ENSEMBLE MODEL TO ESTIMATE INCIDENT CLEARANCE DURATIONS USING SEQUENTIAL PARTITIONING PROCESS AND ROBUST REGRESSION

by Minsu Won and Gang-Len Chang

The University of Maryland, College Park

ACKNOWLEDGEMENT: This study was supported by Maryland-CHART program. We would like to thank MDOT State Highway Administration and their staffs, Egua Igbinosun, Mohammed Raqib, Jason Dicembre, and Sharon Hawkins, for their supports.

ABSTRACT

- This study has proposed an estimation methodology to circumvent various modeling issues and take advantage of unique characteristics revealed in most incident database for yielding a robust estimate of the incident duration.
- With the well-designed partitioning, clustering, and sequential tests to divide all incidents into several distinct groups, the proposed methodology will yield one primary model using all available data and supplemental models for incidents in each group that is specially calibrated to best fit their unique characteristics and statistical properties.
- By using the incident data from Maryland-CHART, the evaluation results confirm that the proposed methodology can indeed improve the estimation accuracy if properly integrated the primary model with each supplemental model.

INTRODUCTION

- Most agencies in response to and managing non-recurrent highway congestion are often requested by general public to provide the estimated delay and impacts of incidents to take proper control strategies, and in that regard the incident duration prediction model is one of the main components of the Traffic Incident Management (TIM) system.
- Despite many contributions of the previous studies to this subject, providing a reliable and robust estimate of the incident duration is still challenging.
  - Highly skewed distribution, complex correlations among the explanatory variables, mixed qualitative and quantitative variables, and heteroscedasticity.
- This study has proposed a methodology that can take advantage of the following unique characteristics of the incident duration data:
  - The resulting duration varies significantly among incidents of different natures (e.g., collision only vs. involving fatality);
  - The required duration to clear each type of incidents is often dominated by one or two factors (e.g., truck over-turned or hazard material related);
  - Some types of incidents, especially those in need of special equipment or assists, often have relatively small samples in the database; and
  - Some qualitative and quantitative factors recorded in the incident database are highly correlated or mutually dependent each other.

MODEL DEVELOPMENT

- This study has proposed the following procedures to yield a reliable ensemble model for estimating the incident duration:
  - **Stage-1**: Partitioning the entire dataset into several subsets via a sequential screening process, based on those factors contributing most to the duration of incidents with some unique natures.
  - **Stage-2**: Applying different modeling methods to each subset, based on available sample sizes, distribution patterns, and key contributing factors.
  - **Stage-3**: Integrating all those models, each developed specifically for one subset of incidents, with proper weights to constitute an ensemble model for the final estimate of a detected incident's required duration.

Stage-1: Sequential partitioning

- **Step-1**: Compute the median value for the entire dataset and for each subset that was classified with each of those identified key contributing factors.
- **Step-2**: Divide the entire dataset into a subgroup (G_1) corresponding to the selected factor (f_1) that has the longest median value (e.g., the group of incidents with fatality in this dataset) and the remaining data (R_1) as shown in FIGURE.
- **Step-3**: Follow the same logic to identify the most critical factor in the remaining dataset (R_1), and then re-divide the remaining dataset into a subgroup (G_2) and the sub-dataset (R_2).
- **Step-4**: Combine the subgroup (G_i) with its previously identified subgroup (G_{i-1}), if their differences in median clearance duration are statistically insignificant (e.g., test with Mann–Whitney U test) and the data points in those two sequentially classified subgroups are highly correlated (e.g., with Spearman's correlation test).
- **Step-5**: Stop the procedures and assign the last remaining data ($R_{i=n}$) to the last remaining group ($G_{i=n+1}$), if no other contributing factor can be used for data classification or no significant median differences exist between the subgroup (G_i) and the remaining data group (R_i).

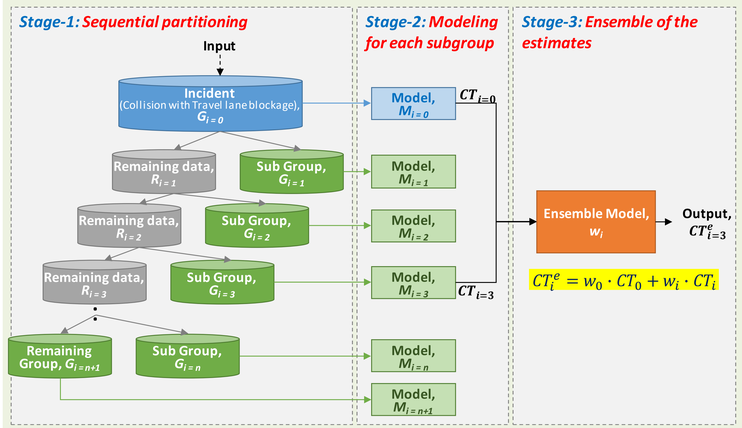
Stage-2: Modeling for each subgroup

- **Step-1**: Use the rule-based method to capture the relations between incident clearance time and its associated factors for those groups of a small sample size.
- **Step-2**: Directly apply the median value and its percentile interval for clearance time estimation for those groups suffering from both the small sample size and the lack of definitive rules.
- **Step-3**: Perform the normality test with both Kolmogorov-Smirnov and Shapiro-Wilk normality tests for those groups with a sufficient sample size.
- **Step-4**: Conduct the estimation of clearance time for those groups with the classical multiple regression if they have sufficient samples and follow the pattern of normal distribution.
- **Step-5**: Apply the hazard-based modeling method for those groups with sufficient samples but not normally distributed clearance times.

Stage-3: Ensemble of all estimates

- The final estimated clearance time (CT_i^e) for each detected incident shall be presented as the weighted combination of two estimates from the primary model(CT_0) and the sub-model(CT_i) using the robust regression

$$CT_i^e = w_0 \cdot CT_0 + w_i \cdot CT_i$$



<FIGURE. Graphically illustration of the modeling concept and its procedures>

<TABLE. Summary of Partitioning and Applied Modeling Approaches in Each Group>

Key contributing factor		Median (minutes)	Standard deviation	# of cases	Applied modeling approach
Primary group (All cases)		34	46	2009	AFT model (Weibull)
Sub group	Incident with Fatality	236	97	21	Rule-based model*
	Truck over-turned	160	94	29	Multiple linear regression
	Truck lost-load	96	9	2	Directly using the median
	6+ lanes blocked	94	92	30	AFT model (Log normal)
	Hazard material related	74	80	9	Rule-based model*
	Vehicle jack-knifed	66	35	11	Rule-based model*
	Vehicle over-turned	59	35	96	AFT model (Weibull)
	Medical service arrived	53	42	35	AFT model (Log logistic)
	TOW service arrived	52	37	505	AFT model (Log logistic)
	Harford region (County)	32	33	137	AFT model (Weibull)
	Incident with Injury	30	23	345	AFT model (Weibull)
	Wet pavement condition	26	22	99	AFT model (Weibull)
	Truck involved	22	27	154	AFT model (Weibull)
	Night time	22	21	111	AFT model (Weibull)
	AOC (Center)	20	20	334	AFT model (Weibull)
	PM-peak hour	14	18	23	Directly using the median
	Remaining cases	9	18	68	AFT model (Weibull)

** The detail process for the rule-based model is presented in Won et al. (2018).*

<TABLE. Calibration and Evaluation Results by Model and Incident Duration>

	Model	Actual incident clearance time (minutes)														
		< 30			30~60			60~120			≥ 120			Overall		
Accuracy (MAE)	1	16.54 (16.74)			13.42 (14.79)			32.39 (30.17)			121.20(76.41)			23.25 (21.82)		
	2	14.04 (15.26)			12.21 (13.11)			27.57 (29.22)			74.38 (67.47)			18.64 (19.94)		
	3	**11.67 (13.12)**			**12.65 (13.56)**			**28.90 (29.23)**			**76.04 (67.29)**			**18.04 (19.20)**		
Precision (SD of Error)	1	9.72 (15.17)			15.31 (14.52)			32.39 (22.98)			120.03(55.66)			38.96 (25.25)		
	2	9.93 (12.41)			10.98 (10.99)			23.20 (20.03)			57.45 (44.18)			22.82 (21.22)		
	3	**8.68 (11.75)**			**9.87 (10.59)**			**22.17 (20.93)**			**55.39 (41.42)**			**22.51 (21.00)**		
	Model	Actual incident clearance time with different error term (minutes)														
		< 30			30~60			60~120			≥ 120			Overall		
		<±15	<±30	<±60	<±15	<±30	<±60	<±15	<±30	<±60	<±30	<±60	<±120	<±15	<±30	<±60
Acceptable Accuracy	1	**45%** (**49%**)	93% (91%)	100% (99%)	71% (62%)	**91%** (**90%**)	98% (98%)	28% (30%)	57% (57%)	**89%** (**91%**)	13% (16%)	27% (40%)	**73%** (**88%**)	49% (48%)	**82%** (**81%**)	94% (94%)
	2	**59%** (**56%**)	94% (89%)	100% (99%)	71% (63%)	**94%** (**91%**)	99% (100%)	33% (29%)	64% (55%)	**92%** (**94%**)	23% (28%)	45% (44%)	**82%** (**92%**)	56% (51%)	**85%** (**80%**)	96% (95%)
	3	**71%** (**65%**)	97% (94%)	100% (99%)	66% (61%)	**96%** (**92%**)	100% (100%)	31% (32%)	60% (60%)	**92%** (**91%**)	22% (28%)	44% (48%)	**84%** (**92%**)	59% (54%)	**87%** (**83%**)	96% (95%)
Total # of Cases		897 (186)			673 (155)			342 (80)			97 (25)			2009 (446)		

** The number in the parenthesis indicates the results from the test set (the year 2017).*

CASE STUDY

- Incident data of I-95 (exit 27-109) in Maryland from Maryland-SHA CHART II Database
  - Years 2012-2015 for model calibration, 2016 for validation, and 2017 for evaluation
  - Incident Clearance Time (CT) as incident duration
  - Incident cases of collisions with travel lane blockage

MODEL EVALUATION

- To evaluate the developed model's performance, we have selected the following three models for comparison:
  - **Model-1**: AFT model developed with data from all cases (CT_0)
  - **Model-2**: Only the set of individual models; each developed specially for each group (CT_i)
  - **Model-3**: The proposed ensemble model with the calibrated weights ($w_0 \cdot CT_0 + w_i \cdot CT_i$)

DISCUSSIONS

- The proposed model can achieve a more reliable result, especially for both tails of the distribution of incident durations (i.e., for <30 & ≥120 categories).
  - One can expect a more accurate estimate for the fatality cases despite its small sample size, because the model is calibrated here first.
  - The remaining group has a quite small variance, despite less contributing factors to its model development. Thus, one can expect a quit accurate estimate without a sophisticated model.
- Due to the unique modeling process of the proposed model, it can improve the performance of any different primary model.

CONCLUSIONS

- This study has proposed an estimation methodology that allows the users to select any best-fit method to develop the primary estimation model and then employ the supplemental model, developed for each class of incidents with unique nature to reflect the dominating impacts of its most critical factors.
- By integrating the results from both the primary and the identified supplemental models, the finally estimated incident duration can not only reflect the compound impacts of all contributing variables, but also capture the dominating effects resulting from the unique role of some factors (e.g., ambulance vehicle) under different incident scenarios.