Vehicle Incident Analysis and Prediction In Nashville Metro

Presented by Geoffrey Pettet
Outline

• Background
• Problem Breakdown
• Dataset Description
• Solution Description
  – Clustering
  – Survival Analysis
  – Bayesian Network
• Validation
• Future Work
Nashville’s Current Dispatch

*Reactive*: Based on euclidean distance

**“Vortex” Problem**

1. Incident Reported in outlying region

Nashville divided into hex cells of 0.5 mile diameter
Nashville’s Current Dispatch

**Reactive:** Based on euclidean distance

**“Vortex” Problem**

1. Incident Reported in outlying region
2. EMS vehicle responds from nearby station

Nashville divided into hex cells of 0.5 mile diameter
Nashville’s Current Dispatch

*Reactive:* Based on euclidean distance

**“Vortex” Problem**

1. Incident Reported in outlying region
2. EMS vehicle responds from nearby station
3. Patient taken to hospital

Nashville divided into hex cells of 0.5 mile diameter
Nashville’s Current Dispatch

*Reactive:* Based on euclidean distance

**“Vortex” Problem**

1. Incident Reported in outlying region
2. EMS vehicle responds from nearby station
3. Patient taken to hospital
4. Ambulance assigned to nearby incident when returning to station

Nashville divided into hex cells of 0.5 mile diameter
Nashville’s Current Dispatch

Reactive: Based on euclidean distance

“Vortex” Problem

1. Incident Reported in outlying region
2. EMS vehicle responds from nearby station
3. Patient taken to hospital
4. Ambulance assigned to nearby incident when returning to station

Nashville divided into hex cells of 0.5 mile diameter
Nashville’s Current Dispatch

**Reactive:** Based on euclidean distance

**“Vortex” Problem**

1. Incident Reported in outlying region
2. EMS vehicle responds from nearby station
3. Patient taken to hospital
4. Ambulance assigned to nearby incident when returning to station

Nashville divided into hex cells of 0.5 mile diameter
Nashville’s Current Dispatch

Average Responses Times per Station

- **Stations**
  - Average Response Time
  - Max Response Time
Nashville’s Current Dispatch

Station Response per Cell

Columns: Cell Regions

Responses

Colors: Unique Stations
Nashville’s Current Dispatch

- This is not optimal
  - Ignores factors such as traffic congestion and possible future incidents
- Goal: Improve
  - Response Time
  - Resource Allocation - reduce spread of EMS vehicles (and operating cost)
Improving *response time* and *resource allocation* requires two steps:

1. **Prediction**
   Understand future resource demand

2. **Dispatch**
   Design an optimal dispatch algorithm based on said future demand
Problem Breakdown

**Prediction**
Understand future resource demand

**Goal:** predict incident likelihood/severity for each region of city within some time interval
Dataset

• Metro Nashville Fire Department incident data from February 2014 to February 2016
• 19,910 usable motor vehicle accident records
Dataset

• Metro Nashville Fire Department incident data from February 2014 to February 2016
• 19,910 usable motor vehicle accident records

• Included Features:
  – GPS Coordinates
  – occurrence time
  – first response time
  – accident description
Dataset

• Metro Nashville Fire Department incident data from February 2014 to February 2016
• 19,910 usable motor vehicle accident records

• Included Features:
  – GPS Coordinates
  – occurrence time
  – first response time
  – accident description

• Added Features:
  – Weather Conditions from DarkSky
  – Intersection Distance from OTP
Incident Prediction - Existing Work

- Popular topic due to large safety and monetary costs
Incident Prediction - Existing Work

• Popular topic due to large safety and monetary costs
• Several techniques researched, including:
  – Binomial distribution analysis [Miaou, Lum 1993]
  – artificial neural networks [Chang 2005]
Incident Prediction - Existing Work

• Popular topic due to large safety and monetary costs
• Several techniques researched, including:
  – Binomial distribution analysis [Miaou, Lum 1993]
  – artificial neural networks [Chang 2005]
• Used to great effect when predicting incident frequency for well defined areas, such as
  – specific intersections
  – segments of a freeway
Incident Prediction - Existing Work

• Popular topic due to large safety and monetary costs
• Several techniques researched, including:
  – Binomial distribution analysis [Miaou, Lum 1993]
  – artificial neural networks [Chang 2005]
• Used to great effect when predicting incident frequency for well defined areas, such as
  – specific intersections
  – segments of a freeway
• **Problem:** make assumptions about the locations being analyzed
Incident Prediction - Existing Work

**Existing Work**

- Works well when location is well defined
- Breaks down when applied to large, heterogeneous area (Nashville Metro)
Incident Prediction - Existing Work

Existing Work

Location Features

Our Method

Incident Features

- Works well when location is well defined
- Breaks down when applied to large, heterogeneous area (Nashville Metro)

- Idea - similar incidents can happen in different city areas
  - not dependent on features of area
Solution Overview

• Find groups of similar incidents
• Hypothesis: similar incidents have similar arrival rates
Solution Overview

- Statistical technique - predict time until “failure”
- Predict time until next incident occurs in each cluster
Solution Overview

- Connect cluster incident probabilities to spatial locations (i.e. hex cells)
Solution Overview

Cluster Identification → Cluster Incident Likelihood → Spatial Correlation

Real-Time Hex Cell Incident Likelihoods
Clustering

- **Hypothesis**: incidents with similar features have similar arrival rates
  - Find groups of similar incidents using *clustering*
Clustering

• **Hypothesis**: incidents with similar features have similar arrival rates
  — Find groups of similar incidents using *clustering*

• Classical clustering techniques (K-means, K-modes, etc.) work on either **numeric** or **nominal** data
Clustering

- **Hypothesis**: incidents with similar features have similar arrival rates
  - Find groups of similar incidents using *clustering*
- **Classical clustering techniques** (K-means, K-modes, etc.) work on either *numeric* or *nominal* data
  - Our dataset has mixed data types

<table>
<thead>
<tr>
<th>Latitude</th>
<th>Longitude</th>
<th>Speed Limit</th>
<th>Road Type</th>
<th>Weather</th>
<th>Severity</th>
<th>Dist. to Intersection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Numeric</td>
<td>Numeric</td>
<td>Numeric</td>
<td>Nominal</td>
<td>Nominal</td>
<td>Nominal</td>
<td>Numeric</td>
</tr>
<tr>
<td>3631249</td>
<td>-86.73</td>
<td>45 mph</td>
<td>Residential</td>
<td>Fog</td>
<td>B</td>
<td>12.5 ft</td>
</tr>
</tbody>
</table>
Clustering

• Hypothesis: incidents with similar features have similar arrival rates
  — Find groups of similar incidents using clustering

• Classical clustering techniques (K-means, K-modes, etc.) work on either numeric or nominal data
  — Our dataset has mixed data types

Need clustering technique that works for mixed typed data
Clustering

- Technique Used: *Similarity Based Agglomerative Clustering (SBAC)* algorithm [Li, Biswas 2002]
Clustering

- **Technique Used**: *Similarity Based Agglomerative Clustering (SBAC)* algorithm [Li, Biswas 2002]
  - Bases grouping on unusual features of incidents
  - More unique value = more weight in similarity calculations
Clustering

- **Technique Used:** *Similarity Based Agglomerative Clustering (SBAC)* algorithm [Li, Biswas 2002]
  - Bases grouping on unusual features of incidents
  - More unique value = more weight in similarity calculations
- **Example:** Incidents Occurring in Snow
  - If only a few incidents occur in snowy weather compared to sunny weather, than the weather feature is more important for these incidents
Clustering

Example Cluster

- **Average Dissimilarity** - how different the values of each feature in this cluster are compared to all other clusters
  - $\text{Dissim} = 1$: values not represented in any other cluster
  - $\text{Dissim} = 0$: values are identical to other clusters
Clustering

Example Cluster

Features

Weather Values in Cluster

<table>
<thead>
<tr>
<th>Avg Dissim with Other Clusters</th>
<th>Weather Values in Cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>Road Type</td>
<td>Snow</td>
</tr>
<tr>
<td>Weather</td>
<td>rain</td>
</tr>
<tr>
<td>Severity</td>
<td>cloudy</td>
</tr>
<tr>
<td>Nature of Accident</td>
<td>clear-day</td>
</tr>
<tr>
<td>Disc. TOD</td>
<td>clear-night</td>
</tr>
<tr>
<td>Day</td>
<td>partly-cloudy-day</td>
</tr>
<tr>
<td>Month</td>
<td>partly-cloudy-night</td>
</tr>
<tr>
<td>Intersection Distance</td>
<td>wind</td>
</tr>
<tr>
<td></td>
<td>fog</td>
</tr>
</tbody>
</table>
Survival Analysis

• Predicts when an event of interest is likely to occur
  – Trained using historical data
  – Used in medicine, component failure analysis, etc.
Survival Analysis

• Predicts when an event of interest is likely to occur
  – Trained using historical data
  – Used in medicine, component failure analysis, etc.

• Applied to each cluster, gives likelihood of incidents occurring in given time frame
  – we use Exponential Model
Survival Analysis

• Predicts when an event of interest is likely to occur
  – Trained using historical data
  – Used in medicine, component failure analysis, etc.

• Applied to each cluster, gives likelihood of incidents occurring in given time frame
  – we use Exponential Model
    • memoryless
Survival Analysis

Cluster Survival Analysis - Results

• Comparison of models applied to clusters and entire dataset
Survival Analysis

Cluster Survival Analysis - Results

- Comparison of models applied to clusters and entire dataset
- Metric: *Log-Likelihood*
  - Comparative accuracy measure - only useful to compare techniques
  - Measures how well generated model fits actual data
  - Want to maximize
- We get this by comparing the fitted survival curves to the actual distribution of incident arrival rates
## Cluster Survival Analysis - Results

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Log-Likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-89,780.2</td>
</tr>
<tr>
<td>2</td>
<td>-52,030.0</td>
</tr>
<tr>
<td>3</td>
<td>-5,041.5</td>
</tr>
<tr>
<td>4</td>
<td>-15,694.4</td>
</tr>
<tr>
<td>5</td>
<td>-11,912.6</td>
</tr>
<tr>
<td>6</td>
<td>-22,706.9</td>
</tr>
<tr>
<td>7</td>
<td>-1,788.1</td>
</tr>
<tr>
<td>8</td>
<td>-719.7</td>
</tr>
<tr>
<td>9</td>
<td>-957.5</td>
</tr>
<tr>
<td>10</td>
<td>-763.5</td>
</tr>
<tr>
<td>11</td>
<td>-164.4</td>
</tr>
<tr>
<td>12</td>
<td>-316.9</td>
</tr>
<tr>
<td>13</td>
<td>-7,434.8</td>
</tr>
</tbody>
</table>
Survival Analysis

Cluster Survival Analysis - Results

Overall Log-Like Results

- Entire Dataset:
  - Survival Model: -180,243
  - Neg. Binomial: -178,488

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Log-Likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-89,780.2</td>
</tr>
<tr>
<td>2</td>
<td>-52,030.0</td>
</tr>
<tr>
<td>3</td>
<td>-5,041.5</td>
</tr>
<tr>
<td>4</td>
<td>-15,694.4</td>
</tr>
<tr>
<td>5</td>
<td>-11,912.6</td>
</tr>
<tr>
<td>6</td>
<td>-22,706.9</td>
</tr>
<tr>
<td>7</td>
<td>-1,788.1</td>
</tr>
<tr>
<td>8</td>
<td>-719.7</td>
</tr>
<tr>
<td>9</td>
<td>-957.5</td>
</tr>
<tr>
<td>10</td>
<td>-763.5</td>
</tr>
<tr>
<td>11</td>
<td>-164.4</td>
</tr>
<tr>
<td>12</td>
<td>-316.9</td>
</tr>
<tr>
<td>13</td>
<td>-7,434.8</td>
</tr>
</tbody>
</table>
Survival Analysis

Cluster Survival Analysis - Results

Overall Log-Like Results

- Entire Dataset: $-180,243$
  - Survival Model: $-180,243$
  - Neg. Binomial: $-178,488$
- Average of Cluster Survival Models: $-16,100$

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Log-Likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-89,780.2</td>
</tr>
<tr>
<td>2</td>
<td>-52,030.0</td>
</tr>
<tr>
<td>3</td>
<td>-5,041.5</td>
</tr>
<tr>
<td>4</td>
<td>-15,694.4</td>
</tr>
<tr>
<td>5</td>
<td>-11,912.6</td>
</tr>
<tr>
<td>6</td>
<td>-22,706.9</td>
</tr>
<tr>
<td>7</td>
<td>-1,788.1</td>
</tr>
<tr>
<td>8</td>
<td>-719.7</td>
</tr>
<tr>
<td>9</td>
<td>-957.5</td>
</tr>
<tr>
<td>10</td>
<td>-763.5</td>
</tr>
<tr>
<td>11</td>
<td>-164.4</td>
</tr>
<tr>
<td>12</td>
<td>-316.9</td>
</tr>
<tr>
<td>13</td>
<td>-7,434.8</td>
</tr>
</tbody>
</table>
Survival Analysis

Cluster Survival Analysis - Results

Overall Log-Like Results

• Entire Dataset:
  – Survival Model: -180,243
  – Neg. Binomial: -178,488
• Average of Cluster Survival Models: -16,100

Cluster Prediction is Order of Magnitude more Accurate

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Log-Likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-89,780.2</td>
</tr>
<tr>
<td>2</td>
<td>-52,030.0</td>
</tr>
<tr>
<td>3</td>
<td>-5,041.5</td>
</tr>
<tr>
<td>4</td>
<td>-15,694.4</td>
</tr>
<tr>
<td>5</td>
<td>-11,912.6</td>
</tr>
<tr>
<td>6</td>
<td>-22,706.9</td>
</tr>
<tr>
<td>7</td>
<td>-1,788.1</td>
</tr>
<tr>
<td>8</td>
<td>-719.7</td>
</tr>
<tr>
<td>9</td>
<td>-957.5</td>
</tr>
<tr>
<td>10</td>
<td>-763.5</td>
</tr>
<tr>
<td>11</td>
<td>-164.4</td>
</tr>
<tr>
<td>12</td>
<td>-316.9</td>
</tr>
<tr>
<td>13</td>
<td>-7,434.8</td>
</tr>
</tbody>
</table>
Mapping to Hex Locations

• We now have incident prediction models *for each cluster*
Mapping to Hex Locations

- We now have incident prediction models for each cluster
- Clusters have no relationship with spacial regions
Mapping to Hex Locations

- We now have incident prediction models for each cluster
- Clusters have no relationship with spatial regions
- Need to map these cluster likelihoods to the hex cell regions
Mapping to Hex Locations

- Use distribution of clusters in each cell
- Example with 2 clusters:
  - Incidents in left cell more likely when cluster 1 more likely
Bayesian Network

- Conditional probability Model
- Trained from historical data
Bayesian Network
Bayesian Network

Input Parameters

- Time Interval
- Day
- Weather
- Month
- Cluster
- Hex Location
Bayesian Network

Input Parameters

Hex Region Probability Output

- Time Interval
- Day
- Weather
- Month

Cluster

Hex Location
Online Procedure

• **Cluster Probability**
  – Find conditional probability of each cluster *given current conditions*
  – From clusters and Bayesian Network
Online Procedure

- **Cluster Incident Likelihood**
  - Probability of incident occurring for each cluster in given time frame
    - Ignore the cluster if it is below a threshold
  - From Survival Models
Online Procedure

- **Spatial Correlation**
  - Likelihood that incident occurs in a hex cell given the cluster likelihoods using learned probabilities
  - From Bayesian Network
Online Procedure

Cluster Identification → Cluster Incident Likelihood → Spatial Correlation

Real-Time Hex Cell Incident Likelihoods
Validation

• Compared Toolchain’s predicted hex cell incident density to validation set
  – 10 months of Nashville incident data from Feb. 6 to Dec. 23 2016
Validation

- Compared Toolchain’s predicted hex cell incident density to validation set
  - 10 months of Nashville incident data from Feb. 6 to Dec. 23 2016

![Probability Difference vs Hex Cell](image)

Normalized Root Mean Squared Error:
1.656425
Validation

- Compared Toolchain’s predicted hex cell incident density to validation set
  - 10 months of Nashville incident data from Feb. 6 to Dec. 23 2016
- Results
  - Majority of predictions within 2% correct
  - Nearly all within 10%

![Normalized Root Mean Squared Error: 1.656425](image)
Validation

• Compared Toolchain’s predicted hex cell incident density to validation set
  – 10 months of Nashville incident data from Feb. 6 to Dec. 23 2016

• Results
  – Majority of predictions within 2% correct
  – Nearly all within 10%

<table>
<thead>
<tr>
<th>Method</th>
<th>Log-Likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average of Cluster Survival Models</td>
<td>-16,100.8</td>
</tr>
<tr>
<td>Entire Dataset - Survival Models</td>
<td>-180,243.8</td>
</tr>
<tr>
<td>Entire Dataset - Neg. Binomial</td>
<td>-178,488.9</td>
</tr>
</tbody>
</table>
Example

- Parameters:
  - Clear weather
  - Thursday
  - Start time 15:00
  - Analysis time 2 hrs
Example

- Parameters:
  - Clear weather
  - Thursday
  - Start time 15:00
  - Analysis time 2 hrs

### Survival Models

<table>
<thead>
<tr>
<th>Rank</th>
<th>Cluster</th>
<th>Incident Likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.6554</td>
</tr>
<tr>
<td>2</td>
<td>13</td>
<td>0.6294</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>0.49461</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>0.4448</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>0.2114</td>
</tr>
</tbody>
</table>

### Hex Probabilities

<table>
<thead>
<tr>
<th>Rank</th>
<th>Hex Cell</th>
<th>Hex Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3523</td>
<td>0.05884</td>
</tr>
<tr>
<td>2</td>
<td>4140</td>
<td>0.05884</td>
</tr>
<tr>
<td>3</td>
<td>5491</td>
<td>0.04682</td>
</tr>
<tr>
<td>4</td>
<td>4703</td>
<td>0.04682</td>
</tr>
<tr>
<td>5</td>
<td>4699</td>
<td>0.04682</td>
</tr>
</tbody>
</table>
Example 2

- **Parameters:**
  - Rainy weather
  - Otherwise the same

- **Results:**
  - Analysis time is the same, so the survival models give the same probabilities
  - Rain changes the likely accident locations, and increases their probabilities

### Survival Models

<table>
<thead>
<tr>
<th>Rank</th>
<th>Cluster</th>
<th>Incident Likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.6554</td>
</tr>
<tr>
<td>2</td>
<td>13</td>
<td>0.6294</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>0.49461</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>0.4448</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>0.2114</td>
</tr>
</tbody>
</table>

### Hex Probabilities

<table>
<thead>
<tr>
<th>Rank</th>
<th>Hex Cell</th>
<th>Hex Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3513</td>
<td>0.13108</td>
</tr>
<tr>
<td>2</td>
<td>4332</td>
<td>0.13108</td>
</tr>
<tr>
<td>3</td>
<td>5290</td>
<td>0.13108</td>
</tr>
<tr>
<td>4</td>
<td>4803</td>
<td>0.13108</td>
</tr>
<tr>
<td>5</td>
<td>3587</td>
<td>0.13108</td>
</tr>
</tbody>
</table>
Example 3

- **Parameters:**
  - Snow
  - January
  - Thursday
  - Start at 17:00
  - 6 hour analysis time

- **Results**
  - High analysis time gives very high accident probabilities
  - Some cells are much more likely to have an incident in snowy conditions

### Survival Models

<table>
<thead>
<tr>
<th>Rank</th>
<th>Cluster</th>
<th>Incident Likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.9591</td>
</tr>
<tr>
<td>2</td>
<td>13</td>
<td>0.9491</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>0.8710</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>0.8288</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>0.5100</td>
</tr>
</tbody>
</table>

### Hex Probabilities

<table>
<thead>
<tr>
<th>Rank</th>
<th>Hex Cell</th>
<th>Hex Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4334</td>
<td>0.201918</td>
</tr>
<tr>
<td>2</td>
<td>3862</td>
<td>0.174190</td>
</tr>
<tr>
<td>3</td>
<td>4350</td>
<td>0.036615</td>
</tr>
<tr>
<td>4</td>
<td>3943</td>
<td>0.036615</td>
</tr>
<tr>
<td>5</td>
<td>3792</td>
<td>0.036615</td>
</tr>
</tbody>
</table>
1. **Prediction**
   Understand future resource demand

2. **Dispatch**
   Design an optimal dispatch algorithm based on said future demand
Current Work

2

Dispatch
Design an optimal dispatch algorithm based on said future demand
Dispatch - Problem

- **Goal:** choose which resource(s) to send to an incident as it is reported
- **Current method:**
  - minimize euclidean distance
- **Proposed:**
  - Create and solve optimization problem that includes predicted incident likelihoods
Toy Problem

- Color Gradient - probability
  - red: incident more likely
Toy Problem

- Color Gradient - probability
  - red: incident more likely
- Which ambulance should respond?
Toy Problem

- Color Gradient - probability
  - red: incident more likely
- Which ambulance should respond?
  - lower right
Toy Problem

- Color Gradient - probability
  - red: incident more likely
- Which ambulance should respond?
  - lower right
- Now less clear
Stationing

- Stationing - optimize where vehicles should be located
Stationing

- Stationing - optimize where vehicles should be located
Dispatch Methods

- Find Incident Likelihoods
- Estimate how dispatching decision effects response time for predicted incidents
- Apply Receding Horizon Control
  - Simulate future incidents and response times
Stationing Methods

• Station periodically
• Similar to dispatch, but must simulate effect of all vehicles moving
• Large search space
  – use heuristic method
    • genetic optimization
    • particle swarm
• S.-P. Miaou and H. Lum, “Modeling vehicle accidents and highway geometric design relationships,” Accident Analysis & Prevention, vol. 25, no. 6, pp. 689–709, 1993