Processing Steps For Creating
Standardized 5-Minute-by-Lane Datasets
(program code in "city_base_yyyy.sas")

This summary is provided to ensure analysis consistency in data processing and calculating data quality measures.

- **Dataset DS1:** Import original source data into SAS using INFILE statement. In SAS, we give this first dataset a "DS1" label (DataSet1). Most, if not all, original source data is submitted as ASCII-text files (fixed-column or space/tab/comma-delimited). Many original source datasets also have extra data that we will not use in the performance reports, such as ramp data or single isolated detectors. In some cases the data will already be summarized in 5-minute time periods; in other cases, the data will be more disaggregate (e.g., 20-second or 1-minute) or less disaggregate (e.g., 15-minute).

- **Dataset DS2:** Merge the DS1 dataset with a detector inventory that contains only detectors that will be used later in performance reports, keeping only those detectors that are both in the DS1 dataset AND the detector inventory. We also do other calculations in this dataset to clean up miscellaneous problems and prepare for quality control (e.g., calculate elapsed time between consecutive detector readings).

- **Dataset DS3:** Perform data validity tests; "tag" and output original source data that fails validity tests to a separate dataset; and then remove failed data values from the good data. The validity tests are described in another file ("qc_for_2001.doc"). If a data value fails a validity test, a test-specific "tag" is attached to that observation designating which test(s) failed. Only observations with data values that fail a test are output to a separate dataset, which has the label "QCF" for Quality Control Failed. We have tags for each test, such that we can record if an observation failed multiple tests. After outputting the original data with the failure tags, we delete the data values that failed the tests. When we delete the data, we also attach a "missing value" tag that explains why the data values are missing (i.e., value failed validity test). We use the missing value tags so that we can determine later whether data are missing because of a validity test failure or because it was never sent with original source data.

- **Dataset DS4 (optional):** Prepare dataset for aggregation to 5-minute-by-lane standard. In this optional step, we create even 5-minute time bins (e.g., 12:00, 12:05, 12:10, etc.) for use in aggregating original source data (if necessary).

- **Dataset DS5:** Aggregate data to 5-minute-by-lane standard and merge with detector-time template. When we aggregate the original source data to a 5-minute standard, we keep track of how many observations within that 5-minute period were available and how many failed the validity tests (using the missing value tag). We factor up the vehicle volumes by the percent of missing data within each 5-minute period. If all volume or speed counts within a 5-minute period are unavailable, the volume or speed value is stored as missing/null (it is estimated later in the performance report code). When we aggregate, we weight the average speed and occupancy values by vehicle volumes. Once the data is aggregated to a 5-minute-

by-lane standard, we merge the dataset with a detector-time template. This template has all possible combinations of detector and 5-minute periods throughout the year. By merging the aggregated 5-minute-by-lane data with this template, we can see where we still have gaps in data for each detector for each 5-minute period in the year. We clean up this dataset, format the variables, and output to a DS5 dataset. This DS5 dataset contains observations for 5-minute-by-lane data for all 5-minute time periods for all detectors used in the performance reports (note that some observations may still have missing/null data values). This dataset is used as the primary input for the performance report code ("city_summaries_yyyy.sas").

Now that we have imported, cleaned, and aggregated the original source data, we calculate summary statistics on data quality. The relevant data quality measures are as follows:

- % of data passing quality control, for both volume and speed values (note that these two percents may be different because some validity tests only discard a suspect volume data value but not the corresponding speed value)
- % of complete data available for:
    - original source dataset (as received from participating agency)
    - dataset after quality/validity tests
    - analysis dataset (as used for performance reports)

The quality control measures are calculated as follows:

% of observations passing quality control – This measure is calculated by counting the number of observations in the DS3 dataset that are tagged as passing quality control, then dividing by the total number of observations in DS3 dataset (see Equation 1). We calculate this measure separately for both volume and speed. Note that we are not counting data we never receive as failing quality control; for this measure, our program code only evaluates the original source data we receive for quality control. For example, 95% of the year's data could be missing, but of the 5% we receive, 99.9% could pass quality control. Also note that we perform a separate analysis that tells us the amount of data that failed each validity test. To calculate these totals, we count the number of failed and tagged observations BY TEST in the QCF dataset, then divide by the total number of observations in the DS3 dataset.

$$\begin{bmatrix}\text{\% Passing Q.C.}\\ \text{volume \& speed}\\ \text{treated separately}\end{bmatrix} = \frac{\text{\# of obs. in DS3 dataset with "QC passed" tags}}{\text{total \# of obs. in DS3 dataset}} \qquad \text{(Eq. 1)}$$

% complete for original source dataset – This measure is calculated by counting the number of non-missing values in the DS2 dataset (after the data that will not be used in performance reports has been removed), then dividing by the total expected observations in the original source data (see Equation 2). The total expected observations is calculated as follows: number of detectors used for performance reports × number of polling cycles in a day (can be different for each city) × 365 days per year. We calculate this measure separately for both volume and speed values.

$$\begin{matrix}\%\ \text{Complete,}\\ \text{Original Source Data}\\ \left[\begin{matrix}\text{volume \& speed}\\ \text{treated separately}\end{matrix}\right]\end{matrix} = \frac{\#\ \text{of non-missing obs. in DS2 dataset}}{\#\ \text{of expected obs.} \left[\begin{matrix}\text{i.e., \# of detectors} \times\\ \text{expected obs. per day} \times \text{365 days per year}\end{matrix}\right]} \qquad \text{(Eq. 2)}$$

% complete after quality/validity tests – This measure is calculated by counting the number of non-missing values in the DS3 dataset (after the data failing validity tests has been removed), then dividing by the total expected observations, which is the same as in the previous step (see Equation 3). Thus, the total expected observations is calculated as follows: number of detectors used for performance reports × number of polling cycles in a day × 365 days per year. We calculate this measure separately for both volume and speed values.

$$\begin{matrix}\%\ \text{Complete,}\\ \text{After Q.C.}\\ \left[\begin{matrix}\text{volume \& speed}\\ \text{treated separately}\end{matrix}\right]\end{matrix} = \frac{\begin{matrix}\#\ \text{of non-missing obs. in DS3 dataset}\\ \left[\text{after "QC failed" data removed}\right]\end{matrix}}{\#\ \text{of expected obs.} \left[\begin{matrix}\text{i.e., \# of detectors} \times\\ \text{expected obs. per day} \times \text{365 days per year}\end{matrix}\right]} \qquad \text{(Eq. 3)}$$

% complete for analysis dataset – This measure is calculated by counting the number of non-missing values in the DS5 dataset (after the data has been aggregated to 5 minutes and merged with the DS5 template), then dividing by the total number of observations in the DS5 template (see Equation 4). The total number of observations in the DS5 template is calculated as follows: number of detectors used for performance reports × 288 5-minute periods in a day × 365 days per year. We calculate this measure separately for both volume and speed.

$$\begin{matrix}\%\ \text{Complete,}\\ \text{Analysis Dataset}\\ \left[\begin{matrix}\text{volume \& speed}\\ \text{treated separately}\end{matrix}\right]\end{matrix} = \frac{\begin{matrix}\#\ \text{of non-missing obs. in DS5 dataset}\\ \left[\text{after merge with DS5 template}\right]\end{matrix}}{\#\ \text{of expected obs.} \left[\begin{matrix}\text{i.e., \# of detectors} \times\\ \text{288 5-minute periods per day} \times \text{365 days per year}\end{matrix}\right]} \qquad \text{(Eq. 4)}$$