Understanding and Communicating Multimodal Transportation Data

Development, Deployment, and Assessment of a New Educational Paradigm for Transportation Professionals and University Students

A Collaboration of the Region X Transportation Consortium

Portland Postgres Users Group
pdxPUG
July 21, 2011

Christopher Monsere, Assistant Professor, Portland State University
Portland State University
Outline

• Course overview
• Connecting R to Postgres
• Sample activities and plots
• Favorite visualizations
• Closing thoughts
Motivations

• Many transportation students...
  – have limited exposure to large-scale, messy, empirical data sets
  – lack basic programming skills
  – missing \textit{basic} statistical concepts
  – have few graphical analyses tools
Course Design Process

1. Audience
2. Vision
3. Learning Outcomes
4. Knowledge Table
5. Learning Skills
6. Assessment
7. Content
Course Vision

• Preparing future transportation professionals to interact, explore, analyze, and explain multimodal transportation data. Students who complete this course will be comfortable with data.
Key Building Blocks

• Open-source software
  – R
  – PostgreSQL

• Reference books
  – Keen, Kevin. Graphics for Statistics and Data Analysis with R
  – Scientific Approaches to Transportation Research Volumes 1 and 2
Course Details

• Delivery Format
  – 10 weeks
  – Two 90 min class sessions per week (30 contact hours)
  – Mostly lab sessions, some lectures
  – 40 activities developed

• Development History
  – Started as 2 credit course (pre-grant)
  – Test 2 credit course (year 1)
  – 4 credit course (this past fall)
  – One final revision for offering Fall 2011
### Data Schemas

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
<th>Records</th>
</tr>
</thead>
<tbody>
<tr>
<td>bicycle</td>
<td>Performance of bicyclists with demographic variables extracted from video taken in Portland, OR</td>
<td><strong>422 records</strong></td>
</tr>
<tr>
<td>loop</td>
<td>ITS freeway data for OR-217 SB at 5-minute resolution (2009).</td>
<td><strong>3,928,120 records</strong></td>
</tr>
<tr>
<td>bluetooth</td>
<td>Blurred MAC id reading from 3 sensors on Powell Blvd (May 10-17, 2010)</td>
<td><strong>4,694,500 records</strong></td>
</tr>
<tr>
<td>incidents</td>
<td>Incidents on OR-217 freeway from CAD (2009).</td>
<td><strong>524 records</strong></td>
</tr>
<tr>
<td>trimet</td>
<td>Stop-level records from Route 19 from BDS data (2009).</td>
<td><strong>1,820,590 records</strong></td>
</tr>
<tr>
<td>wim</td>
<td>Data from 22 weight-in-motion stations in Oregon for 1 month (August 2009).</td>
<td><strong>1,380,000 records</strong></td>
</tr>
</tbody>
</table>
Modules

1. Getting Started With Data
2. Introduction to R
3. Principles of Graphics
4. Basic Statistics Review
5. Using Graphics for Exploratory Data Analysis
6. Putting it All Together: Data Exploration for Understanding
Modules 1-4

1. Getting Started with Data
   – Introduction to relational databases
   – SQL tutorials
   – Excel strawman

2. Introduction to R
   – Steep learning curve
   – Narrated tutorials
   – Plotting
   – Loops and functions

3. Basic Statistics Review
   – Descriptive
   – Random variables
   – Hypothesis testing

4. Principles of Graphics
   – Critiques
   – Brainstorming
   – Readings about visualizations (Tufte)
5. Exploring Data With Graphics

• Single Discrete Variable
  – Dot chart, bar chart, pie chart, stacked bar

• Single Continuous
  – Dotplot, stemplot, boxplot, edf plot, kernel density estimates, histograms, qq plot

• Two Discrete
  – Grouped dot, side-by-side bar, stacked bar, mosiac plots

• Two Continuous
  – Scatter plot, Sunflower plot

• Multivariate
  – Scatterplot with 3D in color, Bubbleplots, contour plots
CONNECTING R TO POSTGRES
Why R?

• Open source
  – Students can take the software with them

• Multi-platform

• Strong community of users

• Flexible, script-based analysis (reproducible)

• Strong statistical and data visualization

http://www.r-project.org
Setup

PACKAGES

R Studio

OUTPUT

PostgreSQL
R Studio

http://rstudio.org/
Connecting R to Postgres

• Postgres:
  – ODBC driver

• R:
  – RODBC package

• Note: I am using Windows 64 bit environment. For other platforms, also DBI package ‘RPostgreSQL’
  – cran.r-project.org/web/packages/RPostgreSQL/
Config Postgres ODBC Driver

**Datasource**= A name you choose for the driver. Keep it simple, no spaces. This is what you call in R.
**Database**= ce510
**Server**= db.cecs.pdx.edu
**User name**= xxxxxx
**Description** = optional
**SSL mode** = allow
**Port**= 5432
**Password** = xxxxxx

http://www.postgresql.org/ftp/odbc/versions/msi/
Packages

• R’s “add-ins”
• Simple process to download, install
• Load in current R session with `library(packagename)`
R Syntax for RODBC

• Download and install RODBC package

```r
library (RODBC)
channel <- odbcConnect("Datasource", uid="Username")
qry <- "SELECT * FROM trimet.route19 WHERE service_day >= '03-04-2007' AND service_day < '03-10-2007'"

#assign query results to data frame trimet
trimet <- sqlQuery(channel, qry)
```
SAMPLE ACTIVITIES FROM COURSE
Sample Activity: Quick Exploration of Data

- Show students how to quickly explore data for outlier checking, missing data, other issues
- January 2009 has 330,864 records
library (RODBC)
channel <- odbcConnect("ce510", uid="xxxx")

# Select the data from the freeway loop data.
qry <- "SELECT * FROM loop.loopdata_5min_217sb_2009 WHERE starttime < '2009-02-01' AND speed>=0 AND occupancy>=0 AND volume >=0"

l <- sqlQuery(channel, qry)  # assign query results to data frame l

# XY Lattice Plot
library(lattice)
xyplot ( l$speed ~ l$volume | as.factor(l$detectorid), xlab="Speed, mph", ylab="Volume, veh/5 min")
Sample Activity: Multivariate Exploration

• Add a third dimension to a plot with color
• Introduce algorithms to devise appropriate breaks and color ramps
• Combine with previous skills
# Subset the data frame l to be just one detector

```r
d <- subset(l, detectorid==1594)
```

# Create vector of color by intervals

```r
library(RColorBrewer)
library(classInt)
plotvar <- d$occupancy
nclr <- 9
plotclr <- brewer.pal(nclr,"PuRd")
brks <- classIntervals(plotvar, n=nclr-1, style="kmeans")
colornum <- findInterval(plotvar, brks$brks, all.inside=T)
colcode <- plotclr[colornum]
```

# Plot the data

```r
plot (d$volume, d$speed, col=colcode, pch=16, xlab="Volume (veh/5min)", ylab="Speed (mph)", main="Speed Flow Curve, Occupancy as Color\nDetector 1594, January 2009")
legend ("topright", pch=16, col=plotclr, legend=round(brks$brks,2) )
```
Sample Activity: Analysis of Means

• Introduce boxplots in detail
• Have students explore effects across groups
• Reviewed hypothesis testing
• Reinforce with graphics
Sample Activity: Distribution Fitting

- Review of random variables
- Introduce kernel density estimates, empirical cdf, q-q plots
- Use R to synthesize data with mean standard dev.
- Compare synthesized data to theoretical distributions with graphical tools
- Explore the effect of sample size
- Apply to “real” data set
Data Quality Activity

• Students given short introduction to WIM systems and data quality
• Introduced to kernel density estimates
• Required to find poor performing station

KDE of Axle 1 Wt
Class 9, by Station, 8-09

KDE of Tandem Drive Axles Spacing
Class 9, by Station, 8-09
Average reported speeds excluding holidays

Each ring is a day of the week, Monday to Sunday
Multnomah cyclist crashes
Multnomah pedestrian crashes
## Collision Type by Functional Class
### Fatal and Injury A Crashes

<table>
<thead>
<tr>
<th>ANGL BACK</th>
<th>FIX</th>
<th>HEAD</th>
<th>MCOL</th>
<th>OTH PARK</th>
<th>PED</th>
<th>REAR</th>
<th>SSM</th>
<th>SS-O</th>
<th>TURN</th>
</tr>
</thead>
<tbody>
<tr>
<td>R INTRST</td>
<td>R LOCAL</td>
<td>R M-JRT</td>
<td>R MAJOR</td>
<td>R MIN-COL</td>
<td>R PR-ART</td>
<td>U COLL</td>
<td>U FYDP</td>
<td>U INTRST</td>
<td>U LOCAL</td>
</tr>
</tbody>
</table>
A. Moore, M. Figliozzi, C. Monsere. Bus Stop Air Quality: An Empirical Analysis of Transit Users’ Exposure to Particulate Matter
Axle Weight Sensors

• Single load cells
• Sensors weigh vehicles traveling at normal highway speeds
• Weight measurement affected by many factors
  – Site characteristics
  – Environmental factors
  – Truck dynamics
Class 9 Steer Axle Weights

<table>
<thead>
<tr>
<th>Lane 1</th>
<th>Lane 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean = 9.75</td>
<td>Mean = 9.07</td>
</tr>
</tbody>
</table>

Expect 9,000-11,000 lbs
Class 9 Steer Axle Weight (March)

Lane 1
Mean = 10.44

Lane 2
Mean = 7.4

Steering Axle (kips)
Density

Steering Axle (kips)
Density

Lane 1
Lane 2

Steering Axle (kips)

Mar 03 Mar 08 Mar 13 Mar 18 Mar 23 Mar 28 Apr 02
CLOSING THOUGHTS
Fussy R

• R is known for steep learning curve
  – Prepare to invest some time

• Endless packages great, but can also be daunting/frustrating to find the right one
  – Good list of popular packages at:
  – [http://unknownr.r-forge.r-project.org/toppkgs.html](http://unknownr.r-forge.r-project.org/toppkgs.html)

• Rough visualizations easy, polished graphs take effort
Managing Large Data Sets

• R can sometimes be slow with large data sets unless you are careful.
  – SQL statements before R or process the entire data frame in R?

• R can be a memory “hog”

• Fastest code takes advantage of R’s vector approach.
Resources

• http://www.r-bloggers.com/
• http://www.statmethods.net/
• http://www.stat.auckland.ac.nz/~paul/RGraph
cics/rgraphics.html
• http://web.cecs.pdx.edu/~monserec/courses/i
ntrosch/index.htm
Acknowledgments

• R.L. Bertini - ITS Lab and PORTAL founder
• Colleagues –
  • Miguel Figliozzi, Kristin Tufte, Ashley Haire, Portland State University
  • Peter Koonce, City of Portland
• Sponsors -
  • National Science Foundation
  • Oregon Department of Transportation
  • Federal Highway Administration
  • TransPort ITS Coordinating Committee
  • City of Portland, Office of Transportation
  • TriMet
  • Oregon Engineering and Technology Industry Council
  • OTREC
  • RTC (Southwest Washington Regional Transportation
• Students
Questions?

Thank You!
www.its.pdx.edu