Need for Data Integration and Statistical Modeling for Various Purposes

Partha Lahiri

Joint Program in Survey Methodology & Department of Mathematics
University of Maryland, College Park

BLS Seminar, Washington DC, September 25, 2019

Statistical data integration to reduce nonsampling errors

Ref: Hansen et al. (1953, pp 483-486)

Estimate the median number of radio stations heard during the day for over 500 counties of the USA (small areas).

Two different survey data used:

Mail Survey

- large sample (1000 families/county) from an incomplete list frame
- response rate was low (about 20%)
- estimates x_i are biased due to non-response and incomplete coverage

Personal Interview Survey

- Smaller sample size
- Sample design:
  - The 500 counties were stratified into 85 primary strata based on geographical region and the type of radio service available.
  - One county was selected from each of these 85 strata with probability proportional to the estimated number of families in the county.
  - A subsample of area segments was selected from each sampled county and the families within the selected area segments were interviewed.

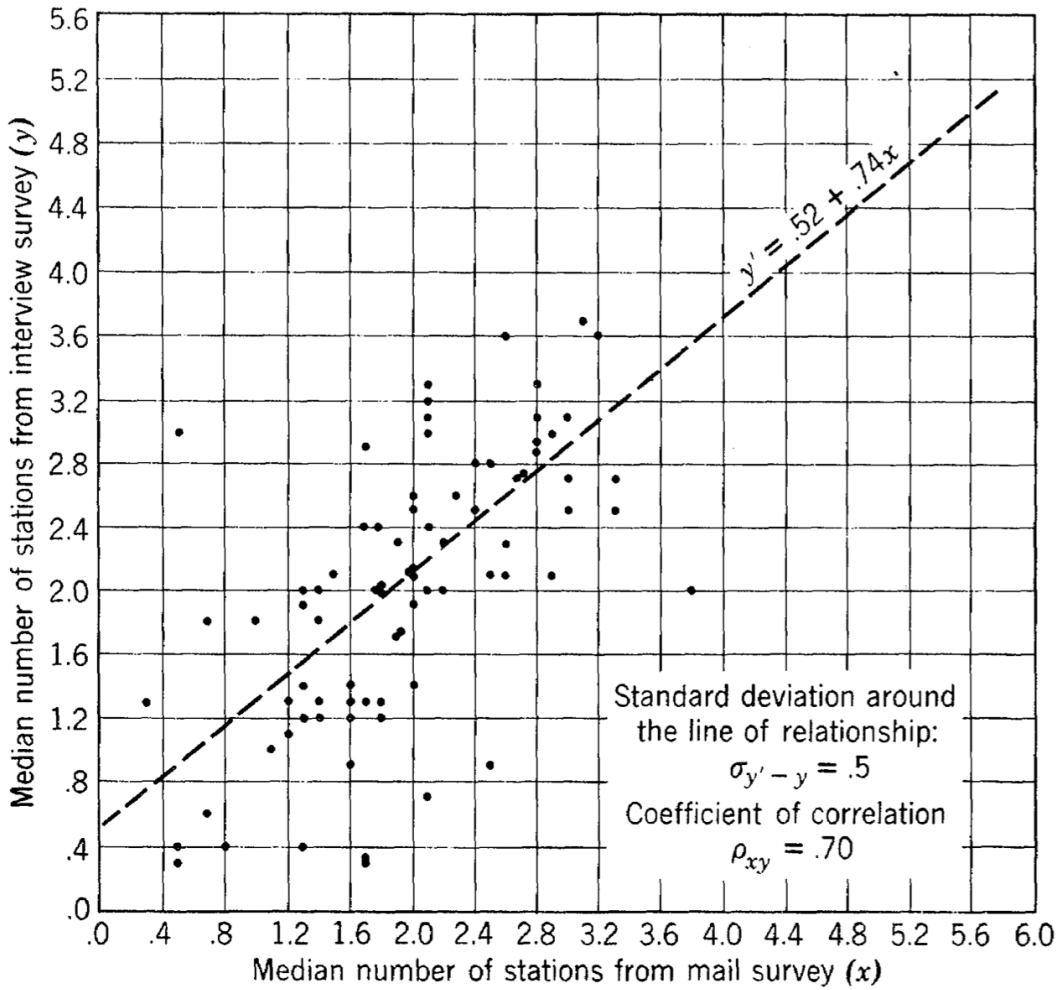


FIG. 2. Comparison of median numbers of stations heard during the day without trouble, as estimated from mail and interview surveys in selected primary sampling units.

Hansen et al. (also see Rao 2003) suggested the following:

- For counties with sample from the personal interview survey, use y_i.
- For counties with sample from only mail survey, use estimate $\hat{y}_i = 52 + .74x_i$

Questions

- Is there a way to improve $\hat{y}_i$ for counties with no data from the personal interview survey?
- How do we measure the uncertainty of $\hat{y}_i$?
- Do we still use y_i if we have small samples from the personal interview survey?

An extension of Hansen et al. (1953)

1. Level 1 (Sampling model):

$$y_i|\theta_{iy} \overset{\text{indep.}}{\sim} N(\theta_{iy}, \sigma^2_{iy}),\ i = 1, 2, ..., 85$$

$$x_i|\theta_{ix} \overset{\text{indep.}}{\sim} N(\theta_{ix}, \sigma^2_{ix}),\ i = 1, 2, ..., 500$$

   Conditionally x_i and y_i are independent. σ^2_{iy} and σ^2_{ix} are assumed to be known.

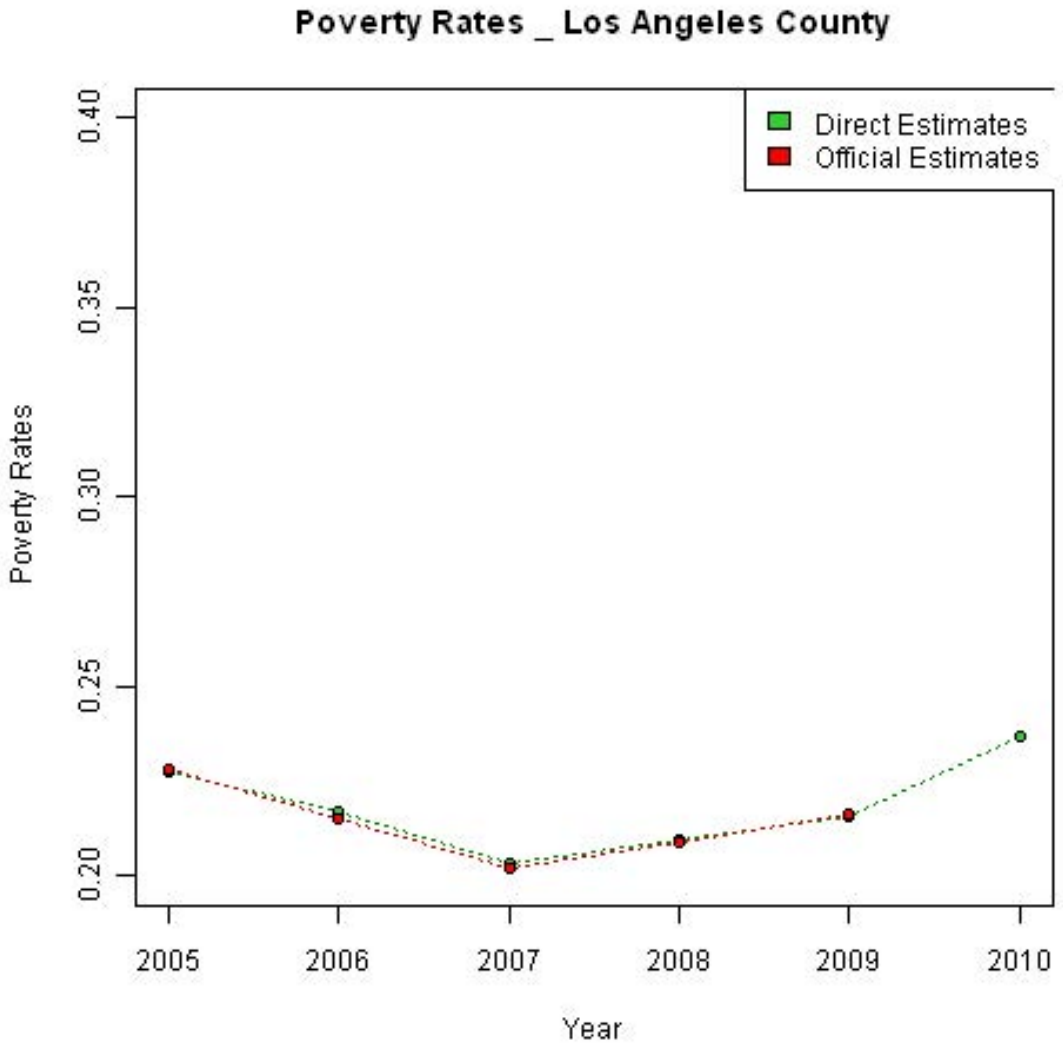
2. Level 2 (Linking model)

$$\theta_{iy}|\theta_{ix} \overset{\text{indep.}}{\sim} N(\beta_0 + \beta_1\theta_{ix} + \beta_2 z_i,\ \tau^2),$$

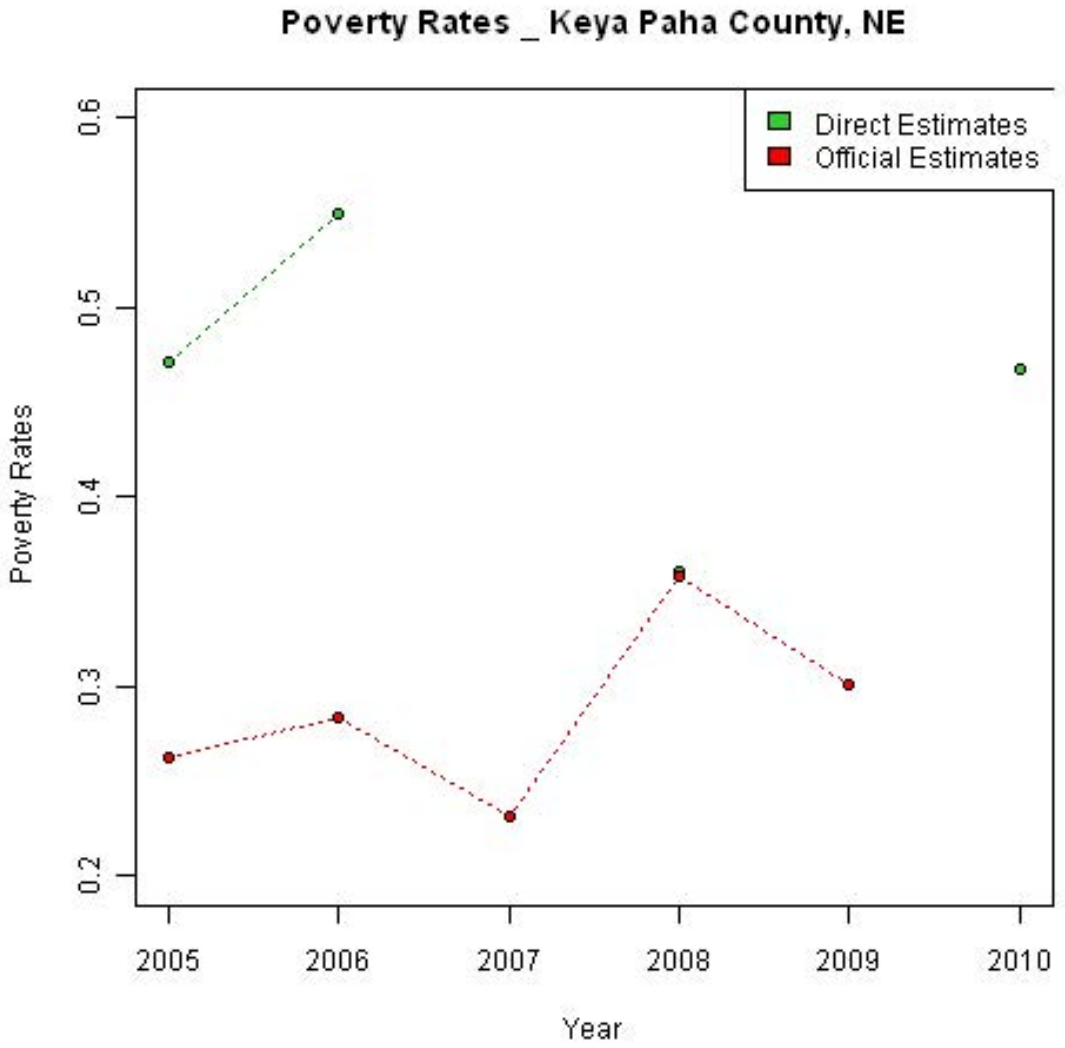
   where z_i is an auxiliary variable.

Question: How do we make inference?
Das and Lahiri (2019) discussed a few other possible extensions and developed methods for statistical inference.

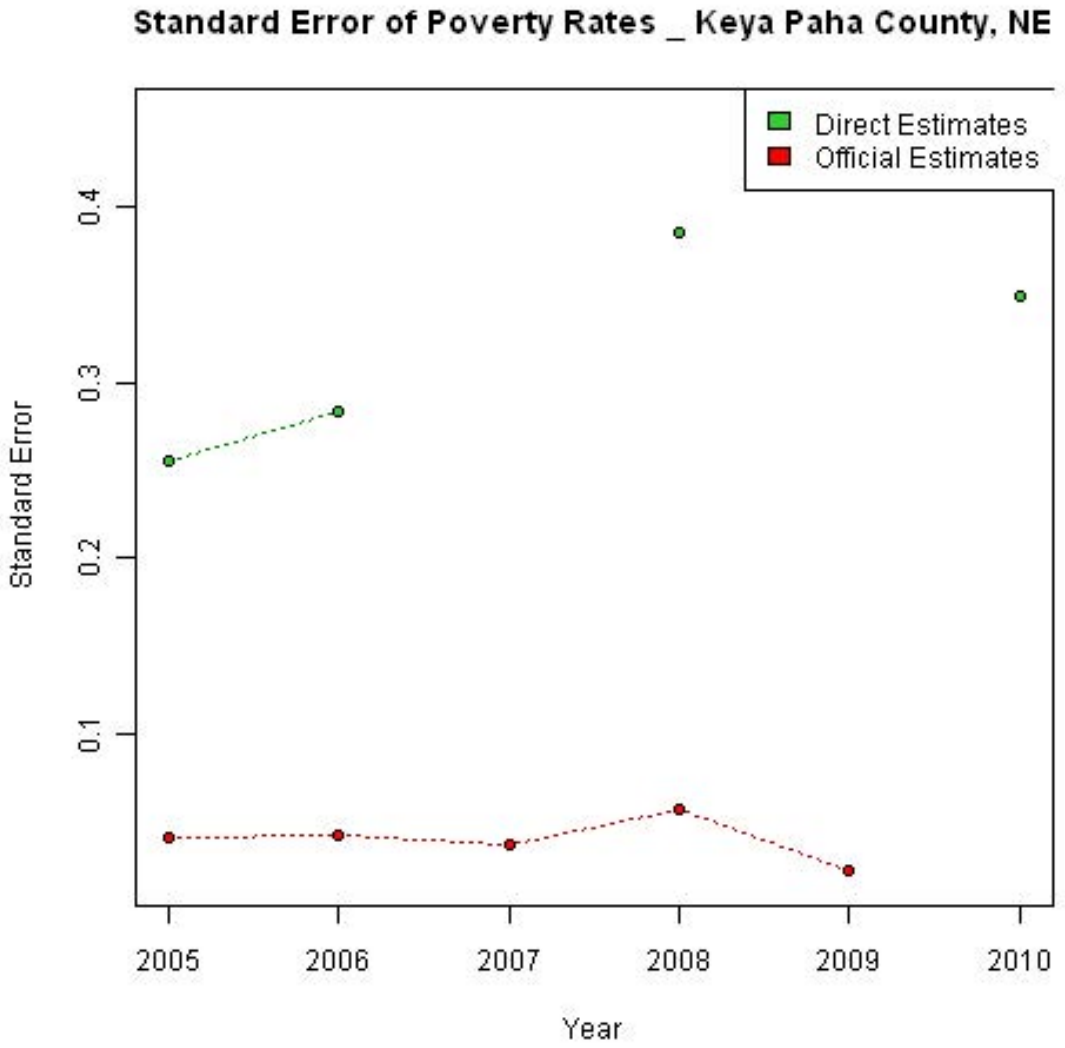
SAIPE Example



SAIPE Example



SAIPE Example



Different purposes and methods of data integration: A few examples

- Linking of two Probability Samples – one nested within the other: Double or two-phase sampling (Cochran 1977)
- Linking of two independent (non-nested) samples (Bose 1943; Hansen et al. 1953; Hidiroglou 2001, Kim and Rao 2012)
- Linking of a probability sample with Big Data (River 2007; Kim 2017)
- Linking census data with probability samples (U.S. Census Coverage Measurement)
- Linking register with sample: register-assisted census (Gabler et al. 2008)
- Use of several administrative databases in non-response follow-up study (Morris et al. 2015)

Examples Continued

- Statistical register: linking several samples with administrative data and registers (Zhang and Nordbotten 2008)
- Use of historical databases to reduce the extent of revision (Lahiri and Li 2008)
- Linking several historical Big Data (Cirillo et al. 2017)
- Linking record by record in absence of unique identifier (probabilistic record linkage: Herzog et al. 2008; Lahiri and Larsen 2005; Han 2018; Han and Lahiri 2018)
- Statistical linking for addressing small area estimation (SAE) (Jiang and Lahiri 2006; Casas-Cordero et al. 2015, Rao and Molina 2015)
- Data integration for the purpose of nowcasting (Das et al. 2019).

Data integration for the purpose of nowcasting

Ref: Das et al. (2019)

Parameter	mean	sd	2.5%	50%	97.5%
(Intercept)	1.157431	0.750100	-0.285946	1.145826	2.674232
MPCE_Rural_09_10	-0.001707	0.000691	-0.003133	-0.001686	-0.000405

Table: The parameter estimates and desired statistics obtained from fitting the data of the year 2009-10 for the rural region.

	Min.	1st.Qu.	Median	Mean	3rd.Qu.	Max.
Logit	-10.98	-7.14	-5.25	-4.90	-1.95	1.06
Bayesian beta Regression	-8.86	-5.19	-2.46	-2.56	0.52	3.76

Table: Summary statistics of *nowcasting errors* for different methods used to estimate the model parameters for the rural region.

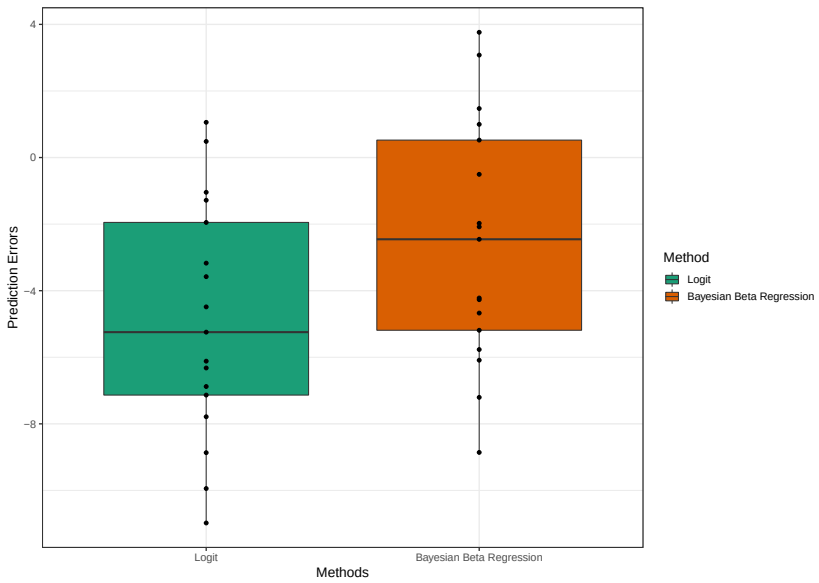


Figure: Boxplots of *nowcasting errors* (in nowcasting the values for the year 2011-12) in the y-axis for the logit based simple linear regression applied on the logit of poverty rate estimates and the Bayesian beta regression. This is for the rural region.

Real-Time Traffic Forecasting Using Big Data

Ref: Cirillo et al. (2018)

- Current traffic systems are reactive
  - Travel route is not updated unless delay on primary route significantly exceeds alternate routes
  - Optimal travel time prediction at a given time is not an optimal travel time prediction 10 minutes later.
- There is a huge demand to change that with traffic predictions
  - A huge amount of literature has been published on traffic forecasting over the past 10 years
- Low hanging fruits 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
- Leveraging the data boom to make robust short-term predictions will spur the next revolution in transportation

Patterns in the Data

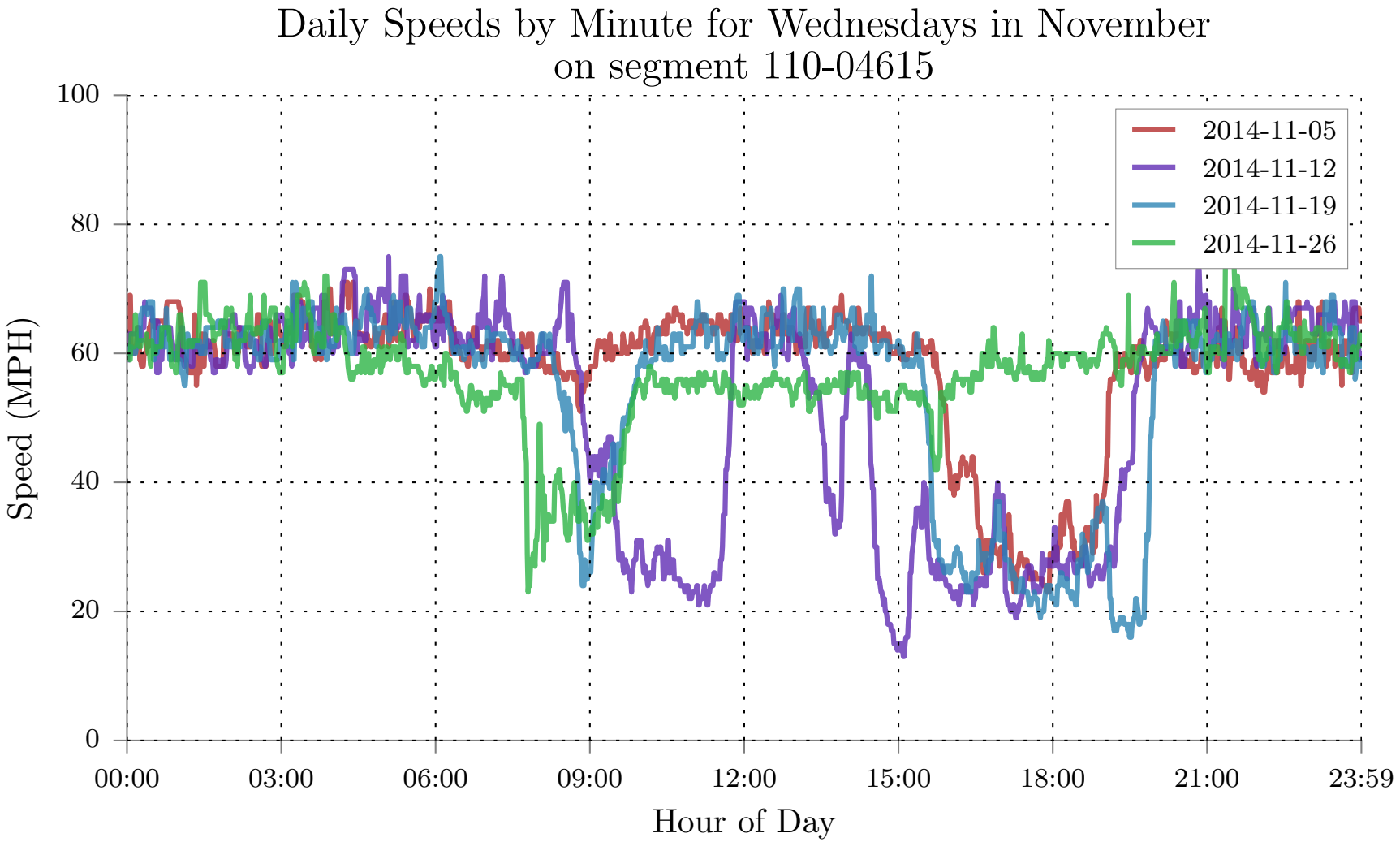


Figure: Data From Four Wednesdays in November on a Segment

SAE vs SAF

- SAE when area has no sample: Real Population
- Forecasting: Conceptual Population
  - Direct Forecasting
  - Small Area Forecasting (SAF)

Synthetic SAF

Using the concept of synthetic SAE, we now propose a radically different idea of completing all the complex model building and estimation well ahead of time, say, a week in advance and then apply the chosen model with estimated parameters for the forecasting in real-time.

Remarks:

- The forecasting is instantaneous and so is appropriate for forecasting in real-time for big data.
- Historical data are used to understand the speed distribution for the *entire* day.

Data Used

- Data from 3 weeks in September 2016 are used to demonstrate the proposed framework, as shown in table 3
  - Only weekday data is used for the study

Table: Data usage structure

Fitting Week	Prediction/Testing Week
September 12, 2016	September 19, 2016
September 19, 2016	September 26, 2016

- Data from 2,654 segments (about 2,000 lane-miles), which form the mobility corridor of Maryland are used
  - Over the 15 days examined, for all 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

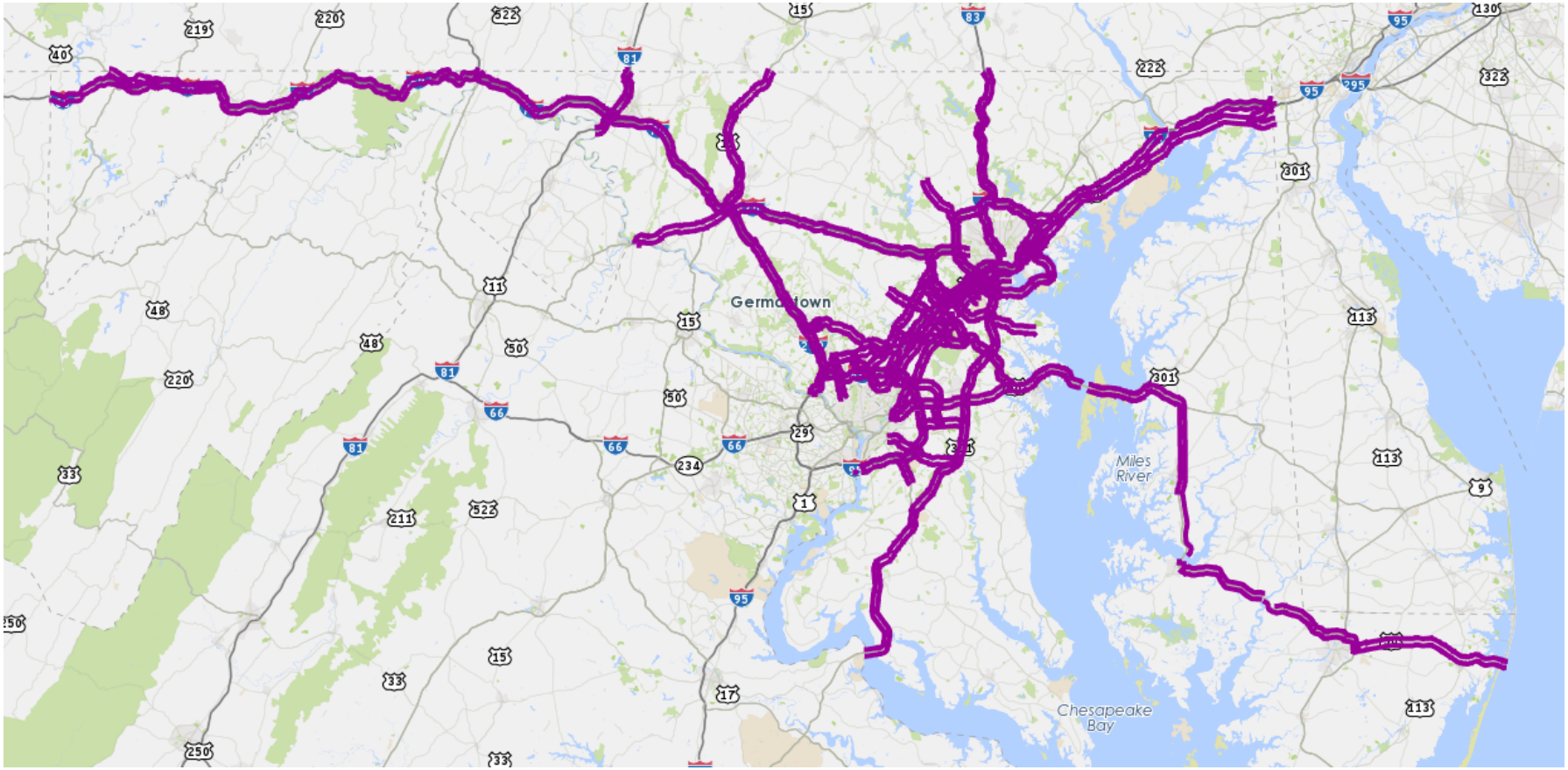
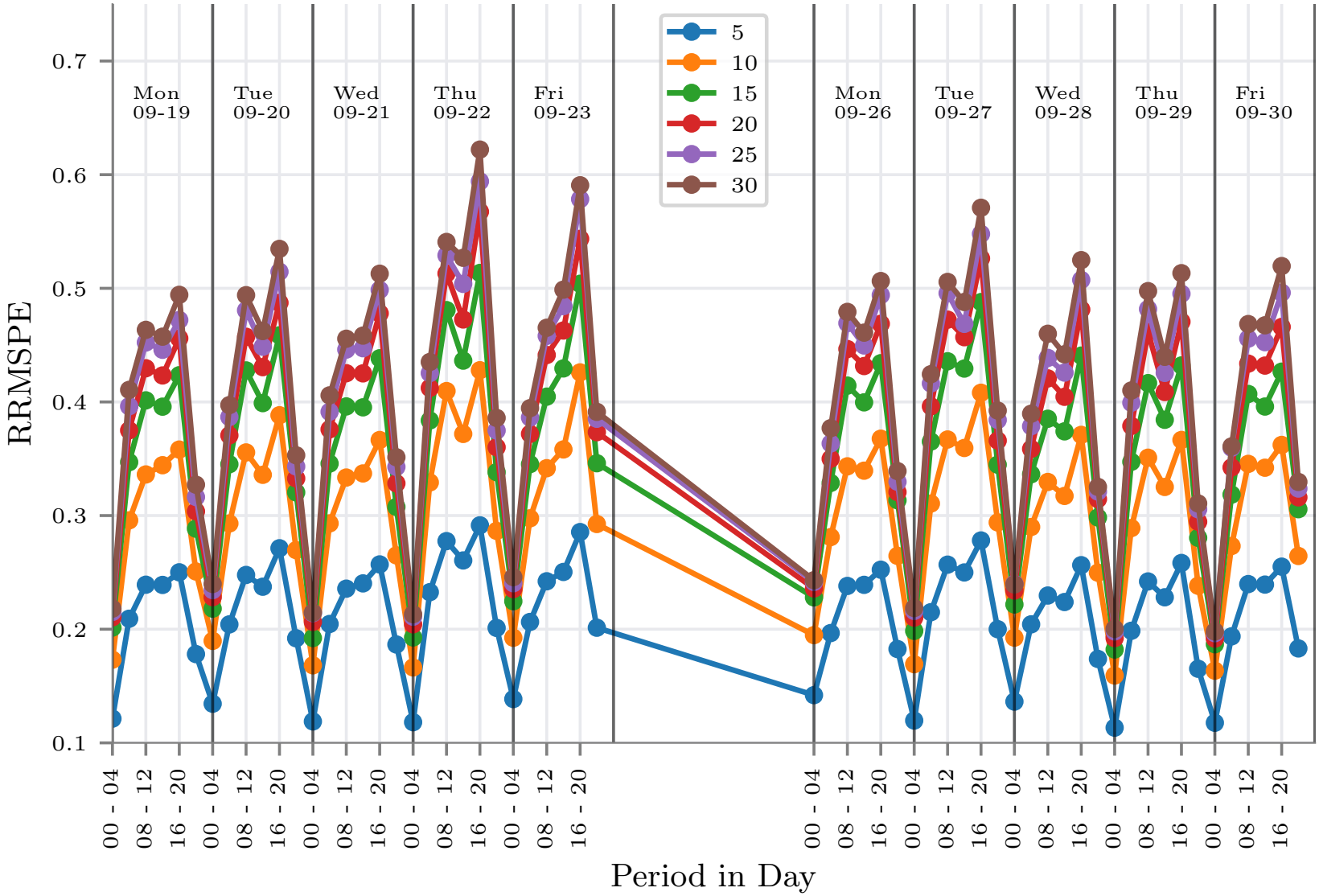


Figure: Map of the Studied Network

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



Advantages of Synthetic Time Series Framework

- The predictions are quite robust compared to literature
  - Especially when compared to similar methods using ARIMA, and even Seasonal ARIMA
- The synthetic framework is easily extendable and flexible
  - It can be used with any model, parametric or data-driven
  - Models with auxiliary variables can also be used within the framework
  - Such auxiliary variables need not be temporally bound to be used
  - Hierarchical models can also be used
  - In fact, the framework itself can be used as levels within a larger modeling environment
- Brings the strengths of the SAF to transportation
- The framework is scalable with network and data size
  - Takes under 3 hours to fit and predict for the demonstrated dataset (using an 8-core Intel i-7 Skylake processor running Python 3.5 on Ubuntu 16.04)

Use of Twitter Data in Estimating Race Distribution for Small Areas

Ref: Cao and Lahiri (2018)

Descriptive Model

Level 1: Individudal Model. The individual parameter $\mathbf{p}_i = (p_{i1}, \ldots, p_{iK})$ is of ultimate interest, $\sum_{k=1}^{K} p_{ik} = 1$.

$$(y_{i1}, \ldots, y_{iK}) | p_{i1}, \ldots, p_{iK} \sim \textit{Multinomial}(n_i, p_{i1}, \ldots, p_{iK}),$$

Level 2: Structural Model. Given the structural parameters,

$$(p_{i1}, \ldots, p_{iK}) | \boldsymbol{\beta}, r \sim \textit{Dirichlet}(rp_{i1}^E, \ldots, rp_{iK}^E),$$

where $p_{ik}^E = \frac{e^{\mathbf{x}_i'\boldsymbol{\beta}_k}}{1+\sum_{j=1}^{K-1} e^{\mathbf{x}_i'\boldsymbol{\beta}_j}}$ for $k = 1, \ldots, K-1$ and $p_{iK}^E = \frac{1}{1+\sum_{j=1}^{K-1} e^{\mathbf{x}_i'\boldsymbol{\beta}_j}}$ so that $\sum_{k=1}^{K} p_{ik}^E = 1$.

Level 3: Distributions on the Structural Parameters.

$$\boldsymbol{\beta}_k \sim \textit{Uniform on } \mathbb{R}^q$$

for $k = 1, \ldots, K-1$, and

$$1/r \sim \textit{Uniform}(0, \infty).$$

Data Example

- Data Collection
  Twitter Streaming API: This dataset contains 161,771,878 Twitter messages sent by 3,670,604 active Twitter users between July 10, 2017 and October 20, 2017 in the continental United States. Contains information about the user (geotag and self-reported name).

- Individual Race Distribution
  Extract the self-reported name from each tweet and use the last name to infer about the race distribution of the Twitter user using Census Bureau's surname list. For example, the surname 'Taylor' is 65.38% Caucasian (non-hispanic), 28.42% African-American, 0.56% Asian or Pacific Islander and 2.46% Hispanic and 3.18% of being other races.

Data Example - continued

- Location
  Location information can be inferred from the geotag contained in each tweet. Feed the geotag to Bing Maps API to get the longitude and lattitude of each user. Use the PUMA shapefile to assign a PUMA to each Twitter user.

- Race Counts
  Let s_i denote the set of observations in the i^{th} PUMA and d_{jk} denote the probability of the race k for the j^{th} Twitter user in s_i. Then the count for race k in the i^{th} PUMA is y_{ik} and is calculated as follows

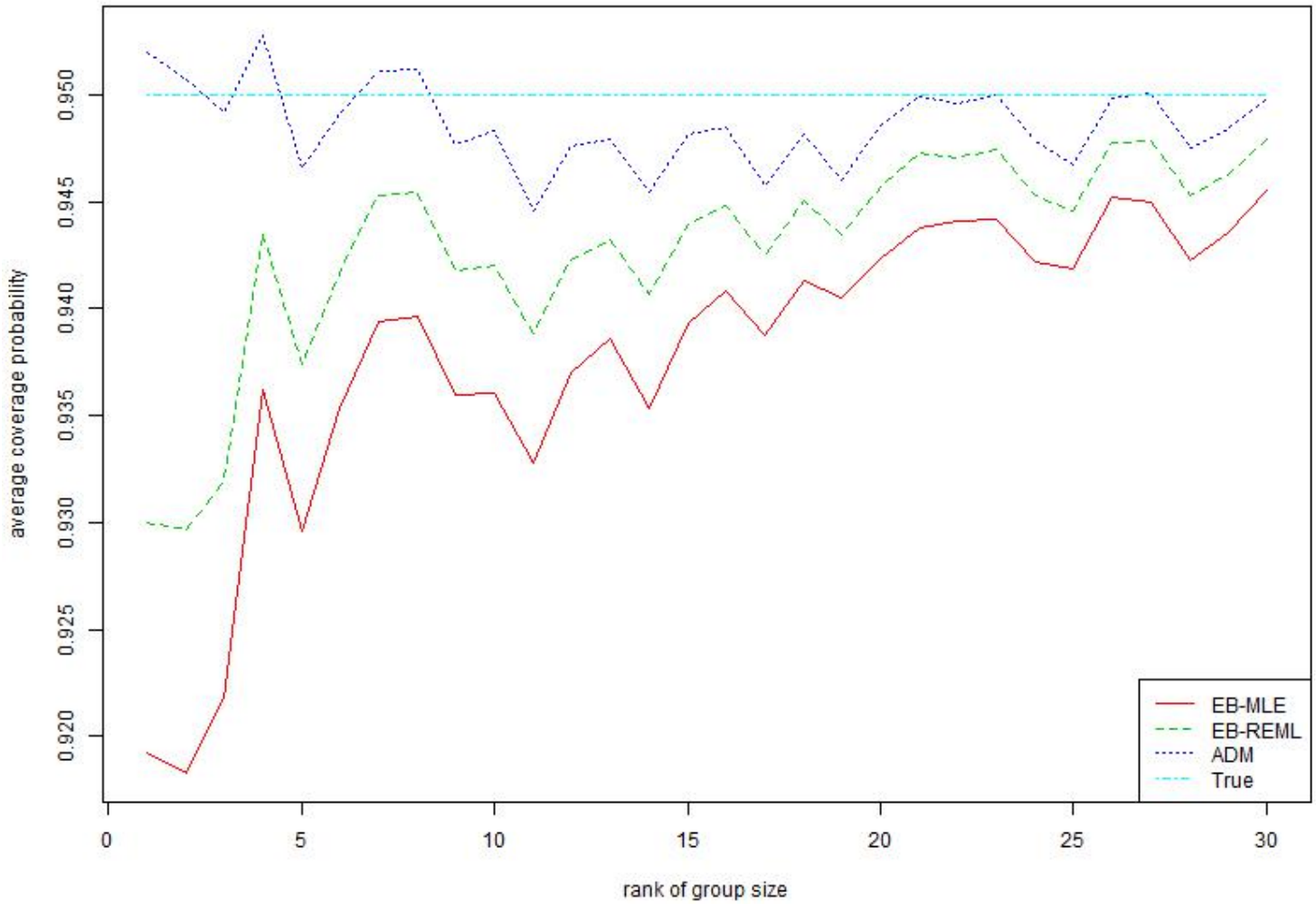
$$y_{ik} = \sum_{j \in s_i} d_{jk}.$$

Comparison of ADM and MCMC: Data Example

center

	Data						MCMC $\hat{\beta} = \begin{pmatrix} 2.62 & -0.21 \\ 1.10 & -0.19 \\ 0.52 & -0.02 \\ 1.54 & 0.07 \end{pmatrix}, \hat{r} = 43.975$					ADM $\hat{\beta} = \begin{pmatrix} 2.54 & -0.20 \\ 1.05 & -0.18 \\ 0.50 & -0.02 \\ 1.47 & 0.07 \end{pmatrix}, \hat{r} = 42.660$				
obs i	y_{i1}	y_{i2}	y_{i3}	y_{i4}	y_{i5}	x_{i1}	$\hat{p}_{i1}$	$\hat{p}_{i2}$	$\hat{p}_{i3}$	$\hat{p}_{i4}$	$\hat{p}_{i5}$	$\hat{p}_{i1}$	$\hat{p}_{i2}$	$\hat{p}_{i3}$	$\hat{p}_{i4}$	$\hat{p}_{i5}$
1	289	62	45	108	19	1	0.55 (0.02)	0.12 (0.01)	0.09 (0.01)	0.21 (0.02)	0.04 (0.01)	0.55 (0.02)	0.12 (0.01)	0.09 (0.01)	0.21 (0.02)	0.04 (0.01)
2	261	65	46	187	19	1	0.46 (0.02)	0.11 (0.01)	0.08 (0.01)	0.32 (0.02)	0.03 (0.01)	0.46 (0.02)	0.11 (0.01)	0.08 (0.01)	0.32 (0.01)	0.03 (0.01)
3	2	0	1	4	1	1	0.47 (0.09)	0.10 (0.05)	0.09 (0.05)	0.28 (0.08)	0.06 (0.04)	0.47 (0.08)	0.10 (0.05)	0.09 (0.04)	0.28 (0.08)	0.06 (0.04)
4	233	45	19	58	13	0	0.62 (0.02)	0.12 (0.02)	0.05 (0.01)	0.16 (0.02)	0.04 (0.01)	0.63 (0.02)	0.12 (0.02)	0.05 (0.01)	0.16 (0.02)	0.04 (0.01)
5	172	41	10	43	9	0	0.62 (0.03)	0.15 (0.02)	0.04 (0.01)	0.16 (0.02)	0.03 (0.01)	0.62 (0.03)	0.15 (0.02)	0.04 (0.01)	0.16 (0.02)	0.03 (0.01)

Monte Carlo Simulation Results



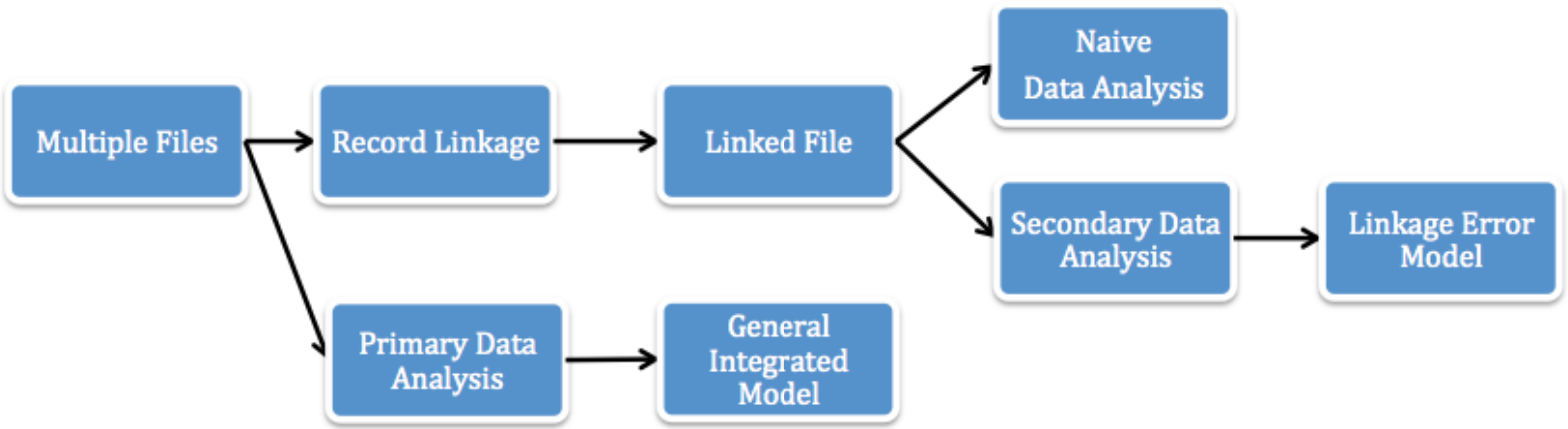
figureAverage coverage rate vs group index, N=30, K=5, q=2, 100 samples of the dataset.

Summary

1. Computation speed: 180 times faster than MCMC without sacrificing the accuracy of the estimates.
2. Estimates are the same when applied to the same dataset using the same model.
3. Capability to handle non-integer counts without rounding.
4. Eliminates the ill-behavior in MLE and REML estimates and corrects the bias in r estimate
5. Better operating characteristics

Probabilistic Record Linkage

Linkage errors are inevitable to occur during the record linkage process. It is important to propagate uncertainty of record linkage into the later estimation process.



figure

Development of statistical analysis with data from multiple files.

Probabilistic Record Linkage

Naive Statistical Analysis:

- Performed directly on linked data by simply ignoring linkage errors.
- Even a small amount of linkage errors can lead to significant biases of estimates (Neter et al. 1965).

Secondary Statistical Analysis, e.g., Chambers (2009):

- Performed directly on linked data but taking account linkage errors.
- Simplifying assumptions on the linkage mechanism has to be made due to limited information about the record linkage process:
  - Equal sizes of files
  - Complete and one-to-one linkage
  - Linkage Completely at Random (LCAR)
  - Requirement of a training sample

A Theoretical Frame: Han and Lahiri (2018)

- Situation: Observations of y and $\boldsymbol{x}$ are separately recorded in two files.
- Data Layout:
  - Sampled units in F_y is a subset of those in F_x (i.e., $S_y \in S_x$).
  - There is no duplicates in either F_y or F_x.

F_y

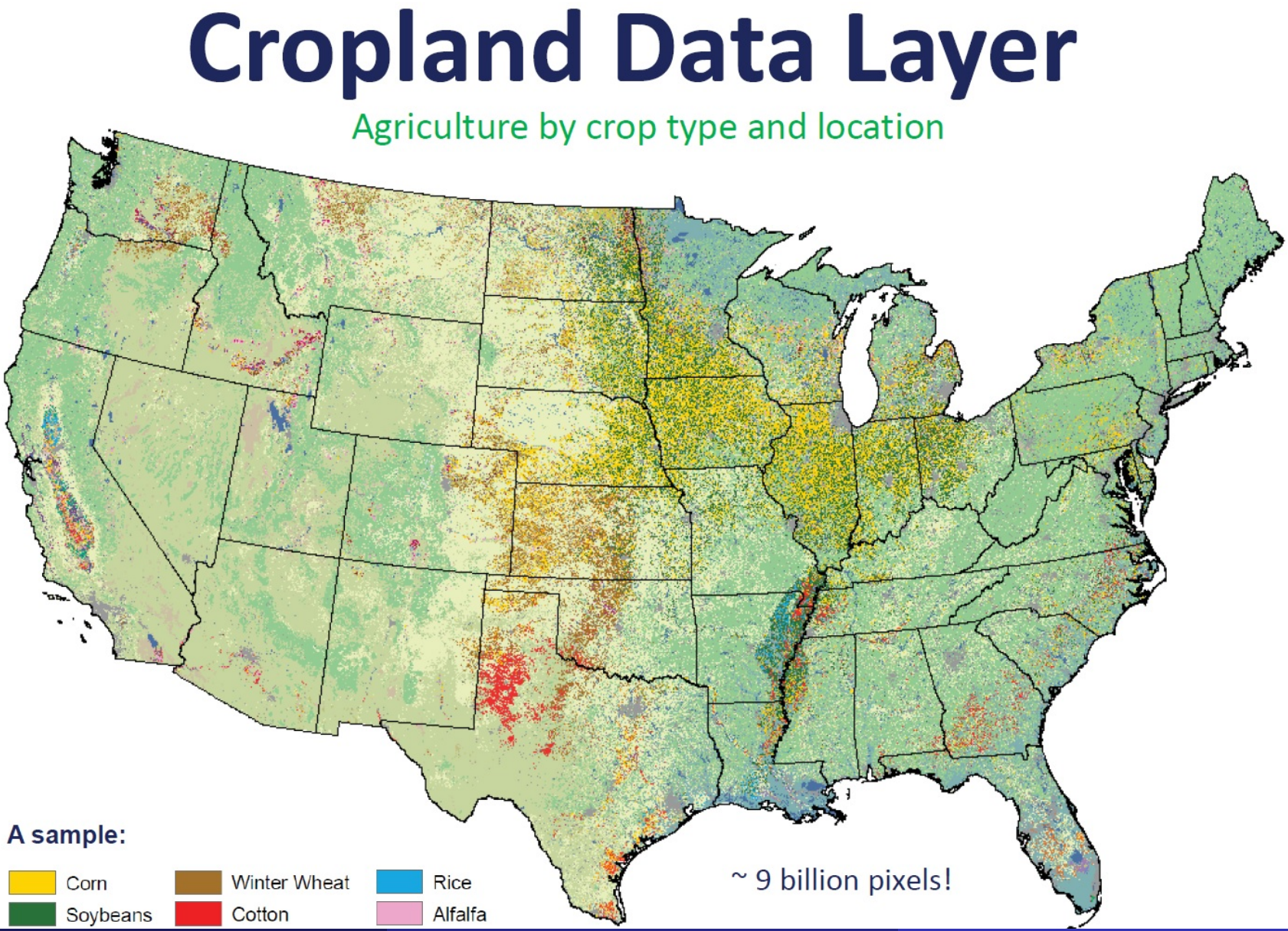
Label	$\tilde{\boldsymbol{w}}'$	$\tilde{y}$	$\tilde{\boldsymbol{x}}'$
1	$\tilde{\boldsymbol{w}}_1'$	$\tilde{y}_1$	$\tilde{\boldsymbol{x}}_1'$
...	...	...	...
j	$\tilde{\boldsymbol{w}}_j'$	$\tilde{y}_j$	$\tilde{\boldsymbol{x}}_j'$
...	...	...	...
n	$\tilde{\boldsymbol{w}}_n'$	$\tilde{y}_n$	$\tilde{\boldsymbol{x}}_n'$

F_x

Label	$\boldsymbol{w}'$	$\boldsymbol{x}'$	y
1	$\boldsymbol{w}_1'$	$\boldsymbol{x}_1'$	y_1
...	...	...	...
j'	$\boldsymbol{w}_{j'}'$	$\boldsymbol{x}_{j'}'$	$y_{j'}$
...	...	...	...
N	$\boldsymbol{w}_N'$	$\boldsymbol{x}_N'$	y_N

- Goal: To estimate the regression coefficient β in a regression model of y on $\boldsymbol{x}$.

A Few Remarks


Cropland Data Layer
Agriculture by crop type and location
A sample:
Corn
Soybeans
Winter Wheat
Cotton
Rice
Alfalfa
~ 9 billion pixels!

Scanner Data: Mango sales in grocery stores

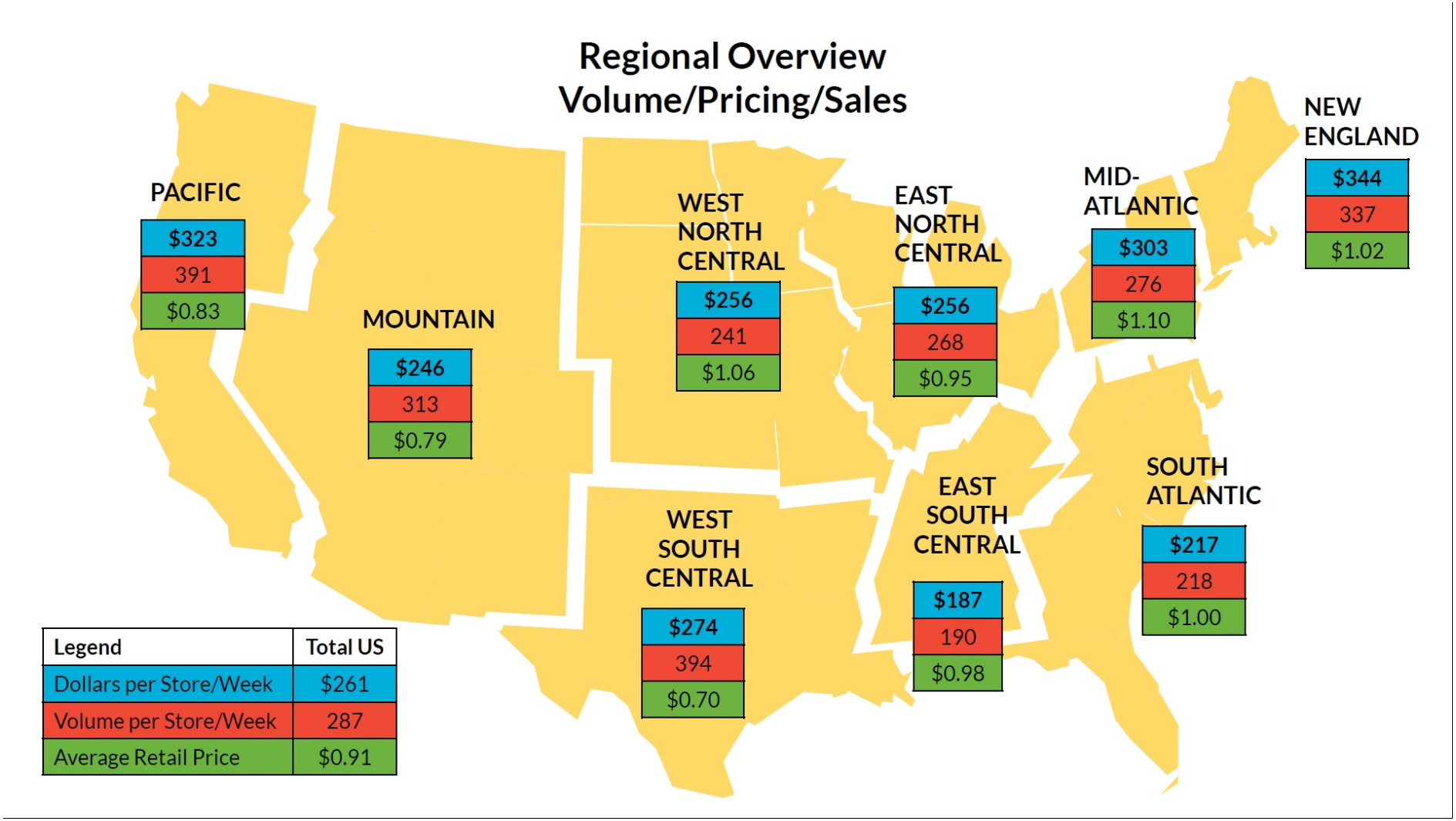


Figure: Scanner data of mango sales in grocery stores over different geographical regions; source: https://www.mango.org/wp-content/uploads/

GPS Probe Data Collection

The following figure (from FHWA, 1998) summarizes the collection of probe data

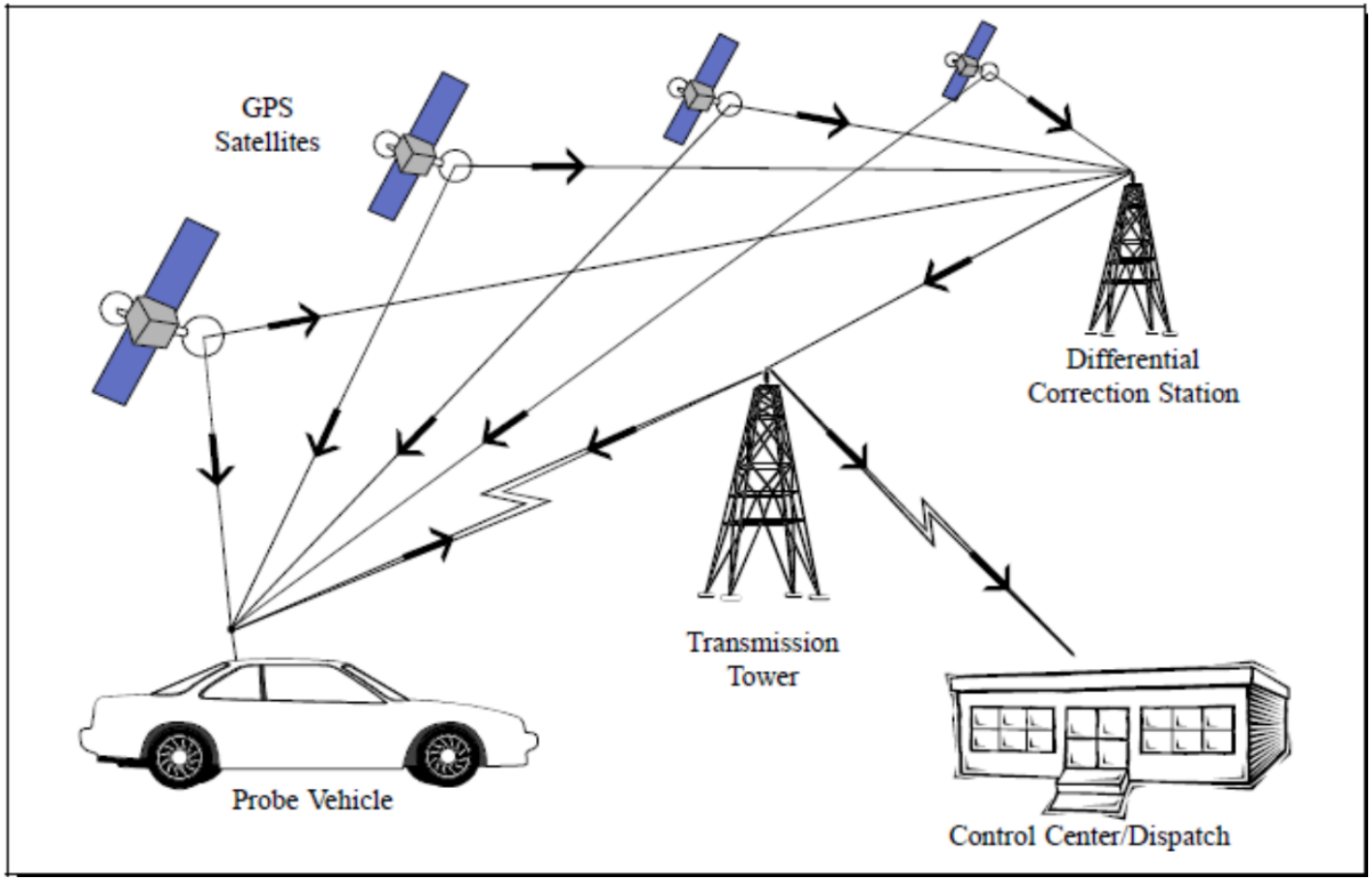


Figure: GPS Data Collection (FHWA, 1998; Source Kartika, C.S.D., 2015)

Sustainable Development Goals (SDG)

UN-SDG WEB Banner

The 17 Sustainable Development Goals (SDG)

- GOAL 1: No Poverty
- GOAL 2: Zero Hunger
- GOAL 3: Good Health and Well-being
- GOAL 4: Quality Education
- GOAL 5: Gender Equality
- GOAL 6: Clean Water and Sanitation
- GOAL 7: Affordable and Clean Energy
- GOAL 8: Decent Work and Economic Growth
- GOAL 9: Industry, Innovation and Infrastructure
- GOAL 10: Reduced Inequality
- GOAL 11: Sustainable Cities and Communities
- GOAL 12: Responsible Consumption and Production
- GOAL 13: Climate Action
- GOAL 14: Life Below Water
- GOAL 15: Life on Land
- GOAL 16: Peace and Justice Strong Institutions
- GOAL 17: Partnerships to achieve the Goal

Thank You!