TRANSPORTATION LEARNING NETWORK

A partnership with MDT•NDDOT•SDDOT•WYDOT and the Mountain-Plains Consortium Universities

Welcome!

Inferencing Traffic Volume using Data-Driven Machine Learning and Graph Theory

Presented by:

Xiaoyue Cathy Liu, Ph.D., P.E.

Our partners:

MOUNTAIN-PLAINS CONSORTIUM REGION 8 · MONTANA MDT DEPARTMENT OF TRANSPORTATION · NORTH Dakota Be Legendary. | Transportation · South Dakota DOT Connecting South Dakota and the Nation · WYOMING TRANSPORTATION DEPARTMENT · NDSU UPPER GREAT PLAINS TRANSPORTATION INSTITUTE TRANSPORTATION LEARNING NETWORK

Project Briefing

1

- MPC-543: Big Transportation Data Analytics

- Jointly sponsored by Mountain-Plains Consortium and the Utah Department of Transportation (UDOT)

- Motivated by the fact that the annual traffic count program represents a significant cost to UDOT, while exposing UDOT staff to the dangers inherent to being exposed to traffic

Introduction

Data

Spatial Dependency

Models

Introduction

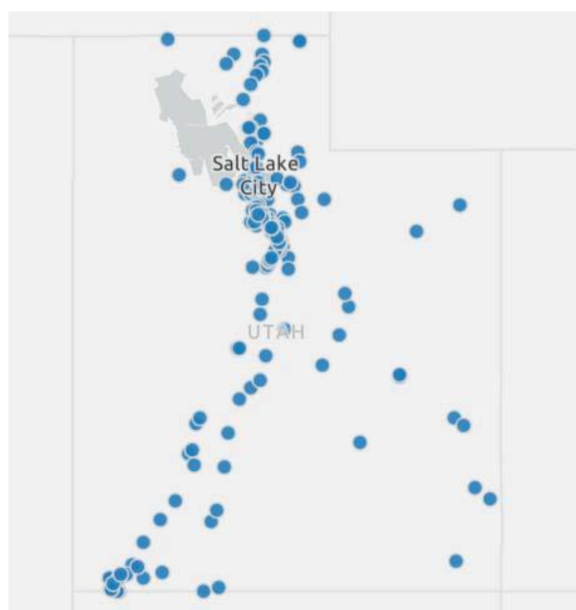
Data

Spatial Dependency

Models

1 Introduction

- Traffic volume data is crucial in many applications, such as transportation operation analysis, congestion management and accident prevention.
- However, deployment of traffic count devices on a large-scale network can be extremely expensive and impractical.
- Agencies are exploring methods to partially or fully supplement the short-duration traffic count program and, in doing so, reduce the effort, cost, and staff exposure of traffic count program.



1 Introduction

How to accurately estimate/predict traffic volume without massive sensor deployment?

Two branches of volume estimation problem:

- Future traffic volume prediction at locations equipped with traffic sensors

- Generally involves estimation of immediate future volume at the same locations within a short time period based on historical information.
- Commonly used methods:
  Auto-regressive integrated moving average (ARIMA) and its variants (e.g. SARIMA)

1 Introduction

Two branches of volume estimation problem:

- Traffic volume estimation: spatial traffic volume prediction

- Spatial prediction usually aims at estimating traffic volumes at locations without sensors.
- Commonly used methods:
  Factoring method
  Regression model
  Machine learning model

1 Introduction

Two branches of volume estimation problem:

- Traffic volume estimation: spatial traffic volume prediction

- Spatial prediction usually aims at estimating traffic volumes at locations without sensors.

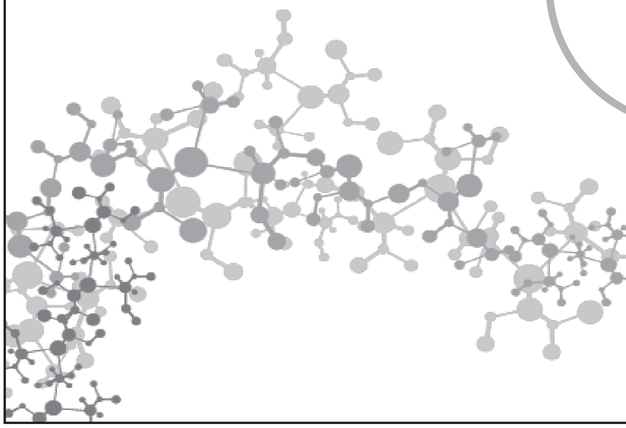
Research oftentimes omits the fact that traffic volume within the road network could be spatially correlated

Commonly used methods:
Kriging methods
Graph-based approaches

1 Objectives

- Explore ML techniques to predict hourly traffic volumes using features that are associated with the variation of traffic volume, and compare against conventional methods

- Enhance the prediction accuracy by exploring methods to quantify the spatial dependency among road segments



Introduction

Data

Spatial dependency

Models

2 Data Description

Traffic count observations:

From 100 CCSs distributed in Utah state

Span from Jan to Dec in 2017

Recorded by hours

Hourly counts in a single direction

CCS	Date	Dir	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	Total
-0301	1/1/2017	Pos	373	302	180	150	149	298	586	941	1,116	1,246	1,548	1,832	2,094	2,151	2,097	1,949	1,680	1,272	1,045	793	669	479	429	233	23,612
-0301	1/1/2017	Neg	503	483	363	206	177	262	373	550	667	831	1,105	1,312	1,556	1,668	1,817	2,164	2,269	2,096	1,628	942	810	661	551	385	23,379
-0301	1/2/2017	Pos	136	120	85	86	166	362	780	1,417	1,870	1,694	1,819	1,912	1,902	1,940	1,982	1,979	1,488	1,289	891	684	519	441	336	188	24,086
-0301	1/2/2017	Neg	218	122	80	95	173	276	528	708	985	1,108	1,211	1,555	1,583	1,710	1,935	2,287	1,971	2,099	1,875	926	644	524	419	283	23,315
-0301	1/3/2017	Pos	115	97	65	109	162	470	1,020	1,697	1,947	1,701	1,383	1,380	1,276	1,467	1,770	2,081	2,377	2,313	1,560	1,003	790	629	479	237	26,128
-0301	1/3/2017	Neg	168	106	80	77	168	350	836	1,447	1,639	1,409	1,197	1,257	1,339	1,470	1,548	1,882	2,045	2,373	1,693	956	744	612	485	318	24,199
-0301	1/4/2017	Pos	149	98	89	116	172	568	1,370	2,164	2,455	1,818	1,511	1,430	1,345	1,355	1,636	1,801	2,109	2,281	1,405	806	570	412	317	147	26,124
-0301	1/4/2017	Neg	169	113	92	90	155	356	767	1,466	1,707	1,159	1,097	1,195	1,227	1,092	1,235	1,612	1,930	2,074	1,320	722	573	467	353	234	21,205
-0301	1/5/2017	Pos	133	78	88	94	124	363	927	1,501	1,941	1,640	1,207	1,278	1,264	1,366	1,518	1,868	2,096	2,342	1,633	954	829	708	471	339	24,762
-0301	1/5/2017	Neg	173	89	56	63	86	234	670	1,348	1,671	1,197	975	1,071	1,174	1,241	1,381	1,716	2,060	2,125	1,783	921	763	688	493	449	22,427
-0301	1/6/2017	Pos	254	158	122	109	169	462	1,269	2,081	2,550	2,135	1,945	1,774	1,893	1,792	2,123	2,433	2,718	2,717	2,087	1,357	1,179	1,034	876	450	33,687
-0301	1/6/2017	Neg	232	145	114	99	121	309	823	1,478	1,787	1,480	1,383	1,477	1,667	1,642	1,728	2,292	2,509	2,603	2,163	1,179	914	838	791	549	28,323
-0301	1/7/2017	Pos	255	188	108	118	174	414	929	1,627	2,006	1,798	1,817	1,895	1,976	2,062	2,102	2,219	1,912	1,744	1,346	961	747	655	580	339	27,972
-0301	1/7/2017	Neg	300	256	217	129	121	221	489	761	922	1,138	1,359	1,575	1,675	1,806	1,991	2,303	2,312	2,308	1,434	924	743	746	870	559	25,159
-0301	1/8/2017	Pos	186	125	88	95	150	298	547	915	1,157	1,100	1,383	1,442	1,411	1,350	1,574	1,491	1,267	1,150	883	700	562	417	302	224	18,817
-0301	1/8/2017	Neg	267	149	110	90	122	208	437	632	595	738	894	1,007	1,208	1,260	1,421	1,694	1,613	1,431	1,046	734	557	441	441	296	17,391
-0301	1/9/2017	Pos	161	106	100	87	176	541	1,309	2,238	2,303	1,843	1,509	1,502	1,502	1,475	1,707	1,846	2,161	2,000	1,254	715	597	439	313	205	26,089
-0301	1/9/2017	Neg	162	96	88	83	156	441	1,003	1,781	2,034	1,429	1,412	1,452	1,412	1,440	1,485	1,659	1,462	1,451	1,771	644	425	415	283	319	22,903
-0301	1/11/2017	Pos	154	70	92	113	179	550	1,257	2,039	2,042	1,609	1,259	1,178	1,294	1,348	1,506	1,748	1,859	1,926	1,320	878	783	621	438	264	24,527
-0301	1/11/2017	Neg	276	135	88	53	99	307	740	1,170	2,057	1,212	889	1,108	1,050	1,133	1,215	1,629	1,662	1,664	1,169	823	534	461	431	286	20,191
-0301	1/12/2017	Pos	211	93	75	85	166	520	1,263	2,313	2,530	1,804	1,505	1,531	1,504	1,688	1,933	2,286	2,539	2,647	1,906	1,142	1,173	940	771	410	31,035
-0301	1/12/2017	Neg	157	140	98	77	95	301	816	1,625	1,860	1,323	1,058	1,186	1,166	1,315	1,346	1,875	1,869	2,241	1,907	1,022	862	833	721	625	24,518
-0301	1/13/2017	Pos	418	205	120	109	195	513	1,384	2,406	2,822	2,239	2,015	2,057	2,111	2,072	2,346	2,718	3,088	3,122	2,276	1,582	1,383	1,113	948	754	37,996
-0301	1/13/2017	Neg	348	265	163	89	130	296	800	1,477	1,750	1,585	1,515	1,503	1,722	1,836	1,983	2,372	2,758	2,840	2,233	1,278	981	1,110	826	730	30,590
-0301	1/14/2017	Pos	451	212	106	132	173	445	1,031	1,755	2,009	2,059	2,051	2,255	2,264	2,380	2,435	2,490	2,347	2,105	1,829	1,270	959	1,015	752	374	32,899
-0301	1/14/2017	Neg	422	271	156	116	127	214	449	682	953	1,293	1,454	1,593	1,819	1,965	2,088	2,400	2,543	2,428	1,976	1,184	1,010	953	760	579	27,435
-0301	1/15/2017	Pos	233	138	125	107	154	322	657	1,278	1,640	1,624	1,806	1,996	2,004	2,044	1,912	1,888	1,741	1,456	1,073	885	673	569	358	282	24,965

100 CCSs

January-December, 2017

Hourly counts

Two directions

2 Data Description

Features

- **Road characteristics**
  - Functional classifications
  - No. of lanes, speed limit
  - Total mileage of roads by functional classifications in 1 mile radius
- **Socio-economic features**
  - Average income, employment density, number of households, average household size, and population density in 3-mile radius of each CCS location.
- **HERE Data**
- **Weather data**

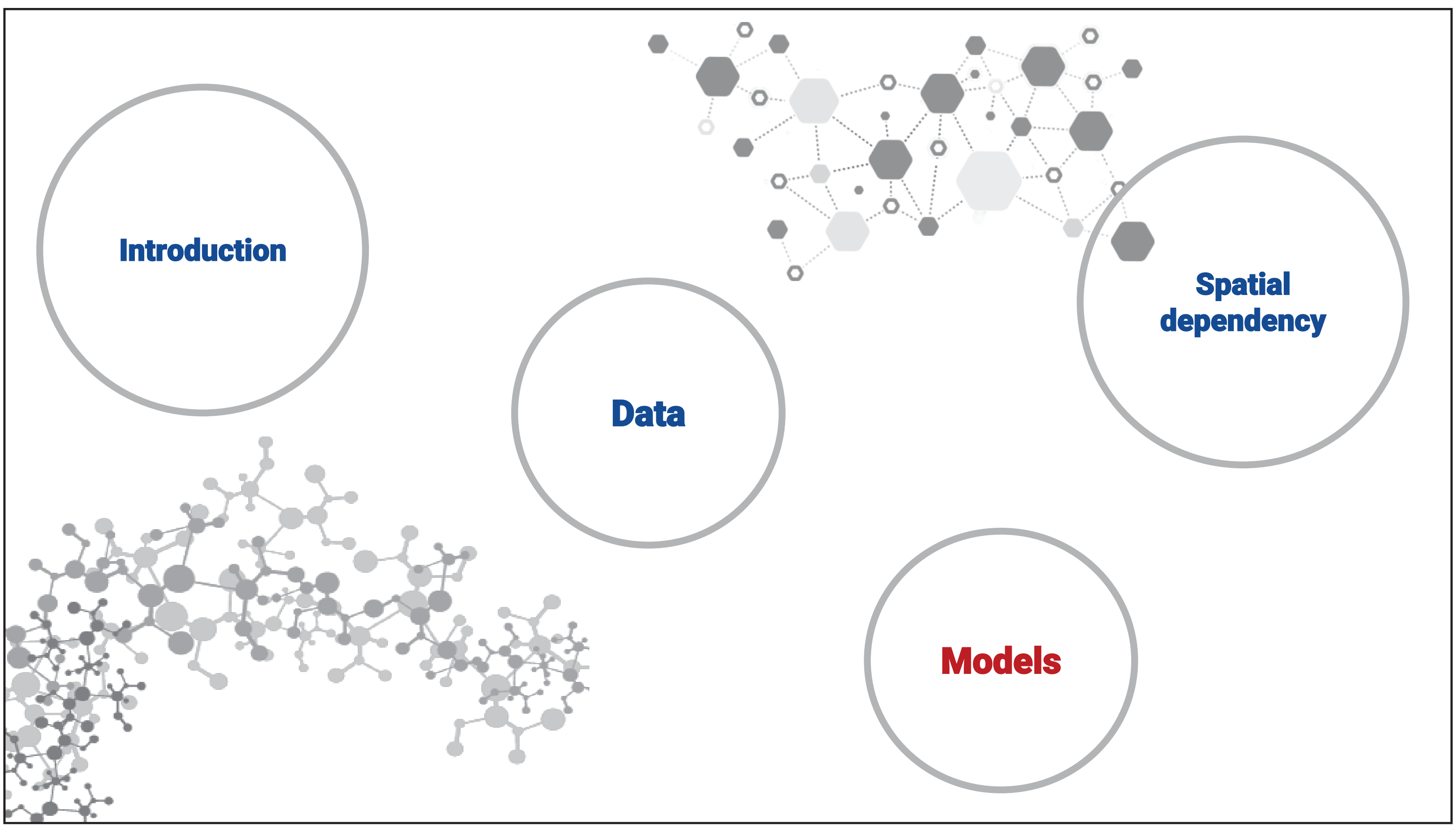
We have roughly:

100 ✖ 12 ✖ 30 ✖ 24 ✖ 2

≈ **1,000,000** observations

✖

73 features



3 Models

Traffic Volume Prediction

- **Traditional method:** **Factoring method**
- **Classic machine learning model:** **Support vector machine (SVM)**
- **Introducing spatial dependency feature:** **XGBoost with breadth-first search (BFS)**

3 Models - Factoring Method

- The basic idea:

A year's AADT was estimated from short-term count (48 hours in continuous) using the formula from Traffic Monitoring Guide (FHWA 2001):

$$AADT_{est,i} = \frac{1}{2}\sum_{d=1}^{2} VOL_i * M_i * D_i * G_i$$

Vol_i: actual 24-hour vehicle count

D_i: day of week factor

M_i: seasonal factor

G_i: traffic growth factor

3 Models - Factoring Method

11

- The basic idea:

A year's AADT was estimated from short-term count (48 hours in continuous) using the formula from Traffic Monitoring Guide (FHWA 2001)

- Advantages:

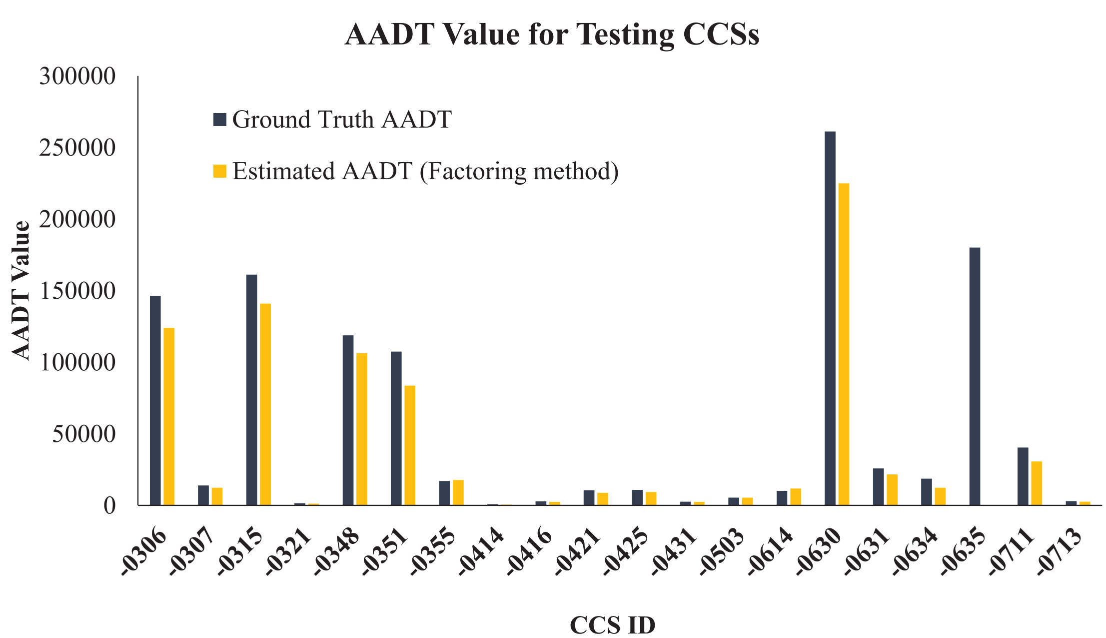
Simple and straightforward

- Disadvantages:

Large random errors

3 Models - Factoring Method

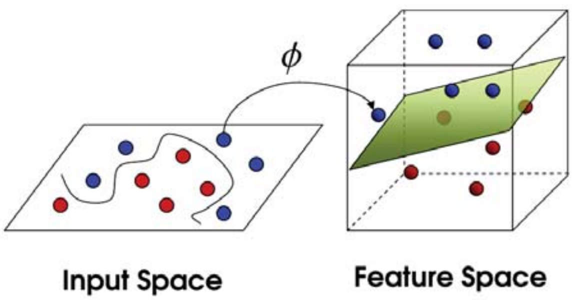
12

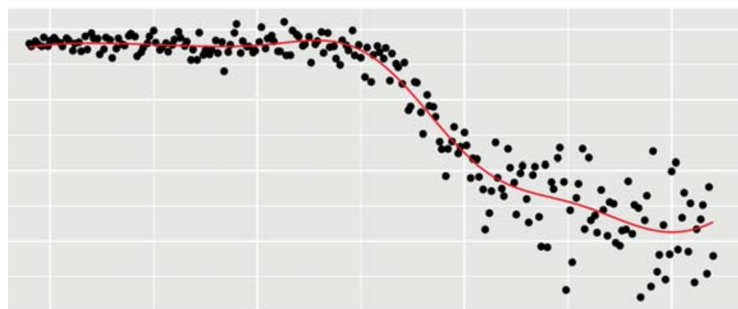


3 Models - SVM

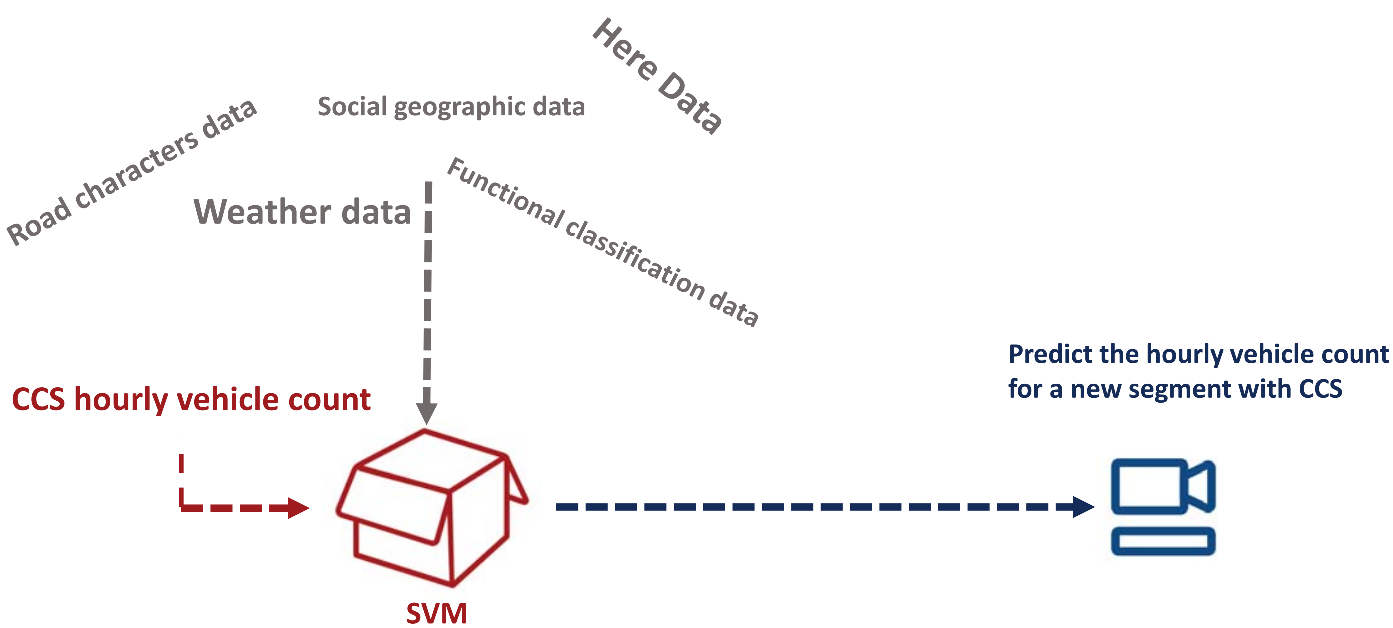
- The basic idea:

A model that constructs hyperplanes on higher dimensional space to solve classification and regression problems (proposed by Cortes and Vapnik in 1995)

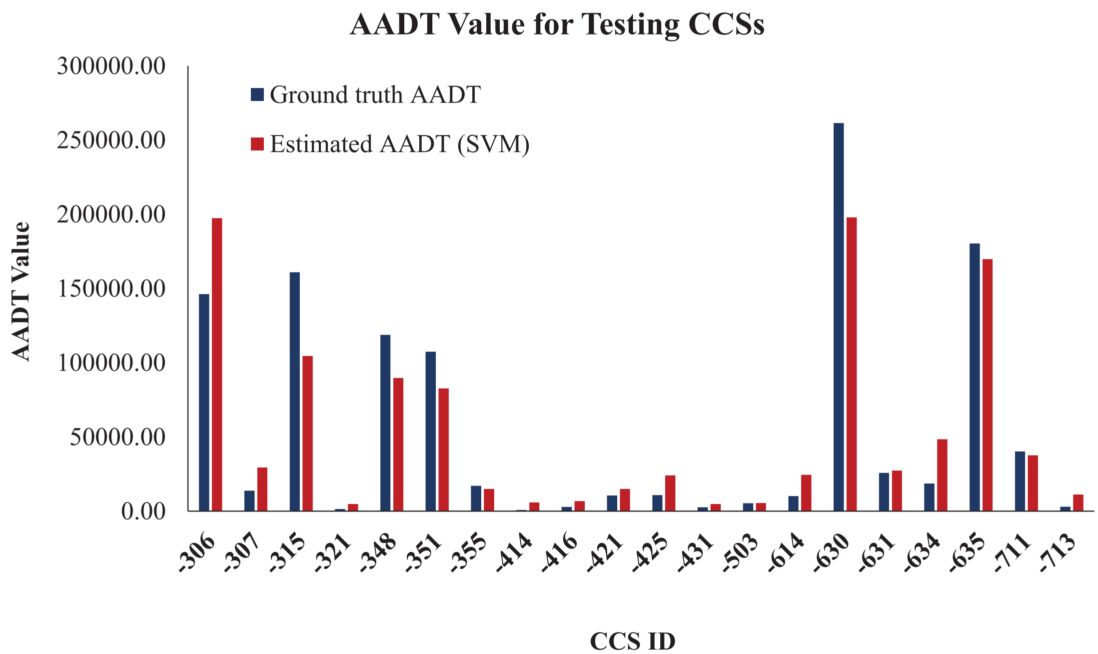




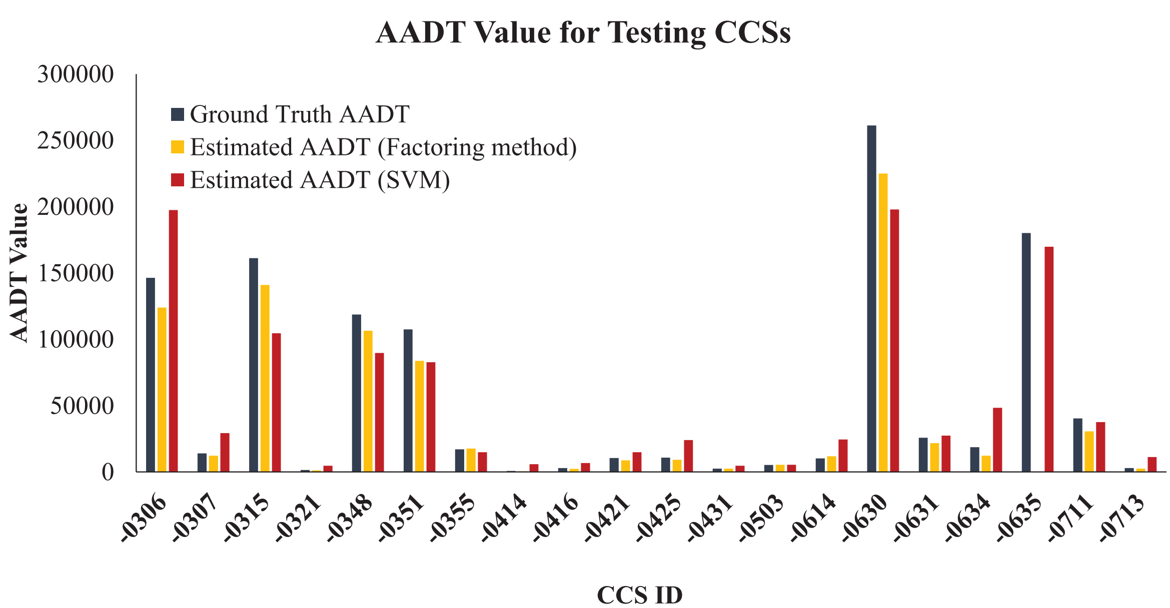
3 Models - SVM



3 Models - SVM



3 Models - Result Comparison



3 Models-SVM

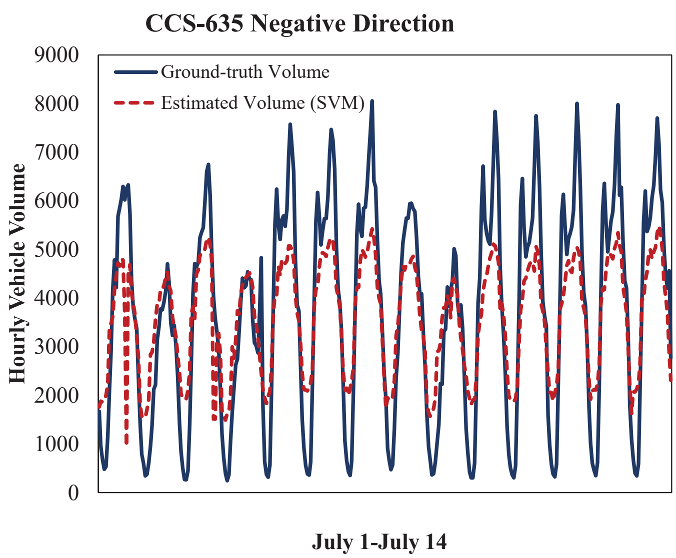
AADT is a traffic volume indicator on a coarse level.

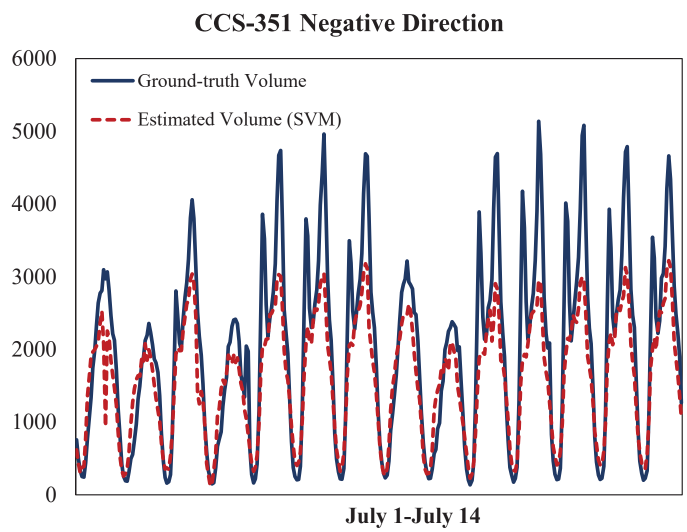
SVM is capable of predicting traffic volume information by hour-a smaller granular level.

We have the data of **100 CCSs** in total. Among them, **80 CCS** are used to train the model and **20 CCSs** are used to test the effectiveness of the model.

3 Models-SVM

To visualize the results, we use the ground-truth hourly counts and the predicted values from SVM for comparison. Here, we visualize the data from two randomly chosen CCSs in consecutive 14 days (CCS 635 and CCS 351).





Introduction

Data

Spatial Dependency

Models

4 Spatial Dependency Feature

19

Compared with factoring method, SVM can better capture information at finer granularity, such as hourly traffic variation.

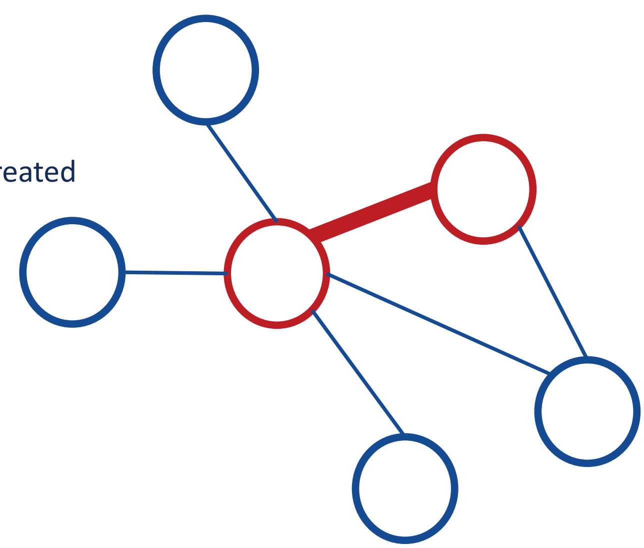
Yet we can do better to improve the prediction!

How?

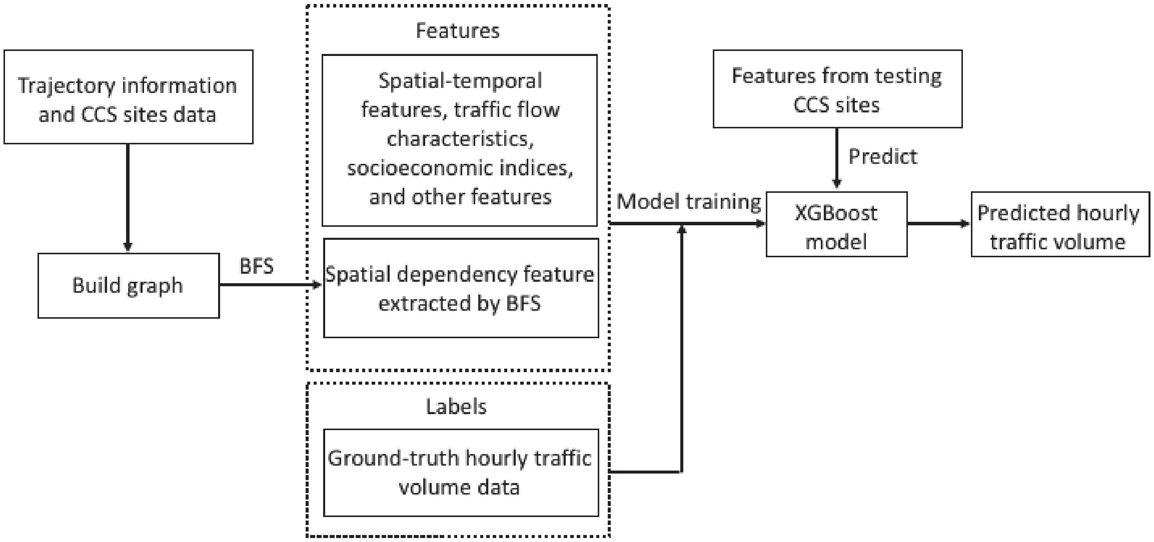
In typical ML techniques, all road segments are treated independently for model construction.

The core idea:

If two roads are strongly connected, their traffic pattern have larger probability to be similar.

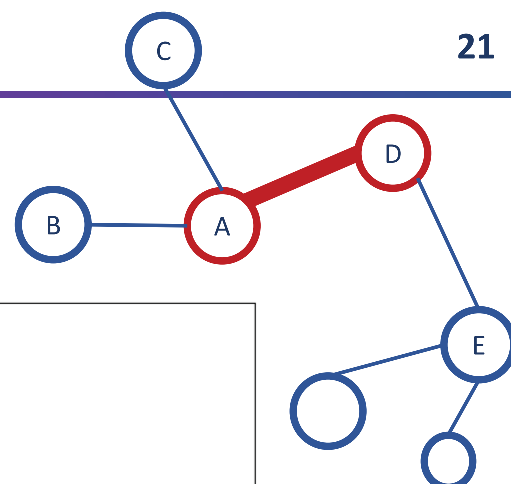


Methodology Framework



Yi, Z., Liu, X. C., Markovic, N., & Phillips, J. (2021). Inferencing hourly traffic volume using data-driven machine learning and graph theory. *Computers, Environment and Urban Systems*, *85*, 101548.

Spatial Dependency Feature



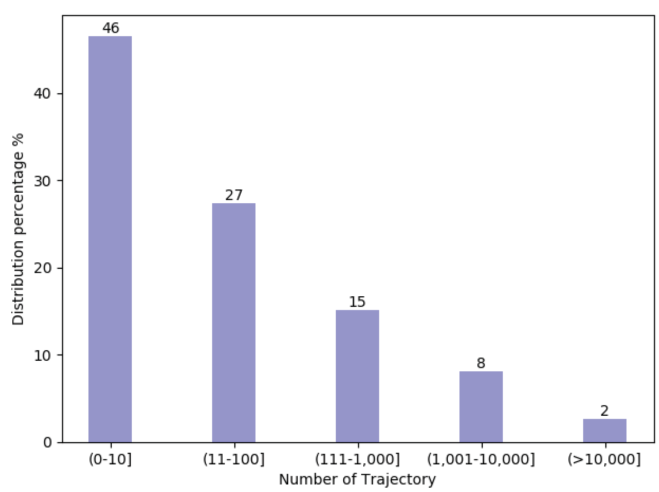
Build a graph:

We use the trajectory data to connect CCSs, and build a graph

The edge weight is calculated by using count of trajectory. Higher weight represents higher connection.

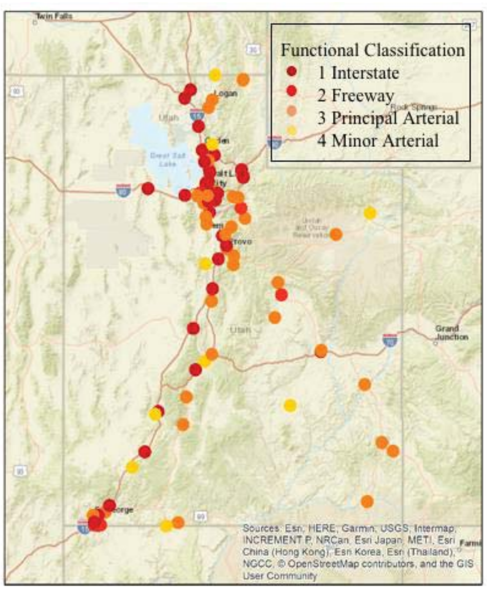
Then we use the Breadth Search First (BFS) algorithm to find neighboring strongly connected CCSs

Use the hourly volume from neighboring CCSs to infer the hourly volume at current location (a new feature)

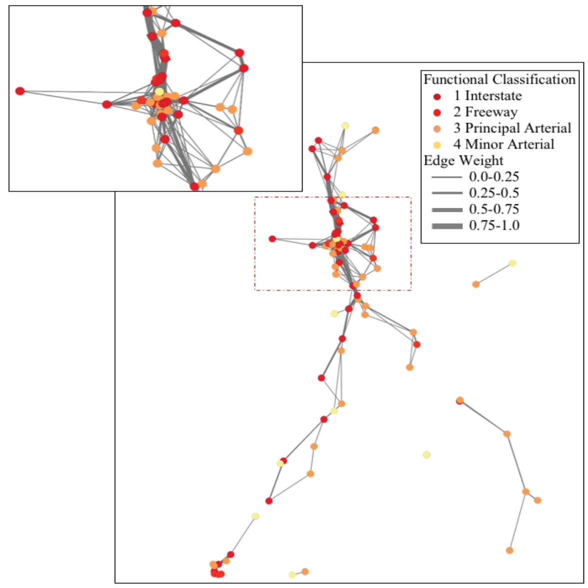


4 Spatial Dependency Feature

Geographical distribution of CCS



Constructed traffic network graph



4 Spatial Dependency Feature

XGBoost (Extreme gradient boosting decision tree)

One of the state-of-the-art machine learning models.

A highly effective tree boosting system that achieves high prediction accuracy on many challenging problems.

Tree ensemble models are capable of analyzing the feature importance.

4 Spatial Dependency Feature

Performance metrics:
R^2, MAE, MAPE

Testing process:
replicated 5 times by different random seeds
In each seed, 15% CCSs are selected as testing site

Model comparison:
M1: XGBoost model without spatial dependency feature

M2: XGBoost model with a spatial dependency feature characterized by Euclidean distance

M3: XGBoost model with a spatial dependency feature characterized by network distance

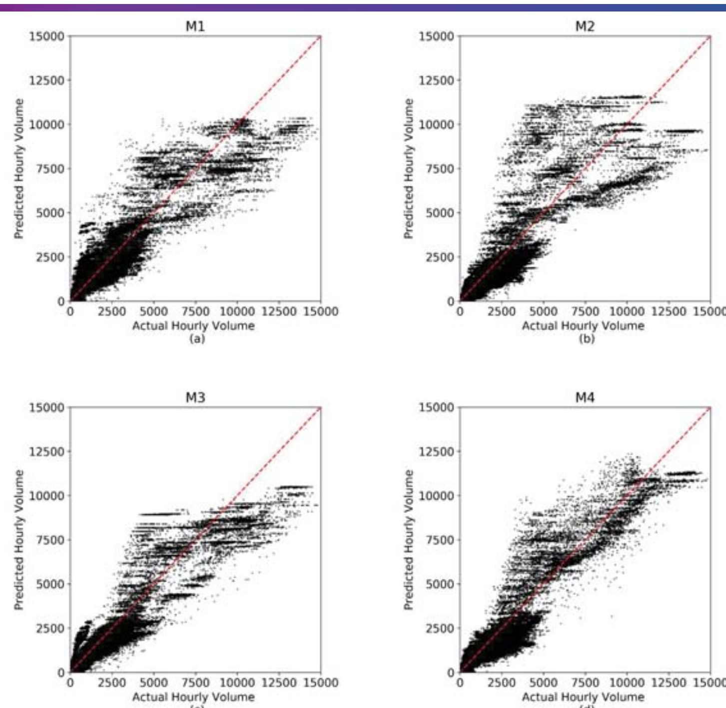
Name of measurement	Mathematical formulation	Brief description
R^2	$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$	R^2 is the proportion of traffic volume variance that is explained by predicting models, and it provides a measure of how well observed outcomes are replicated by the model.
MAE	$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert$	MAE is a measure of difference between actual values and predicted values, which gives a clear interpretation of average magnitude of the errors for predictions.
MAPE	$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left\lvert\frac{y_i - \hat{y}_i}{y_i}\right\rvert$	MAPE is a statistical measure of prediction accuracy, where the prediction error is presented as a percentage. Smaller values indicate better prediction power.

Prediction performance of different models

Seed	R^2 on test set				MAE on test set				MAPE on test set			
	M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
1	0.849	0.780	0.860	0.895	643	720	619	548	0.80	0.56	0.65	0.57
2	0.839	0.862	0.811	0.934	645	554	562	404	1.06	0.93	0.51	0.88
3	0.817	0.875	0.879	0.885	887	773	799	733	1.26	0.57	0.68	0.52
4	0.859	0.866	0.847	0.868	631	397	420	522	1.96	1.03	1.58	1.33
5	0.898	0.831	0.936	0.916	347	681	311	308	2.09	1.66	1.07	1.67
Mean	**0.852**	**0.842**	**0.866**	**0.900**	**631**	**625**	**542**	**503**	**1.43**	**0.95**	**0.90**	**0.79**

4 Spatial Dependency Feature

To better visualize model performance, ground truth hourly counts from test set in Seed 1 are compared against predicted values.



M1: XGBoost model without spatial dependency feature

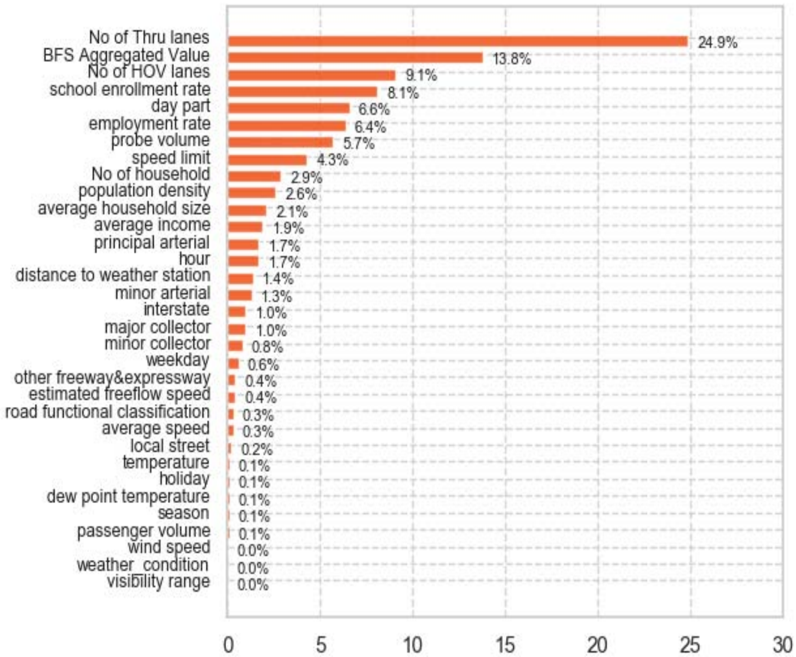
M2: XGBoost model with a spatial dependency feature characterized by Euclidean distance

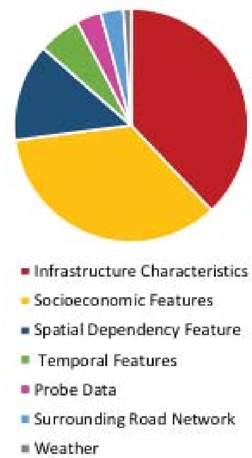
M3: XGBoost model with a spatial dependency feature characterized by network distance

M4: XGBoost model with BFS algorithm

4 Spatial Dependency Feature

26



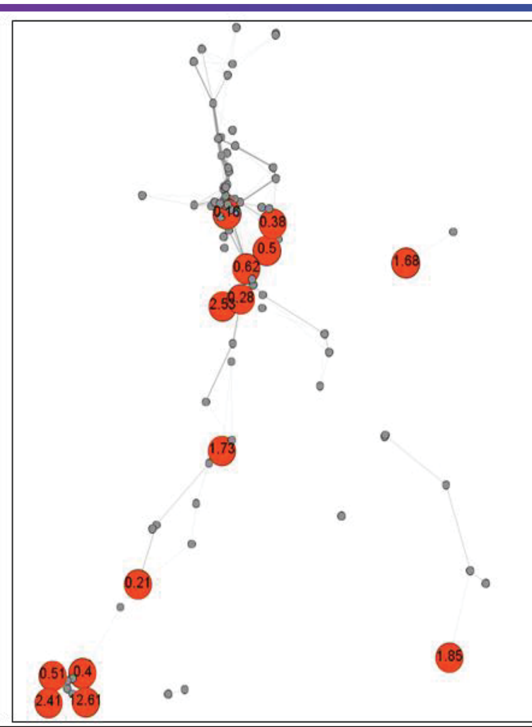


4 Spatial Dependency Feature

27

To further explore spatial dependency, we display the MAPE for CCSs in testing dataset for each random seed.

Seed Number **5**



5 Conclusions

- Both traditional methods and machine learning models generate satisfactory results for predicting AADT.
- Machine learning models (e.g. SVM and XGBoost) are capable of capturing hourly traffic volume.
- Incorporating spatial dependency feature can improve the prediction accuracy.
- Road characteristics, socioeconomic factors, spatial dependency, and temporal features are the most influential feature categories.

TRANSPORTATION LEARNING NETWORK

A partnership with MDT•NDDOT•SDDOT•WYDOT and the Mountain-Plains Consortium Universities

Thank you for participating!

Please take a moment to complete the evaluation included in the reminder email.

We appreciate your feedback.

Contact Information

Chris Padilla
chris.padilla@ndsu.edu
(701) 202-5730

Susan Hendrickson
susan.Hendrickson@ndsu.edu
(701) 238-8646

Shannon Olson
shannon.l.olson@ndsu.edu
(701) 552-0672

https://tln.learnflex.net
https://www.translearning.org

Thank you to our partners: