Applying Machine Learning to Predicting Pavement Conditions with LCMS Data

Arnab Roy, Jie Zhang, Lihui Bai, Zhihui Sun, Xuwen Zhu

University of Louisville

Aaron Collins, Praneeth Nimmatoori, Tracy Nowaczyk, Chad Shive

Operations and Pavement Management Branch

Kentucky Transportation Cabinet

Outline

- Project background
- Visual evaluation system (VES)
- Laser crack measurement system (LCMS)
- Data compatibility and quality
- Methodologies for the mapping
  - Machine learning method
  - Decision tree classification & advanced ensemble model
  - Data rotation
  - Conditional probability
  - AdaBoost
- Results and discussions
- Future research

Pavement Modeling at Kentucky Transportation Cabinet

- **The Kentucky Transportation Cabinet (KYTC)**
  - KYTC is an executive branch agency responsible for supervising the development and maintenance of a safe transportation system throughout the Commonwealth.
  - KYTC manages more than 27,000 miles of highways, including roughly 20,500 miles of secondary roads, 3,600 miles of primary roads, and more than 1,400 interstate and parkway miles.

- **Since 2014, University of Louisville (UofL) has collaborated with KYTC toward data-driven and effective means for pavement management and preservation (PMP) .**

Past PMP Projects

- The KYTC has been collecting pavement condition data for over 15 years.
- There are 9 distress condition indices via visual evaluation pertaining to 5 types of distresses (WPC, RF, OC, OS, APPEAR).
- **The past projects aimed to:**
  - Predict 9 distress condition indices for next year;
  - Develop a prioritization method for selecting pavement projects objectively based on the predicted condition indices and an analytical hierarchical process (AHP).

Pavement Data Collection Methods

- Windshield visual survey (Visual Evaluation System – VES)
  - Rated by experienced technicians.
  - It may has human errors.
  - Rating for same road may vary with different technicians.

- Automated pavement surveys (e.g., LCMS)
  - Featuring high resolution image processing and laser surface profilers.
  - It's more consistent, accurate and reliable.
  - It saves time and cost over visual data.

Transition to LCMS

- Years of windshield visual data collected in the legacy format are of great value for forecasting and analysis, and thus should not be abandoned.

- However, the transition from the windshield visual survey to automated pavement survey is challenging:
  - The compatibility issue between the VES and LCMS databases.
  - VES: 9 variables on Likert-type scale, ordinal data (discrete)
  - LCMS: significantly more variables on numerical scale, interval data (continuous)

- In the current project, UofL-KYTC team aims to establish a mapping process from the LCMS to the legacy VES.

Related Works

- Earlier works in automatic pavement evaluation
  - Groeger et al. (2003): Maryland State HWA, an automated network-level crack detection using automated road analyzer (ARAN) data collection vehicle, Wisecrax crack detection software with QC and QA
  - Timm and McQueen (2004): Alabama, conducted survey on 27 (out of 46) state DOT pavement divisions on their practices of manual and automated data collection. They also performed statistical analyses of manual versus automated data using the Alabama roadway data.
  - "One issue that has stalled the advancement of the automated pavement condition survey is the lack of information about successful transitions from manual to automated data collection."
  - "Making the transition is a major task that few have fully accomplished"
  - Lu et al. (2004) used high-accuracy sensors and an artificial neural network model to statistically estimate crack depth on Florida roadways.
- More recent works
  - Tighe et al. (2008), Ong et al. (2009), Underwood et al. (21010) all study the difference between manual and automatic pavement evaluations
  - Mraz et al. (2006) study the accuracy of the automated surveys under varieties of lighting, speeds, and pavement types by using signal-to-noise ratio.
  - Khadgi et al. in 2016 conducted a small scale pilot study using ANOVA and linear regression to bridge between LCMS and VES, for Kentucky interstate parkways.

VES Data

RT UNIQUE ID	FROM POINT	TO POINT	LANE DIR	WPC JD EXT	WPC JD SEV	RF EXT	RF SEV	OC EXT	OC SEV	OS P EXT	OS P SEV	APP
121-I-0024	45.123	51.9	L	3	2	1	1	0	0	0	0	1
121-I-0024	51.9	65.349	L	3	6	1	1	3	2	2	2	1.5
121-I-0024	45.133	55.629	R	2	3	1	1	2	2	1	2	1.5
121-I-0024	55.629	65.349	R	3	6	1	1	1	3	1	2	1.5
056-I-0265	26.6	30.637	R	4	2	1	1	1	2	0	0	1
...	...	...	...	...	...	...	...	...	...	...	...	...

- Visual Evaluation system (VES) uses nine factors to describe pavement conditions.
  - WPC_EXT, WPC_SEV: 0-9
  - RF_EXT, RF_SEV, OC_EXT, OC_SEV: 0-5
  - OS_EXT, OS_SEV, APPEARANCE: 0-3
  - 0-best condition, 9/5/3-worst condition

LCMS Data

Session Name	Begin MP	End MP	Len (mile)	DIR	FATCRK TYPEA_ LOW	FATCRK TYPEA_ MED	FATCRK TYPEA_ HIGH	FATCRK TYPEA_ SEV	...
056-I-0265	26.6	26.7	0.1	W	1.74	0.08	0	0.25	...
056-I-0265	26.7	26.8	0.1	W	4.3	1.27	0	0.37	...
056-I-0265	26.8	26.9	0.1	W	3.31	0.17	0	0.20	...
056-I-0265	26.9	30	0.1	W	11.08	3.29	0.16	1.87	...

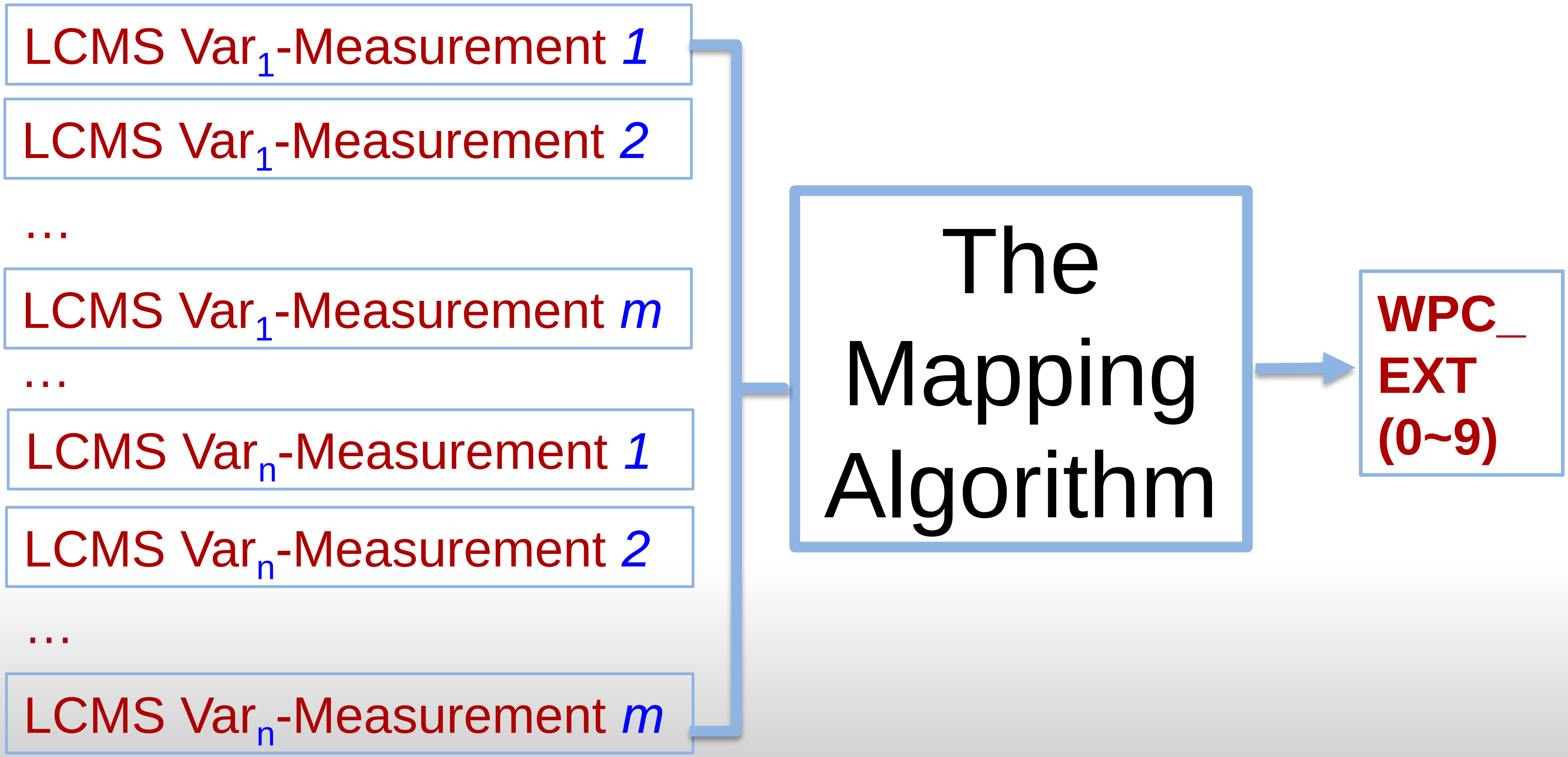
- Laser Crack Measurement System (LCMS) data is from high resolution images generated by laser surface profiler.

- It has approximately 150 different variables.
- It records continuous measurement every 0.1 mile.

- In this talk, we focus on mapping from LCMS variables to WPC_EXT and WPC_SEV ratings.

Problem Statement

- For WPC_EXT, develop the following mapping model.



Data Processing: Variable Identification

- Relevant factors in LCMS were identified by consulting KYTC experts.
- Each VES index has a set of associated LCMS variables.

- 13 LCMS variables correspond to WPC_EXT.
- 7 LCMS variables correspond to WPC_SEV.

Wheel Path Cracking Extent (WPC^{e})	Wheel Path Cracking Severity (WPC^{s})
Fatigue Type A LOW	Fatigue Type A SEV
Fatigue Type A MED	Fatigue Type B SEV
Fatigue Type A HIGH	Fatigue Type C SEV
Fatigue Type B EXT	Non WP Longitudnal SEV
Fatigue Type B Area EXT	Zone 2_long_crack_sev
Fatigue Type C EXT	Zone 3_long_crack_sev
Fatigue Type C Area EXT	Zone 4_long_crack_sev
Non WP Longitudnal LOW	
Non WP Longitudnal MED	
Non WP Longitudnal HIGH	
Zone 2_long_crack_ext	
Zone 3_long_crack_ext	
Zone 4_long_crack_ext	

Data Processing: Resolution Unification

VES

RT UNIQUE ID	FROM POINT	TO POINT	LANE DIR	WPC JD EXT	WPC JD SEV	RF EXT	RF SEV	OC EXT	OC SEV	OS P EXT	OS P SEV	APP
056-I-0265	26.6	30.637	R	4	2	1	1	1	2	0	0	1

LCMS

Session Name	Begin MP	End MP	Len (mile)	DIR	FATCRK TYPEA_ LOW	FATCRK TYPEA_ MED	FATCRK TYPEA_ HIGH	FATCRK TYPEA_ SEV	...
056-I-0265	26.6	26.7	0.1	W	1.74	0.08	0	0.25	...
056-I-0265	26.7	26.8	0.1	W	4.3	1.27	0	0.37	...
056-I-0265	26.8	26.9	0.1	W	3.31	0.17	0	0.20	...
056-I-0265	26.9	30	0.1	W	11.08	3.29	0.16	1.87	...

- VES records are for road segments with varying lengths (e.g., 0.4 mile, 3.2 miles).
- LCMS records measurements for each 0.1 mile.

- For each road segment in VES (e.g., 2.5 mile), we calculate 10th, 20th, ……, 90th percentile, standard deviation, skewness, minimum, maximum value, a total 13 statistics (over 25 entries for the VES segment) in LCMS.
- These 13 statistics are used in the mapping process.

Final Data Set

- 2015 side-by-side LCMS and VES data were used in the test.

- 8429 of 8588 LCMS data entries can be used.
- 220 roads segments out of 500 in VES can have match in LCMS in 2015.
- 47 roads segments from VES are removed because of large discrepancy with LCMS.

- Final data set corresponds to 173 VES roads and their associated LCMS records.
- Later, these 173 will be repeated used as training, validation and testing data.

Final Input Data: 10th Percentile DT for WPC_EXT

- There are 13 of such final input datasets for building the trees and ensemble model for WPC_EXT
- There are 7 of such final input datasets for building the trees and ensemble model for WPC_SEV

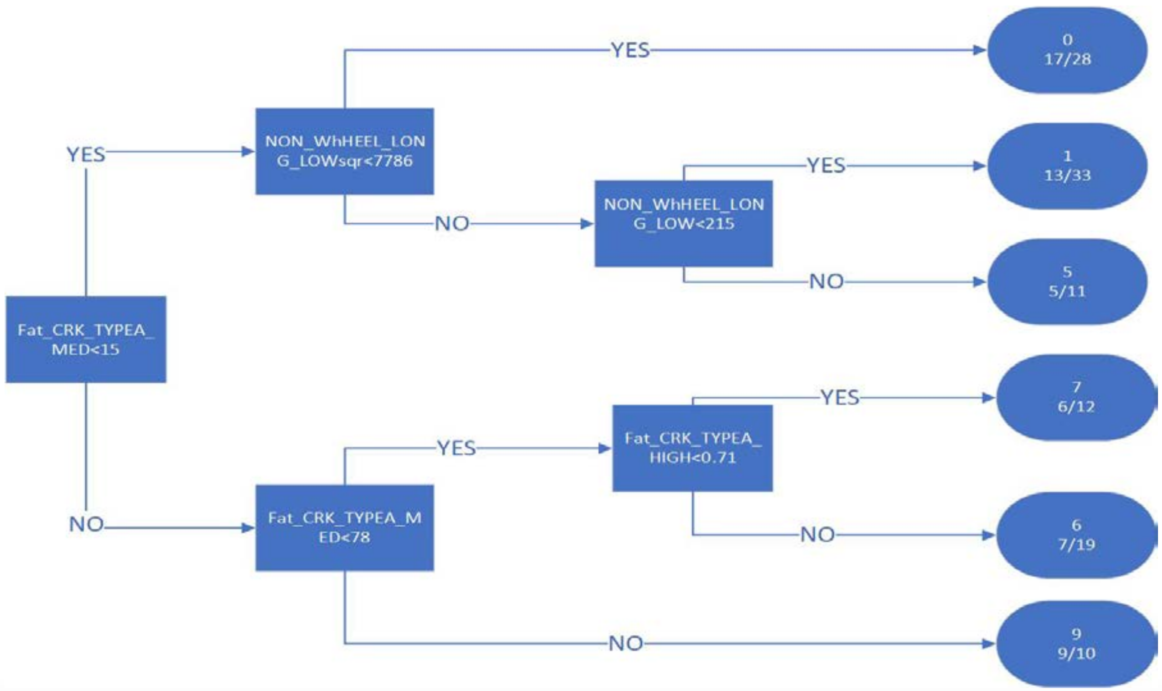
	B	C	D	H	L	N	O	P	R
1									
2	Sample Index	UNID	FAT_CRK_TYPEA_HIGH	FAT_CRK_TYPEB_AREA_EXT	FAT_CRK_TYPEC_EXT	NON_WHEEL_LONG_HIGH	NON_WHEEL_LONG_LOW	NON_WHEEL_LONG_MED	WPC_JD_EXT
3	1	BG9002L47-0-4.9	0	0	0	3.23	61.118	23.59	0
4	2	BG9002L47-4.9-5.82	0	0	0	0	0.416	0.032	7
5	3	BG9002L47-5.82-10.172	0	0	0	0	0.471	0.196	8
6	4	BG9002L90-29.18-35.15	0	0	0	0	0	0	4
7	5	BG9002L90-35.15-39.267	0	0	0	8.428	137.496	108.974	9
8	6	EB9004L24-0-6.77	0	0	0	1.06	130.19	96.168	9
9	7	EB9004L24-6.77-12.13	0	0	0	1.001	164.667	79.908	0
10	8	EB9004L54-39.794-42.446	0	0	0	4.068	175.474	101.97	0
166	164	WK9001R47-123.474-130.948	0	0	0	12.715	438.486	287.711	9
167	165	WK9001R47-130.948-135.1	0	0	0	0	0.658	0	0
168	166	WK9001R47-135.1-136.066	0	0	0	0	0.876	0.142	0
169	167	WK9001R47-136.066-136.796	0	0	0	0.674	119.039	50.712	6
170	168	WN9007L114-0-2.6	0.086	6.714	0	1.981	431.475	177.399	7
171	169	WN9007L114-2.6-9.2	0	0	0	0	10.064	2.59	0
172	170	WN9007L114-9.2-17.8	0	0	0	0.196	30.402	16.044	0
173	171	WN9007R16-17.8-28.5	1.894	28.46	0	49.26	76.148	418.594	9
174	172	WN9007R16-28.5-34.724	0	0	0	0	57.575	21.093	5
175	173	WN9007R30-65.91-72.264	0	0	0	0	6.023	0.341	1

Outline

- Project background
- Visual evaluation system
- Laser crack measurement system (LCMS)
- Data compatibility and quality assurance
- Methodologies for the mapping
  - Machine learning method
  - Decision tree classification & advanced ensemble model
  - Data rotation
  - Conditional probability
  - AdaBoost
- Results and discussions
- Future research

Decision Tree Classifier

- Decision tree is a widely used method in statistics and machine learning.
  - Mirrors human decision making.
  - Requires little data preparation (e.g., normalization is not required)
  - Performs well with large data sets.
  - Simple to understand and interpret.
  - Able to handle categorical data.



Ensemble Model with Decision Trees

- Recall the need for "data unification"
  - 10th, 20th, ……, 90th percentile, standard deviation, skewness, minimum, and maximum value, total 13 statistics in LCMS of each road segments.

VES

RT UNIQUE ID	FROM POINT	TO POINT	LANE DIR	WPC JD EXT	WPC JD SEV	RF EXT	RF SEV	OC EXT	OC SEV	OS P EXT	OS P SEV	APP
056-I-0265	26.6	30.637	R	4	2	1	1	1	2	0	0	1

LCMS

Session Name	Begin MP	End MP	Len (mile)	DIR	FATCRK TYPEA_ LOW	FATCRK TYPEA_ MED	FATCRK TYPEA_ HIGH	FATCRK TYPEA_ SEV	...
056-I-0265	26.6	26.7	0.1	W	1.74	0.08	0	0.25	...
056-I-0265	26.7	26.8	0.1	W	4.3	1.27	0	0.37	...
056-I-0265	26.8	26.9	0.1	W	3.31	0.17	0	0.20	...
056-I-0265	26.9	30	0.1	W	11.08	3.29	0.16	1.87	...

- We grow 13 decision trees based on each of the 13 statistics.
- We then assemble them together with proper weights assigned to each of the 13 trees.
  - Trees with better prediction accuracy receives more weight in the final ensemble.

Determining Weights

- The Accuracy of Tree i can be measured by the following conditional probability:

Pr(i, k) = Probability{Actual VES is within ±1 of the prediction, i.e., [k-1, k+1] | given the prediction is k}

k\Tree	10th	20th	30th	40th	50th	60th	70th	80th	90th	Std	Skewness	Min	max
0	0.578	0.612	0.667	0.780	0.756	0.696	0.791	0.821	0.654	0.611	0.508	0.614	0.467
1	0.692	0.833	0.682	0.500	0.571	0.690	0.452	0.545	0.478	0.489	0.294	0.449	0.600
2	0.543	0.556	0.727	0.556	0.636	0.630	0.600	0.529	0.346	0.010	0.000	0.556	0.000
3	0.010	0.500	0.500	0.333	1.000	0.010	0.500	0.000	0.333	0.000	0.000	0.010	0.000
4	0.083	0.200	0.500	0.250	0.667	0.333	0.500	0.357	0.200	0.000	0.333	0.625	0.000
5	0.000	0.010	0.667	0.010	0.333	0.833	0.000	0.010	0.333	0.000	0.400	0.222	0.500
6	0.435	0.333	0.200	0.500	0.571	0.300	0.500	0.222	0.286	0.200	0.300	0.500	0.125
7	0.200	0.222	0.625	0.333	0.800	0.667	0.500	0.688	0.643	1.000	0.455	0.250	0.333
8	0.333	0.714	0.010	0.010	0.500	0.500	0.500	0.010	1.000	1.000	0.000	0.010	0.010
9	0.737	0.607	0.531	0.564	0.618	0.556	0.563	0.581	0.515	0.355	0.217	0.485	0.467

Data Rotation

- In order to make full use of 173 data, we use a complex data rotation method where each single data is used at least once for validation and once for testing.

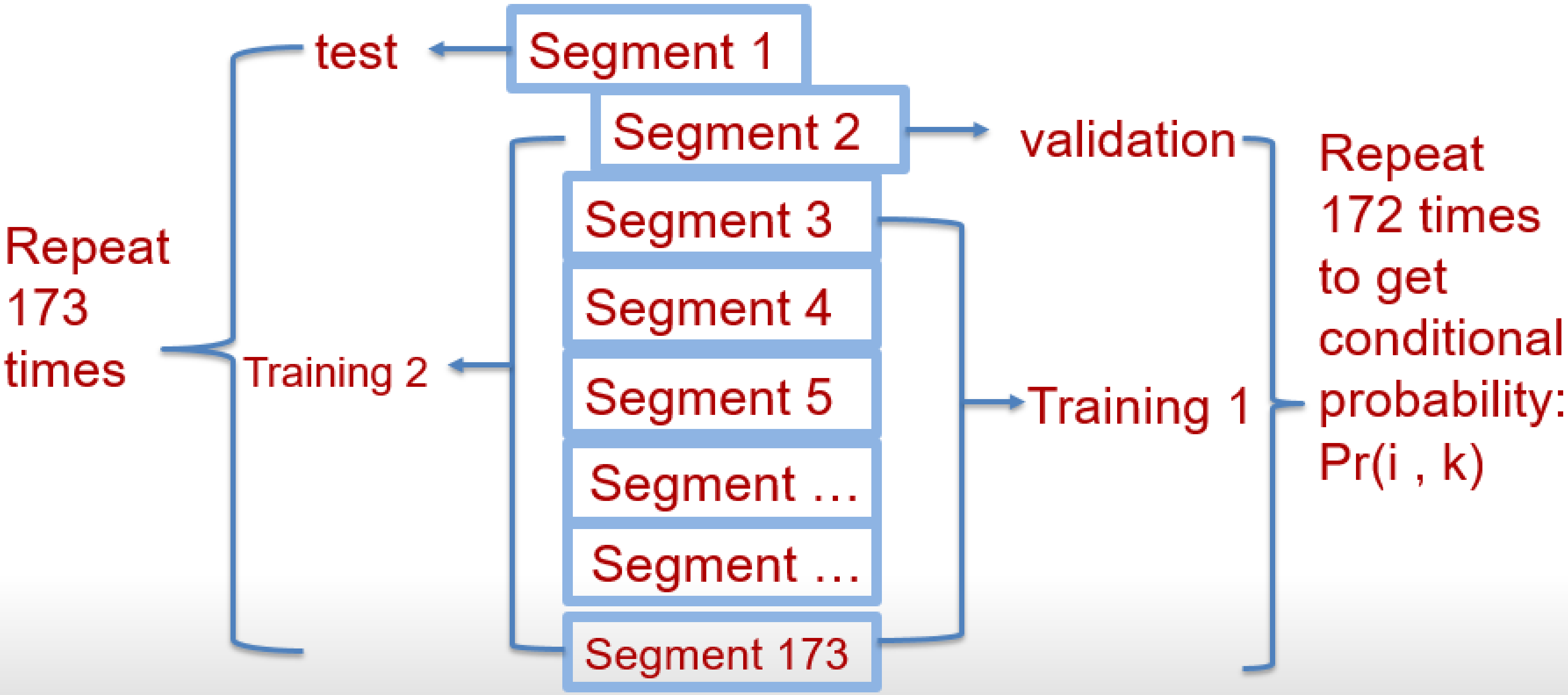
A. Leave 1 road for test.
   1. In remaining 172, leave 1 road for validation.
   2. Use 171 roads to build 13 trees, and predict the 1 validation data in step 1.
   3. Repeat step 1-2 for 172 times.
   4. Evaluate the conditional probability of each tree in predicting the 172 validation data and assign their weights accordingly (trees with better accuracy receives higher weights).
   5. Use 172 roads to build a final model. This will be the DT model based on the current testing data.

B. Predict the 1 test road using the final model in step 5 and it's corresponding conditional probability with:

$$\frac{\sum_{i=1}^{13} prediction_i \times weight_i}{\sum_{i=1}^{13} weight_i}$$

C. Repeat A-B 173 times and get final accuracy on the 173 data points.

Data Rotation (Illustrated)



AdaBoost

- The AdaBoost algorithm is an iterative procedure that combines many base/weak classifiers (essentially a predictor)

A. Start with the unweighted training sample, the AdaBoost algorithm builds a classier, for example a classification tree, that produces class labels
B. If a training data point is misclassified, the weight of that training data point is increased (boosted)
C. A second classifier is built using the new weights, which are no longer equal

AdaBoost (Cont'd)

Predictions from Base Leaners

Base Estimator	Predictions
0	3
1	0
2	1
3	2
4	3
5	0
6	1
7	2
8	4
9	4
10	6
11	3
12	4
13	1
14	0
15	6
16	3
17	2
18	0
19	3

Weights from Base Leaners

Base Estimator	Prediction Weights
0	2.090381482
1	2.053022694
2	2.062486092
3	2.258613447
4	2.282895129
5	2.044023419
6	2.212219558
7	2.406581106
8	2.218521693
9	2.030091209
10	2.297387778
11	2.248681687
12	2.171301487
13	2.060313371
14	2.273830953
15	2.292804024
16	2.173019845
17	2.163732657
18	2.262625653
19	2.283514834

- Class 0 weight = 2.053 + 2.044 + 2.2739 + 2.2626 = 8.3663;
- Class 1 weight = 2.0624 + 2.2122 + 2.060 = 6.335
- Class 3 weight = 2.09 + 2.28 + 2.25 + 2.173 + 2.284 = 11.078
- Finally the class with highest weight is the final prediction.

Computational Experiments

- 10 independent simulation runs were conducted to validate the method.

- Training/testing data have similar composition (i.e., percentage of road segments of each grade 0-9) as in the entire pool of original data.

Class Distribution 2015-2017 data

Class	0	1	2	3	4	5	6	7	8	9
Frequency	193	77	60	43	48	54	73	84	53	62
Percentage	25.83668	10.3079	8.032129	5.756359	6.425703	7.228916	9.772423	11.24498	7.095047	8.299866

The Confusion Matrix: WPC_EXT

		Actual value									
Predicted value		0	1	2	3	4	5	6	7	8	9
	0	18	1	0	0	0	0	0	0	0	0
	1	1	10	1	0	0	0	0	0	0	0
	2	0	0	9	0	0	0	0	0	0	0
	3	0	1	1	1	2	0	0	0	0	0
	4	0	0	1	1	3	0	0	0	0	0
	5	0	0	0	0	0	7	0	0	0	0
	6	0	0	0	0	0	0	10	1	0	0
	7	0	0	0	0	0	0	0	8	0	0
	8	0	0	0	0	0	0	0	2	5	0
	9	0	0	0	0	0	0	0	0	4	8

- For 93 of 95 (97.9%) roads, prediction error is within ±1.
- For 79 of 95 (83.2%) roads, prediction error is 0.

The Confusion Matrix: WPC_SEV

Predicted value		Actual value									
		0	1	2	3	4	5	6	7	8	9
	0	25	4	0	0	0	0	0	0	0	0
	1	1	13	0	0	0	0	0	0	0	0
	2	0	3	11	1	0	0	0	0	0	0
	3	0	0	3	7	0	0	0	0	0	0
	4	0	0	0	4	10	0	0	0	0	0
	5	0	0	0	0	1	5	0	0	0	0
	6	0	0	0	0	1	0	4	0	0	0
	7	0	0	0	0	0	0	0	4	0	0
	8	0	0	0	0	0	0	0	2	2	0
	9	0	0	0	0	0	0	0	0	0	2

- For 102 of 103 (99.0%) roads, prediction error is within ±1.
- For 83 of 103 (80.6%) roads, prediction error is 0.

Performances for Five VES Indices

Performances (among 10 cases)	WPC_EXT		WPC_SEV	
	±1%	0%	±1%	0%
Best	97.90%		99.00%	
Worst	90.70%		95.70%	
Average	95.00%	75.80%	98.00%	78.60%

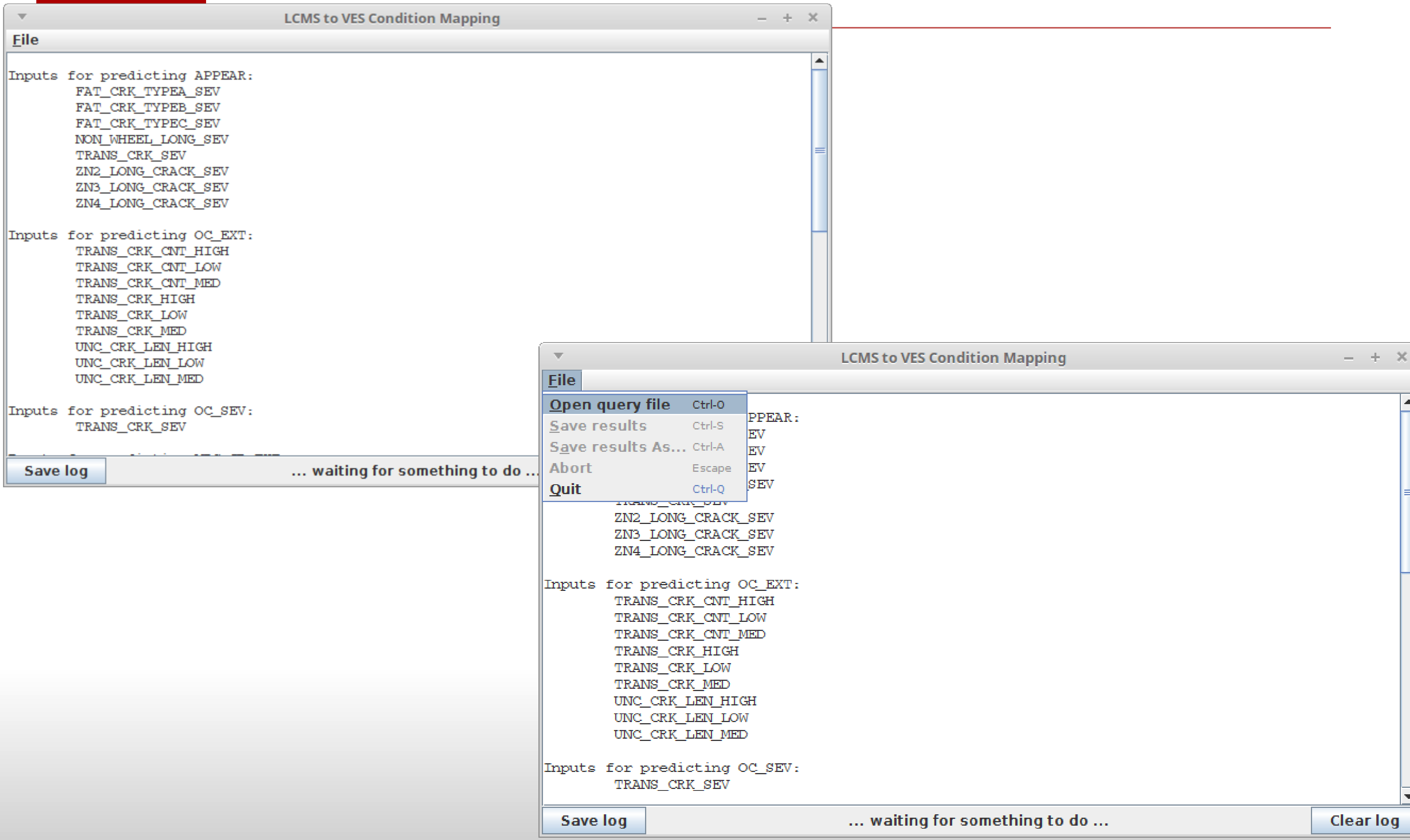
Performances (among 10 cases)	OC_EXT		OC_SEV		Apperance	
	±1%	0%	±1%	0%	±1%	0%
Best	98.00%		86.80%		100%	
Worst	92.60%		77.50%		94.60%	
Average	95.50%	80.40%	82.00%	53.20%	97.30%	76.80%

The Decision Support Systems

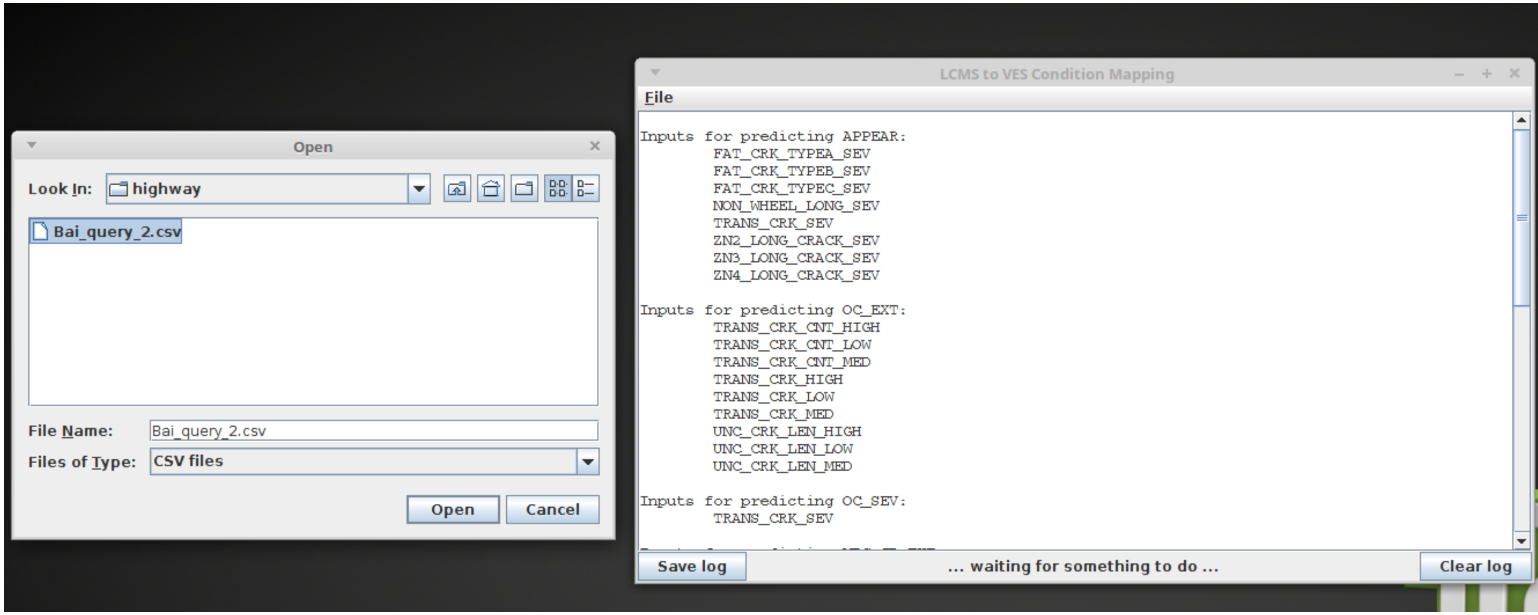
- In the decision support system (DSS), there are four steps to map a series of LCMS records to their corresponding VES ratings.
- The DSS currently is in the development and testing stage.

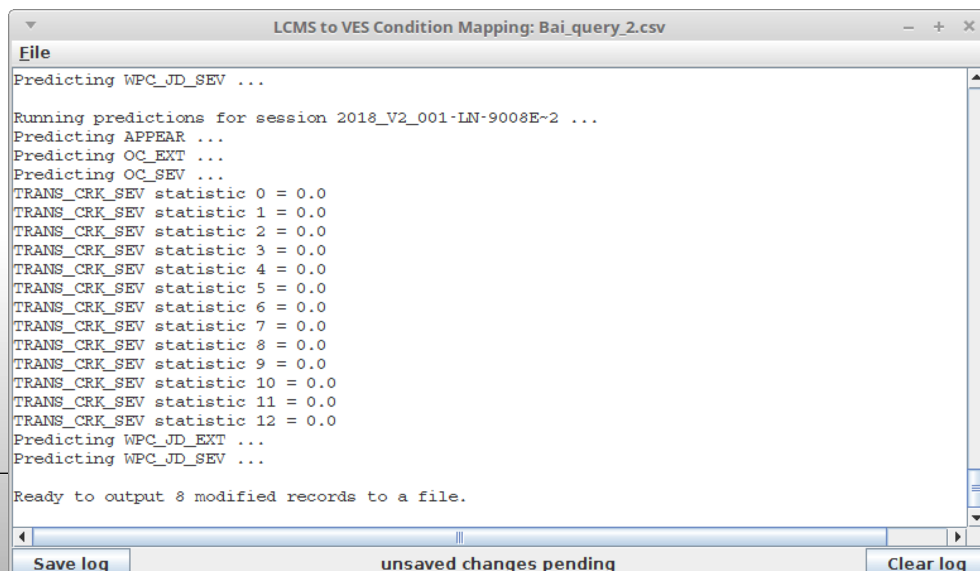
1. Program start up
2. User input and load LCMS queries
3. Predict
4. Save output (predictions) file.

The DSS: Start Up and File Menu

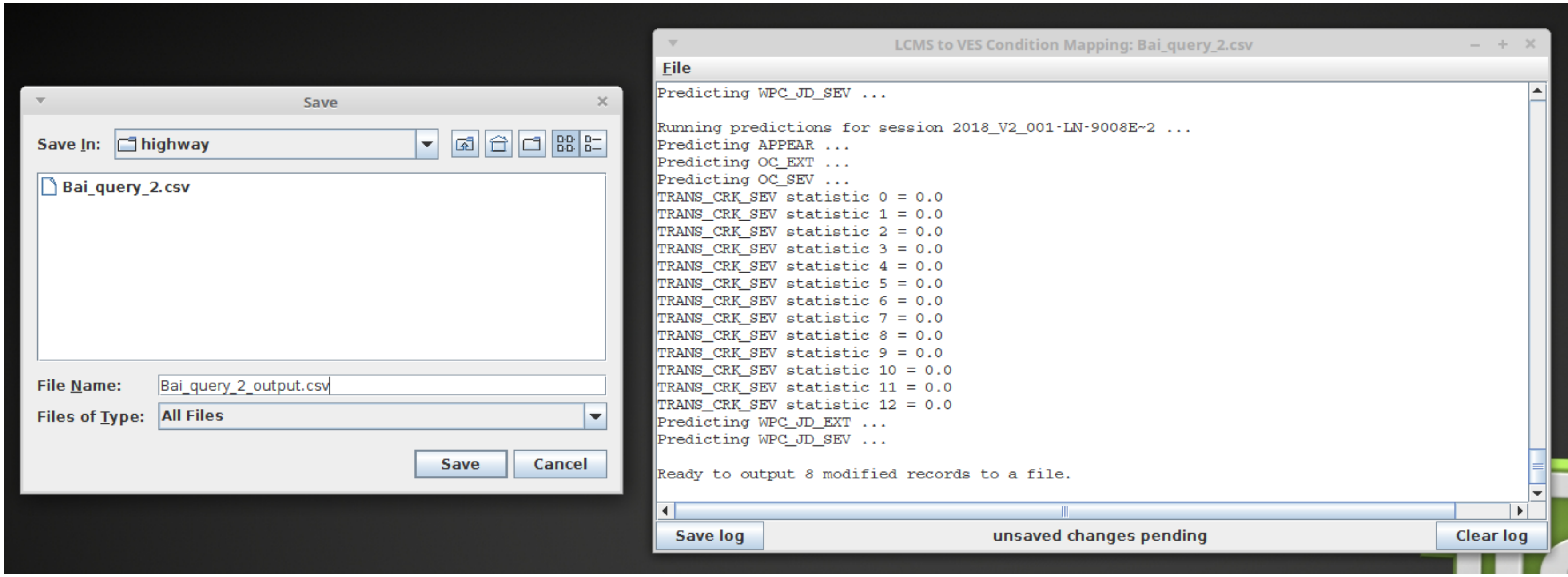


The DSS: Loading LCMS Query & Predicting





The DSS: Output File Option



	A	B	C	D	E	F	G	I
1		URSID	Query	Start MP	End MP	LANE_ID	SESSION_NAME	Predictions
2	0	030-WN-9	030-WN-9	61.557	61.591	2	2018_V2_016-WN-9007N~1	8
3	1	030-WN-9	030-WN-9	61.558	61.587	1	2018_V2_016-WN-9007N	7
4	4	030-WN-9	030-WN-9	70.956	71.055	2	2018_V2_016-WN-9007N~1	8
5	6	008-I -007	008-I -007	169.447	169.472	1	2018_V3_059-I-75N	6
6	7	008-I -007	008-I -007	169.442	169.476	2	2018_V1_041-I-75N	7
7	8	008-I -007	008-I -007	176.022	176.12	1	2018_V3_059-I-75N	6
8	11	008-I -007	008-I -007	179.267	179.366	2	2018_V1_041-I-75N	7
9	13	025-I -006	025-I -006	89.481	89.517	1	2018_V3_120-I-64E	7
10	15	025-I -006	025-I -006	92.908	93.007	1	2018_V3_120-I-64E	4
11	22	088-KY-90	088-KY-90	57.742	57.842	1	2018_V1_119-KY-9009E~2	7
12	23	088-KY-90	088-KY-90	57.742	57.842	2	2018_V4_119-KY-9009E~1	7
13	24	088-KY-90	088-KY-90	59.545	59.646	1	2018_V1_119-KY-9009E~2	5
14	25	088-KY-90	088-KY-90	59.543	59.643	2	2018_V4_119-KY-9009E~1	7
15	28	047-WK-9(	047-WK-9(	120.699	120.798	1	2018_V2_047-WK-9001W	8

Conclusions

- Formalized the engineering statistics problem when agencies transition from legacy windshield pavement surveys to LCMS-based automatic pavement surveys.

- Identified statistically significant LCMS factors for each of the distress indices used by KYTC.
- Developed a framework to ensure data quality and compatibility across two survey databases.
- Developed the capability of mapping LCMS-based pavement measurements to windshield ratings using decision tree method.

- Novelties include:
  - The use of 13 statistics (10^{th}, 20^{th}, …, 90^{th} percentiles and others) to reconcile different data resolutions of LCMS and VES
  - The use of ensemble model for higher robustness
  - The use of conditional probability for higher accuracy
  - The use of adaboosting for robust performance

Ongoing and Future Research

- Extend the model development to other four VES indices.
- Continue to develop the decision support system for easy use of the developed decision tree models.
- Directly using LCMS variables to predict pavement deterioration.

Thanks to our sponsor and hard work of our collaborators from Kentucky Transportation Cabinet!!