A BAYESIAN SURVIVAL APPROACH TO ANALYZING THE RISK OF RECURRENT RAIL DEFECTS

Faeze Ghofrani, Reza Mohammadi, Qing He, Abhishek Pathak, Amjad Aref

INFORMS 2018

Phoenix, Arizona
November 6th, 2018
CONTENTS

1. Introduction
2. Background
3. Contributions of the study
4. Methodology framework
5. Results
6. Findings
7. Conclusion
Background

- Rail defects (and more specifically broken rails) account for the largest portion of total train derailments causes
- Many of rail defects occur in the same location multiple times due to the characteristics of the location.
- Analysis of recurrent rail defects can assist in identifying network hotspots and providing on-time responsive maintenance.


Number of freight train derailments by accident cause on Class I main lines (Liu et al. 2012).
Contributions of the Study

- Designing a comprehensive logical methodology framework for data collection, pre-processing, and modeling based on a collection of datasets from different resources in a Class I railroad

- Applying the correlated event times of survival analysis in the context of railway transportation for recurrent rail defects

- Developing a Bayesian framework by performing Markov Chain Monte Carlo (MCMC) simulation for optimizing the parameters of the model

- Verifying the fit of the model by using Cox-Snell residual plot
The methodology framework

**Type of Data Sources**

- **Defects**
  - Rail Defects
  - Geometry Defects

- **Infrastructure**
  - Curve
  - Grade
  - Rail laid

- **Traffic**
  - Tonnage
  - Speed

- **Maintenance**
  - Inspection
  - Rail Patch

**Inputs**

- Cleaning each dataset
- Creating rail segments of 0.01 mile
- Creating time intervals based on tonnage
- Aggregating rail defects of same type
- Linking data sources according to location and date
- Providing new factors based on the existing factors
- Attribute or record elimination

**Processing**

- Descriptive Statistical Analysis
- Developing Bayesian hazard-based duration models

**Modeling**

**Decision-Making**

Capital Planning and Responsive Maintenance Planning
# Data Pre-processing

- Dividing the network into short sections (0.01 mile)
- Creating tonnage intervals for each segment
- Creating new variables for each interval
- Linking all datasets based on spatial units with rail defect dataset
- Record elimination

### Table: Sample Data

<table>
<thead>
<tr>
<th>Prefix</th>
<th>Track Type</th>
<th>Milepost</th>
<th>Defect Type</th>
<th>ID</th>
<th>Interval_Number</th>
<th>Time_from</th>
<th>Time_to</th>
<th>Status</th>
<th>Start_Tonn</th>
<th>End_Tonn</th>
<th>Tonn_Gap</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>390.89</td>
<td>TDD</td>
<td>351</td>
<td>1</td>
<td>1/3/2011</td>
<td>2/12/2015</td>
<td>1</td>
<td>798768707.1</td>
<td>798768707.1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>390.89</td>
<td>TDD</td>
<td>351</td>
<td>2</td>
<td>2/12/2015</td>
<td>5/14/2015</td>
<td>1</td>
<td>798768707.1</td>
<td>873199155</td>
<td>744304477.9</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>390.89</td>
<td>TDD</td>
<td>351</td>
<td>3</td>
<td>5/14/2015</td>
<td>5/14/2015</td>
<td>1</td>
<td>873199155</td>
<td>875232759.1</td>
<td>2033604.1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>390.89</td>
<td>TDD</td>
<td>351</td>
<td>4</td>
<td>5/14/2015</td>
<td>12/30/2016</td>
<td>0</td>
<td>875232759.1</td>
<td>1221722965</td>
<td>346490205.9</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>390.89</td>
<td>TDD</td>
<td>351</td>
<td>5</td>
<td>1/3/2011</td>
<td>8/8/2011</td>
<td>1</td>
<td>18186814.97</td>
<td>18186814.97</td>
<td>0</td>
</tr>
<tr>
<td>0 CC</td>
<td>SG</td>
<td>249.14</td>
<td>CH</td>
<td>7791</td>
<td>1</td>
<td>8/8/2011</td>
<td>10/13/2011</td>
<td>1</td>
<td>18186814.97</td>
<td>25629467.96</td>
<td>7442652.99</td>
</tr>
<tr>
<td>0 CC</td>
<td>SG</td>
<td>249.14</td>
<td>CH</td>
<td>7791</td>
<td>2</td>
<td>10/13/2011</td>
<td>12/13/2011</td>
<td>1</td>
<td>25629467.96</td>
<td>35059404.49</td>
<td>9429336.53</td>
</tr>
<tr>
<td>0 CC</td>
<td>SG</td>
<td>249.14</td>
<td>CH</td>
<td>7791</td>
<td>3</td>
<td>12/13/2011</td>
<td>2/17/2012</td>
<td>1</td>
<td>35059404.49</td>
<td>43349418.76</td>
<td>8290014.27</td>
</tr>
<tr>
<td>0 CC</td>
<td>SG</td>
<td>249.14</td>
<td>CH</td>
<td>7791</td>
<td>4</td>
<td>2/17/2012</td>
<td>12/30/2016</td>
<td>0</td>
<td>43349418.76</td>
<td>240865278.6</td>
<td>197515859.8</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>2.023</td>
<td>SD</td>
<td>101</td>
<td>1</td>
<td>1/3/2011</td>
<td>9/9/2016</td>
<td>1</td>
<td>75235419.02</td>
<td>75235419.02</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>2.024</td>
<td>SD</td>
<td>101</td>
<td>2</td>
<td>9/9/2016</td>
<td>10/7/2016</td>
<td>1</td>
<td>75235419.02</td>
<td>77930334.73</td>
<td>26949153.71</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>2.024</td>
<td>SD</td>
<td>101</td>
<td>3</td>
<td>10/7/2016</td>
<td>12/30/2016</td>
<td>0</td>
<td>77930334.73</td>
<td>81864102.31</td>
<td>3933767.58</td>
</tr>
</tbody>
</table>
Descriptive Statistical Analysis

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Def_Type</td>
<td>Type of defects (BHB, EBF, FBW, TDD, etc.)</td>
</tr>
<tr>
<td>Geo_Def</td>
<td>Number of geometry defects in the last three years</td>
</tr>
<tr>
<td>Freight_Speed</td>
<td>Freight train speed at the location of the defect</td>
</tr>
<tr>
<td>Def_Freq</td>
<td>The frequency of defects occurred in the same location</td>
</tr>
<tr>
<td>Inspection_Freq</td>
<td>Number of inspections in the last three years</td>
</tr>
<tr>
<td>Curve/Tang</td>
<td>Whether the occurred defect is on the low side of a curve, high side of a curve or on tangents</td>
</tr>
<tr>
<td>Weight</td>
<td>The weight of rail at the location of the defect</td>
</tr>
<tr>
<td>MGT</td>
<td>The tonnage of freight (in MGT) at the location of the defect</td>
</tr>
</tbody>
</table>

Segments with one, two and more than two recurrent defects

Type of defects
## Modeling Approach

<table>
<thead>
<tr>
<th>Model Types</th>
<th>When Used?</th>
<th>Intensity Function</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Event Counts Models</td>
<td>- Individuals frequently experience the events of interest</td>
<td>[ \lambda(t</td>
<td>H(t)) = \lim_{\Delta t \downarrow 0} \frac{\Pr {\Delta N(t) = 1</td>
</tr>
<tr>
<td></td>
<td>- Events are incidental</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gap Time Models</td>
<td>- Events are relatively infrequent</td>
<td>[ \lambda(t</td>
<td>H(t)) = h(t - T_{N(t^-)}) ]</td>
</tr>
<tr>
<td></td>
<td>- Some type of individual renewal occurs after an event</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Prediction of the time to the next event is of interest.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Correlated event times

Rail segments might have different features, varied experiences, material traits, etc.

Sources of Correlation

the risk for a rail defect is a function of the occurrence of previous defects

Occurrence dependence

Heterogeneity across track segments

Shared Frailty Model
Shared Frailty Model Specification

Baseline hazard function: *Weibull* distribution with parameter $\mu$ and $\gamma$

$$h(t) = \mu \gamma t^{\gamma - 1}$$

hazard function

$$S(t) = e^{-\mu t^\gamma}$$

survival function

$$f(t) = \mu \lambda t^{\gamma - 1} e^{-\mu t^\gamma}, \mu > 0, \gamma > 0$$

density function for Weibull distribution

Likelihoods for Proportional Hazard (PH) model under the frailty approach

$$L_{PHFM} = \prod_{i=1}^{n} \left[ \mu \gamma t_i^{\gamma - 1} e^{\beta' x + z_i} \right]^{\delta_i} e^{-\mu t^\gamma e^{\beta' x + z_i}}$$

$z_i$: shared frailty variable with normal prior ($z_1, \ldots, z_m \sim N(0, \tau^2)$) and density function:

$$f(Z) = \frac{1}{\tau \sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{Z}{\tau}\right)^2} ; -\infty < Z < \infty, \tau > 0$$

Prior Specification

$\tau^{-2} \sim \Gamma(a_\tau, b_\tau), \beta \sim N(\mu, \mu), \mu \sim \Gamma(\rho, \rho), \gamma \sim \Gamma(a, b)$

Posterior Calculation and parameter configuration

$$P_{PH} = \prod_{i=1}^{n} \left[ \mu \gamma t_i^{\gamma - 1} e^{\beta' x + z_i} \right]^{\delta_i} e^{-\mu t^\gamma e^{\beta' x + z_i}} \pi(Z) \pi(\mu) \pi(\gamma) \pi(\beta) \pi(\tau^2)$$

MCMC calculations have been implemented with burn-in period of 5,000 iterations and the Markov chain subsample to get a final chain size of 4,000 iterations
## Results

### Posterior Inference of Regression Coefficients and frailty variance

<table>
<thead>
<tr>
<th>Variables</th>
<th>Mean</th>
<th>Median</th>
<th>Std. Dev.</th>
<th>95% CI-Low</th>
<th>95% CI-Upp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight of Rail</td>
<td>-0.0006</td>
<td>-0.0005</td>
<td>0.0026</td>
<td>-0.0058</td>
<td>0.0044</td>
</tr>
<tr>
<td>Frequency of Geometry Defects</td>
<td>0.0123</td>
<td>0.0124</td>
<td>0.0082</td>
<td>-0.0036</td>
<td>0.0283</td>
</tr>
<tr>
<td>Location of Rail (L)</td>
<td>-0.0320</td>
<td>-0.0306</td>
<td>0.0527</td>
<td>-0.1426</td>
<td>0.0689</td>
</tr>
<tr>
<td>Location of Rail (T)</td>
<td>-0.0551</td>
<td>-0.0546</td>
<td>0.0320</td>
<td>-0.1187</td>
<td>0.0083</td>
</tr>
<tr>
<td>Freight Speed Limit</td>
<td>0.0014</td>
<td>0.0015</td>
<td>0.0012</td>
<td>-0.0010</td>
<td>0.0038</td>
</tr>
<tr>
<td>Type of Defect (TDD)</td>
<td>0.1693</td>
<td>0.1691</td>
<td>0.0674</td>
<td>0.0369</td>
<td>0.3074</td>
</tr>
<tr>
<td>Type of Defect (TW)</td>
<td>0.1999</td>
<td>0.1994</td>
<td>0.0721</td>
<td>0.0603</td>
<td>0.3370</td>
</tr>
<tr>
<td>Type of Defect (VSH)</td>
<td>0.2956</td>
<td>0.2976</td>
<td>0.1202</td>
<td>0.0549</td>
<td>0.5222</td>
</tr>
</tbody>
</table>

### Posterior inference of frailty variance

| Posterior inference of frailty variance | 0.491 | 0.271 | 0.040 | 0.006 | 0.891 |

There is a positive relationship between the occurrence of the defects on the same rail segment;
Trace plots for the regression coefficients and frailty variance

- The parameters have sufficient state changes as the MCMC algorithm runs
- This implies that priori distribution is well calibrated
Density plots for posterior distribution of the regression coefficients and frailty variance

- The estimates of the posterior marginal distribution for the coefficients have smooth and unimodal shapes.
Model Diagnostics

Estimated cumulative hazard function versus the Cox and Snell residual

Cox Snell Plot for the fitted model

Residuals from a correctly fitted model follow unit exponential distribution
Findings

The following factors increase the risk interval for defect recurrences (decrease the hazard):

- Weight of rail
- Location of defect (if low side of the curve or tangent) compared to high side of the curve

The following factors decrease the risk interval for defect recurrences (increase the hazard):

- Number of geo defects in the last three years
- Freight train speed limit
- “TDD, TW and VSH” types of defect compared to “SD”
Conclusions

- Data collection from a Class I railroad
- Cleaning, fusion, pre-processing and restructuring of multiple datasets
- Study of models for recurrent survival analysis
- Provide a model for predicting the tonnage until the re-occurrence of a rail defect
- According to findings railroad could record the data on lighter segments of rail as well as on the high side of the curves on rail segments to be more certain on the risk of defect recurrence.
- The impact of past geometry defects can be reduced by minimizing their occurrence either by accurately predicting the defects or by decreasing the inspection intervals
- The impact of the speed can be reduced by decreasing the speed limit on certain segments.
- The rail segments with higher risk of frequent rail defects shall be treated more frequently.
Thank you!

faezegho@buffalo.edu
References


