Transportation Data Management and Analysis (TDMA)

An Introduction

Southeast Florida FSUTMS Users Group

June 13, 2014

Manish Jain
• Overview

• Select Datasets

• Data Processing Tools

• Examples

• Q&A
Transportation Data – Multiple Dimensions

• Modes
  – Highway, Arterials, Freight, Transit, etc.

• Devices
  – Roadway sensors/detectors, Transit AVL and APC, Bluetooth, GPS, etc.

• Sources
  – Arterial and Freeway Management Systems, Transit Operators, DOTs, TMCs, Freight, Third Party Data (INRIX/HERE), etc.

• Data Types
  – Traffic speed/volume/occupancy, incidents, transit passenger ONs/OFFs, Freight tonnage, etc.

• Miscellaneous
  – Activity Based Models, Dynamic Traffic Assignment, Specialized Transportation Services, etc.
Transportation “Big-Data”

Adapted from ADMS presentation made by Dr. Genevieve Giuliano at USC
Select Datasets
Speed Dataset – INRIX

- Real-time speed data collected from nearly 100 million anonymous mobile phones, trucks, delivery vans, and other fleet vehicles equipped with GPS locator devices
- Temporal resolution: 1, 5, 15, 30 and 60 minutes
- Data elements: TMC, Speed, Travel Time, Reference Speed, Historical Speed, Confidence value
- Size: Current archive 19TB; growing @ 21GB/Day
- Geographic coverage: Freeways and arterials
- Traffic Message Channel (TMC) shape file for visualization

<table>
<thead>
<tr>
<th>TMC</th>
<th>Date</th>
<th>Time</th>
<th>Speed</th>
<th>ReferenceSpeed</th>
<th>Average Speed</th>
<th>Score</th>
<th>TravelTimeMinutes</th>
<th>C_Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>110+04131</td>
<td>6/22/2009</td>
<td>5:45</td>
<td>41</td>
<td>46</td>
<td>46</td>
<td>30</td>
<td>0.10</td>
<td>99</td>
</tr>
<tr>
<td>110+04131</td>
<td>6/22/2009</td>
<td>5:50</td>
<td>41</td>
<td>46</td>
<td>46</td>
<td>30</td>
<td>0.10</td>
<td>100</td>
</tr>
<tr>
<td>110+04131</td>
<td>6/22/2009</td>
<td>5:55</td>
<td>41</td>
<td>46</td>
<td>46</td>
<td>30</td>
<td>0.10</td>
<td>100</td>
</tr>
<tr>
<td>110+04131</td>
<td>6/22/2009</td>
<td>6:00</td>
<td>41</td>
<td>46</td>
<td>46</td>
<td>30</td>
<td>0.10</td>
<td>100</td>
</tr>
<tr>
<td>110+04131</td>
<td>6/22/2009</td>
<td>6:05</td>
<td>46</td>
<td>46</td>
<td>46</td>
<td>10</td>
<td>0.09</td>
<td></td>
</tr>
</tbody>
</table>
### Speed Dataset: INRIX - Florida

Collected on 33,700 Traffic Message Channel (TMC) links during the period 7/1/2010 to 6/30/2011

<table>
<thead>
<tr>
<th>District/State</th>
<th>Centerline Miles</th>
<th>Speed-Records (5-minute interval)</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1</td>
<td>3,333</td>
<td>105,875,469</td>
</tr>
<tr>
<td>D2</td>
<td>3,587</td>
<td>101,748,753</td>
</tr>
<tr>
<td>D3</td>
<td>1,767</td>
<td>40,137,130</td>
</tr>
<tr>
<td>D4</td>
<td>2,627</td>
<td>127,347,582</td>
</tr>
<tr>
<td>D5</td>
<td>3,077</td>
<td>142,516,885</td>
</tr>
<tr>
<td>D6</td>
<td>1,205</td>
<td>90,790,223</td>
</tr>
<tr>
<td>D7</td>
<td>2,414</td>
<td>112,953,993</td>
</tr>
<tr>
<td>Florida</td>
<td>18,010</td>
<td>711,351,697</td>
</tr>
</tbody>
</table>
Speed Dataset – NPMRDS / HERE

- National Performance Measure Research Data Set (NPMRDS), provided by HERE
- Similar probe data as INRIX but separates Cars and Trucks
- Temporal resolution: 5 minutes with no data imputations
- Data elements: TMC, Date, Epoch, Travel Time (All vehicles, passenger vehicles, freight trucks)
- Size: Monthly data files between 500MB and 4GB in size
- Geographic coverage: National Highway System
- TMC shape file for visualization
- Provided by FHWA to public agencies for free (HERE as vendor)

<table>
<thead>
<tr>
<th>TMC</th>
<th>DATE</th>
<th>EPOCH</th>
<th>Travel_TIME_ALL_VEHICLES</th>
<th>Travel_TIME_PASSENGER_VEHICLES</th>
<th>Travel_TIME_FREIGHT_TRUCKS</th>
</tr>
</thead>
<tbody>
<tr>
<td>115N04098</td>
<td>10052013</td>
<td>262</td>
<td>97</td>
<td>97</td>
<td></td>
</tr>
<tr>
<td>115N04098</td>
<td>10012013</td>
<td>34</td>
<td>101</td>
<td>101</td>
<td></td>
</tr>
<tr>
<td>115N04098</td>
<td>10012013</td>
<td>66</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>115N04098</td>
<td>10012013</td>
<td>98</td>
<td>112</td>
<td>111</td>
<td>111</td>
</tr>
</tbody>
</table>
VDOT’s Northern Region Operations (NRO) has 1,337 signalized intersections with 15,765 detectors. Also has 1,319 detectors on freeways and ramps.

- Volume, Occupancy and Speed data collected every 15-minutes
- Arterial data: 96 * 15,765 = 1,513,440 detector records/day
- Freeway data: 96 * 1,319 = 126,624 detector records / day
- Average data size per month: 2.4 GB

<table>
<thead>
<tr>
<th>detector_id</th>
<th>data_date_time</th>
<th>ignore</th>
<th>Volume</th>
<th>occupancy</th>
<th>speed</th>
<th>ignore</th>
<th>status</th>
<th>data_period</th>
<th>ignore</th>
</tr>
</thead>
<tbody>
<tr>
<td>300003511</td>
<td>10/7/2010 20:30</td>
<td>0</td>
<td>24</td>
<td>33</td>
<td>0</td>
<td>0</td>
<td>ONLINE</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>300003512</td>
<td>10/7/2010 20:30</td>
<td>0</td>
<td>24</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>ONLINE</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>300003513</td>
<td>10/7/2010 20:30</td>
<td>0</td>
<td>480</td>
<td>6</td>
<td>51</td>
<td>0</td>
<td>ONLINE</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>300003514</td>
<td>10/7/2010 20:30</td>
<td>0</td>
<td>608</td>
<td>9</td>
<td>44</td>
<td>0</td>
<td>ONLINE</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>300003515</td>
<td>10/7/2010 20:30</td>
<td>0</td>
<td>268</td>
<td>2</td>
<td>52</td>
<td>0</td>
<td>ONLINE</td>
<td>15</td>
<td>0</td>
</tr>
</tbody>
</table>
Count Dataset: VDOT ADMS

Detector Locations in Northern Virginia

Legend

- Detector Locations
- Centerline
Miscellaneous Datasets: ABM

- Activity Based Model (ABMs) output tours and trips at record level
  - SERPM: Over 9 million tours and 21 million trips (Base year)
  - FOCUS: Over 4 million tours and 11 million trips (Year 2010)
- Typically model outputs stored in series of tables
  - Tours, Trips, Half-tours, Household data, Person data, etc.
## SERPM ABM Database

<table>
<thead>
<tr>
<th>Table</th>
<th>Table Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synthetic Households</td>
<td>SYNHH</td>
<td>Input synthesized households from the PopSyn.</td>
</tr>
<tr>
<td>Synthetic Persons</td>
<td>SYNPERSO</td>
<td>Input synthesized persons from the PopSyn.</td>
</tr>
<tr>
<td>TAZ Data</td>
<td>TAZ</td>
<td>Input TAZ data.</td>
</tr>
<tr>
<td>MGRA Data</td>
<td>MGRA</td>
<td>Input MGRA (MAZ) data such as employment, etc.</td>
</tr>
<tr>
<td>TAP Data</td>
<td>TAP</td>
<td>Input TAP data.</td>
</tr>
<tr>
<td>MGRA to TAP Data</td>
<td>MGRATOTAP</td>
<td>MGRA to TAP distances, etc.</td>
</tr>
<tr>
<td>MGRA to Stop Data</td>
<td>MGRATOSTOP</td>
<td>MGRA to all transit stops distances, etc.</td>
</tr>
<tr>
<td>MGRA to MGRA Data</td>
<td>MGRATOMGRA</td>
<td>MGRA to MGRA distances, etc.</td>
</tr>
<tr>
<td>TAZ to TAP Data</td>
<td>TAZTOTAP</td>
<td>TAZ to TAP distances, etc.</td>
</tr>
<tr>
<td>Accessibilities</td>
<td>ACCESSIBILITIES</td>
<td>Model results for accessibilities.</td>
</tr>
<tr>
<td>Household Data</td>
<td>HHDATA</td>
<td>Model results for household level choice models.</td>
</tr>
<tr>
<td>Person Data</td>
<td>PERSONDATA</td>
<td>Model results for person level choice models.</td>
</tr>
<tr>
<td>Work and School Location</td>
<td>WSLOCATION</td>
<td>Model results for usual work and school location choice models.</td>
</tr>
<tr>
<td>Individual Tours</td>
<td>INDIVTOUR</td>
<td>Modeled individual tours.</td>
</tr>
<tr>
<td>Joint Tours</td>
<td>JOINTTOUR</td>
<td>Modeled joint tours.</td>
</tr>
<tr>
<td>Individual Trips</td>
<td>INDIVTRIP</td>
<td>Modeled individual trips.</td>
</tr>
<tr>
<td>Joint Trips</td>
<td>JOINTTRIP</td>
<td>Modeled joint trips.</td>
</tr>
<tr>
<td>CBD Vehicle Trips</td>
<td>CBDVEHICLES</td>
<td>Number of vehicle trips to CBD by time period.</td>
</tr>
<tr>
<td>All microsimulated trips</td>
<td>TRIP</td>
<td>All microsimulated trips.</td>
</tr>
<tr>
<td>District/County Definitions</td>
<td>DISTRICTDEFINITIONS</td>
<td>Mapping of TAZs to counties and districts.</td>
</tr>
<tr>
<td>PECAS Occupations</td>
<td>PECASCODES</td>
<td>Mapping of PECAS codes to occupations.</td>
</tr>
</tbody>
</table>
Miscellaneous Datasets (cont’d)

• Freight Flow Data
  – Transearch commodity flow data in series of tables. Requires relational database to query and summarize information
  – Surface Transportation Boards Way Bill Data in flat ASCII files

• Transit
  – APC data at each stop for each transit trip. Includes ONs, OFFs, Arrival Time, Departure Time, etc.
  – AVL data
  – Fare gate to fare gate data for each passenger
  – Specialized Transportation Service Data: Includes passenger information, trip purpose, boarding location, alighting location, trip time, fare, etc.

All these datasets can become extremely large over time
Data Processing and Examples
• Hardware and software requirements:
  – Excel can only handle up to a million records
  – Access has 2GB per table limit; will be quickly exceeded
  – Requires database and programming/scripting skills
  – GIS expertise required for mapping
    • TMCs in INRIX/HERE data have many to many relationship with shape file links
      – One LINK can reference many TMCs
      – One TMC can reference many TMCs

• Open Source tools and databases
  – MySQL, PostgreSQL, Python, R, TRANSIMS SysLib, etc.

• Commercial tools and databases
  – Microsoft SQL, Oracle, SPSS/SAS, C++, FORTRAN, Matlab, etc.
• Integrating INRIX/HERE data with GIS will require resolving relationships between TMCs and LINKS. Displaying data by direction could be challenging too!
  – Recommend managing the spatial data in a relational database system
  – Downside is that the spatial table would become huge

• Programming languages such as Matlab, R, and Python extensions provide powerful graphic capabilities
Data Filtering – An Example

Illustration of how data filtering and statistical analysis can be used to improve speed estimates:

- Data from detector station 4 on I-80 westbound at Berkeley Highway Laboratory (BHL), California
- Six step approach for data-filtering:
  - Step 1: Identify erroneous samples
  - Step 2: Raw estimate of speed
  - Step 3: Speed-Flow filter
  - Step 4: Speed-Occupancy filter
  - Step 5: Speed filter
  - Step 6: Moving median of three samples
- Used MATLAB for analysis

Plot showing progression of improvement in speed estimates. The plot A shows the estimates at the end of Step 1, B shows the estimates at the end of Step 2, C shows the estimates at the end of Step 3, D shows the estimates at end of Step 4, E at the end of Step 5 and F at the end of Step 6.

Reference: Jain, M and Coifman, B, “Improved Speed Estimates from Freeway Traffic Detectors”, the Ohio State University.
Freeway Sensor Data – Example 1

- Data extraction and management in Python and SQLite3
- 800 sensors with about 700 operational at any point in time
- ~28 million records per year
- Plot shows data for:
  - Two month weekday (Mon-Fri) 15 minute averaged dataset
  - Single sensor on I-55, Chicago
  - Speed in meters/second

Courtesy: Dr. Hubert Ley, Argonne National Laboratory
Count Data Summary – Example 2

- Data management and extraction using TRANSIMS SysLib CountSum program
- Customized filtering based on days / facility-types / detector-types / signal-types / station-types
- Outputs data in a variety of file-formats (dbase, binary, sqlite3, csv, etc.)
- Automatic or user guided tagging of detectors to links

VDOT NRO Arterial Data; 1,513,440 detector records/day; Multiple Days

<table>
<thead>
<tr>
<th>detector_id</th>
<th>data_date_time</th>
<th>ignore</th>
<th>Volume</th>
<th>occupancy</th>
<th>speed</th>
<th>ignore</th>
<th>status</th>
<th>data_period</th>
<th>ignore</th>
</tr>
</thead>
<tbody>
<tr>
<td>300003511</td>
<td>10/7/2010 20:30</td>
<td>0</td>
<td>24</td>
<td>33</td>
<td>0</td>
<td>0</td>
<td>ONLINE</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>300003512</td>
<td>10/7/2010 20:30</td>
<td>0</td>
<td>24</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>ONLINE</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>300003513</td>
<td>10/7/2010 20:30</td>
<td>0</td>
<td>480</td>
<td>6</td>
<td>51</td>
<td>0</td>
<td>ONLINE</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>300003514</td>
<td>10/7/2010 20:30</td>
<td>0</td>
<td>608</td>
<td>9</td>
<td>44</td>
<td>0</td>
<td>ONLINE</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>300003515</td>
<td>10/7/2010 20:30</td>
<td>0</td>
<td>268</td>
<td>2</td>
<td>52</td>
<td>0</td>
<td>ONLINE</td>
<td>15</td>
<td>0</td>
</tr>
</tbody>
</table>
ABM Data Extraction – Example 3

- Example SQL Query to extract transit tours by Income, Purpose, Mode of Access, Time of Day, and Attraction District
- Queries can be designed via a wizard and does not require extensive knowledge of SQL syntax.
- SERPM ABM provides some standard reporting queries; additional custom queries can be developed using T-SQL.

Example SQL Query:
```
FROM dbo.Households INNER JOIN
dbo.Persons ON dbo.Households.HHID = dbo.Persons.HHID INNER JOIN
dbo.Tours ON dbo.Tours.PersonID = dbo.Tours.PersonID INNER JOIN
dbo.TourType ON dbo.Tours.TourTypeID = dbo.TourType.TourTypeID INNER JOIN
dbo.Purpose ON dbo.Tours.TourPurposeID = dbo.Purpose.PurposeID CROSS JOIN
IRM.dbo.Tours.CrossAlt.ArrivalTimeTransitPeriod CROSS JOIN
dbo.Zones
WHERE (dbo.Tours.TourModeChoice IN ('Walk to Transit', 'Drive to Transit'))
```
INRIX Data Coverage – Example 4

- 5-Minute INRIX Speed Data for D4 processed using Cube scripts to extract data for Oakland Park Blvd TMCs.
- Resultant data analyzed in Excel

<table>
<thead>
<tr>
<th>TMC</th>
<th>SPEED</th>
<th>DATE</th>
<th>TIME</th>
</tr>
</thead>
<tbody>
<tr>
<td>102+11717</td>
<td>36.00</td>
<td>10/05/2010</td>
<td>12:55</td>
</tr>
<tr>
<td>102+11717</td>
<td>26.00</td>
<td>10/05/2010</td>
<td>16:25</td>
</tr>
<tr>
<td>102+11717</td>
<td>26.00</td>
<td>10/05/2010</td>
<td>16:30</td>
</tr>
<tr>
<td>102+11717</td>
<td>41.00</td>
<td>10/05/2010</td>
<td>16:35</td>
</tr>
<tr>
<td>102+11717</td>
<td>39.00</td>
<td>10/05/2010</td>
<td>16:40</td>
</tr>
<tr>
<td>102+11717</td>
<td>33.00</td>
<td>10/05/2010</td>
<td>16:45</td>
</tr>
<tr>
<td>102+11717</td>
<td>35.00</td>
<td>10/06/2010</td>
<td>09:20</td>
</tr>
<tr>
<td>102+11717</td>
<td>39.00</td>
<td>10/06/2010</td>
<td>13:20</td>
</tr>
<tr>
<td>102+11717</td>
<td>27.00</td>
<td>10/06/2010</td>
<td>18:40</td>
</tr>
<tr>
<td>102+11717</td>
<td>29.00</td>
<td>10/06/2010</td>
<td>18:45</td>
</tr>
<tr>
<td>102+11717</td>
<td>29.00</td>
<td>10/06/2010</td>
<td>18:50</td>
</tr>
</tbody>
</table>
Brian Gregor and Ben Stabler at Oregon DOT used R script to automate census data extraction:

- Script downloads Census 2000 county to county commuting data for any number of states
- Creates and saves the data on instate commutes
- Creates and saves an origin-destination matrix of the instate commutes
- Summarizes the instate commutes and saves the results

```r
# Define function for retrieving a commuting file from the Census
get.commute <- function(state = "OR", type = "RES") {
do
  # Check for valid state and type codes
  state.abb <- c("AK", "AL", "AR", "AZ", "CA", "CO", "CT", "DC", "DE", "FL",
                  "GA", "HI", "IA", "ID", "IL", "IN", "KS", "KY", "LA", "MA", "MD",
                  "ME", "MI", "MN", "MS", "MT", "NE", "NH", "NJ", "NM", "NV",
                  "NY", "OH", "OK", "OR", "PA", "RI", "SC", "SD", "TN", "TX",
                  "UT", "VA", "VT", "WA", "WI", "WV", "WY")
  if(!(state %in% state.abb)) stop("Must use valid state abbreviation")
  if(!(type %in% types)) stop("Type must be RES or WRK")

do
  # Make the file name and the url to get the data from
  commute.file <- paste("http://www.census.gov/population/cen2000/commuting/",
                        commute.file, sep = "")

do
  # Connect to the Census 2000 county to county commute file by residence or work place
  http://www.census.gov/population/cen2000/commuting/coxcolayout.txt

do
  # Read downloaded file into a data.frame
  specify field widths (varies from Census documentation because a space
  is located between each field
  census.width <- c(2, 4, 5, 5, 42, 3, 4, 5, 42, 7)

do
  # set field names
  census.name <- c("res.state", "res.county", "res.msa", "res.pmsa",
                   "res.name", "wrk.state", "wrk.county", "wrk.msa", "wrk.pmsa",
                   "wrk.name", "count")

do
  # Summarize OD matrix in several ways
  commute.origins <- rowSums(commute.od)
do
  # compute computes by origin county
  commute.destinations <- colSums(commute.od)
do
  # compute commutes by destination county
  internal <- commute.od[row(commute.od) == col(commute.od)]
do
  # internal commutes are on the diagonal of of the od matrix
  internal.pct <- 100 * internal / commute.origins
do
  # compute the internal percentage
  internal.pct <- round(internal.pct, 1)
do
  # round the internal percentage to the first decimal place
  outflow <- commute.origins - internal
do
  # compute the number of commuters leaving the county
  inflow <- commute.destinations - internal
do
  # compute the number of commuters entering the county
```

Average Number of Trips Produced by District

![Chart showing average number of trips produced by district]
Summary

• Transportation Data archived from operations presents a tremendous opportunity for planning activities.

• Traditional data processing tools used by planners not capable of managing and analyzing “Big-Data”.

• Various data management and analysis tools exist and can be customized to agency needs.

• Three key phases for leveraging “Big-Data”
  i. Data Management
  ii. Data Analysis
  iii. Data Application

• Analysis and Application are next steps.
Questions?

Thank you!

Manish.Jain@AECOM.com
+1 703.340.3049

Courtesy: http://dilbert.com/strips/