

Calculating Vehicular Emissions in Manhattan

Prince Abunku, Shreya Bamne, Gerardo Rodriguez

Introduction:

Air Quality has become a topic of significant importance as cities continue to urbanize. While there are many cities that track air quality, it is still a difficult problem to understand its sources. Air quality is given such importance as it has severe impacts on health. As 80% percent of population lives in the cities, it is further more important to address this issue. Out of all the sources identified so far, vehicular emissions is one such source which is mentioned by many studies. The amount of vehicular emissions is dependent on the amount of vehicles, their types, and traffic conditions in a city. On its own estimating traffic mobility is important to cities to be able to better plan cities. Inferring emissions from this data could help New York City solve policy challenges related to air quality and transportation.

Data:

1. Traffic Volume Counts

We are using the Traffic Volume Counts data which is collected by the Department of Transportation (DOT) for the year 2012-13. The dataset has 5945 records and 31 columns. It has the segment ID which represents a road segment, date on which the data was recorded, direction which indicates the direction of the vehicles on that road segment going north or south bound and counts of vehicles by each hour. the following figure shows the dataset screenshot. The road segment id corresponds to the lion road segment ID in the Department of City Planning's LION base map file.

Segment ID	Roadway Name	From	To	Direction	Date	12:00-1:00 AM	1:00-2:00AM	2:00-3:00AM	...	2:00-3:00PM	3:00-4:00PM	4:00-5:00PM
2153	HUGUENOT AVE	WOODROW RD	STAFFORD AVE	NB	02/02/2013	106	74	45.0	...	371	398	324
2153	HUGUENOT AVE	WOODROW RD	STAFFORD AVE	NB	02/03/2013	109	74	55.0	...	308	291	313
2153	HUGUENOT AVE	WOODROW RD	STAFFORD AVE	NB	02/04/2013	36	28	11.0	...	426	425	419

Figure 1: Screenshot of Traffic Counts Dataset

2. NYC LION base map file

This dataset contains all the road segments in NYC. There are 218,349 different road segments contained in the dataset. Each SegmentID is composed of 7 digits. This file also contains the location of the road segment in the city with the geometry field as well as the borough and zip. The file also contains the physical attributes of the road segments including the number of lanes, a binary curved_street field, and road condition.

3. Yellow Taxi Trip data

The yellow taxi data contains over 14 million records. Each record indicates a trip taken throughout the city. The data provides the pickup and drop off location, distance traveled, the amount of time of the trip and the arrival and end time of the trip.

vendor_id	pickup_datetime	dropoff_datetime	passenger_count	trip_distance	pickup_longitude	pickup_latitude	rate_code	store_and_fwd_flag	dropoff_longitude
VTS	2012-09-01 05:35:00	2012-09-01 05:41:00	1	2.27	-73.995642	40.725272	1	NaN	-73.992367
VTS	2012-09-01 05:31:00	2012-09-01 05:41:00	1	3.94	-73.973277	40.792908	1	NaN	-73.976047
VTS	2012-09-01 05:16:00	2012-09-01 05:40:00	2	16.75	-73.937563	40.801260	2	NaN	-73.783300

Figure 2: Screenshot of Yellow Taxi Data

4. Vehicle Classification Counts

This is a similar data set than that of the traffic volume counts; although significantly smaller in records, it contains vehicle type classification for each traffic count. The data, for October 2012, will be used to determine the ratio between the different vehicle types which will serve as input for a better prediction outcome by the model.

Methods:

Before vehicle emissions could be predicted, the amount of traffic in the city has to be estimated. The traffic would be estimated by comparing the number of taxis on a road segment to the actual traffic count of vehicles on that road segment. (“Predicting Vehicular Emissions in High Spatial Resolution Using Pervasively Measured Transportation Data and Microscopic Emissions Model.” 2016) To do this the volume counts data was first merged with the LION base map in order to determine where the traffic counts were being taken. The segment ID field in the volume count data had many values that had less than 7 digits. Our assumption was that the volume count data removed the trailing zeroes. We added trailing zeroes to the segment ID and then looked at several IDs to see if they were also contained in the base map. Looking at the roadway name in the traffic counts and the street name in the base map we saw that they were the same and thus concluded that the new segment IDs were correct. After creating the segment IDs we observed a significant problem with our traffic count data set. Much of the data recorded by the DOT was missing. Only several months actually had any data about the traffic counts for the

year 2012. In addition to that, the month of October had 4422 rows. Nearly 75% of the data belonged to the month of October. This means in order to make sure our dataset was sufficiently large we could only work with this month. The expected records we would have for a month would be the number of detectors * the number of days in the month. There were 364 unique segment IDs in the month of October and so we would expect to see 11,284 records. Twice that if we considered northbound and southbound traffic counts individually. Thus even within the month of October we still had much missing data. To overcome this issue we looked for a time period within October that had a large and consistent amount of data. The amount of recorded data is given in figure 2.

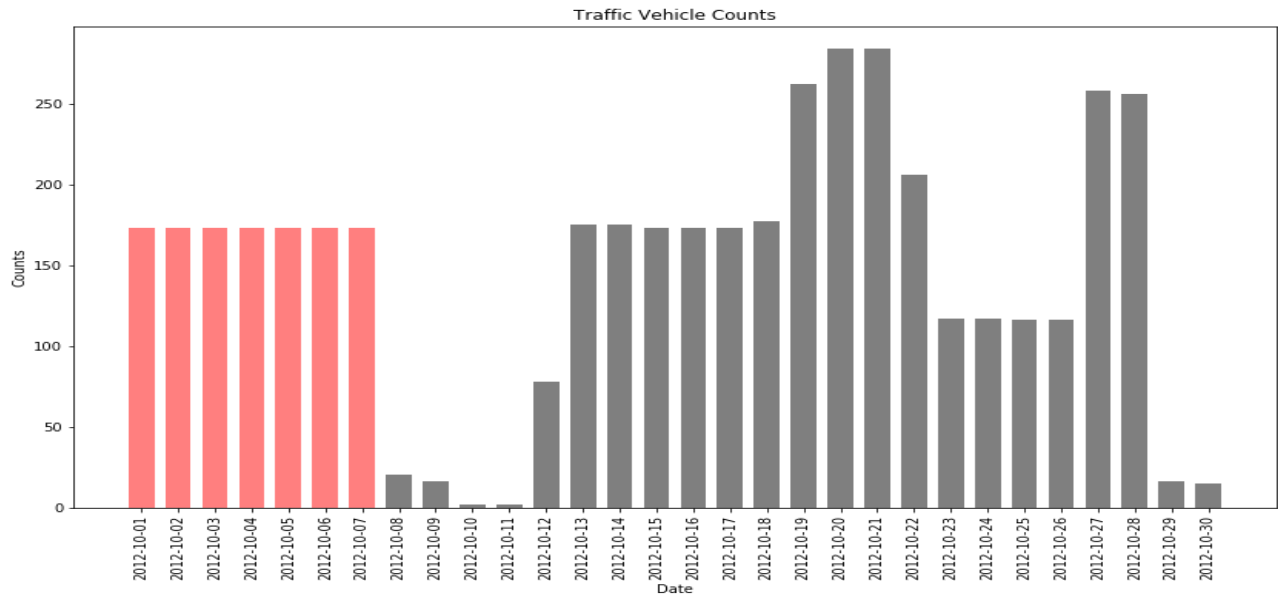


Figure 3: Traffic Vehicle Counts – for different months in dataset

The most consistent data was the first week so the dates of October 1st through October 7th were selected. This had a total number of records of 1092. The number of unique detectors were 114. This data seems the most reliable to use and merge with the base map. Before we could merge the datasets the base map was first reduced to only the areas interested in. The base map has two columns indicating if either side is in a specific borough. Our target area for the data is Manhattan and thus other borough streets were removed. Also, only streets where motor vehicles could operate were selected. This reduced the number of road segments to about 23,000. Combining these two datasets we have 1309 rows that include 7 days of traffic count data and their locations. Traffic counts recorded within Manhattan are shown in Figure 3.

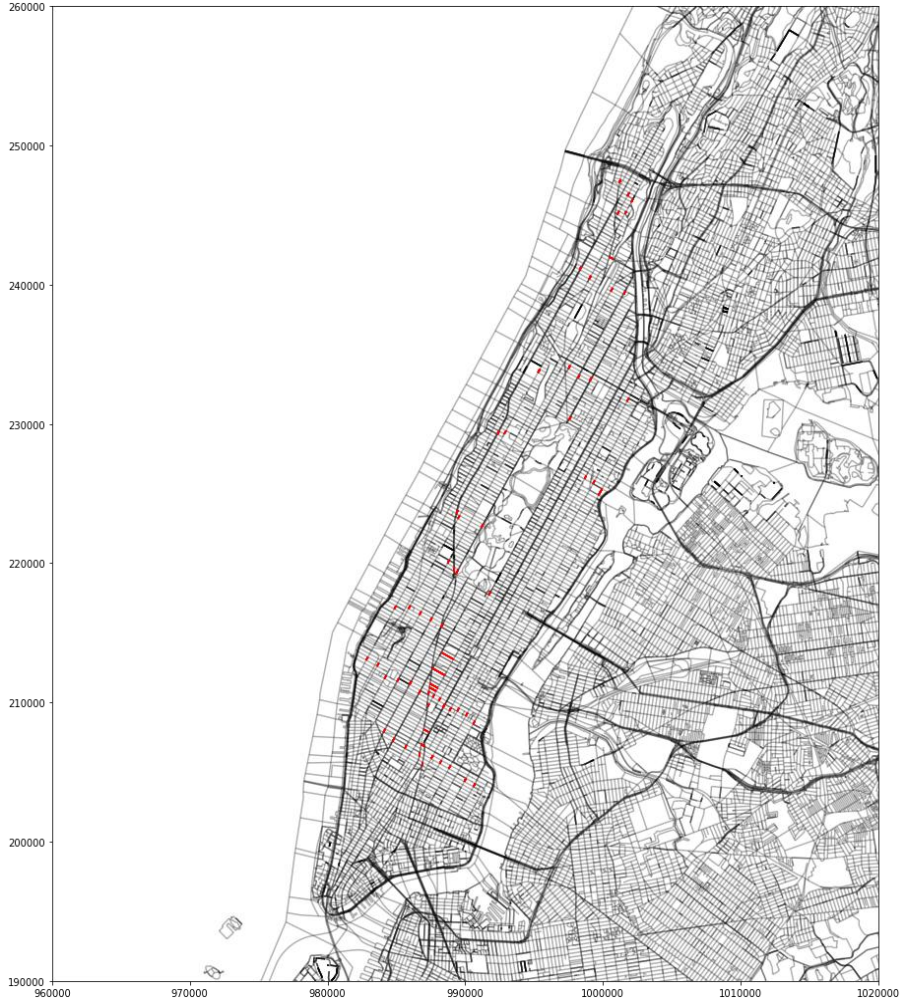


Figure 4: Traffic sensors location in Manhattan (shown in red) for October

The next dataset to process was the yellow taxi dataset in the same time period as traffic data. There were several assumptions we had to make about the data. There was no unique identifier so we had to assume each taxi was unique. A unique id was given to each taxi corresponding to its position in the dataset. There were also no route attributes given so another assumption was that a trip would take the optimal route to its destination. The Google Directions API was used to determine the optimal route. Unfortunately the earliest data we could use is from October 2012 and so roads could have changed since then. The Google API also has a limit of 2500 requests per day. In order to stay under the limit we used a random sample of 1000 yellow taxi trips. We also made sure that all trips selected were located in Manhattan. The API gave us the routes in order to reach a trip's destination. Each step in the route was a road that was traveled. These steps we treated as line segments. We then took the line segments and created a new dataset that

made each row one that contained the unique taxi id and the line segment for a trip. This dataset is then combined with the traffic count data in ArcGIS. This gives the number of taxis that crossed a particular road segment, the time of the trip, and the ground truth data. We could then use the taxis to predict the number of vehicles on a road segment.

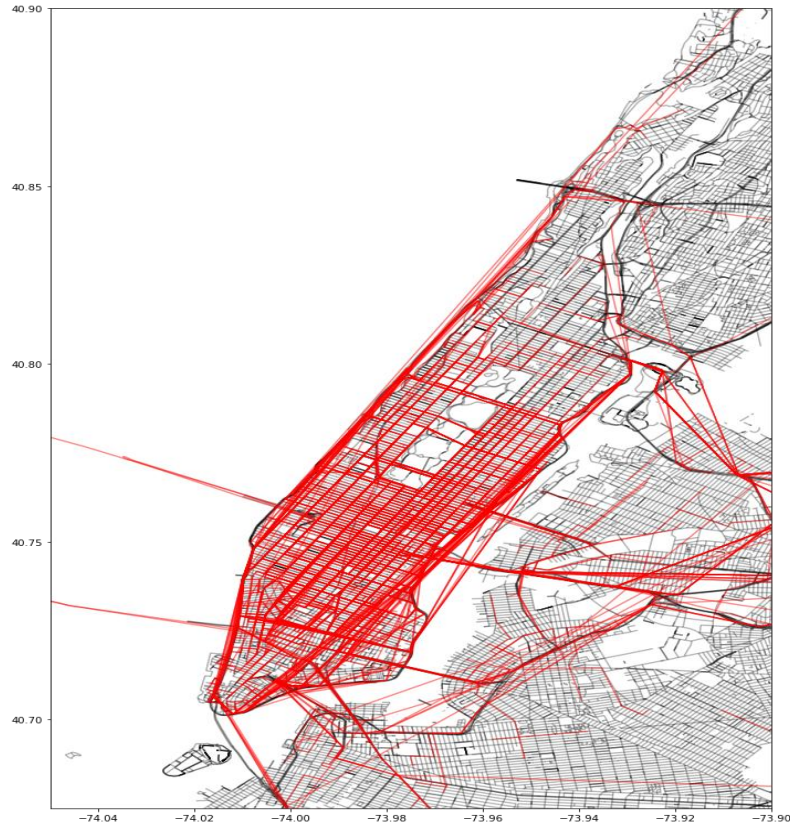


Figure 5: Taxi trips simulated routes for the first week of October 2012

In addition to the overlapping of segments, we also plotted counts of vehicles and taxis for the month of October. The number of traffic counts for road segments was taken from Traffic Volume Counts Data.

The taxi trips and the mean traffic counts were plotted for the month of October, highlighting the weeks of interest. We can infer by the shape of the distribution that there is a certain trend that peaks during Saturdays and hits a low point on Mondays. This is congruent with what Seejon Lim et al. found (Aslam et al. 2012) and can be considered common practice in the field of urban science, to distinguish between workday and weekend.



Figure 6: Taxi Trips and Traffic Counts - Compared

The taxi counts were calculated by filtering data by October month and then using group by to get the counts. We then plotted both of them and found a similar trend which can be seen in figure 4. This adds to the previous statement about using taxi data to represent traffic.

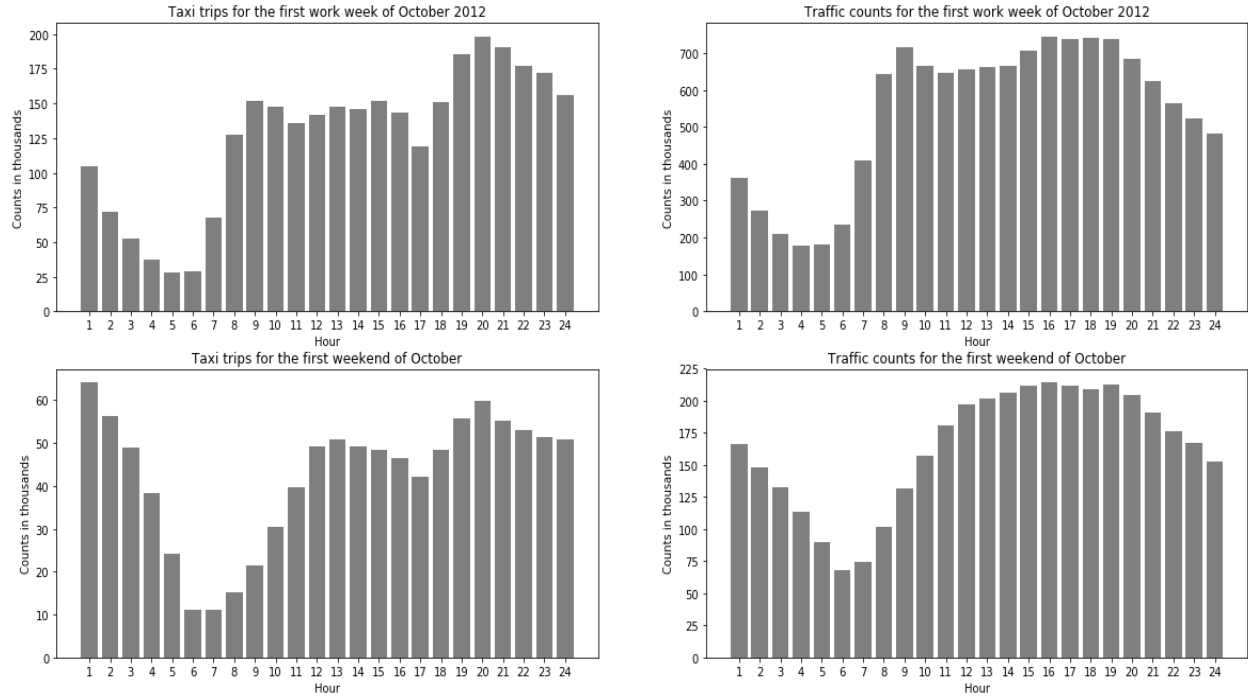


Figure 7: Distribution comparison for number of taxi trips and traffic counts for the first week in October 2012

The following graph shows the ratio between different vehicle types for common intersections between the traffic counts data set and the vehicle classification counts. Note that the counts are significantly smaller due to the fact that only 9 detectors matched the chosen intersections.

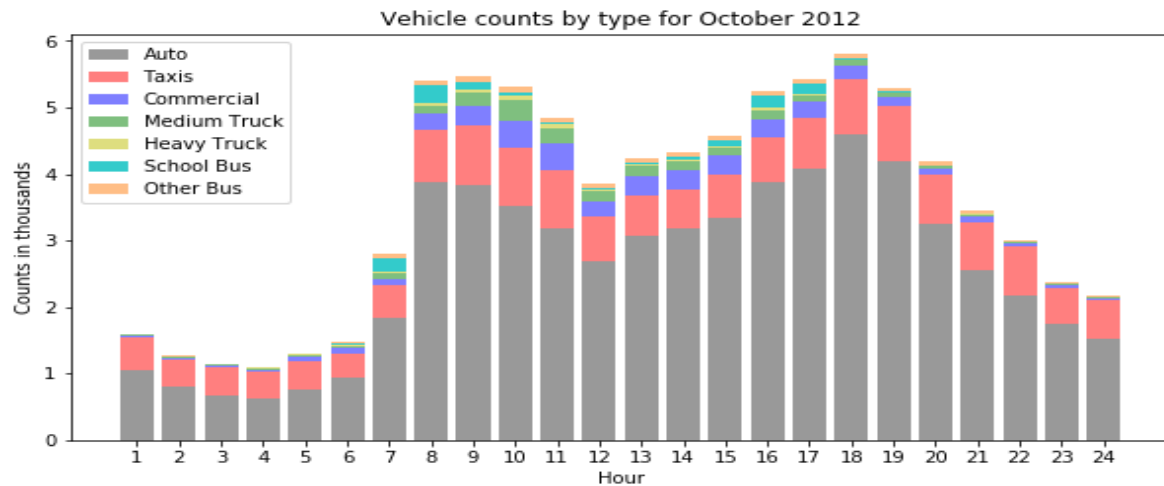
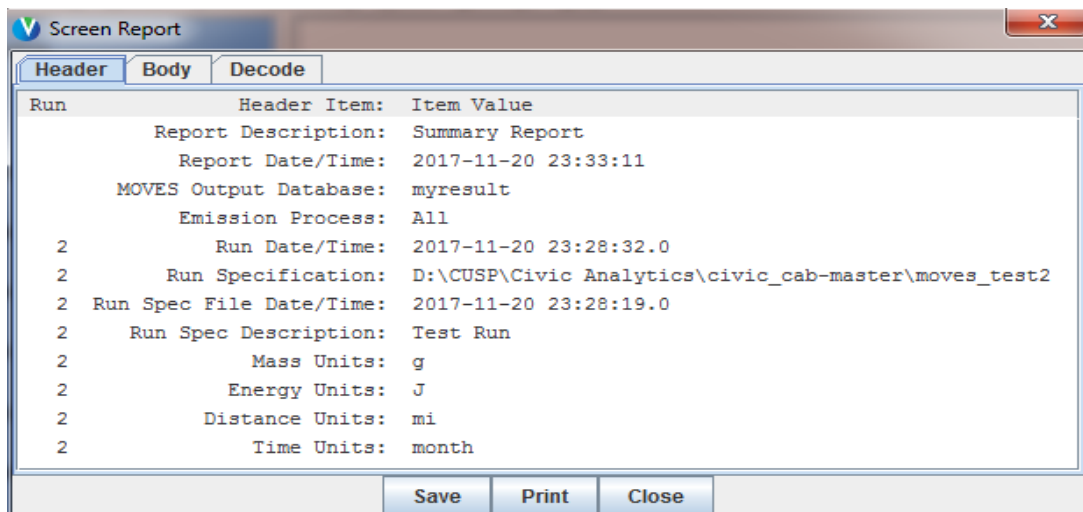


Figure 8: Vehicle Counts by Type

The model:

- DRACULA will serve as our traffic micro-simulation model. This is a software developed at Leeds ITS that is able to represent the progress of and interactions between individual vehicles as they pass through a road network..(“DRACULA - Home Page”, n.d.)
- In order to evaluate our predictions of emissions, we plan to use MOVES by Environmental Protection Agency (EPA), that estimates emissions for mobile sources.(“MOVES and Related Models | US EPA”, n.d.) It does so by taking various inputs which are as follows:
 - Scale: In this we need to select whether the model is Onroad or NonRoad. Then we need to select which database we want to use. There is a default national database which has some default state and allocation factor. Then we need to select the calculation type whether it is emissions for a region or emissions per unit activity.
 - Time span: This includes selection of different aggregation levels of time by year, month and day of week (weekdays or weekends or both).
 - Geographic bounds: Here we need to specify whether analysis at what level – nation, state, country, zone, etc.
 - Vehicle equipment: This requires user to select combinations of fuels (like diesel, CNG) and source types (like Passenger car, Motorcycle).
 - Road Types: There are bunch of options here like urban, rural, etc.
 - Pollutants and processes: Here we need to select the pollutants for which emissions are calculated along with processes (like Brakewear, start exhaust)
 - Input Database: We can either use the default database or we can create our own. MOVES uses MYSQL database.
 - Output: We need to select units for output. Also location for storing results needs to be specified.

After the model is run, the output can be viewed in a summary report. We will use this report as a reference to compare our results. Following figures show MOVES output for a test run on their inbuilt database.

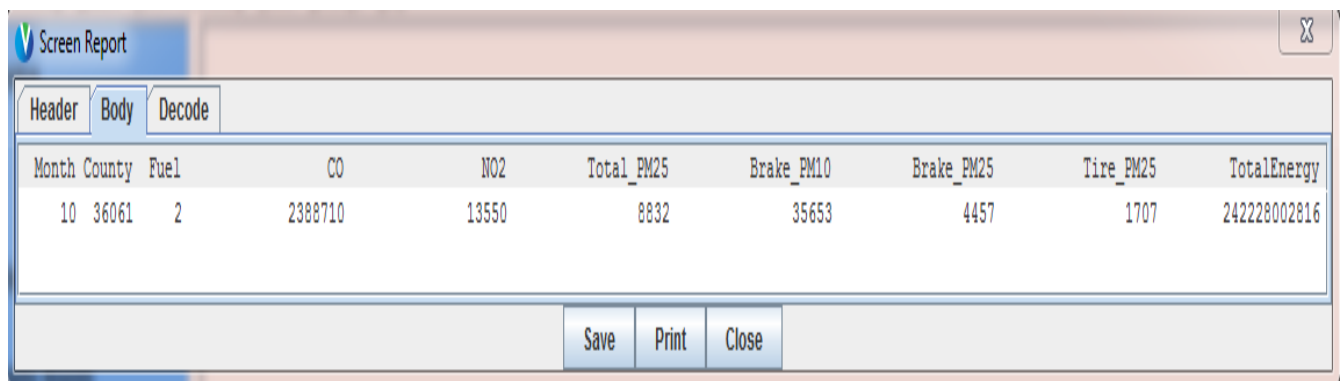


The screenshot shows the 'Screen Report' window with the 'Header' tab selected. The report contains the following information:

Run	Header Item:	Item Value
	Report Description:	Summary Report
	Report Date/Time:	2017-11-20 23:33:11
	MOVES Output Database:	myresult
	Emission Process:	All
2	Run Date/Time:	2017-11-20 23:28:32.0
2	Run Specification:	D:\CUSP\Civic Analytics\civic_cab-master\moves_test2
2	Run Spec File Date/Time:	2017-11-20 23:28:19.0
2	Run Spec Description:	Test Run
2	Mass Units:	g
2	Energy Units:	J
2	Distance Units:	mi
2	Time Units:	month

Buttons at the bottom: Save, Print, Close.

Figure 9: MOVES Report part 1 - Units and Result database details

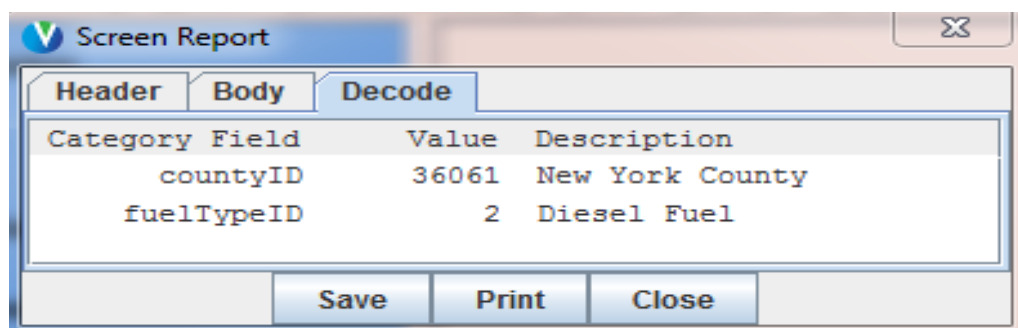


The screenshot shows the 'Screen Report' window with the 'Body' tab selected. The report displays emission data for a Diesel Passenger Car in New York county.

Month	County	Fuel	CO	NO2	Total_PM25	Brake_PM10	Brake_PM25	Tire_PM25	TotalEnergy
10	36061	2	2388710	13550	8832	35653	4457	1707	242228002816

Buttons at the bottom: Save, Print, Close.

Figure 10: Emissions calculated for Diesel Passenger Car - New York county



The screenshot shows the 'Screen Report' window with the 'Decode' tab selected. The report displays details of the county and fuel type used for emission calculation.

Category	Field	Value	Description
	countyID	36061	New York County
	fuelTypeID	2	Diesel Fuel

Buttons at the bottom: Save, Print, Close.

Figure 11: Details of county and fuel type used for emission calculation

References

2016. *Atmospheric Environment* 140.

Aslam, Javed, Sejoon Lim, Xinghao Pan, and Daniela Rus. 2012. “City-Scale Traffic Estimation from a Roving Sensor Network”. In *Proceedings of the 10th ACM Conference on Embedded Network Sensor Systems - SenSys 12*. ACM Press. doi:10.1145/2426656.2426671.

n.d. <https://www.its.leeds.ac.uk/software/dracula/>. <https://www.its.leeds.ac.uk/software/dracula/>.

n.d. <https://www.epa.gov/moves>. <https://www.epa.gov/moves>.