APPENDIX C

Data Quality Control Examples Data Quality Control Examples

This appendix provides more information about data quality control. This appendix describes some general quality control insights for traffic volume data but is primarily focused on probe speed datasets. An example of completeness and validity analysis on the FHWA NPMRDS is included as well as general guidance to practitioners regarding quality control at the very end of this appendix. Much of this appendix text was excerpted from a recent FHWA project documentation. (78)

Prior to data analysis, it is important that the analyst perform quality control of the datasets to ensure certain specifications are met. As described elsewhere, the quality control process typically includes one or more of the following actions:

1. Reviewing the traffic data format and basic internal consistency;
2. Comparing traffic data values to specified validation criteria;
3. Marking or flagging traffic data values that do not meet the validation criteria;
4. Reviewing marked or flagged traffic data values for final resolution; and
5. Imputed marked, flagged, or missing traffic data values with "best estimates" (while still retaining original data values and labeling imputed values as estimates).

The American Association of State Highway and Transportation Officials Guidelines for Traffic Data Programs(79) describes these quality control processes in more detail and the interested reader is referred there for further information. Of particular interest are the definitions for traffic data quality measures, including accuracy, completeness (also referred to as availability), validity, timeliness, coverage, and accessibility (also referred to as usability). More specifically, AASHTO spells out validation criteria for vehicle count, classification, and weight data from intelligent transportation systems (ITS) detector sources. Additional examples of quality control checks ("business rules") for ITS detector sources are described elsewhere.

In some cases, quality control by visual inspection is valuable. Visual inspection is helpful when it is not easy to automate the quality control with business rules. Sometimes the human eye is more adept at identifying reasonableness in data-time series. For example, graphing speed or volume plots by time for a variety of days in the month on the same graphic or looking at lane-by-lane speed and volume relationships on the same graph. Visual inspection of graphics like this allow the analyst to identify places where more "drill down" analyses may be warranted if something suspicious is found. More examples are documented elsewhere. (80)

Probe Speed Data Quality Control

Probe speed data are a very cost-effective source for systemwide data collection. With the increased and widespread use of probe speed data for bottleneck analysis (including truck bottleneck analyses), quality control of these data sources is of particular interest and is the focus of this section. The Federal

Highway Administration (FHWA) National Performance Management Research Data Set (NPMRDS) is used as an example in this section to illustrate quality control considerations for a probe speed dataset.

NPMRDS provides travel-time data in five-minute-time aggregations. Due to the recent release of NPMRDS, there has been limited investigation of the data source and State DOT and metropolitan planning organization (MPO) practitioners are asking how the dataset can be used for performance measurement and truck analysis on freeways and arterials, and about the general quality of the data for these uses.

As part of an ongoing FHWA Pooled Fund Project (TPF-5[198], Mobility Measurement in Urban Transportation [MMUT]), TTI investigated several states' worth of the NPMRDS to characterize the roadway coverage, completeness (temporal and spatial), and validity of the five-minute travel-time data. Data for the 14 states in the FHWA pooled fund project at the time of the analysis were included (California, Colorado, Florida, Kentucky, Maryland, Minnesota, New York, North Carolina, Ohio, Oregon, South Carolina, Texas, Virginia, Washington). Findings are provided here as well as additional analyses related to the truck speed data completeness and validity performed as part of this project. Completeness of the dataset included identification of missing records by traffic message channel (TMC) (the industry standard mapping geography) and by functional classification. For the validity checks, the travel-time data were first converted to speeds, because they can be more intuitively understood and investigated in terms of speed.

Practitioners using large datasets such as NPMRDS for any analyses (such as truck bottlenecks) should understand the data set prior to performing analyses for their specific application. For planning applications, annual averages of five-minute travel-time data may be acceptable rather than day-to-day information that might be more appropriate for operational analyses. Depending upon the application, small nuances in the speeds may not matter, but systematic nuances could cause unacceptable errors.

NPMRDS Coverage

The FHWA pooled fund project analysis investigated the coverage of the NPMRDS. Aggregate findings for the United States are summarized in Table C-1. Analysts investigated the 2011 and 2012 Highway Performance Monitoring System (HPMS) data, and the total directional miles in 2012 on the National Highway System (NHS) increased by 42 percent over 2011 (reflecting the "enhanced NHS" in 2012). Analysts investigated the data from all 13 states in the FHWA pooled fund project and also the entire United States network. Table C-1 shows that the U.S. directional-miles of coverage in the 2012 HPMS (NHS) (479,178 directional miles) compares favorably to the NPMRDS network coverage in directional-miles (486,953 directional miles).

Analysts took the analysis a step further and put the NPMRDS network on a GIS map because that is what is required to perform conflation of speeds and volumes and, again, the results are favorable (475,407 directional miles mapped). The full TMC-encoded network used by industry has approximately 80 percent more coverage (877,882 directional miles—including more arterial/collector coverage) than the NPMRDS network across the United States.

The coverage results of the individual states were generally the same as those documented in Table C-1 for the United States.

Table C-1. United States TMC Roadway Coverage in NHS and NPMRDS

2012 HPMS NHS (directional miles)	NPMRDS Network (Directional Miles)	NPMRDS GIS/ Map Network (Directional Miles)	Full TMC Network (Directional Miles)
479,178	486,953	475,407	877,882

Source: Adapted from *Information Sharing on FHWA's NPMRDS*, Webinar. Texas A&M Transportation Institute, FHWA Pooled Fund Project: Mobility Measurement in Urban Transportation, April 2014.

NPMRDS Completeness

The NPMRDS is five-minute travel-time data for each day and data are present ONLY when probe vehicles are present (i.e., data are not estimated when missing). Therefore, it is important for analysts to understand how incomplete data may affect analyses and measure calculation. Analyst should ask himself or herself what level of completeness is needed for their performance measure application. If performance is needed on specific days, the analyst may need to impute missing values on those specific days. If overall performance for extended periods of time is needed, aggregated statistics to monthly, quarterly, or even annually for day-of-week may be acceptable.

NPMRDS provides three travel-time values for each five-minute-time period and TMC: 1) mixed vehicle; 2) passenger car; and 3) truck. Analysts investigated the completeness of mixed-vehicle and truck travel-time values in the dataset. A three-month dataset from November 2013 to January 2014 was used for the analysis. Daylight hours were used for analysis (6 a.m. to 8 p.m.) because most analysis are most concerned with daytime traffic conditions, rather than overnight hours. Analysts looked at three aggregation-time periods of results: 1) individual day-to-day; 2) one-month average day-of-week; and 3) three months day-of-week.

Completeness in #3 (three months day-of-week) was satisfied if any five-minute travel-time value was present for the given TMC for the given five-minute-time period over the three-month period. Similarly, the one-month average day-of-week was satisfied if a travel-time value was present for the given TMC over the one-month time period. The individual day-to-day value represents the percentage of time the data were available for the specific 5-minute-time period of interest.

Analysts also aggregated the 5-minute data to 15-minute-time periods, urban versus rural and roadway functional classification. The average completeness values for mixed traffic and trucks are shown as rows in Table C-2. The results in Table C-2 are average completeness estimates across the 13 states represented in the FHWA pooled fund project.

The following observations are made about the completeness results in Table C-2:

- Aggregation from individual day to monthly or quarterly increases completeness;
- Aggregation from 5-minute data to 15-minute data increases completeness;
- Truck data completeness percent is substantially less than mixed vehicle;
- Completeness decreases with decreasing functional classification; and
- Completeness in urban areas is generally slightly higher than in rural areas (documented elsewhere). (81)

Table C-2. Completeness Percentages for the NPMRDS Dataset for Mixed Traffic and Trucks

Travel-Time Type	Urban/ Rural	Functional Classification	Individual Day-to-Day		One Month – Average Day-of-Week		Three Months – Average Day-of-Week	
			5-min	15-min	5-min	15-min	5-min	15-min
Mixed-Vehicle	Urban and Rural	All NHS	28%[a]	48%	58%	76%	78%	90%
Mixed-Vehicle	Urban	Class 1 – Interstate	54%	76%	82%	93%	95%	98%
Mixed-Vehicle	Urban	Class 2 – Other Freeway/ Expressway	36%	61%	71%	88%	91%	97%

Travel-Time Type	Urban/ Rural	Functional Classification	Individual Day-to-Day		One Month – Average Day-of-Week		Three Months – Average Day-of-Week	
			5-min	15-min	5-min	15-min	5-min	15-min
Mixed-Vehicle	Urban	Class 3 and 4 – Principal/ Minor Arterial	23%	44%	55%	77%	81%	93%
Truck	Urban and Rural	All NHS	6%[b]	15%	30%	33%	48%	67%
Truck	Urban	Class 1 – Interstate	19%	33%	58%	67%	81%	91%
Truck	Urban	Class 2 – Other Freeway/ Expressway	6%	13%	3%	39%	60%	79%
Truck	Urban	Class 3 and 4 – Principal/ Minor Arterial	3%	5%	16%	21%	37%	60%

Note: All NHS/NPMRDS (Daytime only, 6 a.m. to 8 p.m.); Average completeness values of 14 states represented in the FHWA MMUT Pooled Fund Project; three months of data from November 2013 to January 2014.

[a] Twenty-eight percent is based on 583,070,592 observations out of a total of 2,053,298,688.

[b] Six percent is based on 128,723,738 observations out of a total of 2,053,298,688.

Source: Adapted from *Information Sharing on FHWA's NPMRDS*, Webinar. Texas A&M Transportation Institute, FHWA Pooled Fund Project: Mobility Measurement in Urban Transportation, April 2014 and Margiotta, R., B. Eisele, and J. Short. *Freight Performance Measure Approaches for Bottlenecks, Arterials, and Linking Volumes to Congestion Report,* Federal Highway Administration, Report No. FHWA-HOP-15-033, Washington, D.C., August 2015. Available: http://www.ops.fhwa.dot.gov/publications/fhwahop15033/fhwahop15033.pdf.

These NPMRDS completeness results suggest the analyst should recognize that the five-minute travel-time data are thin in some cases, particularly for the truck data and particularly on principal and minor arterials. Analysts should use caution and be careful not to "slice the data too thinly" for certain performance activities such as planning-level analysis, including many truck bottleneck studies.

It should be noted that aggregation is one method for practitioners to handle missing data – imputation is another. Imputation methods could look at time slices before/after the missing data and/or look at adjacent days.

NPMRDS Validity

Analysts performed a validity test of NPMRDS meant to verify how often speeds were in a specific range or how often there were notable differences between the car and truck travel-time data. As mentioned previously, the five-minute travel-time data from NPMRDS were converted to speeds and the following validity tests were investigated by functional classification:

- Percent of mixed-vehicle speeds and truck speeds less than 5 miles per hour by functional classification (Table C-3);
- Percent of mixed-vehicle speeds and truck speeds greater than 75 miles per hour by functional classification (Table C-3); and
- "Car minus truck speed" difference cumulative percentage distribution by functional classification (Figure C-1 through Figure C-3).

Table C-3. Mixed Speed and Truck Speed Percentage Validation Results by Functional Classification

Validity Check	Speed Data Type	Functional Classification	Percentage
Speed < 5 mph	Mixed-vehicle	Interstate	0%
		Other Freeway and Expressway	0%
		Principal and Minor Arterials	3%
	Truck	Interstate	0%
		Other Freeway and Expressway	1%
		Principal and Minor Arterials	2%
Speed > 75 mph	Mixed-vehicle	Interstate	1%
		Other Freeway and Expressway	1%
		Principal and Minor Arterials	0%
	Truck	Interstate	0.08%a
		Other Freeway and Expressway	0.03%b
		Principal and Minor Arterials	0.05%c

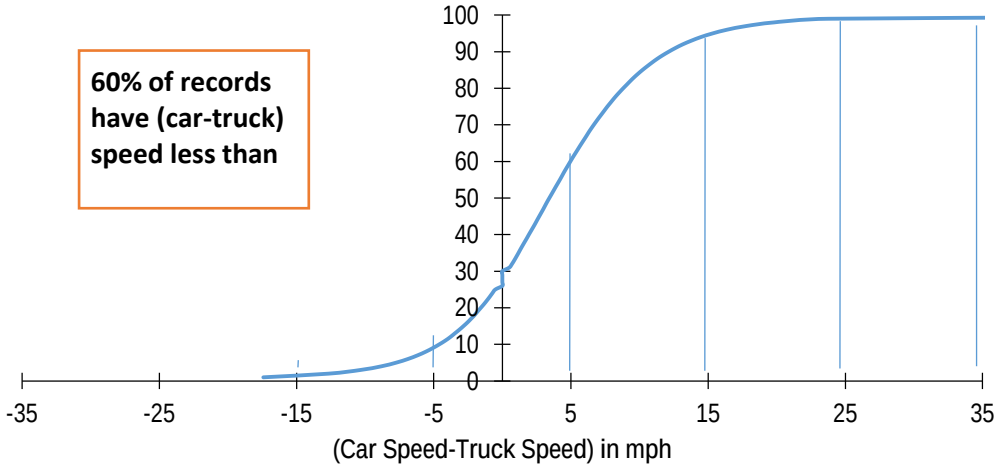
Note: All NHS/NPMRDS (Daytime only, 6 a.m. to 8 p.m.); Average validity values of 14 states represented in the FHWA Pooled Fund Project; one month of data used in the analysis (January 2014).

[a] 0.08% is based on 18,510 observations out of a total of 23,474,056 observations.

[b] 0.03% is based on 1,153 observations out of a total of 4,568,560 observations.

[c] 0.05% is based on 5,526 observations out of a total of 10,542,144 observations.

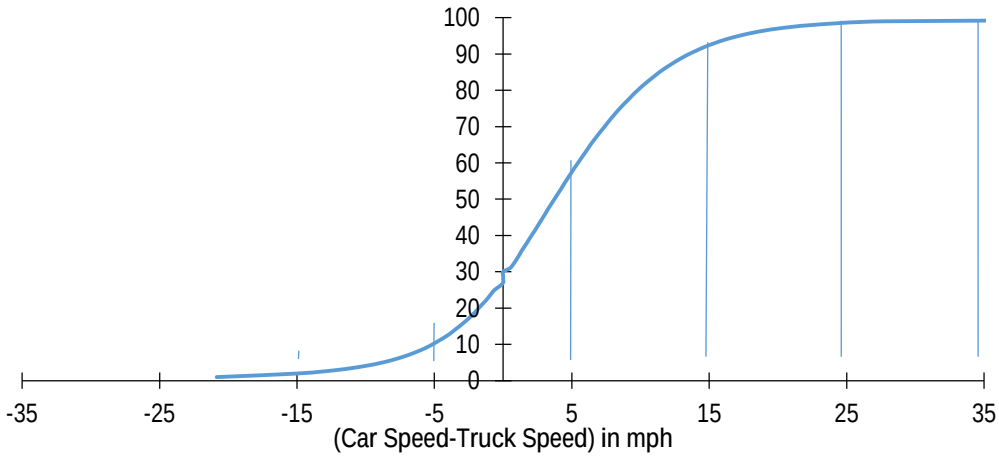
Source: Adapted from *Information Sharing on FHWA's NPMRDS*, Webinar. Texas A&M Transportation Institute, FHWA Pooled Fund Project: Mobility Measurement in Urban Transportation, April 2014 and Margiotta, R., B. Eisele, and J. Short. *Freight Performance Measure Approaches for Bottlenecks, Arterials, and Linking Volumes to Congestion Report,* Federal Highway Administration, Report No. FHWA-HOP-15-033, Washington, D.C., August 2015. Available: http://www.ops.fhwa.dot.gov/publications/fhwahop15033/fhwahop15033.pdf.



Source: Original analysis of FHWA Pooled Fund Project: Mobility Measurement in Urban Transportation data for Margiotta, R., B. Eisele, and J. Short. *Freight Performance Measure Approaches for Bottlenecks, Arterials, and Linking Volumes to Congestion Report,* Federal Highway Administration, Report No. FHWA-HOP-15-033, Washington, D.C., August 2015. Available: http://www.ops.fhwa.dot.gov/publications/fhwahop15033/fhwahop15033.pdf.

Note: All NHS/NPMRDS (Daytime only, 6 a.m. to 8 p.m.); Average validity values of 14 states represented in the FHWA Pooled Fund Project; one month of data used in the analysis (January 2014).

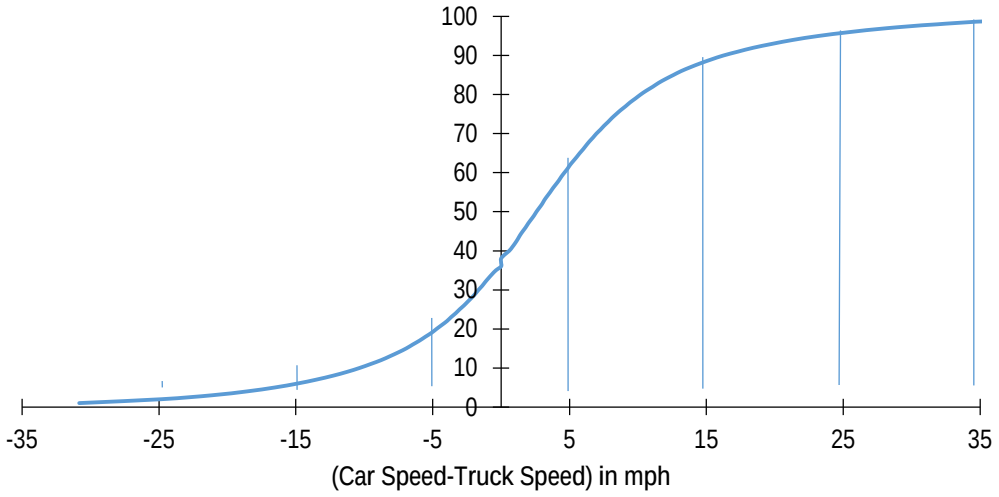
Figure C-1. "Car Speed Minus Truck Speed" Difference Cumulative Percentage Distribution Interstates



Source: Original analysis of FHWA Pooled Fund Project: Mobility Measurement in Urban Transportation data for Margiotta, R., B. Eisele, and J. Short. *Freight Performance Measure Approaches for Bottlenecks, Arterials, and Linking Volumes to Congestion Report,* Federal Highway Administration, Report No. FHWA-HOP-15-033, Washington, D.C., August 2015. Available: http://www.ops.fhwa.dot.gov/publications/fhwahop15033/fhwahop15033.pdf.

Note: All NHS/NPMRDS (Daytime only, 6 a.m. to 8 p.m.); Average validity values of 14 states represented in the FHWA Pooled Fund Project; one month of data used in the analysis (January 2014).

Figure C-2. "Car Speed Minus Truck Speed" Difference Cumulative Percentage Distribution Other Freeway and Expressway



Source: Original analysis of FHWA Pooled Fund Project: Mobility Measurement in Urban Transportation data for Margiotta, R., B. Eisele, and J. Short. *Freight Performance Measure Approaches for Bottlenecks, Arterials, and Linking Volumes to Congestion Report,* Federal Highway Administration, Report No. FHWA-HOP-15-033, Washington, D.C., August 2015. Available: http://www.ops.fhwa.dot.gov/publications/fhwahop15033/fhwahop15033.pdf.

Note: All NHS/NPMRDS (Daytime only, 6 a.m. to 8 p.m.); Average validity values of 14 states represented in the FHWA Pooled Fund Project; one month of data used in the analysis (January 2014).

Figure C-3. "Car Speed Minus Truck Speed" Difference Cumulative Percentage Distribution Principal and Minor Arterials

The validity results in Table C-3 generally indicate low occurrences of the validity checks being satisfied. Other results appear less intuitive (i.e., differences between car speeds and truck speeds,

particularly when trucks are faster). For example, on interstates and other freeways and expressways (Figures C-1 and C-2), approximately 5 percent of data have truck speeds 10 miles per hour faster or more than cars. For principal and minor arterials, approximately 10 percent of the data have truck speeds 10 miles per hour faster or more than cars. Depending upon functional classification, between 25 percent and 35 percent of the data indicate truck speeds faster than car speeds.

The relatively low occurrences documented in Table C-3 and Figures C-1 through C-3 may not impact overall results, but analysts should run such validity tests to verify if/when they do occur and whether they occur during a time period that could impact the results for their specific application (e.g., truck bottleneck analysis).

Because NPMRDS travel-time data are aggregated to the five-minute level, the analyst should verify that adequate travel-time data sample is available for the analyses desired. As described in the prior section, in some cases the travel-time samples at five-minutes are limited (particularly the truck-only travel-time data). To obtain adequate NPMRDS travel-time data sample for a particular analysis, the analyst may need to do the following (for TMCs of interest):

- Aggregate the 5-minute travel-time data to a 15-minute or hourly travel-time estimate;
- Aggregate the daily travel-time data to a monthly, seasonal or yearly travel-time estimate; and/or
- Impute data from adjacent-time periods and/or adjacent days.

Concluding Thoughts for Practitioners

The following are two important questions the analysts should determine the most appropriate aggregation level of the NPMRDS travel-time data:

- *What temporal aggregation is necessary for decision-makers using the results of this analysis?* For example, for a planning study, it is possible that seasonal or annual statistics summarized from 15-minute or hourly data will be adequate.
- *What spatial aggregation is necessary for decision-makers using the results of this analysis?* The travel-time data from NPMRDS begin at the TMC level. In rural areas, TMCs are typically very long while in urban areas they can be shorter (e.g., ramp-to-ramp on an interstate). It is likely that the analyst will want to perform segmentation of their roadway network for analysis differently than the TMC level.

After determining the appropriate temporal and spatial aggregation of the NPMRDS travel-time data (or by imputation), it is recommended that the analyst convert the travel-time information to speeds for quality control prior to aggregation. Within the NPMRDS, there is an inventory file with TMC segment information, including length (miles), a road label (description) and GPS x-y coordinates for the start and endpoints of the TMC. The TMC length (miles) divided by the travel time (in seconds) multiplied by 3,600 (seconds in an hour) gives the speed (miles per hour) for the TMC of interest.

Reviewing speeds is more intuitive for recognizing suspicious speed data and performing quality control. The analyst may want to remove (or cap) speeds that are unreasonably high. An appropriate speed cap, if desired, should consider the functional classification of the roadway (freeway or arterial) and the speed data being investigated ("truck" or "passenger car" or all vehicles"). After quality control of the data, average speeds are computed for the temporal and spatial aggregation levels desired.