Conducting spatial predictions of driver behavior using multi-parametric data

PIONEERing
Solutions for the Smart City Challenge
Online Student Workshop
15 April 2021

Apostolos Ziakopoulos
George Yannis

Department of Transportation Planning and Engineering,
National Technical University of Athens

www.nrso.ntua.gr/apziak
apziak@central.ntua.gr
Research scope and questions

Spatial analysis of harsh event frequencies (harsh brakings/accelerations) in road segments

1. How can high-resolution naturalistic driving smartphone data and road segment geometric and road network characteristic data be combined (map-matched) and examined in road safety investigations?

2. How can harsh event frequencies be analyzed spatially in urban networks?

3. Which road geometry and road network characteristics affect harsh event frequencies in urban road network environments?

4. How transferable are the previous results in a different study area? Can reliable predictions be conducted?
Merits of harsh event examination

Harsh events: **harsh brakings** and **harsh accelerations** recorded by smartphone sensors for telematics-based vehicle insurance primes

- Parameters measuring **road safety levels** (correlations with spatial and temporal headways)
- Inherently linked with **driver risk** (Tselentis et al, 2017)
- **Different phenomena**, correlations with different variables (Ziakopoulos et al, 2020)

Considerable **comparative advantages** for their investigation:

1. Applications in driver **evaluation** and **classification** (Bonsall et al., 2005; Gündüz et al., 2018).
2. **Proactive** road safety indicators anticipating safety-critical events (Zohar et al., 2014; Jansen & Wesseling, 2018); evaluations **before** crashes occur
3. Non-aggressive driving reduces **emissions** by up to 40% (Alessandrini, 2012)
4. Investigated by the **insurance** industry (Paefgen et al., 2012; 2014)
5. Apparent **research gaps** in the investigation of harsh event frequencies
Data collection (1/2): Digital map road geometry data

Data of road segment geometry and road network characteristics on a microscopic level from digital maps

**OpenStreetMap**: Open source digital map platform
Hierarchical elements:
1. **Nodes**
2. **Ways** from node groups
3. **Relations** from node and way groups

Obtaining a wealth of data in WGS84 through API queries (Overpass Turbo API through Overpass Query Language)

**NASA SRTM** topography
Altitude data provided by NASA:
- **Freely** available
- **Altitude** resolution per 10 cm – compared with OSM altitudes for verification – some accuracy issues
- Majority of populated areas available

---

Apostolos Ziakopoulos | Conducting spatial predictions of driver behavior using multi-parametric data
Data collection (2/2): Naturalistic driving data from smartphones

Naturalistic driving data from real-world conditions obtained from smartphones (per trip-second), primarily recorded for telematics-based vehicle insurance primes

Utilization of the application/platform of OSeven Telematics

- APIs utilization for data reading from smartphone sensors
- Exploited sensors: GPS, accelerometer, gyroscope, device orientation
- Transmission from smartphone to central storage database
- Data cleaning and processing via a series of Machine Learning (ML) algorithms
- Several data are provided, indicatively: trip position, speed, acceleration, harsh brakings/accelerations, event intensity, speeding, mobile phone use
- Total anonymity during all data handling phases (GDPR)

High resolution big data from driver trips including behaviour indicators
Data processing: Geometric characteristics (1/2)

Calculation of geometric characteristics based on OSM node coordinates

**Roadway segment length**
- Calculation based on modern geoids/ellipsoid models through available libraries
- Sum of elementary lengths (2 nodes each)

**Determination of road segment centroids**

**Gradient**
- Sum of elementary gradients (2 nodes each)
- Road segment average, weighted by elementary lengths

**Curvature**
- Menger’s formula per elementary triangle (3 nodes each)
- Road segment average, weighted by elementary lengths
Data processing: Geometric characteristics (2/2)

Neighborhood complexity calculation

- Measurement of density and complexity of immediate road segment environment: (i) in reality (ii) on the digital maps
- Logarithm of nodes within a window of 470m * 470m from each road segment centroid

Obtaining of additional road segment characteristics from OSM:

1. Presence of pedestrian crossing
2. Presence of traffic lights
3. Lane number
4. Road type (exclusion of walkways/footpaths/surfaces without vehicles)
5. Direction number (one-way or two-way)

Calculation with original purpose-made algorithms and sub-routines created in R-studio, iteratively for each road segment
Data processing: Map-matching (1/2)

Map-matching: Plotting of naturalistic driving data on maps after determination of the corresponding segment

Matching of GPS trace to each road segment per second

Identification of:
1. Nearest node (point-to-point distance)
2. Minimum distance way – MDW (point-to-polyline distance)
   - Moving polygon serving to reduce candidate ways
   - Time-consuming and computationally demanding process
   - Corrections are essential in dense road segments with parallel axes through a specialized vote-count algorithm

Recording and assignment per road segment:
1. Pass count
2. Harsh brakings/accelerations
3. Speeding seconds
4. Mobile use seconds
Data processing: Map-matching (2/2)

Map-Matching subroutine start

Set number of total trip-secs in the dataset = n

Set i = 0

False

i < n

i = i + 1

True

Build a small moving polygon around position lat/lon.

Determine nodes in the moving polygon

Find the node with the minimum std. distance

Find all ways related to that node

Find way with minimum point-to-polyline distance

Pass_countSecs = Pass_countSecs + 1

False

Harsh acceleration on trip-second t?

True

Harsh_accAccel = Harsh_accAccel + 1

False

Harsh braking on trip-second t?

True

Harsh_brbrkAccel = Harsh_brbrkAccel + 1

False

Speeding on trip-second t?

True

Speeding_speedSecs = Speeding_speedSecs + 1

False

Mobile use on trip-second t?

True

Mobile_use_mobileSecs = Mobile_use_mobileSecs + 1

Map-Matching subroutine end

First trip-second on this MDW:
Number of total passes: +1
Number of total seconds driven: +1
If ha/hb: total ha/hb number: +1
If speeding: total speeding speedSecs: +1
If using mobile: total mobile mobileSecs: +1

Second trip-second on this MDW:
Number of total passes: +0
Number of total seconds driven: -1
If ha/hb: total ha/hb number: +1
If speeding: total speeding speedSecs: +1
If using mobile: total mobile mobileSecs: -1

Apostolos Ziakopoulos | Conducting spatial predictions of driver behavior using multi-parametric data
869 road segments (removal of 14 footways) with 4293 nodes

- 49 road segments with traffic lights
- 80 road segments with pedestrian crossings

**Naturalistic driving** data:
- Trips between 01-10-2019 & 29-11-2019 – **2 months**
- **A total of** 3294 trips from 230 drivers
- 1,000,273 **driving seconds**: average trip duration 304 s
- 1348 harsh **brakings**
- 921 harsh **accelerations**

90% of road segments feature **at least 1 trip**

Variable distributions
- **Positive** skewness (larger right tails)
- **High** kurtosis (non-normal distributions)
Sample description (2/2) – Omonoia urban road network

1237 road segments (removal of 78 footways) with 6115 nodes
- 319 road segments with traffic lights
- 317 road segments with pedestrian crossings

Naturalistic driving data:
- Trips between 01-10-2019 & 29-11-2019 – 2 months
- A total of 2615 trips from 257 drivers
- 964,693 driving seconds: average trip duration 369 s
- 1036 harsh brakings
- 938 harsh accelerations

86% of road segments feature at least 1 trip

Variable distributions
- **Positive** skewness (larger right tails)
- **High** kurtosis (non-normal distributions)
**Arsenal of spatial statistical & ML models**

Integration of spatial heterogeneity

Event frequencies: Log-normal Poisson framework

1. Geographically Weighted Poisson Regression (GWPR)
   - Frequentist functional models: local micro-regressions are conducted, b coefficients can vary locally

2. Conditional Autoregressive Prior Regression (CAR)
   - Bayesian functional models: Bayesian regressions are conducted with spatially structured and unstructured terms, b coefficient distributions are obtained

3. Extreme Gradient Boosting (XGBoost) – AI methods
   - Machine learning: Multiple additive regression trees (ensemble), obtained information regarding variable contribution (gain)
   - Random Cross-Validation – RCV
   - Spatial Cross-Validation – SPCV

Source: Lovelace et al. (2019)
Harsh braking spatial analyses in urban road networks

<table>
<thead>
<tr>
<th>Independent variables</th>
<th>GWPR Coefficients</th>
<th>CAR Mean posterior values</th>
<th>RCV XGBoost Gain values</th>
<th>SPCV XGBoost Gain values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.4636</td>
<td>-1.4134</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Gradient</td>
<td>-2.4864</td>
<td>-9.7538</td>
<td>0.0806</td>
<td>0.0860</td>
</tr>
<tr>
<td>Curvature</td>
<td>—</td>
<td>—</td>
<td>0.0444</td>
<td>0.0626</td>
</tr>
<tr>
<td>Neighborhood complexity</td>
<td>-0.2919</td>
<td>-0.1787</td>
<td>0.0344</td>
<td>0.0684</td>
</tr>
<tr>
<td>Segment length</td>
<td>0.0039</td>
<td>0.0075</td>
<td>0.1436</td>
<td>0.1400</td>
</tr>
<tr>
<td>Pass count</td>
<td>0.0040</td>
<td>0.0086</td>
<td>0.6788</td>
<td>0.6271</td>
</tr>
<tr>
<td>Traffic lights: Yes [Ref.: Traffic lights: No]</td>
<td>0.2563</td>
<td>-0.0902</td>
<td>0.0037</td>
<td>0.0010</td>
</tr>
<tr>
<td>Pedestrian crossing: Yes [Ref.: Pedestrian crossing: No]</td>
<td>-0.1463</td>
<td>0.3820</td>
<td>0.0024</td>
<td>0.0024</td>
</tr>
<tr>
<td>Lanes: 2 [Ref.: Lanes: 1]</td>
<td>-0.2435</td>
<td>-0.1713</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lanes: 3 [Ref.: Lanes: 1]</td>
<td>0.3669</td>
<td>-0.5719</td>
<td>0.0072</td>
<td>0.0048</td>
</tr>
<tr>
<td>Lanes: 4 [Ref.: Lanes: 1]</td>
<td>0.3578</td>
<td>1.9169</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Road type: secondary [Ref.: Road type: primary]</td>
<td>1.0520</td>
<td>-0.1094</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Road type: tertiary [Ref.: Road type: primary]</td>
<td>-0.0070</td>
<td>-1.6389</td>
<td>0.0049</td>
<td>0.0078</td>
</tr>
<tr>
<td>Road type: residential [Ref.: Road type: primary]</td>
<td>-1.0084</td>
<td>-2.5578</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sigma-phi² [Spatially structured effects]</td>
<td>N/A</td>
<td>700.3172</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Sigma-theta² [Spatially unstructured effects]</td>
<td>N/A</td>
<td>2.3455</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

Performance metrics

<table>
<thead>
<tr>
<th>Metric</th>
<th>GWPR</th>
<th>CAR</th>
<th>RCV XGBoost</th>
<th>SPCV XGBoost</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSE</td>
<td>3.2954</td>
<td>1.2830</td>
<td>1.4215</td>
<td>1.8293</td>
</tr>
<tr>
<td>MAE</td>
<td>1.3048</td>
<td>0.4115</td>
<td>0.4971</td>
<td>0.4994</td>
</tr>
<tr>
<td>RMSLE</td>
<td>0.5569</td>
<td>0.1727</td>
<td>0.3140</td>
<td>0.2390</td>
</tr>
<tr>
<td>CA</td>
<td>80.90%</td>
<td>96.32%</td>
<td>90.56%</td>
<td>91.71%</td>
</tr>
</tbody>
</table>

Positive correlation:
- Segment length
- Pass count

Negative correlation:
- Gradient
- Neighborhood complexity
- Road type [Residential]

Marginally positive correlation:
- Road type [Secondary]
- Traffic lights
- Pedestrian crossing

Marginally negative correlation:
- Road type [Tertiary]
Harsh braking prediction & transferability

Predictions on Omonoia test area

1. Geographically Weighted Poisson Regression (GWPR)
   - Local b-coefficient fluctuations are not transferable
   - Predictions using global Poisson regression

2. Bayesian Conditional Autoregressive Prior Regression (CAR)
   - Spatially structured and unstructured effects are not transferable
   - Predictions using new Bayesian Poisson regression

3. Extreme Gradient Boosting (XGBoost)
   - Seamless transferability of machine learning ensemble trees/rules using both RCV and SPCV

SPCV XGBoost has the best individual performance from all implemented methods

<table>
<thead>
<tr>
<th>Performance metrics</th>
<th>GWPR global Poisson</th>
<th>Bayesian Poisson (XGBoost)</th>
<th>RCV XGBoost</th>
<th>SPCV XGBoost</th>
<th>Combined Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSE</td>
<td>1.9792</td>
<td>1.9804</td>
<td>1.9834</td>
<td>1.8418</td>
<td>1.6114</td>
</tr>
<tr>
<td>MAE</td>
<td>1.0265</td>
<td>1.0290</td>
<td>0.8415</td>
<td>0.7542</td>
<td>0.6645</td>
</tr>
<tr>
<td>RMSLE</td>
<td>0.5508</td>
<td>0.5520</td>
<td>0.5484</td>
<td>0.5189</td>
<td>0.4514</td>
</tr>
<tr>
<td>CA</td>
<td>82.64%</td>
<td>82.74%</td>
<td>83.40%</td>
<td>85.27%</td>
<td>87.55%</td>
</tr>
</tbody>
</table>
Using **combined average**, spatial models **mitigate** their weaknesses and lead to a **balanced** predictive outcome for harsh brakings.
Case study findings

1. It is **possible to combine** high resolution **multi-parametric** naturalistic driving and geometric data that can be exploited to conduct meaningful spatial analyses on a road segment basis.

2. The implementation of both **functional** spatial methods (GWPR, CAR) and **innovative ML** methods (RCV & SPCV, XGBoost) is feasible for spatial analyses of harsh braking frequencies on a road segment basis.

3. **Precise predictions (87.6% accuracy)** of harsh braking frequencies can be **successfully conducted**. Several correlations were obtained.

4. Using **combined average**, spatial models **mitigate** their weaknesses and lead to a **balanced** predictive outcome for harsh brakings.

5. The analyses were mirrored for harsh accelerations (89% acc). A **more complete** image of harsh event hotspots is obtained.
Future tasks – extension to industrial practices

1. **Correlation with crash data**
   Conducting spatial analyses including crash data per road segment – examination of possible hotspot overlap

2. **Investigation of further aspects**
   Temporal dimension, additional spatial/ML models, additional road environments, driver aggressiveness categories

3. **Creation of a seamless and constantly updating system**
   From smartphone data collection to heatmap rendition on a recurring basis using integrated AI algorithms

4. **Expanding benefits for road users and authorities**
   Road safety hotspot identification before crashes occur – Added information for pedestrians, professional drivers, mobility-impaired individuals

5. **Additional maps can be created for any indicator**
   E.g.: speeding, mobile phone use, emissions etc.
Conducting spatial predictions of driver behavior using multi-parametric data

PIONEERing Solutions for the Smart City Challenge
Online Student Workshop
15 April 2021

Apostolos Ziakopoulos
George Yannis

Department of Transportation Planning and Engineering, National Technical University of Athens

www.nrso.ntua.gr/apziak
apziak@central.ntua.gr