

# 24 May 2021 Report

Brad Burkman

24 May 2021

## Contents

<b>1</b>	<b>Progress</b>	<b>1</b>
<b>2</b>	<b>Jargon to Understand</b>	<b>1</b>
<b>3</b>	<b>IRB, SHRP Database</b>	<b>2</b>
<b>4</b>	<b>NGSIM Database</b>	<b>2</b>
<b>5</b>	<b>Mining Bibliographies</b>	<b>2</b>
<b>6</b>	<b>Ideas for a Paper</b>	<b>3</b>
6.1	Comparing Image Transformation Techniques . . . . .	3
6.2	Comparing Data Cleaning Techniques . . . . .	3
6.3	Missing Data Analysis . . . . .	3
6.4	Fill Holes in Osman et al . . . . .	3
<b>7</b>	<b>References</b>	<b>4</b>

## 1 Progress

Besides the things below, I reviewed more *AAEP* articles, and read the two articles in the References. Reading the articles is slow, because I'm taking notes on, and looking up, things I don't understand.

## 2 Jargon to Understand

- Naturalistic Driving Data - Data collected from sensors installed in the driver's own car, trying to get as close as possible to the driver's "natural" behavior.
- Heterogeneity. I understand vaguely what "data heterogeneity" means, but I'm going to watch for the term to see how it's used in the context of these papers.

### 3 IRB, SHRP Database

Eleven of the papers in *Accident Analysis and Prevention* used the Strategic Highway Research Program 2 (SHRP2) Naturalistic Driving Study (NDS), which put sensors in 3400 cars and recorded five million trips, including crashes. To get “Qualified Researcher Status” with “full access to data that has been made available through the SHRP 2 NDS Data Access Website,” I had to submit a certificate of training on research with human subjects. I did the training through the UL Institutional Review Board (IRB). I now have access.

### 4 NGSIM Database

Three papers use the Next Generation Simulation dataset from the US Dept of Transportation, and it’s available for download with no restrictions.

### 5 Mining Bibliographies

There’s an open-source project called Anystyle. If you have a folder with (hundreds, thousands) of PDF files, Anystyle will extract the *References* sections of all of the files as a BibTEX file.

I have a file of the 190 articles from *Accident Analysis and Prevention* that I’ve reviewed. I used Anystyle to extract the references, and converted it to a Pandas dataframe. Of the 10,822 references, 8,685 are unique. Most interesting is to find the most-cited articles. Seventeen of the 190 articles cited Mannering’s 2016 article.

	title	author	date	counts
8231	Unobserved heterogeneity and the statistical a...	Mannering, F.L. and Shankar, V. and Bhat, C.R.	2016	17
6238	Random forests	Breiman, L.	2001	16
7854	The statistical analysis of crash-frequency da...	Lord, D. and Mannering, F.	2010	14
2080	Comparison of four statistical and machine lea...	Iranitalab, A. and Khattak, A.	2017	12
8374	Utilizing support vector machine in real-time ...	Yu, R. and Abdel-Aty, M.	2013	11
1201	Analytic methods in accident research: methodo...	Mannering, F.L. and Bhat, C.R.	2014	11
7860	The statistical analysis of highway crash-inju...	Savolainen, P.T. and Mannering, F.L. and Lord,...	2011	10
4120	Highway accident severities and the mixed logi...	Milton, J.C. and Shankar, V.N. and Mannering, ...	2008	9
8345	Using support vector machine models for crash ...	Li, Z. and Liu, P. and Wang, W. and Xu, C.	2012	8
6600	SMOTE: synthetic minority over-sampling technique	Chawla, N.V. and Bowyer, K.W. and Hall, L.O. a...	2002	8
5961	Predicting crash likelihood and severity on fr...	Xu, C. and Tarko, A.P. and Wang, W. and Liu, P.	2013	8
4301	Identifying significant predictors of injury s...	Delen, D. and Sharda, R. and Bessonov, M.	2006	8

I had to spend some time playing with the arguments, because default Anystyle doesn’t handle two-column PDF files well.

## 6 Ideas for a Paper

### 6.1 Comparing Image Transformation Techniques

In “A deep learning based traffic crash severity prediction framework” by Rahim (LSU) [1], they transformed each data point into an image, and used the images as input for a Convolutional Neural Network (CNN) for classification. There are several ways to make such a transformation. Write a paper where I compare different image transformation methods, and their effect on different metrics (precision, recall, accuracy, sensitivity, f1, false alarm rate) of the classification of the test set.

### 6.2 Comparing Data Cleaning Techniques

In “A deep learning based traffic crash severity prediction framework” by Rahim (LSU) [1], they just deleted any records with missing or inconsistent data. The *Titanic* Kaggle sites you showed me use several other methods for filling in incomplete data. Write a paper where I compare different methods for dealing with missing data, and their effect on different metrics (precision, recall, accuracy, sensitivity, f1, false alarm rate) of the classification of the test set.

Rahim’s article took out 37% of the records for missing or inconsistent data, but only 21% of the fatal crashes; could that imbalance in the data cleaning skew the model prediction? It makes sense that police would be more meticulous in their record keeping for fatal crashes, but 21% and 37% are huge.

Would we get a better model if we found a good way to fill in missing data?

### 6.3 Missing Data Analysis

This isn’t a ML paper, but somebody should write it. In what situations is the crash data most likely to be incomplete or inconsistent? Rather than a, “How can we fix the roads to make them safer?,” this would be a “How can we get better records, so we can know how to fix the roads to make them safer?”

### 6.4 Fill Holes in Osman et al

A group at LSU has a 2019 paper [2] using the SHRP-2 data to predict, based on the driving behavior whether the driver was making a phone call, texting, or having a conversation with a passenger, but “the effect of roadway and geometric characteristics is not within the scope of this paper.” They say that in the introduction and devote the entire last paragraph of the conclusion to laying out that hole that needs to be filled.

The data is not just crashes, but any time a person was using a phone, texting, or talking.

## 7 References

- [RH21] Md Adilur Rahim and Hany M. Hassan. “A deep learning based traffic crash severity prediction framework”. In: *Accident Analysis & Prevention* 154 (2021), p. 106090. ISSN: 0001-4575. DOI: <https://doi.org/10.1016/j.aap.2021.106090>. URL: <https://www.sciencedirect.com/science/article/pii/S0001457521001214>.
- [Osm+19] Osama A. Osman, Mustafa Hajij, Sogand Karbalaieali, et al. “A hierarchical machine learning classification approach for secondary task identification from observed driving behavior data”. In: *Accident Analysis & Prevention* 123 (2019), pp. 274–281. ISSN: 0001-4575. DOI: <https://doi.org/10.1016/j.aap.2018.12.005>. URL: <https://www.sciencedirect.com/science/article/pii/S000145751831114X>.