Annotated Bibliography for SDI Waze Pilot Project

Dan Flynn | Volpe Center

February 2018

Table of Contents

Annotated bibliography of tools and approaches used in existing analyses of road safety data, relevant for the Safety Data Initiative Waze/EDT pilot project.

Machine learning approaches in transportation safety

1 Modeling the dynamics of driver's dilemma zone perception using agent based modeling techniques

- In a driving simulation study, used Agent Based Models (ABM) to investigate how drivers impacts of driving in a 'dilemma zone', such as too close to an intersection to safely stop. Models include the MATsim dynamic agent-based traffic simulation model.

2 Influence of injury risk thresholds on the performance of an algorithm to predict crashes with serious injuries

- Optimization for first responders, using logistic regression on National Automotive Sampling System / Crash-worthiness Data System (NASS/CDS) to determine how variation in injury risk thresholds affects crash predictions. Standard logistic regression approach, where outcomes are binary for each type of crash (separate models, not ordinal).

3 A two-stage mining framework to explore key risk conditions on one-vehicle crash severity

- This research combines data mining and a logistic regression approach to identify crash severity in one-vehicle crashes. Genetic mining rule (GMR) model developed, to identify 'rules' which correspond to variables most associated with risk of a crash. The variables were then used in a hierarchical logistic regression (mixed logit model) to identify road conditions associated with serious crashes.
- Similar to proposed SDI Waze project approach, where random forests used to identify combinations of variables highly associated with EDT-level crashes, and then logistic regression used to assign probability of a crash to Waze events and test statistical significance. Use a training/validation approach for the rule-mining, 70% of data for training, 30% for validation.

4 Estimating likelihood of future crashes for crash-prone drivers

- Logistic regression on 8 years of traffic crash data in Louisiana. Use road characteristics, human factors, collision type, and weather in the model; use model diagnostics to assess true positives, sensitivity, and false positive rate for model predictions. Use area under receiver-operator curve (AUC) to assess model fit. Can correctly identify responsibility of crash of 62% of crashes, with the response variable being "at-fault" true or false.

5 Investigating injury severity risk factors in automobile crashes with predictive analytics and sensitivity analysis methods

- Predictive analytics used for injury severity models. refer to multinomial logistic regression (namely, ordinal logistic regression) as commonly used for injury severity analysis. Refer to previous work on

FARS data to use logistic regression for estimating if a crash would be fatal (Liu and McGee 1988). Some studies have used combination of ordered probit, ordered logit, and multinomial logit in combination (Park et al. 2012). Here use machine learning methods: artificial neural networks, support vector machines, and decision trees as an ensemble to develop a ranking of risk factors for crash injury severity.

- Data from the National Automotive Sampling System General Estimates System (NASS GES), with 1% of all national automobile crashes, for 2011 and 2012. Approximately 25 predictors used. K-fold cross-validation used in model development, and models evaluated with AUC.
- Focus is on developing a ranking of risk factors, rather than estimating crash severity from new input data, differing from the goals of the SDI Waze project.

6 Empirical Bayes approach for estimating urban deer-vehicle crashes using police and maintenance records

- 150 highway sections in Iowa. Use the Empirical Bayes approach with zero-inflated negative binomial regression for frequency of deer-vehicle crashes. Average annual daily traffic (AADT) used as exposure for highway sections, following AASHTO 2010 Highway Safety Manual recommendations.
- Model produces rankings of which highway sections are most suitable for focused safety improvement, based on crashes per mile-year.

7 Development of a Prediction Model for Crash Occurrence by Analyzing Traffic Crash and Citation Data

- Focus on human factors, such as traffic violation and crash history, in developing model of likelihood of crash occurrence at the driver level. Logistic regression approach, using Minitab, with model selection by AIC and assessment by AUC.

8 Application of classification algorithms for analysis of road safety risk factor dependencies

- Severity of injury for accidents modeled from historical incident data in California, 2004-2010. Naive Bayes and decision tree (CART) used to identify risk factors of greatest importance; use logistic regression to compare the output of the two classification approaches. AUC for model assessment.
- Refer to other studies using decision tree (CART) approaches for injury severity modeling: Kashani and Mohyamany 2011, Montella et al. 2011a,b, and others. Sohn and Shin 2001 compared ANN, logistic regression and CART for severity classification, finding each has similar classification accuracy.
- Differs from goals of SDI Waze in focusing on ranking risk factors, rather than producing estimated counts of crashes based on geospatial data. Found that the decision tree approach had best combination of true positive rate and false positive rate (AUC).

9 Data science application in intelligent transportation systems: An integrative approach for border delay prediction and traffic accident analysis

- Thesis focusing on intelligent transportation systems (ITS) in general. Uses Seasonal Autoregressive Integrated Moving Average Model (SARIMA) and Support Vector Regression (SVR) to model traffic accident data. Use k-nearest neighbor (KNN) as well.

10 Estimating the safety performance of urban road transportation networks

11 Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory

- Lord et al.[10,11] review of three common models for modelling crash count data: Poisson, Zero-inflated Poisson, and Zero-inflated negative binomial (also called Poisson-gamma) models. They point out that models which can account for zero-inflation, which can arise because of overly narrow time and space scale selection and rarity of crashes, often provide the best statistical fit, but may not characterize the underlying crash process completely.
- Provide detailed reviews of statistical theory behind these crash count models, and lay out how a zero-inflated model makes a simplifying assumption that a roadway can exist in either a 0-crash, ‘perfectly safe’ condition, or a non-zero crash, ‘imperfectly safe’ condition. They argue that having a too-small spatial or temporal scale can lead to over-estimation of the ‘perfectly safe’ condition.

12 The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives

- Excellent review of models used for crash frequency data. Discusses the commonly-used zero inflated negative binomial, as well as Poisson regression more generally. Discusses challenges with modeling crash frequency data, including overdispersion (variance exceeding the mean), correct choice of time window, and temporal and spatial correlation.
- Refers to common approaches for dealing with temporal and spatial correlation. GEE, GAM, and random effects (hierarchical) models also discussed.

- Machine learning models are briefly discussed, including neural networks and support vector machine models.

13 Performance Measures for Prioritizing Highway Safety Improvements Based on Predicted Crash Frequency and Severity

- Thesis on crash frequency modeling based on incident features, roadway infrastructure, demographic, and roadway network flow data. Estimate crash severity in scenarios of differing infrastructure and demographic change.
- Ordered probit model for crash frequency, which is an unusual application.

14 Factors influencing specificity and sensitivity of injury severity prediction (ISP) algorithm for AACN

- Use NASS CDS database of US vehicle accidents, 2005-2012, using a 'branching logistic regression' approach for modeling occurrence of minor or serious injury for crashes. Similar in some respects to a decision tree approach.
- Crash-level estimations of severity are the focus, rather than the number, pattern, and severity of crashes. Crash-level features include speed, impact direction, seat belt use, age, and gender.

15 Proactive Assessment of Accident Risk to Improve Safety on a System of Freeways

- Four freeway corridors selected, and historical crash data 2010-2011 assessed in combination with real-time traffic patterns. Logistic regression and decision trees (CART) used to assess crash or non-crash outcomes.
- Data aggregation and preparation discussed. Includes a useful literature review.
- Similar in some respects to goals of SDI Waze project, but using different data sets and with a different geographic and temporal scope.

16 Prioritizing Highway Safety Manual's crash prediction variables using boosted regression trees

- Decision tree (BRT, similar to random forests, based on CART) approach used to evaluate the impact of individual roadway characteristics on crash predictions. The goal here was to rank roadway characteristics, to prioritize which variables should be the focus of data collection, when resources are limited for roadway monitoring. Roadway characteristics are the input for the Highway Safety Manual (HSM) empirical Bayes approach to estimating crash frequency with negative binomial regression.
- Five years of data (2008-2012) in Florida used. Boosted regression tree (BRT) are similar to random forests, in using an ensemble of decision trees, and can be useful when most individual decision trees produce weak statistical predictions. Implemented in *gbm* package in R.

17 An Exploratory Computational Piecewise Approach to Characterizing and Analyzing Traffic Accident Data

- Six years of data (2008-2013) in North Dakota from state sources. Large number of crash-level data used. "Data analysis" is only fitting polynomial functions to the bivariate patterns, no statistical inference.

18 Risk Factors Analysis for Drivers with Multiple Crashes

- Identifying high-risk drivers using demographic characteristics, historical violations, and specific violation types with negative binomial regression. Crash estimation model identifies the set of predictors most strongly associated with high-risk drivers. Standard regression approach, models evaluated by AIC.

19 Crash Prediction Method for Freeway Facilities with High Occupancy Vehicle (HOV) and High Occupancy Toll (HOT) Lanes

- Segment-based crash frequency modeling, separate models for fatal/injury crashes and all crashes. Negative binomial regression approach, with AADT as exposure variable, segment length and number of lanes as additional important variables. Data from three states, CA, WA, and FL, from the Highway Safety Information System (HSIS). Models were run in SPSS, and spreadsheet tool developed using the fitted coefficients.

20 Developing Crash Models with Supporting Vector Machine for Urban Transportation Planning

- Support vector machines (SVM) unsupervised learning approach to discover patterns in crash frequency. Data from Louisiana urban roadways in 2011-2013, with crash frequency, roadway geometry, and AADT as main inputs. Little detail on model specification or application provide, largely a demonstration that SVM can be used.

21 Exploration of Advances in Statistical Methodologies for Crash Count and Severity Prediction Models

- PhD thesis from University of Connecticut, focusing on the simultaneous estimation of injury severity and vehicle damage using regression models. Simultaneous estimation is done by "copula based models"", and finds high correlation between injury and vehicle damage. Spatial analysis of road intersections and segments using socio-economic variables. Thirdly, carried analysis of crash type and crash severity on rural two-lane highways, using a multivariate Poisson lognormal model.
- Crash type and severity were better predicted by the multivariate Poisson lognormal than by negative binomial or univariate Poisson lognormal models.

22 Analyzing Traffic Crash Severity in Work Zones under Different Light Conditions

- Focus on work zones in Tennessee, 2003-2015, to assess factors determining crash severity (not count). Use Classification and Regression Trees (CART) to show importance of light conditions in crash severity, as well as roadway geometry factors, driver factors, and environmental factors (e.g., Weather, clear or not clear).
- Highest proportion of injury crashes for head-on collisions, along roadways, with greater than two lanes. Create three decision trees, one for each of the light conditions, and compare results. For instance, traffic control devices were effective in reducing crash severity in daylight and dark-lighted, but not dark-not-lighted conditions.

23 Predicting motor vehicle collisions using Bayesian neural network models: An empirical analysis

- Analysis of rural roads in Texas, comparing two types of neural network machine learning models, and a negative binomial regression model. Suggest that the Bayesian neural network is a useful approach for estimating crash counts in rural highways.
- Reviews the limitations of regression model approaches: need for clearly defined function relating crash frequencies and explanatory variables. Neural networks do not require *a priori* specification of a functional form relating these variables. Such models have however been criticized for over-fitting data and resulting in models without interpretable coefficients for explanatory variables. The Bayesian approach to a neural network can alleviate the former concern.
- Using a training/testing framework, neural networks outperformed negative binomial regressions for crash counts, with predictors of segment length, vehicles per day, shoulder width, and lane width.

Crowdsourced data analysis approaches

24 Learning from the crowd: Road infrastructure monitoring system

- Data collected automatically from new vehicles used as input for a decision tree analysis of road condition. Data collection relies on GPS, Wi-Fi, and sensors of vertical acceleration and pitch rate to detect features such as potholes.

25 Predicting Traffic Flow Regimes From Simulated Connected Vehicle Messages Using Data Analytics and Machine Learning

- Simulated data from a highway corridor in Seattle, to model traffic flow regimes under different conditions for connected vehicles. Three machine learning approaches were taken for traffic flow estimation: logistic regression, individual decision trees (CART), and random forests. Models were run in a Microsoft Azure cloud computing environment, using Apache Spark machine learning libraries.

- Focus is on connected vehicle configuration, operational conditions, market penetration, and estimating traffic flow rather than estimating crash counts. Useful detail on feature extraction, relying on principal component analysis (PCA) in R.

Spatial regression for road safety

26 Comparison of adjacency and distance-based approaches for spatial analysis of multimodal traffic crash data

- Model of spatial correlation for traffic crash counts at the county level. Develop two Bayesian models to look at how much adjacency explains in crash counts. 58 counties in California for 2012. Exposure variable of daily vehicle miles traveled (DVMT) from Highway Performance Monitoring System (HPMS) of FHWA.
- Poisson model for crash counts, with errors drawn from a normal distribution. Spatial autocorrelation is built in via the hyperparameter Σ, the covariance matrix which is used as the standard deviation of the error u_{ij}, using multivariate conditional auto-regressive (MCAR) model. Do not specify the modeling tool, but Stan likely used.

27 Spatial regression analysis of traffic crashes in Seoul

- Road segment based analysis for traffic crashes in Seoul, Korea, in 2010, using geographically weighted regression to account for spatial autocorrelation. Discuss conditional autoregressive (CAR) model, but end up using geographically weighted regression. Use Moran's I to assess strength of spatial autocorrelation. Use AIC to evaluate competing models.

28 Use of Roadway Attributes in Hot Spot Identification and Analysis

- Analysis of Utah "hot spots" for crashes, adding detailed roadway attribute layers such as vertical sag and grade to traditional variables such as lane width, number of lanes, shoulder width, and horizontal curvature. Use a hierarchical Bayesian Poisson mixture model.
- Use Bayesian horseshoe method for variable selection. This approach can take in a large number of possible variables, and assign a coefficient of zero to those which are unimportant. Lasso and ridge regression techniques serve a similar purpose in logistic regression models. Once variables were selected, a Bayesian Poisson regression was done on segments, using non-informative priors.
- Areas where many segments have observed crashes much greater than predicted crashes are considered hot spots. A number of specific hot spots are examined in detail.

29 Modeling crash and fatality counts along mainlines and frontage roads across Texas: The roles of design, the built environment, and weather

- Analysis of Texas highways, using spatial data on traffic, demography, land use, population and job density, rainfall, income, and education. Compare zero-inflated negative binomial, zero-inflated Poisson, and negative binomial models, finding the first preferred.
- Fully-spatial analysis (e.g., conditional autoregressive analysis) can be intractable for very large data sets, so segment-based analysis is typically used.
- Use 50-year average rainfall as the weather variable. Separate analysis for mainlanes and frontage roads. Population density and job densities found to be the strongest predictors of crash counts, along with urbanization. Age and income have negative effects; average rainfall slightly positive.

30 Bayesian spatial joint modeling of traffic crashes on an urban road network

- Poisson, negative binomial, and conditional autoregressive (CAR) models used to model crash counts at intersections and along road segments. A combination of spatial approaches to join intersections and segment models, with the segment models having traditional crash frequency modeling. Presents one way to approach fully spatially-explicit modeling of crash frequency on a road network, but too data-intensive to be useful for SDI Waze project.

References

1. Abbas, M. M. & Machiani, S. G. Modeling the dynamics of driver's dilemma zone perception using agent based modeling techniques. *International journal of transportation* **4,** 1–14 (2016).

2. Bahouth, G., Digges, K. & Schulman, C. Influence of injury risk thresholds on the performance of an algorithm to predict crashes with serious injuries. *Annals of Advances in Automotive Medicine* **56,** 223 (2012).

3. Chiou, Y.-C., Lan, L. W. & Chen, W.-P. A two-stage mining framework to explore key risk conditions on one-vehicle crash severity. *Accident Analysis & Prevention* **50,** 405–415 (2013).

4. Das, S., Sun, X., Wang, F. & Leboeuf, C. Estimating likelihood of future crashes for crash-prone drivers. *Journal of Traffic and Transportation Engineering (English Edition)* **2,** 145–157 (2015).

5. Delen, D., Tomak, L., Topuz, K. & Eryarsoy, E. Investigating injury severity risk factors in automobile crashes with predictive analytics and sensitivity analysis methods. *Journal of Transport & Health* **4,** 118–131 (2017).

6. Gkritza, K., Souleyrette, R. R., Baird, M. J. & Danielson, B. J. Empirical bayes approach for estimating urban deer-vehicle crashes using police and maintenance records. *Journal of Transportation Engineering* **140,** 04013002 (2013).

7. Gonzalez-Velez, E. & Gonzalez-Bonilla, A. *Development of a prediction model for crash occurrence by analyzing traffic crash and citation data.* (Transportation Informatics Tier I University Transportation Center, University at Buffalo, 2017).

8. Kwon, O. H., Rhee, W. & Yoon, Y. Application of classification algorithms for analysis of road safety risk factor dependencies. *Accident Analysis & Prevention* **75,** 1–15 (2015).

9. Lin, L. Data science application in intelligent transportation systems: An integrative approach for border delay prediction and traffic accident analysis. (State University of New York at Buffalo, 2015).

10. Lord, D. & Persaud, B. N. Estimating the safety performance of urban road transportation networks. *Accident Analysis & Prevention* **36,** 609–620 (2004).

11. Lord, D., Washington, S. P. & Ivan, J. N. Poisson, poisson-gamma and zero-inflated regression models of motor vehicle crashes: Balancing statistical fit and theory. *Accident Analysis & Prevention* **37,** 35–46 (2005).

12. Lord, D. & Mannering, F. The statistical analysis of crash-frequency data: A review and assessment of

methodological alternatives. *Transportation Research Part A: Policy and Practice* **44,** 291–305 (2010).

13. Morgan, N. S. Performance measures for prioritizing highway safety improvements based on predicted crash frequency and severity. (Auburn, 2013).

14. Pal, C. *et al.* Factors influencing specificity and sensitivity of injury severity prediction (isp) algorithm for aacn. *International journal of automotive engineering* **7,** 15–22 (2016).

15. Pande, A., Nuworsoo, C. & Shew, C. *Proactive assessment of accident risk to improve safety on a system of freeways.* (Mineta Transportation Institute; MTI Report 11-15, 2012).

16. Saha, D., Alluri, P. & Gan, A. Prioritizing highway safety manual's crash prediction variables using boosted regression trees. *Accident Analysis & Prevention* **79,** 133–144 (2015).

17. Saleem, F., Asa, E. & Membah, J. An exploratory computational piecewise approach to characterizing and analyzing traffic accident data. *International Journal of Scientific and Technical Research in Engineering* (2016).

18. Shawky, M. & Al-Ghafli, A. Risk factors analysis for drivers with multiple crashes. *International Journal of Engineering and Applied Sciences* **3,** 42–48 (2016).

19. Srinivasan, S. *et al. Crash prediction method for freeway facilities with high occupancy vehicle (hov) and high occupancy toll (hot) lanes.* (Florida Department of Transportation, 2015).

20. Sun, X., Das, S. & Broussard, N. Developing crash models with supporting vector machine for urban transportation planning. in *17th international conference road safety on five continents (rs5c 2016), rio de janeiro, brazil, 17-19 may 2016.* (Statens väg-och transportforskningsinstitut, 2016).

21. Wang, K. Exploration of advances in statistical methodologies for crash count and severity prediction models. (University of Connecticut, 2016).

22. Wei, X., Shu, X., Huang, B., Taylor, E. L. & Chen, H. Analyzing traffic crash severity in work zones under different light conditions. *Journal of Advanced Transportation* **2017,** (2017).

23. Xie, Y., Lord, D. & Zhang, Y. Predicting motor vehicle collisions using bayesian neural network models: An empirical analysis. *Accident Analysis & Prevention* **39,** 922–933 (2007).

24. Masino, J., Thumm, J., Frey, M. & Gauterin, F. Learning from the crowd: Road infrastructure monitoring system. *Journal of Traffic and Transportation Engineering (English Edition)* **4,** 451–463 (2017).

25. Vasudevan, M., Curtis, C., Lowman, A. & O'Hara, J. *Predicting traffic flow regimes from simulated connected vehicle messages using data analytics and machine learning.* (ITS Joint Program Office, Department of Transportation, 2016).

26. Gill, G., Sakrani, T., Cheng, W. & Zhou, J. Comparison of adjacency and distance-based approaches for spatial analysis of multimodal traffic crash data. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences* **42,** (2017).

27. Rhee, K.-A., Kim, J.-K., Lee, Y.-I. & Ulfarsson, G. F. Spatial regression analysis of traffic crashes in seoul. *Accident Analysis & Prevention* **91,** 190–199 (2016).

28. Schultz, B., Grant G. *Use of roadway attributes in hot spot identification and analysis.* (Utah Department of Transportation; Brigham Young University-Provo, 2015).

29. Xu, J., Kockelman, K. M. & Wang, Y. Modeling crash and fatality counts along mainlanes and frontage roads across texas: The roles of design, the built environment, and weather. *93rd Annual Meeting of the Transportation Research* **22,** 24 (2014).

30. Zeng, Q. & Huang, H. Bayesian spatial joint modeling of traffic crashes on an urban road network. *Accident Analysis & Prevention* **67,** 105–112 (2014).