# NCST Caltrans project on sensor data error estimation Dataset

**Dataset available at:** https://doi.org/10.25338/B8TP5Q

(This dataset supports report **Improving Transportation Information Resilience: Error Estimation for Networked Sensor Data,** https://doi.org/10.7922/G2610XK9)

**Metadata from the Dryad Repository record:**

Abstract: The aim of this script is to automate the process of directed network graph formation, i.e., creation of incidence matrix, node adjacency matrix, and map the sensors to appropriate links. The data used for creating the freeway network is obtained from open street maps, whereas, the data for sensors is obtained from PeMS. The results of the algorithm will later be fed to the network sensor error estimation algorithm, that quantifies the erroneous sensors using network statistics.

Methods

- **Major libraries used:**
    - **Osmar** - To import data from open street maps **Leaflet**, **mapview**, **sp** - Mapping the network **dplyr**, **plyr** - To manipulate data frames
- **Data used:**
    - The data used for creating the freeway network is obtained from **open street maps**, whereas, the data for sensors is obtained from **PeMS**.
- **Methodology**
    - **Creating the map required** We start by downloading the bulk osm data for California (approx. 18GB). Next, using osmar library, we extract the required data using a bounding box, demarcating the latitude and longitude boundaries. *bbox = corner_bbox(-118.0042, 33.6363, -117.7226, 33.9194)* Then extracting only the freeway information (links and nodes) from the resulting data.  The problem here is that most of the links contain more than two nodes, which would make the incidence matrix unnecessarily large. Thus, we extract the nodes that connect two links, and map them over the links.
    - **Creating the incidence matrix** To create the incidence matrix, each link was looked up for the first and last node incident on it, keeping in mind the direction. The rows represent nodes, whereas the columns represent links. +1 was assigned at the head and -1 to the tail of the link, all other entries to 0. The incidence matrix has a dimension of 2973 by 2640.

- **Assigning links to appropriate freeways** The osm data does not explicitly mention about which freeway is a particular link part of. Thus, for the freeway links a combination of "name" and "ref" tag was used to obtain the information about the freeway.
- **Adding ramps to the Fwy data frame** So far, we only have freeway links assigned to the appropriate freeways. Now, we would like to add the immediate links that go off or on the freeway, aka ramps, to the data frame Fwy. This is done by checking for each node on the freeway segment, out of the links it is incident upon, which one is tagged as "motorway-link" in the osm data. If there is such a link, it was added to the Fwy data set with appropriate freeway values.
- **Overlapping sensors over the network** The sensor data is now used to add a layer over the existing graph to help visualize the complete network. Different types of sensors are grouped separately and can be viewed as per user's choice by clicking the check boxes. The legend shows the color used for each sensor type. Hovering upon the sensor, link or node highlights their IDs. By default, main line (ML) sensors are checked.
- **Mapping sensors to the appropriate links** Currently, by visualization we can figure out the link that contains a particular sensor. As PeMS sensor metadata does not interact with the osm data, the link-sensor relation is unknown. We need to create an algorithm such that each sensor automatically gets mapped to the link using the geographical properties. The algorithm for mapping sensors to the links is as follows:
    1. For each sensor location, extract all the links on the freeway segment in the direction sensor is installed
    2. For the nodes on each link, calculate 3 distances
        1. Distance between the nodes (d1)
        2. Distance between first node and sensor (d2)
        3. Distance between last node and sensor
    3. Calculate d1 - (d2+d3), call it d4
    4. Calculate d4 for each link, and arrange d4 in ascending order
    5. The link for which d4 is smallest and lesser than a threshold (1e-4 in this case), assign it the sensor
    6. Repeat the above steps for each sensor location

  Finally, the results are stored as a form of a list (linkId). For illustration purpose, 5 ML sensors on I-5, highlighted on the map, are shown below with the appropriate link chosen by the algorithm. One can verify the IDs by hovering above the links in the map and cross checking with the table that appears below.
- **Re-mapping ramp and freeway-freeway sensors** Looking closely, one would figure out that the ramp sensors (OR/FR) are located on the freeways rather than ramps. Same for freeway-Freeway (FF) sensors. This was one of the tedious challenges I encountered in this project. But with a combination of a simple algorithm and manual work, the sensors were remapped. The details are omitted in this document. In the map below, all the remapped sensors are shown on their appropriate new links.

  - **Creating the adjacency matrix** To create adjacency matrix, for each link, nodes having 1 or -1 were searched in the incidence matrix. The cell corresponding to these nodes in the adjacency matrix was assigned 1, else 0.
- **Conclusion**
  - The results of this project, namely, incidence matrix, adjacency matrix and link-sensor relation data frame were used for the network sensor error estimation algorithm. This project led to the application of the error estimation algorithm on large networks, which is expected to result in an important contribution to the field of sensor bias estimation.

Usage Notes

For more details and some illustrations to facilitate a better understanding of the data and script, please refer to the following github link: https://github.com/sbhmaheshwari/Projects/blob/master/Network%20graph%20automation/Graph_Contraction.md

Funding

California Department of Transportation, Award: NCST Caltrans TO 003
National Science Foundation,
National Science Foundation, Award: CMMI 1825873

**Recommended citation:**

Fan, Yueyue; Maheshwari, Saurabh (2020), NCST Caltrans project on sensor data error estimation, Dryad, Dataset, https://doi.org/10.25338/B8TP5Q

**Dataset description:**

This dataset contains 1 .zip file collection described below.

**doi_10.25338_B8TPQ_v1.zip:**

This collection contains 1 .R file listed below. For more information on this file type please visit https://www.file-extensions.org/search/?searchstring=.R&searchtype=2 for more information.

- Orange_network_graph_(1).R