Big Data, Big Promise, Big Challenge: Can Small Area Estimation Play a Role in the Big Data Centric World?

Partha Lahiri

JPSM and Department of Mathematics, University of Maryland, College Park

plahiri@umd.edu

University of Pisa Seminar

November 27, 2017

What is Big Data?

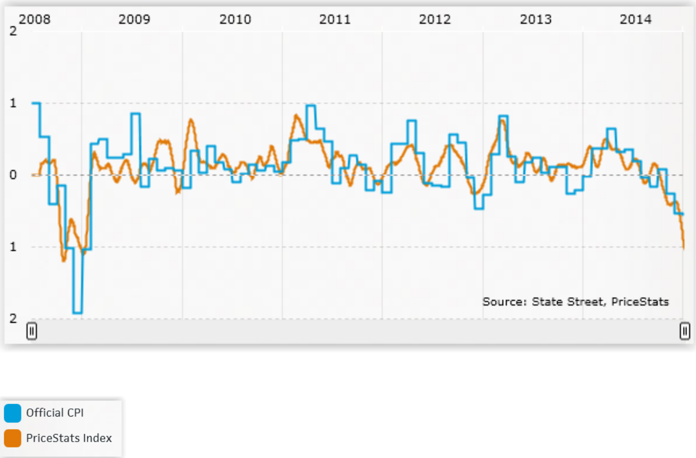
Characteristics of Big Data

- **Volume:** the sheer amount of data available for analysis.
- **Velocity:** the speed at which these data collection events can occur and the pressure of managing large streams of real-time data.
- **Variety:** complexity of formats in which Big Data can exist.
- **Variability**: inconsistency of the data across time,
- **Veracity**: ability to trust the data is accurate
- **Complexity**: need to link multiple data sources
- **Found/Organic Data**: not being initially made through the intervention of some researcher.
- **Confidentiality Concerns**

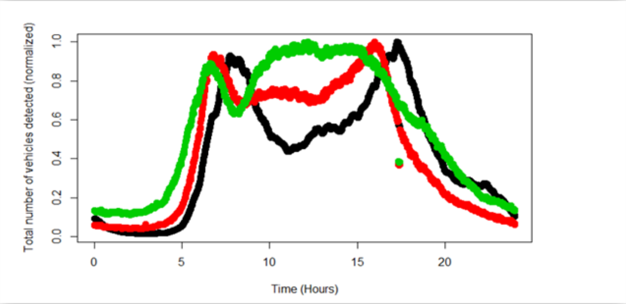
Different Types of Big Data Sources

- Social media data
- Personal data (e.g. data from tracking devices)
- Sensor data
- Transactional data
- Administrative data

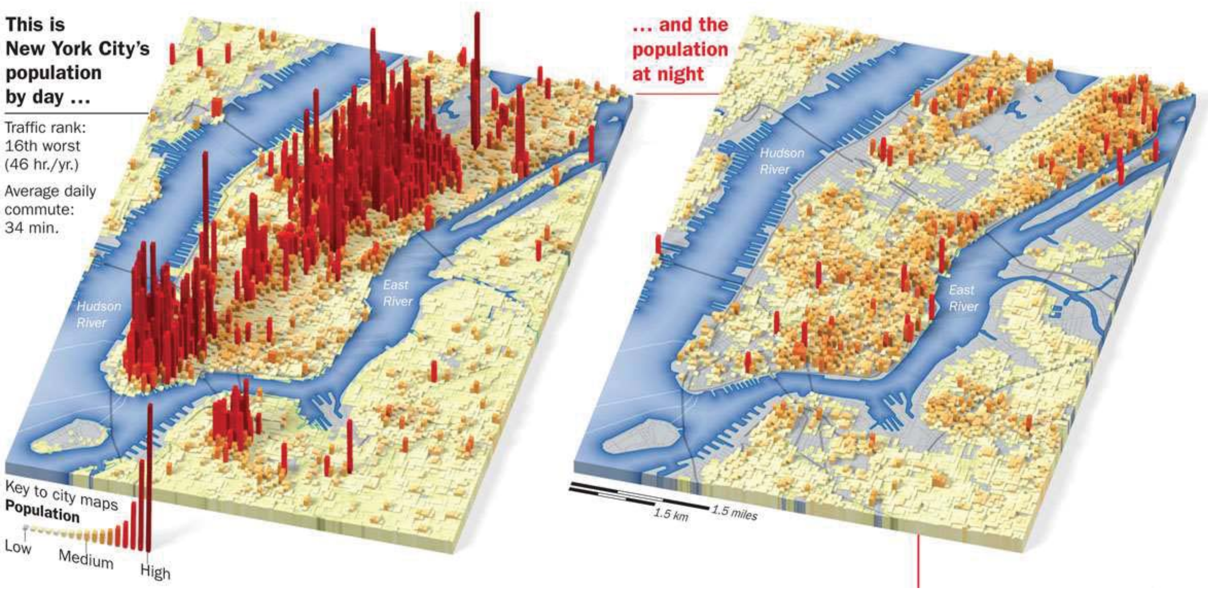
Example 1: Online Prices (AAPOR Report)



Example 2: Traffic and Infrastructure (AAPOR Report)



Location data from mobile phones



Source: Pfeffermann (2017)

A Few Points to Remember

- May not contain the variable(s) of interest
- Missing-data
- Errors due to measurement, classification, self selection, etc.
- Massive complex data for local area
- Computational issue

Three Examples of Local Area Statistics

- Estimation of income and poverty statistics for the administration of federal programs and the allocation of federal funds to local jurisdictions.
- Estimation of crop acreage, crop production, crop yield for the purpose of local agricultural decision making, payments to farmers if crop yields are below certain levels.
- Estimation of transportation related variables such as purpose of the trip (work, shopping, social, etc.), means of transportation (car, walk, bus, subway, etc.), travel time of trip to assist transportation planners and policy makers who need comprehensive data on travel and transportation patterns.

Problem 1: BIGDATA from Administrative Records

- Internal Revenue Service Data
- Supplemental Nutrition Assistance Program (SNAP) data

Problem 2: Remote Sensing BIGDATA

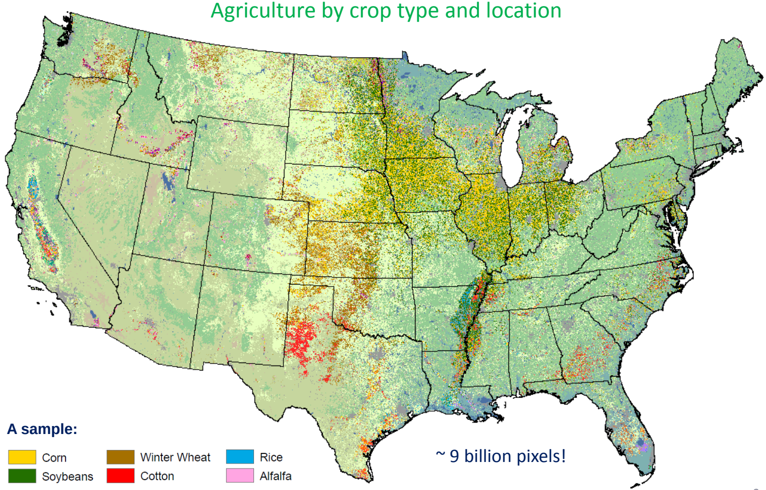
- Can earth resources satellite data provide useful ancillary data source for county estimates of crop acreage?
- Satellite information is recorded for *pixels* (a term for *picture elements*). A pixel is about .45 hectares;
- Based on satellite readings in early Fall, it is possible to classify the crop cover all pixels. This generates big data.

A Quote from Bellow et al.

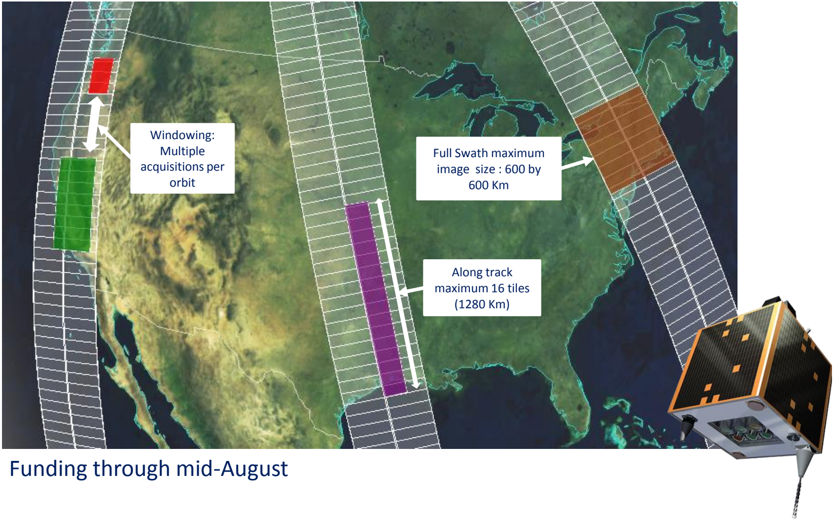
The polar-orbiting Landsat satellites contain a multi-spectral scanner (MSS) that measures reflected energy in four bands of the electromagnetic spectrum for an area of just under one acre. The spectral bands were selected to be responsive to vegetation characteristics. In addition to the MSS sensor, Landsats IV and V have a Thematic Mapper (TM) sensor which measures seven energy bands and has increased spatial resolution. The large area (185 by 170 km) and repeat (16 day per satellite) coverage of these satellites opened new areas of remote sensing research: large area crop inventories, crop yields, land cover mapping, area frame stratification, and small area crop cover estimation.

Cropland Data Layer

Agriculture by crop type and location



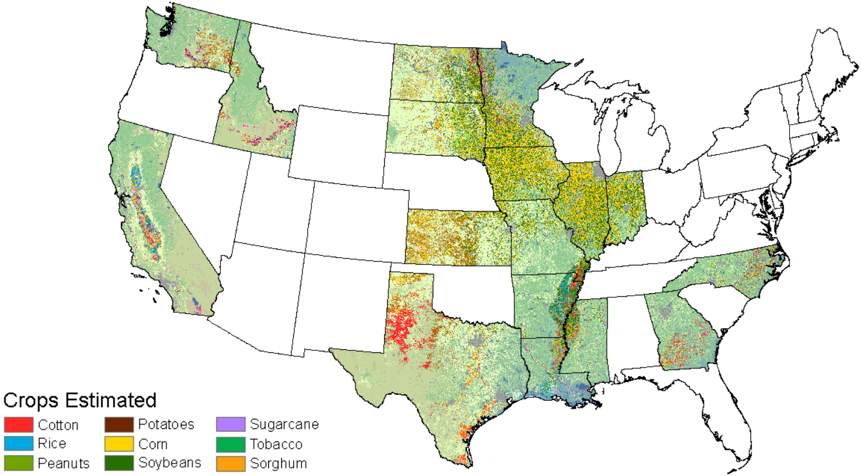
2014 Deimos-1/UK2 Satellite Tasking



Funding through mid-August

September

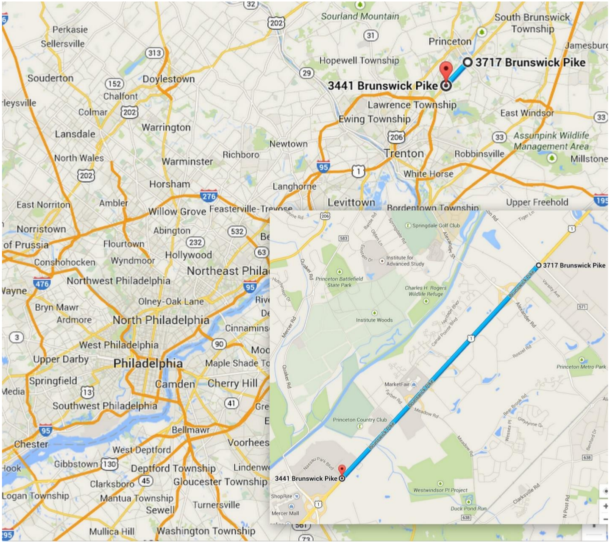
17 States Classified
9 Crops Estimated
Imagery from April - August



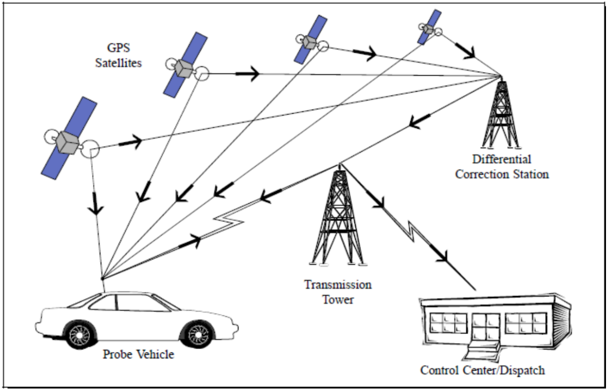
Problem 3: Vehicle Probe Project (VPP) BIGDATA

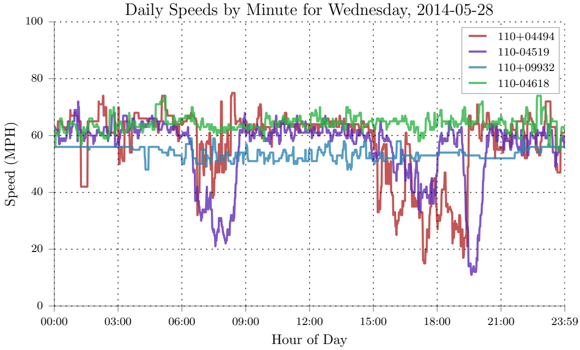
- VPP data was first contracted in July 2008.
- Contractually, the vendors are required to report data at one minute intervals for VPP.
- Archived in the VPP Suite maintained by CATT Laboratory at UMD.
- Currently, the VPP contractually reports traffic conditions on over 7,000 miles of freeways and 32,000 miles of arterials.
- Original goal: to enable a wide-variety of transportation operations and planning applications that require a high-quality data source.
- Data contains travel time, speed, historic speed, etc. for different road segments called Traffic Message Channels (TMC).
- Applications include congestion management systems, traveler information systems, travel-time on changeable message signs.
- If data for a whole year, for all 12,295 TMC segments in Maryland were to be downloaded, the estimated number of records is 6.46 billion. The physical disk size of this data is estimated to be 375GB.

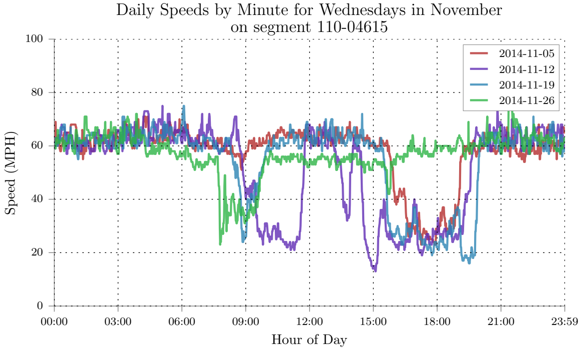
FIGURE: Location of NJ11-0009 segment in New Jersey, near Philadelphia.



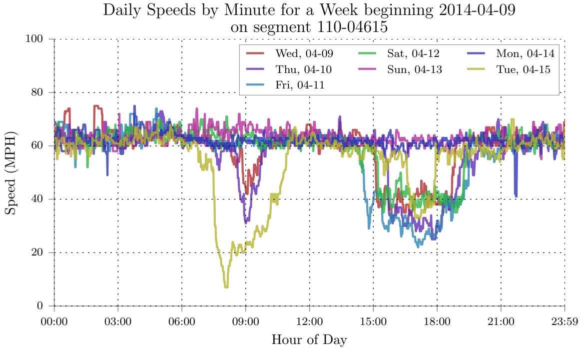
Communication from GPS (FHWA, 1998) [Ref: Kartika, C.S.D (2015)




Daily Speeds by Minute for Wednesday, 2014-05-28
110+04494
110-04519
110+09932
110-04618
Speed (MPH)
100
80
60
40
20
0
00:00
03:00
06:00
09:00
12:00
15:00
18:00
21:00
23:59
Hour of Day



Daily Speeds by Minute for Wednesdays in November
on segment 110-04615
Speed (MPH)
100
80
60
40
20
0
00:00
03:00
06:00
09:00
12:00
15:00
18:00
21:00
23:59
Hour of Day
2014-11-05
2014-11-12
2014-11-19
2014-11-26



Daily Speeds by Minute for a Week beginning 2014-04-09
on segment 110-04615
Wed, 04-09
Thu, 04-10
Fri, 04-11
Sat, 04-12
Sun, 04-13
Mon, 04-14
Tue, 04-15
Speed (MPH)
100
80
60
40
20
0
00:00
03:00
06:00
09:00
12:00
15:00
18:00
21:00
23:59
Hour of Day

Table 3: County-wise Number of TMC Segments

County	Number of TMC Segments
ALLEGANY	114
ANNE ARUNDEL	1,128
BALTIMORE	3,666
BALTIMORE CITY	8
BALTIMORE COUNTY	64
CALVERT	52
CAROLINE	120
CARROLL	305
CECIL	299
CHARLES	263
DORCHESTER	78
FREDERICK	617
GARRETT	86
HARFORD	491
HOWARD	634
KENT	22
MONTGOMERY	1,905
PRINCE GEORGE'S	1,694
QUEEN ANNE'S	148
SOMERSET	30
ST. MARY'S	66
TALBOT	30
WASHINGTON	261
WICOMICO	107
WORCESTER	107
Total	**12,295**

How do we correct Big Data?

Look for existing sample survey data or conduct a new survey

Some features of sample surveys

- Finite populations
- Representativeness
- Large samples for large areas, but small or no sample for small areas
- Variable(s) of interest can be included
- Chance selection: equal/epsem
- Stratification to improve precision and administrative control

Ref: Cochran (1977); Kalton (1983); Lohr (2010)

Sample Survey Data

- **Problem 1:** ACS
- **Problem 2:** June Enumerative Survey
- **Problem 3:** National Household Travel Survey (NHTS) and American Community Survey (ACS)

How do we combine Big Data with Sample Survey Data?

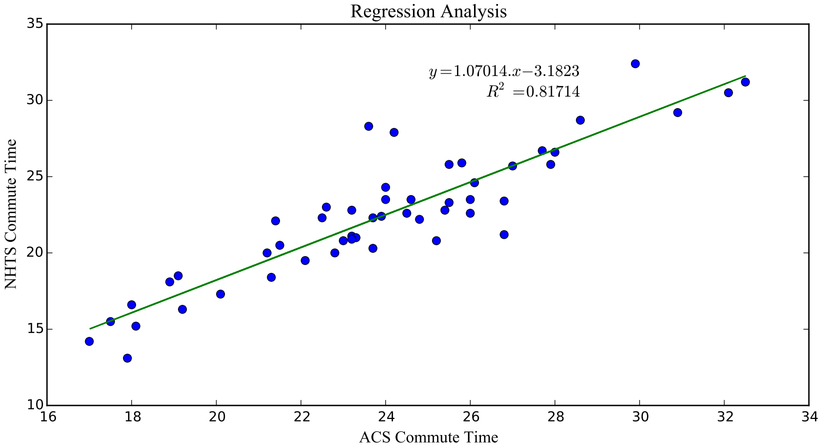
Data Fusion

- **Sample Survey Data**
  - National Household Travel Survey (NHTS)
  - American Community Survey (ACS)

- **Aggregated Administrative Data**
  - Supplemental Nutrition Assistance Program (SNAP) data (county level)
  - Internal Revenue Service Aggregate data (state level)

- **BIGDATA**
  - Vehicle Probe Project (VPP)
  - National Performance Management Research Data Set (NPMRDS)

A Proof of Data Fusion Concept



Two Cases

- **Case 1:** No or little overlap between the two data sources
- **Case 2:** Most of the survey data can be linked with Big Data

Case 1: Statistical Matching

Small Area Level Model

Ref: Fay and Herriot (JASA 1979)

For $i = 1, \cdots, m$,

$$\textit{Level 1: (Sampling Distribution)}: \; y_i|\theta_i \sim N(\theta_i, \psi_i);$$
$$\textit{Level 2: (Prior Distribution)}: \; \theta_i \sim N(\boldsymbol{x}_i'\boldsymbol{\beta}, A)$$

where

- m : number of small area;
- y_i : direct survey estimate of θ_i;
- θ_i : true mean for area i;
- $\boldsymbol{x}_i$: $p \times 1$ vector of known auxiliary variables;
- ψ_i: known sampling variance of the direct estimate;
- The $p \times 1$ vector of regression coefficients $\boldsymbol{\beta}$ and model variance A are unknown.

Estimation Method

Parameter of Interest: θ_i

Inferences based on the posterior distribution of θ_i:

$$\theta_i | y; \beta, A \overset{ind}{\sim} N(\hat{\theta}_i^B,\ \sigma_i^2(A)),$$

where

- $\hat{\theta}_i^B = (1 - B_i) y_i + B_i \boldsymbol{x}_i' \boldsymbol{\beta}$
- $B_i = \frac{\psi_i}{A + \psi_i}$
- $\sigma_i^2(A) = (1 - B_i)\psi_i$

EB: Treat β and A fixed and estimate them by consistent estimators (e.g., ANOVA, ML, REML, adjusted ML)

HB: Put priors, possible non-informative flat priors, on β and A. The inference is based on the posterior distribution of the target parameter.

The James-Stein Estimator

$$\hat{\theta}_i^{JS} = (1 - \hat{B}_{JS})y_i, \text{ where } \hat{B}_{JS} = \frac{m-2}{\sum_{j=1}^m y_j^2}.$$

Results:

- Total MSE (TMSE) of direct estimator: $\sum_{j=1}^m E[(y_i - \theta_i)^2|\theta] = m$
- TMSE of JS estimator: $\sum_{j=1}^m E[(\hat{\theta}_i^{JS} - \theta_i)^2|\theta] \le m - \frac{(m-2)^2}{m-2+\sum_i \theta_i^2}$. (Efron)

Remarks:

- If $\theta_i = 0, \ (i = 1, \cdots, m)$, then TMSE of JS$\le [m - (m-2)] = 2$. Thus, the largest reduction is obtained when $\theta_i = 0 \ (i = 1, \cdots, m)$ and m large.
- If any $|y_j| \to \infty$, the JS converges to the direct.

Measurement Error Issue in Big Data

Two Situations:

- **Situation 1:** The sources of measurement error can be reasonably identified and we have enough data to explain them.
- **Situation 2:** The sources cannot be easily detected or we do not have data to explain the measurement error even if the sources of error are identified.

Situation 1: An Example

$$\text{Level 1 (Sampling model): } \begin{pmatrix} y_i \\ \boldsymbol{x}_i \end{pmatrix} | \theta_i, \boldsymbol{X}_i \overset{ind}{\sim} N\left(\begin{pmatrix} \theta_i \\ \boldsymbol{X}_i \end{pmatrix}, \left(\begin{array}{c|c} \psi_{iy} & 0 \\ \hline 0 & \boldsymbol{\Psi}_{ix} \end{array} \right) \right)$$

$$\text{Level 2 (Linking model): } \quad \theta_i | \boldsymbol{X}_i \overset{ind}{\sim} N(\boldsymbol{X}_i' \boldsymbol{\beta}, A)$$

Remark: The above model reduces to the FH model when $\Psi = \mathbf{0}$.

The Bayes estimator of θ_i under FH:

$$\hat{\theta}_i^B = (1 - B_i) y_i + B_i \boldsymbol{x}_i' \boldsymbol{\beta},$$

where

$$B_i = \frac{\psi_{iy}}{A + \psi_{iy}}$$

The Bayes estimator of θ_i under FH with ME:

$$\hat{\theta}_i^{B*} = (1 - B_i^*) y_i + B_i^* \boldsymbol{x}_i' \boldsymbol{\beta},$$

where

$$B_i^* = \frac{\psi_{iy}}{A + \psi_{iy} + \boldsymbol{\beta}' \boldsymbol{\Psi}_{ix} \boldsymbol{\beta}}$$

Remarks

Under the FH-ME,

$$\text{MSE}(\hat{\theta}_i^B) = (1 - B_i)\psi_{iy} + B_i^2 \boldsymbol{\beta}' \boldsymbol{\Psi}_{ix} \boldsymbol{\beta},$$

which i s greater than ψ_{iy} i f $\boldsymbol{\beta}' \boldsymbol{\Psi}_{ix} \boldsymbol{\beta} > A + \psi_{iy}$ but

$$\text{MSE}(\hat{\theta}_i^{B^\star}) = (1 - B_i^*)\psi_{iy} < \psi_{iy}$$

Ref: Datta et al. (1999; 2002); Ybarra and Lohr (2008); Marchetti et al. (2015), Mosaferi (2015).

Situation 2: A Partial Solution (Ref: Datta and Lahiri 1995)

An Outlier Resistent Model

For $i = 1, \cdots, m$,

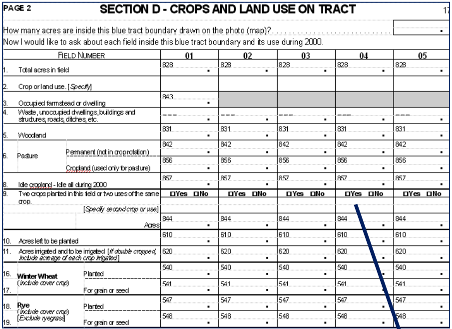
$$\textit{Level 1: (Sampling Distribution)}: \; y_i|\theta \overset{ind}{\sim} N(\theta_i,\; \psi_i);$$

$$\textit{Level 2: (Prior Distribution)}: \; \theta_i|\boldsymbol{\beta},\; A \overset{ind}{\sim} \frac{1}{\sqrt{A}} p_i \left(\frac{\theta_i - \boldsymbol{x}_i' \boldsymbol{\beta}}{\sqrt{A}} \right)$$

where $p_i(x) = \int_0^\infty r^{1/2} \psi(x r^{1/2}) g_i(r) dr$, $\phi(x)$ being the pdf of a standard normal distribution.

To retain shrinking in presence of an outlier in residual, use a heavy tail distribution (e.g., Cauchy) for the mixing distribution $g_i(\dot{)}$

Case 2: Record Linkage

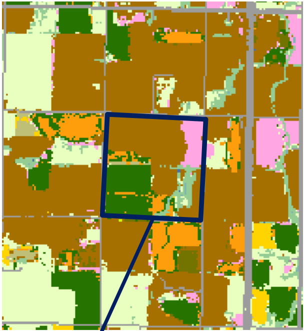


PAGE 2 SECTION D - CROPS AND LAND USE ON TRACT 17

How many acres are inside this blue tract boundary drawn on the photo (map)? . . . .

Now I would like to ask about each field inside this blue tract boundary and its use during 2000.

FIELD NUMBER	01	02	03	04	05
1. Total acres in field	828 .	828 .	828 .	828 .	828 .
2. Crop or land use. [Specify]					
3. Occupied farmstead or dwelling	843 .				
4. Waste, unoccupied dwellings, buildings and structures, roads, ditches, etc.	___ .	___ .	___ .	___ .	___ .
5. Woodland	831 .	831 .	831 .	831 .	831 .
6. Pasture — Permanent (not in crop rotation)	842 .	842 .	842 .	842 .	842 .
Pasture — Cropland (used only for pasture)	856 .	856 .	856 .	856 .	856 .
8. Idle cropland - Idle all during 2000	857 .	857 .	857 .	857 .	857 .
9. Two crops planted in this field or two uses of the same crop.	☐Yes ☐No	☐Yes ☐No	☐Yes ☐No	☐Yes ☐No	☐Yes ☐No
[Specify second crop or use]					
Acres	844 .	844 .	844 .	844 .	844 .
10. Acres left to be planted	610 .	610 .	610 .	610 .	610 .
11. Acres irrigated and to be irrigated [If double cropped, include acreage of each crop irrigated.]	620 .	620 .	620 .	620 .	620 .
16. Winter Wheat (include cover crop) — Planted	540 .	540 .	540 .	540 .	540 .
17. For grain or seed	541 .	541 .	541 .	541 .	541 .
18. Rye (include cover crop) [Exclude ryegrass] — Planted	547 .	547 .	547 .	547 .	547 .
19. For grain or seed	548 .	548 .	548 .	548 .	548 .



REGRESSION VARIABLES:

	Dependent Y	Independent X
	Enumerated JAS Segments	CDL Classified Acres
Soybeans	227	273
Wheat	337	541

USDA

28

Battese, Harter and Fuller (1988 JASA)

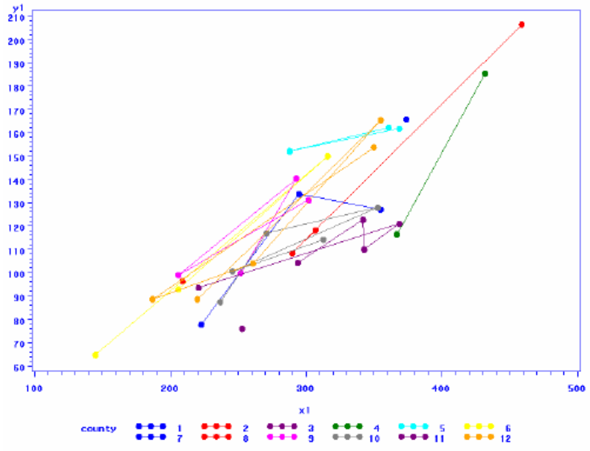
Table 1. Survey and Satellite Data for Corn and Soybeans in 12 Iowa Counties

	No. of segments		Reported hectares		No. of pixels in sample segments		Mean number of pixels per segment*	
County	Sample	County	Corn	Soybeans	Corn	Soybeans	Corn	Soybeans
Cerro Gordo	1	545	165.76	8.09	374	55	295.29	189.70
Hamilton	1	566	96.32	106.03	209	218	300.40	196.65
Worth	1	394	76.08	103.60	253	250	289.60	205.28
Humboldt	2	424	185.35	6.47	432	96	290.74	220.22
			116.43	63.82	367	178		
Franklin	3	564	162.08	43.50	361	137	318.21	188.06
			152.04	71.43	288	206		
			161.75	42.49	369	165		
Pocahontas	3	570	92.88	105.26	206	218	257.17	247.13
			149.94	76.49	316	221		
			64.75	174.34	145	338		
Winnebago	3	402	127.07	95.67	355	128	291.77	185.37
			133.55	76.57	295	147		
			77.70	93.48	223	204		
Wright	3	567	206.39	37.84	459	77	301.26	221.36
			108.33	131.12	290	217		
			118.17	124.44	307	258		
Webster	4	687	99.96	144.15	252	303	262.17	247.09
			140.43	103.60	293	221		
			98.95	88.59	206	222		
			131.04	115.58	302	274		
Hancock	5	569	114.12	99.15	313	190	314.28	198.66
			100.60	124.56	246	270		
			127.88	110.88	353	172		
			116.90	109.14	271	228		
			87.41	143.66	237	297		
Kossuth	5	965	93.48	91.05	221	167	298.65	204.61
			121.00	132.33	369	191		
			109.91	143.14	343	249		
			122.66	104.13	342	182		
			104.21	118.57	294	179		
Hardin	6	556	88.59	102.59	220	262	325.99	177.05
			88.59	29.46	340	87		
			165.35	69.28	355	160		
			104.00	99.15	261	221		
			88.63	143.66	187	345		
			153.70	94.49	350	190		

* The mean number of pixels of a given crop per segment in a county is the total number of pixels classified as that crop, divided by the number of segments in that county.

How to make BIGDATA useful?

Fig 2: Plot of Corn Hectares versus Corn Pixels by County



This plot also reflects the strong relationship between the reported hectares of corn and the number of pixels of corn for counties separately. But the slopes and/or intercepts

How do we combine information?

- y_{ij} : value of the study variable for the jth unit of the i small area population $(i = 1, \cdots, m; \;\; j = 1, \cdots, N_i)$
- We are interested in estimating the finite population means:

$$\bar{Y}_i = N_i^{-1} \sum_{j=1}^{N_i} y_{ij}.$$

Nested Error Regression Model

For $i = 1, \cdots, m; \;\; j = 1, \cdots, N_i$

$$y_{ij} = x_{ij}'\beta + v_i + e_{ij},$$

where x_{ij} is a $p \times 1$ column vector of known auxiliary variables; $\{v_i\}$ and $\{e_{ij}\}$ are all independent with $v_i \overset{iid}{\sim} N(0, \sigma_v^2)$ and $e_{ij} \overset{iid}{\sim} N(0, \sigma_e^2)$

An Example

- Estimation of the number of hectares of corn for 12 Iowa counties based on the 1978 June Enumerative Survey and satellite data.
- y_{ij}: the number of hectares of corn in the jth segment of the ith county as reported in the June Enumerative Survey.
- $x'_{ij} = (1, x_{1ij}, x_{2ij})$, where x_{1ij} (x_{2ij}) is the number of *pixels* classified as corn (soybean) in the jth segment of the ith county.
- $\bar{X}' = (1, \bar{X}_{1i}, \bar{X}_{2i})$, where $\bar{X}_{1i}$ ($\bar{X}_{2i}$) is the mean number of pixels per segment classified as corn (soybean) for county i.

Unit Level Model with BIGDATA (Ref: Gershunskaya and Lahiri 2011)

Model:

For $i = 1, \cdots, m, ; j = 1, \cdots, N_i$,

$$y_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + v_i + e_{ij},$$

where

- $v_i \overset{iid}{\sim} N(0, \tau^2)$
- $e_{ij} \overset{iid}{\sim} (1 - z_{ij})N(0, \sigma_1^2) + z_{ij}N(0, \sigma_2^2)$
- z_{ij} is the mixture part indicator random variable with

$$z_{ij}|\pi \overset{iid}{\sim} \text{Bin}(1, \pi)$$

Real Time Traffic Prediction

Smart City Context to Traffic Data
Ref: Cirillo et al. (2017)

- Smart cities are composed of many networks, and to each of them it is possible to associate one or several datasets.
  - The White House issued a press statement announcing a new Smart Cities initiative to help communities tackle local challenges, improve city services and quality of life.
- Transportation is one such physical network, that is increasingly being powered by large amounts of collected data.
  - Traffic data help users avoid congested and slow areas and transport operators reduce and manage congestion.
  - Such decision support systems are collectively called Advanced Traveler Information Systems (ATIS).

Requirement for Traffic Prediction

- Robust Traffic predictions is in high demand
  - The large amount of literature published recently dealing with traffic predictions is a testament to the demand: Transportation Research Part C recently published a special issue focusing just on traffic predictions (Zhang, 2014).
- The benefits of prediction are quite numerous, especially because it allows proactive reaction to developing conditions.
  - Faster response to changing conditions allows the system to react quickly, reducing wasted time, energy and resources.
- The data revolution in transportation is making real-time data more ubiquitously available both in space and time.
- Leveraging this data to make robust short-term predictions will spur the next revolution in transportation.

Traffic Data Sources

- Traditionally, traffic data collection was very expensive: it required roadside counters (often people with counters) and detectors (embedded loop, radar, microwave, camera, etc.)
- Also detector collection is geographically very limited, required constant calibration and maintenance.
- Since mid 2000s, however, ubiquitous use of GPS devices capable of mobile telemetry — especially by the freight industry — made it possible to collect data continuously and over a large area for a fraction of the cost of traditional methods.
- Since all vehicles do not transmit at all times, this is considered as "probe" data, where data is collected from only a sample (probe) of vehicles on the roadway.
- GPS probe data can be collected anywhere an equipped, transmitting vehicle can travel.
- Therefore, it gives potential visibility over the state of the whole network.

Vehicle Probe Project at University of Maryland

- As seen in figure 1 vehicles transmit data to a central control/dispatch center
- Usually, vehicles transmit location, direction of travel and current speed
- This data is then collected by companies specializing in probe data (Inrix Inc., Here Inc., TomTom, etc.) and aggregated to roadway segments using the location and direction information
- Roadway segmentation is traditionally based on Traffic Message Channel (TMC) codes, which divide a roadway from intersection to intersection
- This data is usually aggregated to a predefined reporting window
- States in the I-95 Corridor Coalition have been purchasing this data from the providers at one minute frequencies since 2008
- The Center for Advanced Transportation Technologies (CATT) at UMD is tasked with archiving this data, and creating analytic tools for state agencies to use
- This suite of tools, including the data archival is called the Vehicle Probe Project (VPP)

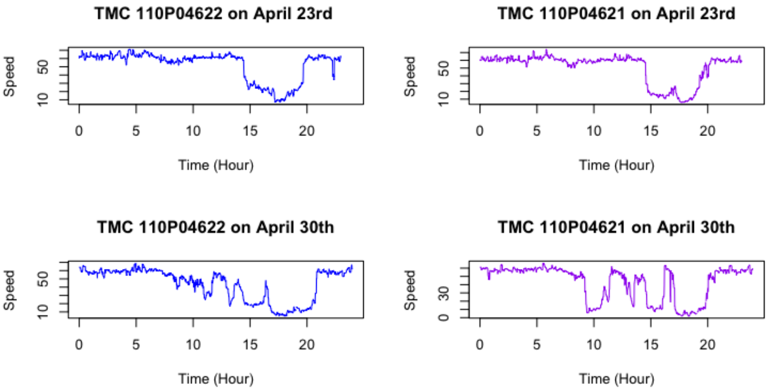


Figure: Time series plots of speed for two TMCs and two consecutive weeks

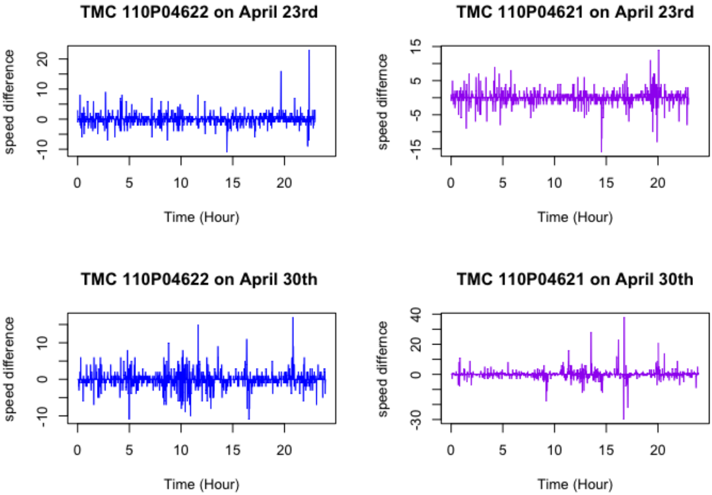


Figure: First order difference of speed data for TMC 110P04622 and TMC 110P04621 on April 23rd and April 30th.

Mathematical Formulation of ARIMA with Auxiliary Variables

$$\psi_w(B)(1-B)^d y_{t,w} = x_{t,w}^T \gamma_w + \eta_w(B) z_{t,w},$$

where

$\psi_w(B) = 1 - \psi_{1w} B - \cdots - \psi_{pw} B^p$,
$\eta_w(B) = 1 - \eta_{1w} B - \cdots - \eta_{qw} B^q$,
B is a back shift operator: $B^d y_{t,w} = y_{t-d,w}$,
$z_{t,w}$ are white noises that follow normal distributions with zero means and constant variance σ^2,
$x_{t,w}$ is a $s \times 1$ vector of known auxiliary variables,
γ_w is a $s \times 1$ vector of unknown fixed coefficients,
$\psi_{1w}, \cdots \psi_{pw}$ and $\eta_{1w}, \cdots \eta_{qw}$ are unknown model parameters.

Key Assumptions

The key assumption is that the traffic patterns do not change between the time period used for model fitting and the time period when predictions are made.

- Weekly modeling is used, i.e. assume traffic conditions repeat on a given day of week
  - As per figure 4, this is a robust assumption, as the majority of traffic patterns repeat across the day over different weeks
- We define w as the week in which predictions are required, and models are fit to week $w-1$, as shown:

$$\psi_w(B) = \psi_{w-1}(B)$$
$$\eta_w(B) = \eta_{w-1}(B)$$
$$\gamma_w = \gamma_{w-1}$$

Algorithm

- The first step is model selection, for each given segment, and day
  - 27 ARIMA orders (p, d, q) are tested for each segment, where each of p, d and q can take values from $\{0, 1, 2\}$
  - The model with the lowest Bayesian Information Criterion (BIC) is selected
- The most reasonable selected model is used to make predictions for the next week
  - Predictions are done online, on the incoming stream of data
  - Based on the ARIMA order specification, data points from the required time steps before the most recent are used
- Predictions are made every minute, up to 30 minutes into the future
  - Predictions are stopped at the end of the day
  - The first prediction is only made after sufficient number of data points have been received (informed by the ARIMA order; for example 2 observation for $ARIMA(0, 1, 0)$)
  - Similarly predictions of future minutes can be based entirely on interim predictions (for example, predictions of minute 20 is based on predicted values of minutes 18 and 19 for $ARIMA(0, 1, 0)$)

Data Used

- Data from 3 weeks in September 2016 are used to demonstrate the proposed framework
- Only weekday data is used for the study
- The first set of models are fit to data from the week of September 12
- These models are used to predict for each corresponding day in the week of September 19
- Similarly, the second set of models are fit to real data collected in the week of September 19
- Predictions from the second set of models are made for the week of September 26
- Data from 2,654 segments that form the mobility corridor network of Maryland are used
- Over the 15 days examined, for the 2,654 segments the total size of data is slightly over 57 million records
- Predictions up to 30 minutes in the future for each segment for all 15 days result in about 1.7 billion records

Network Map

The complete map of the studied network is presented in the figure below

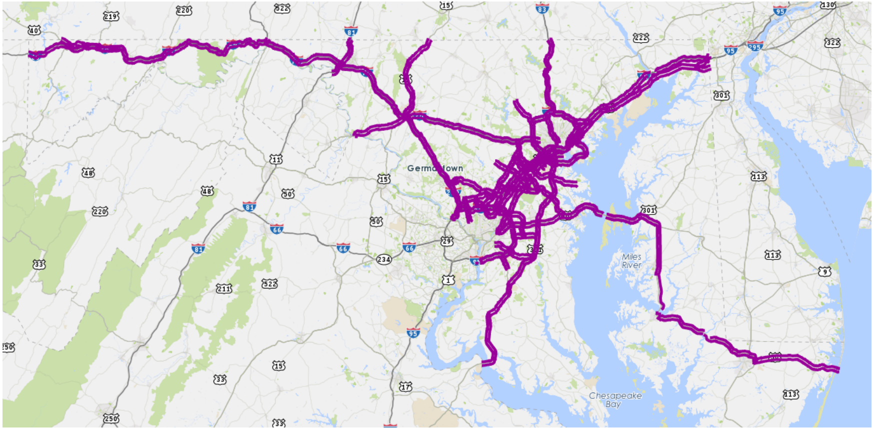


Figure 5: Map of the Studied Network

Imputations and Interpolations

- The VPP does not report data at rounded minutes
- Consequently, there is uneven interval between two data points
- Speed readings at the exact minute are computed by linear interpolation between the observations received before and after the minute
- Sometimes data over short periods is not received or goes missing due to transmission or other failures
- Such short duration data losses are also covered by linearly interpolating between the available data points
- The data is imputed from source with historic speeds when real-time observations are completely lacking

Selected ARIMA Models

The following table gives the ARIMA order and the number of times it was selected as the most reasonable model for each segment, each day. The sum total of selected models is 26,540 (2,654 segments, 10 days), out of 716,580 total fitted models. The most selected orders are highlighted.

Order	Count	Order	Count	Order	Count
(0, 0, 0)	NA	(1, 0, 0)	5,700	(2, 0, 0)	817
(0, 0, 1)	0	(1, 0, 1)	185	(2, 0, 1)	239
(0, 0, 2)	0	(1, 0, 2)	886	(2, 0, 2)	381
(0, 1, 0)	3,077	(1, 1, 0)	589	(2, 1, 0)	154
(0, 1, 1)	405	(1, 1, 1)	10,995	(2, 1, 1)	1,254
(0, 1, 2)	196	(1, 1, 2)	567	(2, 1, 2)	1,095
(0, 2, 0)	0	(1, 2, 0)	0	(2, 2, 0)	0
(0, 2, 1)	0	(1, 2, 1)	0	(2, 2, 1)	0
(0, 2, 2)	0	(1, 2, 2)	0	(2, 2, 2)	0

Table 1: Selected ARIMA Orders

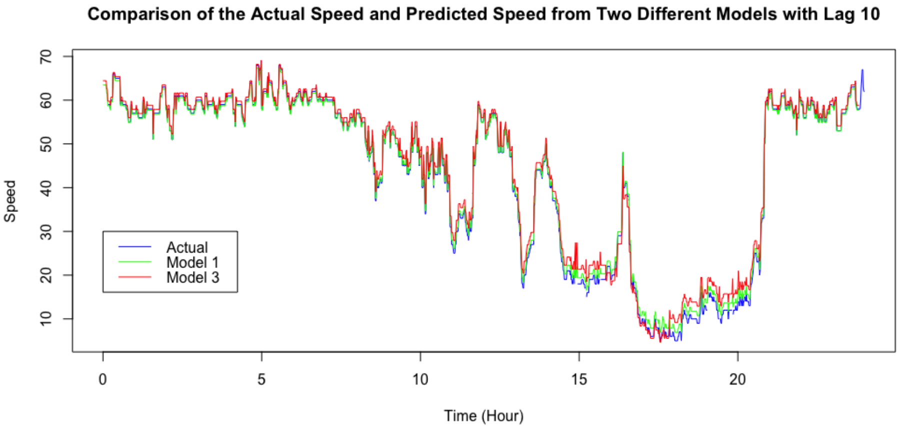


Figure: The actual and predicted values of speed data for TMC 110P04622 on April 30th.

Relative Root Mean Squared Prediction Error

- To robustly quantify the errors we propose the RRMSPE as defined below:

$$RRMSPE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\frac{\hat{y}_t - y_t}{y_t}\right)^2},$$

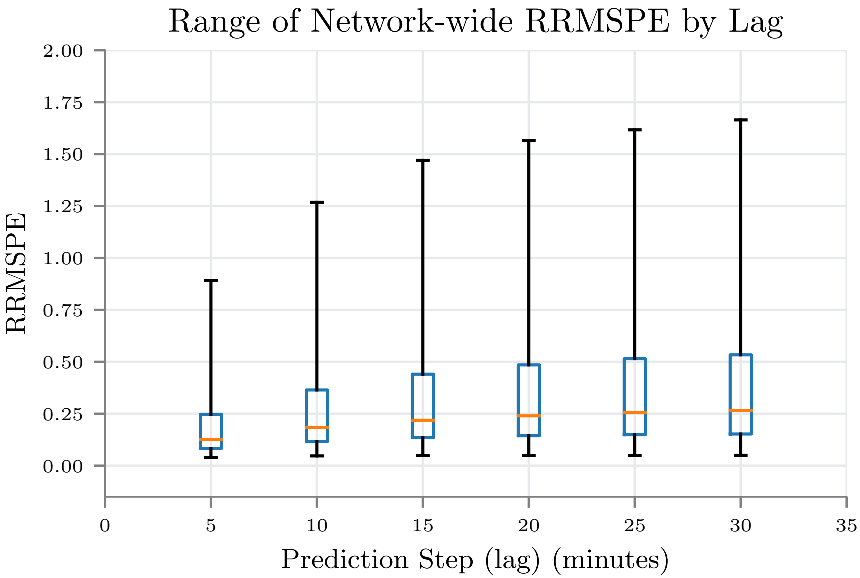
where

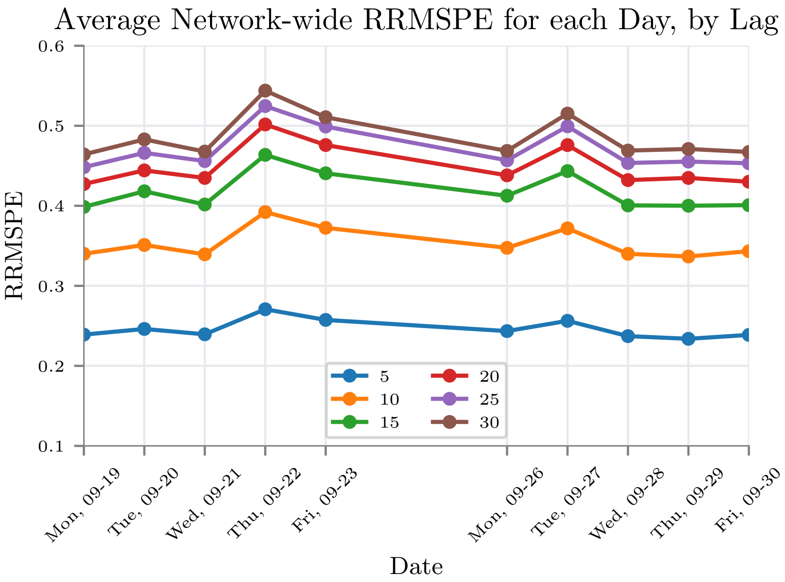
$\hat{y}_t$ is the predicted speed at time t,
y_t is the real observed speed at time t,
T is the total number of time steps. For a day, $T = 1440$, or $T = 60$ for an hour.

- Note that the RRMSPE is a relative error, and can be interpreted as the percent deviation of the predicted value from the true value
- Further, the error is calculated over the whole network for given prediction intervals (lag). Thus it includes freeways and arterials
- Due to limitations of the probe data, it is not as robust on heavily signalized arterials as compared to freeways (Kaushik et al., 2015, 2014)

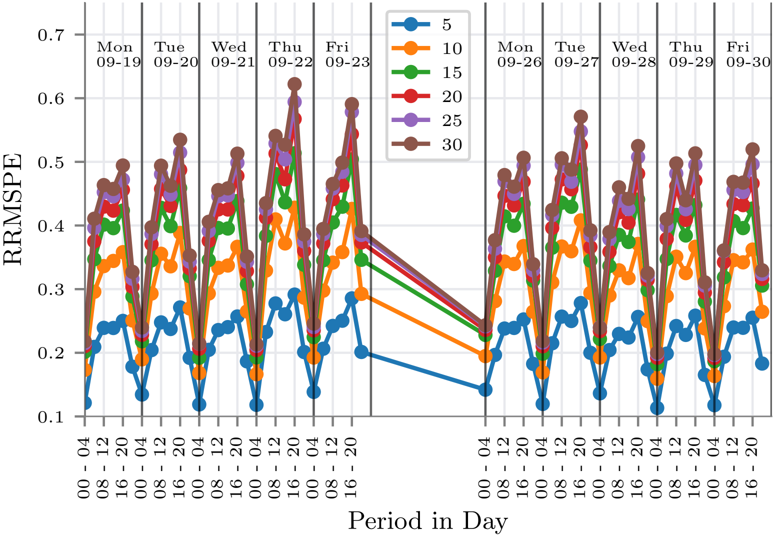
Error Plots

- We define predictions intervals as lags: minutes prior to current minute that were used to predict the speeds at current minute
  - A lag of 5 means speed at the current minute was predicted from data received 5 minutes ago
- The following slides show the RRMSPE calculated for important lag intervals
  - Only lags of 5, 10, 15, 20, 25 and 30 minutes are shown
  - For more complex plots, lags of 20 and 25 minutes are not shown


Range of Network-wide RRMSPE by Lag
RRMSPE
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0
5
10
15
20
25
30
35
Prediction Step (lag) (minutes)


Average Network-wide RRMSPE for each Day, by Lag
RRMSPE
0.6
0.5
0.4
0.3
0.2
0.1
Date
Mon, 09-19
Tue, 09-20
Wed, 09-21
Thu, 09-22
Fri, 09-23
Mon, 09-26
Tue, 09-27
Wed, 09-28
Thu, 09-29
Fri, 09-30
5
10
15
20
25
30

Average Network-wide RRMSPE for Period in Day, by Lag



Relative Estimate Residuals

- In order to find out if the models are optimistic or pessimistic, relative residuals are computed
- Relative residuals is similar to RRMSPE, with the difference that the residuals are not squared
- This allows one to directly examine the signed percent error in the predictions
- We compute relative residuals as shown:

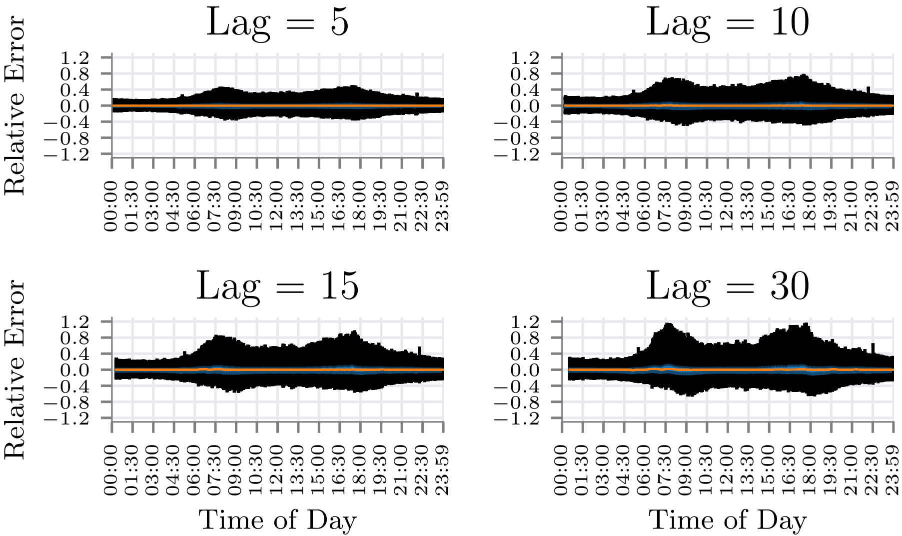
$$Rr_t = \frac{\hat{y}_t - y_t}{y_t}, \tag{1}$$

where

Rr_t is the relative residuals at time t,
$\hat{y}_t$ is the predicted speed at time t,
y_t is the real observed speed at time t.

- The following figure plots box plots with the relative residuals for each minute of the day
- There are, therefore 26,540 points in each of 1,440 boxes, one box for each minute of the day

Range of Network-wide RRMSPE Each Minute in Day



David Salsburg, ASA Connect Discussion

"...D.J. Finney once wrote about the statistician whose client comes in and says, "Here is my mountain of trash. Find the gems that lie therein." Finney's advice was to not throw him out of the office but to attempt to find out what he considers "gems". After all, if the trained statistician does not help, he will find some one who will...."

SAE Conferences

- SAE 2015: First Latin American ISI Satellite Conference on Small Area Estimation, Santiago, Chile (`http://www.encuestas.uc.cl/sae2015/program_sae.html`)
- SAE 2014: Small Area Estimation Conference (Poznan, Poland, 2014)
- SAE 2013: The First Asian ISI Satellite Meeting on Small Area Estimation (Bangkok, Thailand, 2013)
- SAE 2011: Conference on Small Area Statistics (Trier, Germany, 2011)
- SAE 2009: Rhine River Cruise Conference 2009 on Recent Advances in Small Area Estimation (Germany, 2009)
- SAE 2009: SAE 2009 Conference on Small Area Estimation (Elche, Spain, 2009)
- SAE 2007: IASS Satellite Conference on SAE (Pisa, Italy, 2007)
- SAE 2001: International Conference on SAE and Related Topics (Maryland, USA, 2001)

THANK YOU!