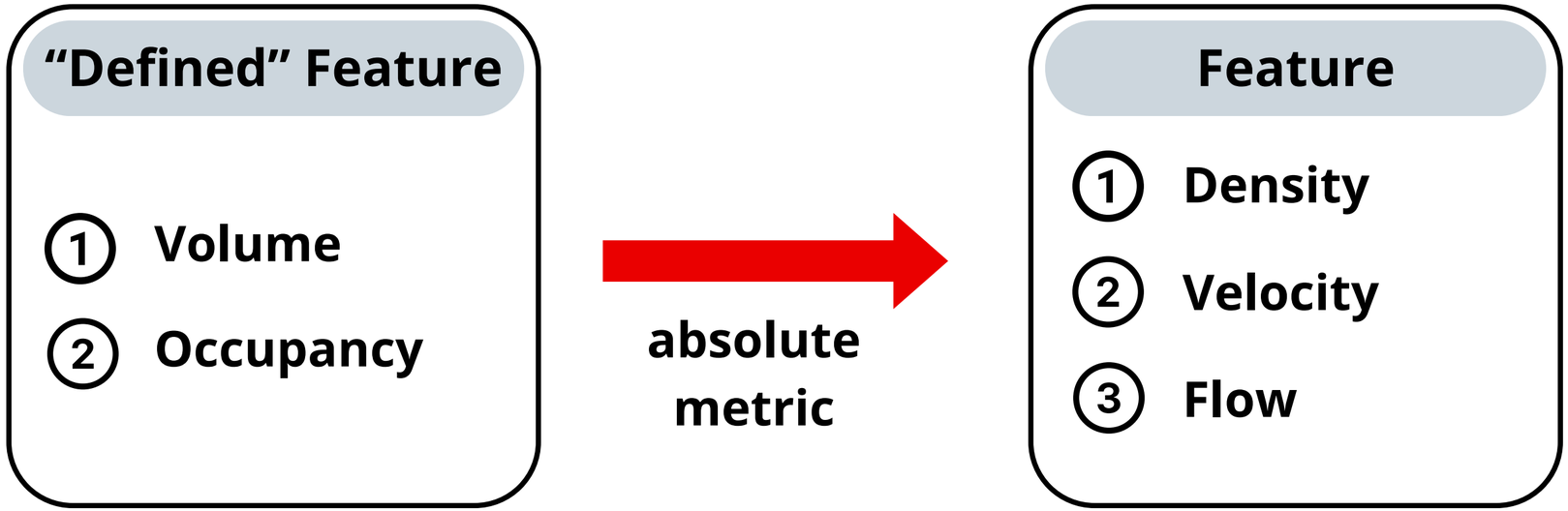
1. Extending the **Time Period** Further

2. Utilizing **LSTM** for Sequential Time Consideration

Conclusion

Defined Feature

Feature we want to measure → $E = kT$



Conclusion

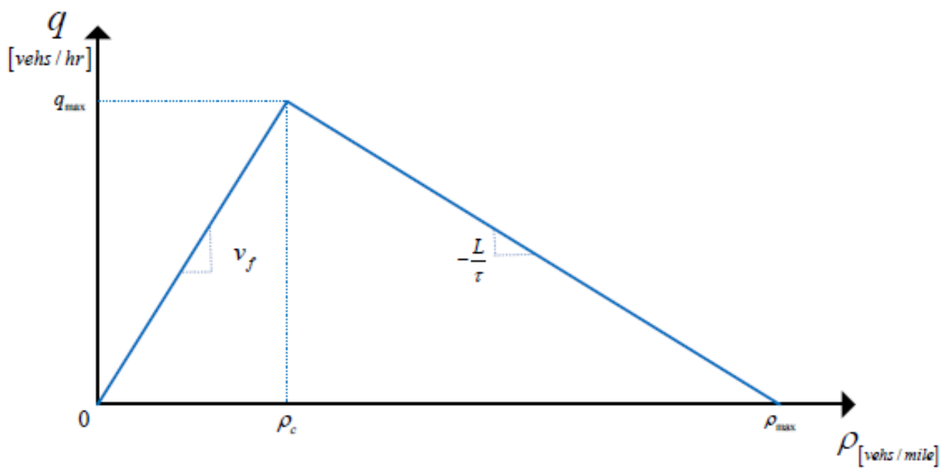
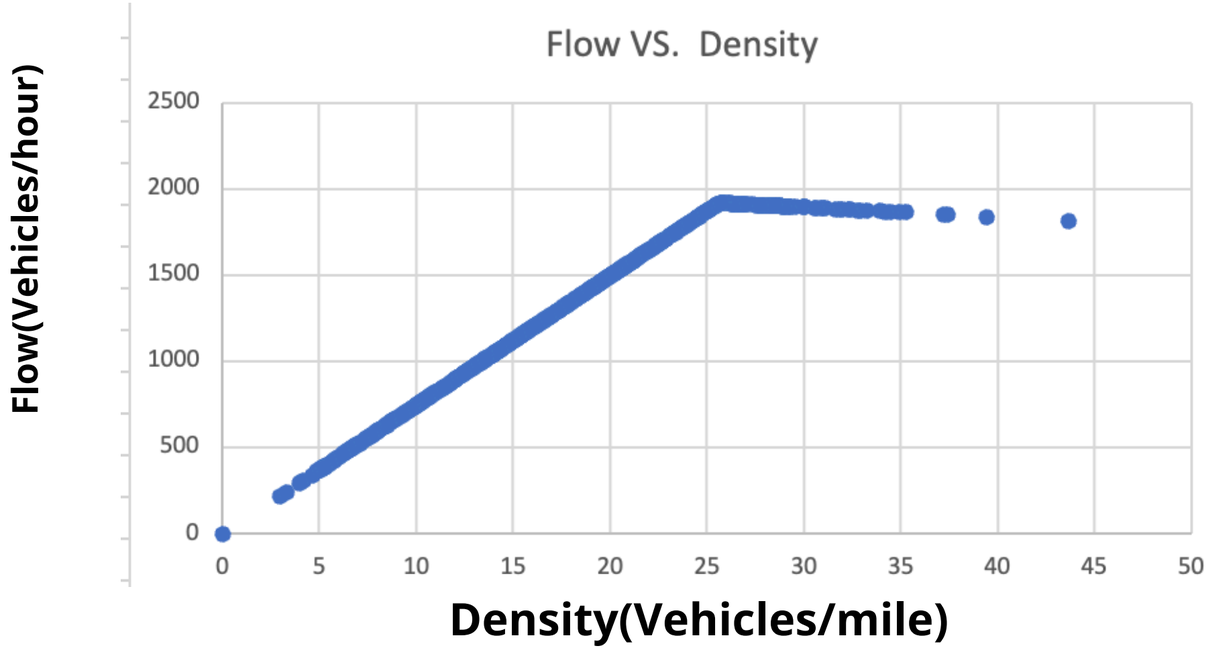


Fig. 1. Traffic Flow on a Road

Paper's Fundamental Traffic Flow Diagram



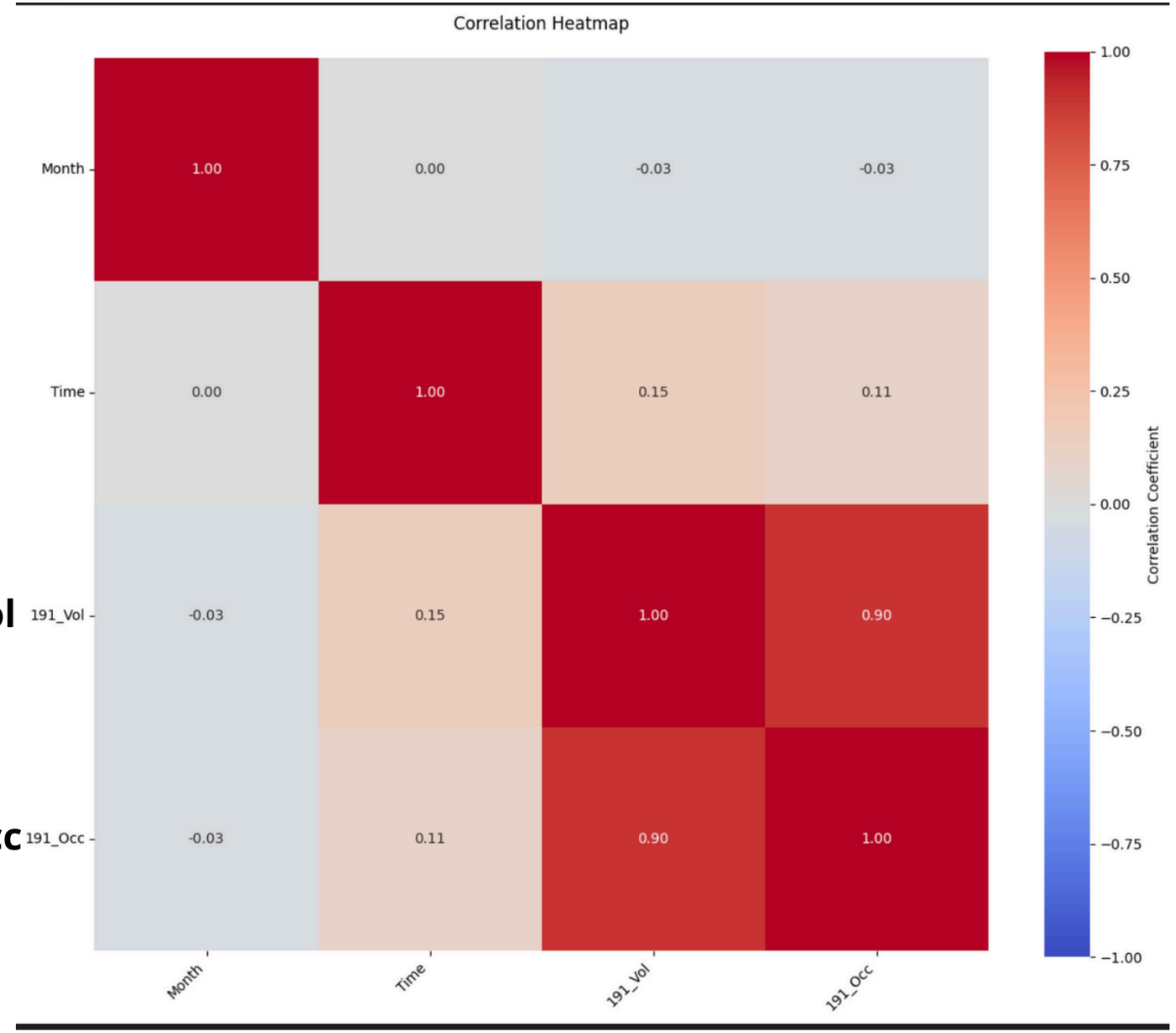
Our Fundamental Traffic Flow Diagram Using Predicted Vol and Occ

Thank You

Table 2: Missing Value Handling

Method	R^2	MAE	RMSE
Linear interpolation	0.770	4.729	6.315
Maximum	0.772	4.752	6.349
Mean	0.771	4.727	6.314
Median	0.771	4.728	6.314
Minimum	0.771	4.729	6.321
Next value	0.771	4.732	6.320
Previous value	0.771	4.732	6.319

Group 1



191_vol

191_Occ

Group 1

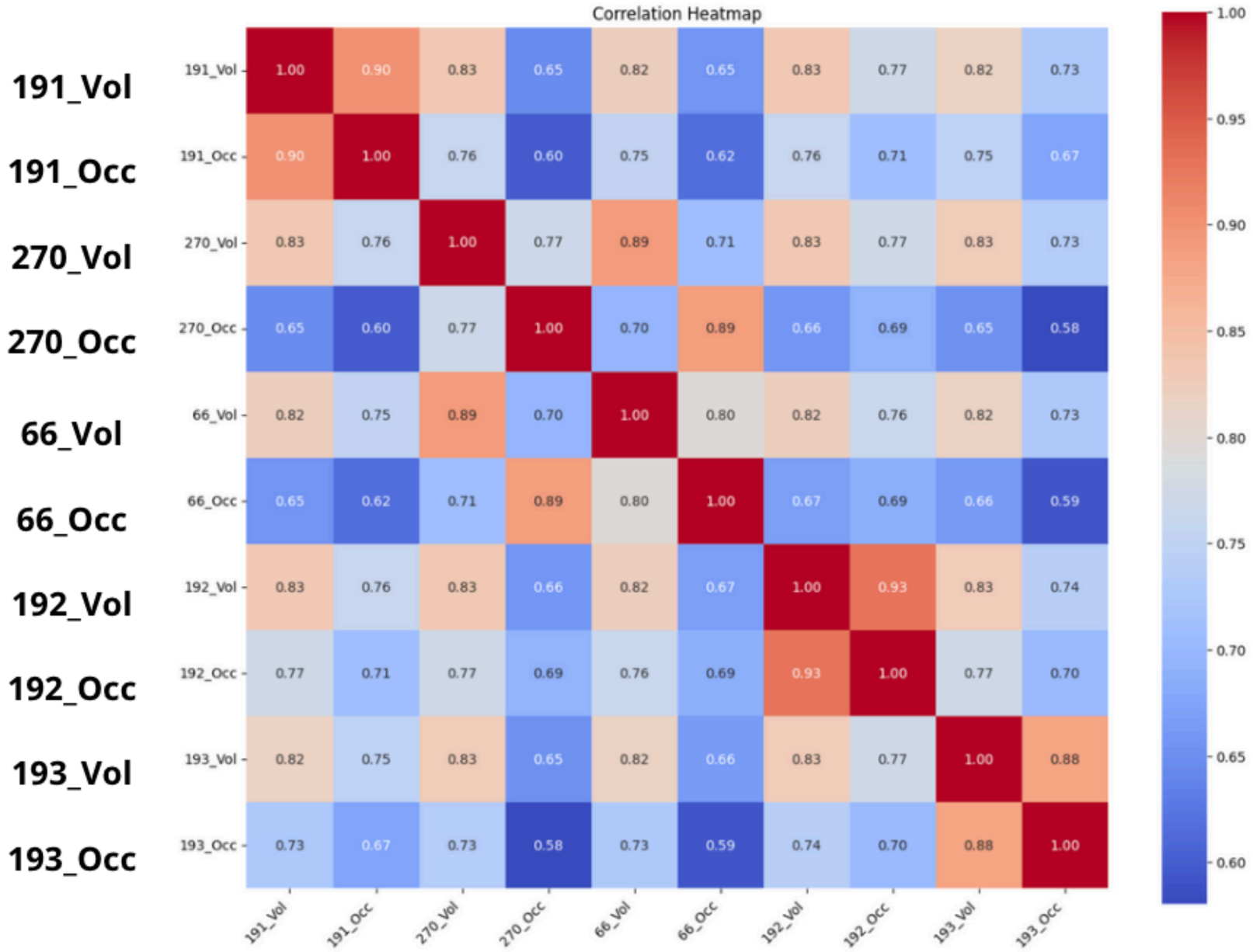

Correlation Heatmap
191_Vol
191_Occ
270_Vol
270_Occ
66_Vol
66_Occ
192_Vol
192_Occ
193_Vol
193_Occ
1.00 0.90 0.83 0.65 0.82 0.65 0.83 0.77 0.82 0.73
0.90 1.00 0.76 0.60 0.75 0.62 0.76 0.71 0.75 0.67
0.83 0.76 1.00 0.77 0.89 0.71 0.83 0.77 0.83 0.73
0.65 0.60 0.77 1.00 0.70 0.89 0.66 0.69 0.65 0.58
0.82 0.75 0.89 0.70 1.00 0.80 0.82 0.76 0.82 0.73
0.65 0.62 0.71 0.89 0.80 1.00 0.67 0.69 0.66 0.59
0.83 0.76 0.83 0.66 0.82 0.67 1.00 0.93 0.83 0.74
0.77 0.71 0.77 0.69 0.76 0.69 0.93 1.00 0.77 0.70
0.82 0.75 0.83 0.65 0.82 0.66 0.83 0.77 1.00 0.88
0.73 0.67 0.73 0.58 0.73 0.59 0.74 0.70 0.88 1.00
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60

Table 4: Performance comparison by Number of Upstream detectors

# of Upstream Detector	MSE	MAE	R^2
1	58.080	5.871	0.666
2	49.747	5.347	0.714
3	42.675	4.906	0.755
4	39.871	4.727	0.771