Set Up the Environment

Software Tools

Majority of the data processing work was completed in R. The latest version of R can be downloaded from the software website (https://www.r-project.org/). Before executing any of the R scripts, the first step is to set the R working directory to the root folder which contains all the data, result, scripts and the User Guide (this document). It is also required to first install all R packages used in the calculation. R packages can be installed using the code `install.packages("package name")`. Below is a list of all required R packages

- sqldf
- sas7bdat
- geosphere

The Polk Vehicle Registration data was processed by Python due to its efficiency in working with big datasets. The latest version of Python can be downloaded from https://www.python.org/downloads/. The Python script is provided in the file `04_count_bus_by_zip.py`.

Data Preparation

All the required data have been preprocessed and provided in the `Data/` folder. This section summarizes key steps in preparing the data for analysis and calculation.

National Transit Database (NTD)

The NTD datasets can be downloaded from the Federal Transit Administration (FTA) website (https://www.transit.dot.gov/ntd/ntd-data). This project uses five years (i.e., 2012-2016) of NTD data. The data tables are placed in the `Data/NTD/` folder. For each year, two datasets have been downloaded, Service table to obtain occupancy information and Agency Information table to map transit agencies to different locations.

Some column names have been revised for the ease of data processing. In the Service table, these changes are summarized in the following table.

Original Column Name	New Column Name
NTD ID	ntdid
Agency Name	name
Reporter Type	reporter
Mode	mode
Time Period	time
Vehicles/Passenger Cars Available for Maximum Service	veh_avai
Actual Vehicles/Passenger Car Miles	vmt
Actual Vehicles/Passenger Car Revenue Miles	vmt_rev
Passenger Miles	pmt

Changes of column names in the Agency table are summarized in the table below.

Original Column Name	New Column Name
NTD ID	ntdid

City	city
State	state
Zip Code	zipcode

School Bus Fleet Data

The U.S. State by State Transportation Statistics 2015-16 reported by SchoolBusFleet.com (Data Source: http://files.schoolbusfleet.com/stats/SBFFB18StateByState.pdf) is employed to calculate the school bus occupancy for state level. The report provides a breakdown of information for each of the 50 states, including the number of K-12 public and private school students transported daily, the number of school buses in each state and the total annual route mileage. The data table is located in the `Data/` folder as `SchoolBusFleet.csv`.

Student Enrollment Data

The student enrollment data provided by the Governing.com (https://www.governing.com/gov-data/education-data/school-district-totals-average-enrollment-statistics-for-states-metro-areas.html) are used to estimate school bus occupancy rate for states with missing data and urbanized areas. Specifically, the local factor data is used to estimate school bus occupancy for 14 states with missing, located in the `Data/` folder as `localFactor.csv`; the student enrollment data at state and metro area level are used to calculate school bus occupancy for urbanized areas, located in the `Data/` folder as `state_school_enroll.csv and urban_area_school_enroll.csv`. To map mero area data into the urbanized areas, the relationship between the urbanized areas and metro areas (https://www.census.gov/geo/maps-data/data/ua_rel_download.html) provide by the U.S. Census Bureau website was referred.

Port Authority Bus Terminal (PABT) Data

The PABT data table is used to calculate motorcoach occupancy rates. The data table is located in the `Data/` folder as `PANYNJ_2015.csv`. The dataset is requested from the Port Authority of New York and New Jersey (PANYNJ). In addition to the original data, we also manually mapped each motorcoach route with states and urbanized areas along the route. Such data are store in the additional columns named state1, state2, …, urban_area1, urban_area2, …

Polk Vehicle Registration Data

The Polk data are used to calculate the number of buses in each type so that occupancy rates from different bus types can be aggregated to produce an overall bus occupancy rate. Note that the Polk dataset is of a huge size thus not included in the deliverables. An empty file named `trucks.sas7bdat` is placed in the `Data/` folder as a placeholder.

Supplementary Datasets

Five supplementary datasets are also used to help estimate occupancy rates or create geospatial relationship between different data tables. All these datasets are located in the `Data/` folder.

`states.csv` includes state code, state name, state abbreviation for each of the State and the District of Columbia. Additionally, longitude and latitude of the center point of each state are obtained from the website DistanceFromTo (https://www.distancefromto.net/).

`gdp_state.csv` includes the All Industry GDP and GDP for Transit and Ground Passenger Transportation for each of the State and the District of Columbia from 2012 to 2016. The data can be downloaded from the U.S. Bureau of Economic Analysis (BEA) website (https://apps.bea.gov/itable/iTable.cfm?ReqID=70&step=1#reqid=70&step=1&isuri=1).

`pop_state.csv` includes the population, area, and population density for each of the State and the District of Columbia from 2012 to 2016. The data can be downloaded from the U.S. Census Bureau website (https://www.census.gov/data/tables/2017/demo/popest/state-total.html).

`urban_areas_200k.csv` summarizes 183 urbanized areas with population higher than 200,000 as defined by the U.S. Census Bureau. The data can be downloaded from the U.S. Census Bureau website (https://www.census.gov/geo/reference/ua/urban-rural-2010.html).

`urban_county.csv` and `urban_ZCTA.csv` define relationships between the urban areas with counties and zip code tabulation areas. The data tables can be downloaded from the U.S. Census Bureau website (https://www.census.gov/geo/maps-data/data/ua_rel_download.html).

Calculate the Occupancy Rates

Execute Software Codes

The R and Python scripts can be executed to reproduce the bus and truck occupancy rates as summarized in the project report. It is recommended that the scripts being executed following their name order. However, most scripts can be run independently except `05_aggregate_bus.R`, which relies on the output from scripts 01-04. All the scripts have been pre-executed with the results generated in the `Result/` folder. Detailed comments can be found in the scripts accompanying the software codes. Here we list the input and output of each script so that users can better understand the code structure.

Script name: `01_transit_bus.R`

Input	
File	**Description**
`states.csv`	U.S. state and federal district names, codes, abbreviations and coordinates
`gdp_state.csv`	2012-2016 state level all industry GDP and GDP in transit and ground passenger transportation
`pop_state.csv`	2012-2016 state level population and population density
`urban_areas_200k.csv`	U.S. Census Bureau defined urbanized areas with population higher than 200,000
`urban_county.csv`	2010 urban area to county relationship file
`urban_ZCTA.csv`	2010 urban area to zip code tabulation area (ZCTA) relationship file
`2016_Service.csv`	2016 National Transit Database service table
`2016_Agency.csv`	2016 National Transit Database agency information table
`2015_Service.csv`	2015 National Transit Database service table
`2015_Agency.csv`	2015 National Transit Database agency information table
`2014_Service.csv`	2014 National Transit Database service table
`2014_Agency.csv`	2014 National Transit Database agency information table
`2013_Service.csv`	2013 National Transit Database service table
`2013_Agency.csv`	2013 National Transit Database agency information table
`2012_Service.csv`	2012 National Transit Database service table
`2012_Agency.csv`	2012 National Transit Database agency information table
Output	
File	**Description**
`transit_state.csv`	Average transit bus occupancy and VMT in each of the 50 U.S. states and District of Columbia

Script name: `02_school_bus.R`

Input	
File	**Description**
`SchoolBusFleet.csv`	School bus transportation Statistics 2015-16 reported by SchoolBusFleet.com
`localFactor.csv`	State local factors data including average district enrollment and total schools
`urban_area_school_enroll.csv`	Average district enrollment and total districts for urbanized areas
`state_school_enroll.csv`	Average district enrollment for states
`urban_school.csv`	Urbanized area name, belong state, and population

Output	
File	**Description**
school_state.csv	Average school bus occupancy and VMT in each of the 50 U.S. states and District of Columbia

Script name: 03_motorcoach_bus.R

Input	
File	**Description**
states.csv	U.S. state and federal district names, codes, abbreviations and coordinates
gdp_state.csv	2012-2016 state level all industry GDP and GDP in transit and ground passenger transportation
pop_state.csv	2012-2016 state level population and population density
urban_areas_200k.csv	U.S. Census Bureau defined urbanized areas with population higher than 200,000
urban_county.csv	2010 urban area to county relationship file
urban_ZCTA.csv	2010 urban area to zip code tabulation area (ZCTA) relationship file
PANYNJ_2015.csv	2015 Port Authority Bus Terminal data
Output	
File	**Description**
motorcoach_state.csv	Average motorcoach occupancy and VMT in each of the 50 U.S. states and District of Columbia

Script name: 04_count_bus_by_zip.py

Input	
File	**Description**
trucks.sas7bdat	Polk Vehicle Registration data
Output	
File	**Description**
bus_cnt_zip.csv	Number of buses by carrier type in each zip code tabulation area

Script name: 05_aggregate_bus.R

Input	
File	**Description**
states.csv	U.S. state and federal district names, codes, abbreviations and coordinates
urban_areas_200k.csv	U.S. Census Bureau defined urbanized areas with population higher than 200,000
urban_county.csv	2010 urban area to county relationship file
urban_ZCTA.csv	2010 urban area to zip code tabulation area (ZCTA) relationship file
bus_cnt_zip.csv	Number of buses by carrier type in each zip code tabulation area
Output	
File	**Description**
bus_state.csv	Average bus occupancy and VMT in each of the 50 U.S. states and District of Columbia