Steering Clear of Data Potholes

V-SPOC Data Quality Improvement

Elizabeth Schneider
Wisconsin Department of Transportation
Xiao Qin
University of Wisconsin-Milwaukee

ITE Traffic Engineering Workshop
April 9, 2019
4,000+ Detectors
Devices

- Message Signs
- Cameras
- Ramp Meters
- Traffic Detectors
Volume SPeed OCcupancy

Query & retrieve freeway traffic detector data from the WisDOT TMC. Contains archive of detector data from 1996. Updated every 24-hours.

- Planning models
- Ramp metering
- Bottleneck analysis
- Capacity analysis
- Work zone analysis
- Incident management
Events and Closures
Data Quality

Hardware Failed

Closed Road
User Survey

Which data quality tools have you used in V-SPOC?

- ATMS Detector data availability report: 16%
- Quality assurance report: 26%
- V-SPOC Data availability calendar: 58%
“A way to see what detector info is available (and accurate) at any given time would be beneficial so that you don't spend the time pulling volumes only to see there is no data/inaccurate data.”

“Favorite detectors - if I can add favorite detectors into its own option, it will be easier for me to download a lot of data Detectors.“

“There is no way we know from just looking at the detector list/controller list, whether the detectors shown there were still available or not.”

“We've had a few concerns with the results of the speed data, especially at low-volume locations.“
DATA
POTHOLE PATROL
Methodologies

Archived Data Management System (ADMS)
- California PeMS (Performance Measurement System),
- CATT (Center for Advanced Transportation Technology), Maryland,
- Central Florida Data Warehouse,
- Kentucky ADMS,
- Virginia ADMS,
- Phoenix RADS (Regional Archived Data Server),
- PORTAL (Portland Oregon Regional Transportation Archive Listing),
- WisTransportal V-SPOC.
- FHWA Mobility Monitoring Program

Research papers on detector data quality in recent 10 years

Recommendations
- Current application
- Preferred traffic variables
- Data specifications (e.g., time frame and resolution)
- Data quality checking methods
- Extensive tests
- Sensitivity analysis
- Case studies
- A combination of rule-based and data-driven approaches
What did we learn from literature review?

• Typical data QA/QC procedures include:
  
  a) **univariate & multivariate range**: validate a single traffic variable or a combination of traffic variables against a predetermined threshold.
  
  b) **temporal consistency**: monitor changes within a time period or compare with historical trends.
  
  c) **spatial consistency**: compare with the traffic patterns from nearby detectors.

• Data validity tests in practice are closely alike, and most are less complex and probably, less effective.

• Potential to leverage new research to improve detector data quality.
An online survey was designed and sent to 118 V-SPOC users, and a 47% (54) response rate was received.

- New validity tests for volume and speed and those with a five-minute (5-min) resolution are preferred.
- Users generally prefer multivariate range checks and temporal consistency checks.
- Spatial consistency checks are therefore not recommended.
Recommended new quality checks

- Abnormally high VSO spikes: 11
- High speed during congestion: 9
- High free-flow volume: 9
- Change from daily pattern: 7
- Lane 2 lane differences: 5
- High free-flow occupancy: 4
- Others: 2
Candidate Data Validity Tests

- **Basic**
  - missing data
  - univariate range
  - zero consistency

- **Advanced**
  - Alternative multivariate range
    - free-flow volume
    - speed in congestion
    - average effective vehicle length
  - Temporal consistency
    - Repeating zero counts
    - Non-zero occupancy struck
    - change in volume
    - change in speed
    - daily traffic pattern
Phase I:
- Station 1 on I-94 EB (55mph),
- Station 2 on the Beltline Highway (55mph) and
- Station 3 on US 51 NB (25mph), a two-lane two-way undivided highway

Phase II
- Milwaukee I-94, 152 detectors
- Madison Beltline highway, 87 detectors
- Sept-Dec, 2017 and 2018
Traffic Flow Analysis

High Free Flow Volume:
if $VOL > 1,200$ vphpl & $OCC < 5\%$, then invalid

Infeasible Speed in Congestion:
95 percent confidence region:
$$\left[ \frac{798}{OCC} - 10, \frac{1658}{OCC} - 16 \right]$$
Average Effective Vehicle Length

\[ AEVL \left( \frac{ft}{veh} \right) = 52.8 \times \frac{SPD \times \%OCC}{VOL} \]

proposed valid range is [9 ft., 60 ft.]
Abrupt Change Detection

\[ \Delta V_t = V_t - \frac{V_{t-1} + V_{t+1}}{2} \]

\[ \Delta S_t = S_t - \frac{S_{t-1} + S_{t+1}}{2} \]
## Temporal Pattern Analysis

<table>
<thead>
<tr>
<th>Repeating Zero Volume</th>
<th>Non-zero Occupancy Stuck</th>
<th>Anomalous Daily Traffic Pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\sum_{i}^{4} i = -4, i \neq 0 \mid (\text{VOL}(t) = \text{VOL}(t + i)) &gt; J)</td>
<td>(\sum_{i}^{6} i = -6, i \neq 0 \mid (\text{OCC}(t) = \text{OCC}(t + i)) &gt; 3)</td>
<td>This test checks whether the traffic pattern of volume in one weekday/weekend is similar to same day of week in the same month.</td>
</tr>
</tbody>
</table>

During 8 time intervals (or 40 minutes), if more than J zeros are observed.
- 6am-10pm
- 10pm-6am

If a non-zero occupancy repeats more than 3 times (or 15 minutes) during a one-hour period.
# Summary

<table>
<thead>
<tr>
<th>Validity Test</th>
<th>Description</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Basic Check</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Missing data check</td>
<td>If any traffic variable has a missing value, then invalid</td>
<td>Valid ranges: [0, 3100 vphpl] (VOL), [0, 100 MPH] (SPD), [0%, 100%] (OCC)</td>
</tr>
<tr>
<td>Univariate range checks</td>
<td>If value of any traffic variable is out of the feasible range, then invalid</td>
<td>If VOL, SPD, and OCC are not all zeros when any of them is zero, then invalid</td>
</tr>
<tr>
<td>Zero consistency checks</td>
<td>If not all three traffic variables are zeros when there is no passing vehicle, then invalid</td>
<td></td>
</tr>
<tr>
<td><strong>Multivariate Range Check</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High free flow volume</td>
<td>If the volume is too high in the free flow traffic, then invalid</td>
<td>If VOL &gt; 1,200 vphpl and OCC &lt; 5%, then invalid</td>
</tr>
<tr>
<td>Infeasible speed in congestion</td>
<td>If the speed is out of the feasible range in congestion, then invalid</td>
<td>Valid range: 798/OCC-10 &lt; SPD &lt; 1658/OCC-16</td>
</tr>
<tr>
<td>Infeasible Average Effective Vehicle length (AEVL)</td>
<td>If AEVL is out of the feasible range, then invalid</td>
<td>Valid AEVL range: [9 ft, 60 ft]</td>
</tr>
<tr>
<td><strong>Temporal Consistency Check</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Repeating zero volume</td>
<td>If there are too many repeated zero volumes, then invalid</td>
<td>If $\text{VOL}(t) = 0 &amp; \sum_{i=-4}^{4} I(\text{VOL}(t) = \text{VOL}(t + i)) &gt; J$ in 6 am-10 pm/11 pm-5 am if the average historical 5-min volume is $\mu$, then invalid.</td>
</tr>
<tr>
<td>Non-zero occupancy stuck</td>
<td>If there are too many repeated non-zero occupancies, then invalid</td>
<td>If $1% &lt; \text{OCC}(t) &lt; 100% &amp; \sum_{i=-6}^{6} I(\text{OCC}(t) = \text{OCC}(t + i)) &gt; 3$, then invalid.</td>
</tr>
<tr>
<td>Abrupt change in volume</td>
<td>If the change in volume exceeds the max feasible value, then invalid</td>
<td>If $</td>
</tr>
<tr>
<td>Abrupt change in speed</td>
<td>If the change in speed exceeds the max feasible value, then invalid</td>
<td>If $</td>
</tr>
<tr>
<td>Anomalous daily traffic pattern</td>
<td>If the daily traffic profile is substantially deviated from that in similar days based on a similarity measure, then further investigation is needed</td>
<td>If the correlation coefficient is below 0.8, the traffic records of the corresponding day are flagged. (Correlation coefficient: $r_t = \frac{\sum_{t=1}^{T}(V_{t,i} - \bar{V})(V_t - \bar{V})}{\sqrt{\sum_{t=1}^{T}(V_{t,i} - \bar{V})^2 \sum_{t=1}^{T}(V_t - \bar{V})^2}}$)</td>
</tr>
</tbody>
</table>
Conclusions

• The recommended detector data validity tests are selected from a broad range of tests, surveyed by users, and demonstrated through case studies using VSPOC data.

• This study underscores the importance of keeping the basic and adding more advanced tests to detect less obvious yet important data issues.

• Methodologies for establishing rule-based or data-driven criteria for validity tests can be applied directly or adjusted using local data.

• Testing more sites may result in different threshold values for individual validity tests, so data from more sites can be pooled to form rules that are more universally applicable.
Thank you!

Contact Information:

Elizabeth Schneider, PE,
State Freeway Operations Engineer
BTO, Wisconsin Department of Transportation
Elizabeth1.Schneider@dot.wi.gov

Xiao (Shaw) Qin, Ph.D., PE
Associate Professor
Director, Safe and Smart (S²) Traffic Laboratory
University of Wisconsin-Milwaukee
qinx@uwm.edu