Big Data Validation, Augmentation with k-Anonymity & Differential Privacy for Multi-year Driving Study

B Wallace¹, R Goubran¹,², F Knoefel¹,², S Marshall³, M Porter¹, M Harlow¹,⁴, A Puli¹

¹ Carleton University, ² Bruyere Research Institute, ³ Ottawa Hospital Research Institute, ⁴ University of Manitoba

Objective
- Anonymity of participants must be maintained
- Dataset shared with differing trust models
- Candrive: multi-year study, 1000 older drivers
- GPS and OBDII sensor deployed in participant vehicles
- Automation required for massive (1TB) dataset

Method

Data Set processing requirements
- Data Validation
  - GPS reception failures
  - Incomplete samples
- Data Augmentation
  - GIS map data (posted limits, road hazards)
  - Solar cycle (day/night) data
  - Weather information
- Data Anonymization challenge
  - Participants identified by number only but:
    - Location – Provides details on where a user lives/works/visits → their identity
    - k- anonymity – combinations of data elements can’t be combined to re-identify participant
    - Differential Privacy – Provide differing data access based on trust models

Data set statistics for participants

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metropolitan Area Size</td>
<td>~100km East - West</td>
</tr>
<tr>
<td></td>
<td>~50km North - South</td>
</tr>
<tr>
<td>Variation in Sunrise/Sunset time to East/West limits</td>
<td>+/- 4 mins</td>
</tr>
<tr>
<td>Variation in Solar Day time to North/South limits</td>
<td>+/- 1 min</td>
</tr>
</tbody>
</table>

Example sensor data captured at 1Hz

<table>
<thead>
<tr>
<th>Sensor</th>
<th>Measured Value</th>
<th>Units / format</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPS</td>
<td>Date &amp; Time</td>
<td>Date string</td>
</tr>
<tr>
<td>Longitude/Latitude</td>
<td>Degrees</td>
<td>Degrees</td>
</tr>
<tr>
<td>Velocity</td>
<td>km/hr</td>
<td></td>
</tr>
<tr>
<td>GIS</td>
<td>Posted Limit</td>
<td>km/hr</td>
</tr>
<tr>
<td>Alerts</td>
<td>Text String</td>
<td></td>
</tr>
<tr>
<td>RFID</td>
<td>RFID tag number</td>
<td>Serial number</td>
</tr>
<tr>
<td>OBDII</td>
<td>Velocity</td>
<td>km/hr</td>
</tr>
<tr>
<td>Throttle Position</td>
<td>Percentage</td>
<td></td>
</tr>
</tbody>
</table>

Automated data processing algorithm

Results and Conclusions

Example trip trace showing posted speed limit and location of road hazards

Legend:
- blue: 40km/hr
- green: 60km/hr
- red: 80 km/hr
- black: 100km/hr
- yellow: road hazard

Anonymity model
- 3 levels of trust
- K-anonymity
- Differential Privacy
- Ensures anonymity of participant
- Guidelines for publication

Acknowledgements
Natural Sciences and Engineering Research Council (NSERC) and industrial and government partners, through the Healthcare Support program through Information Technology Enhancements (hSITE) Strategic Research Network. Candrive prospective study (www.candrive.ca) funded by the Canadian Institutes of Health Research (CIHR) Innovation Fund of the Alternative Funding Plan for the Academic Health Sciences Centres of Ontario.