Piecing Together a Gasoline Station Frame Using Third Party and Administrative Data
DRAFT

For

FCSM Policy Conference

December 6, 2016 | Washington, DC

By

Amerine Woodyard, Survey Statistician and Jeramiah Yeksavich, Survey Statistician

Background - Survey

- Motor Gasoline Price Survey (EIA-878)
  - Weekly
  - Mandatory
  - Confidential Information Protection and Statistical Efficiency Act (CIPSEA)
  - Multi-mode
- Same day collection, processing, and dissemination
- 276 published price estimates

Background

- Frame Update and Sample Redesign

- Frame (wish list of variables)
  - Name, address, phone, and county of gasoline stations
  - Sales volume by station
  - Latitude and longitude
  - Grades sold and ethanol content
  - Hypermarket or Big Box stations identified
  - Branded vs. unbranded
  - Station owner information

Background - Sources

- Reviewed 17 various federal, state, and private data sources
- Obtained via federal and state agencies, web scraping
- Inquired about data with private vendors, federal, state, and academic sources

Sources	
AggData	National Petroleum News
Census County Business Patterns	NAVTEC
Data.com	New Image Marketing
First Data / Palantir	Nielsen (NACS)
Gasbuddy	Oil Price Information Service (OPIS)
HSIP	SUNY-Albany / NETS
Infogroup	Underground Storage Tanks (UST)
Institute of Transportation Engineers	Yellow Pages
Mastercard	

Criteria for Evaluating Sources

Main Issue	Variable Evaluated
Measure of Size	Availability of volumetric data of gasoline sold by grade
Coverage	Availability of fields
	State counts
	Big box stations
Expense	Cost
	Level of effort
Age of the Data	Frequency of updates

Measure of Size

Source	Total Outlets	Pros	Cons
National Petroleum News	156,065	Report with state tax rates and net motor fuels revenue	Data too aggregated
Underground Storage Tanks (UST)	NA	• Tank capacity • Station owner information	• Cannot find sales volume correlation • No uniformity of variables • Station owner info for 30 states
Institute of Transportation Engineers	NA	Total number of trips anticipated to be generated by gas station	Uses multiple years of data and is not stratified by year or by jurisdiction type
Mastercard	150,764	Gasoline demand	• Gasoline demand too aggregated • Cannot disclose station level info
First Data/Palantir	129,551	Total fuel sales at possibly county level	• Fuel sales include other fuels • Undercoverage of cash sales • Cannot disclose station level info

Criteria for Evaluating Sources

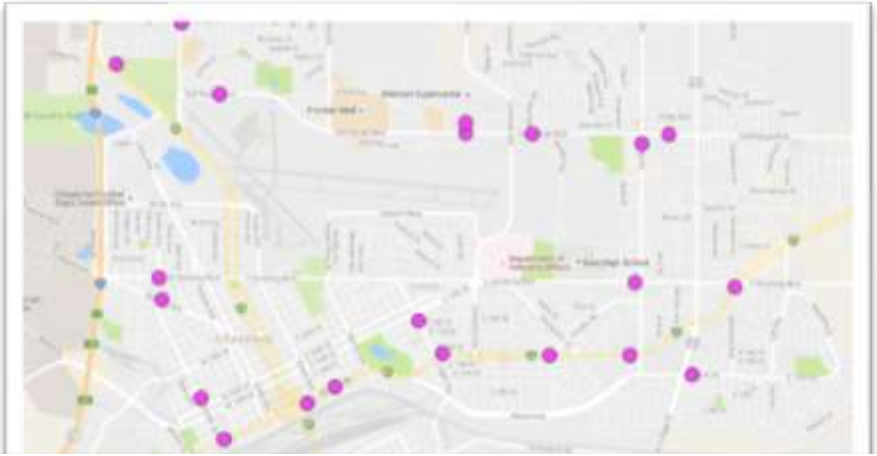
Main Issue	Variable Evaluated
Measure of Size	Availability of volumetric data of gasoline sold by grade
Coverage	Availability of fields
	State counts
	Big box stations
Expense	Cost
	Level of effort
Age of the Data	Frequency of updates

Coverage

- Compiled information from Federal and industry sources
- Used counts of gasoline stations by state
- Contacted vendors, scraped, or copy/paste from websites
  - Availability of our wish list variables
  - State counts
  - Sample file for 4 states
  - List of stations from a handful of companies
- Geospatial analysis on Wyoming stations

Coverage - Geospatial Analysis

- Entries from 8 Databases for Wyoming as a test case were geocoded using Google's protocol
- Most accurate locations ("Hits") were tallied for comparison
- Accurate location hit rates ranged from 80% to 94%



Wyoming Databases
Geocoded Addresses Summary Results By Database

	Yellow Pages		Gas Buddy		Data.Com		Infogroup		New Image Market		Navteq		OPIS		UST
	Orig. # Obs: 360		Orig. # Obs: 505		Orig. # Obs: 370		Orig. # Obs: 428		Org.# Obs: 395		Orig. # Obs: 354		Org.# Obs: 381		Orig. # Obs: 1,797
	Percent of Obs. Coded: 99.4%		Percent of Obs. Coded: 99.6%		Percent of Obs. Coded: 99.7%		Percent of Obs. Coded: 100%		Percent of Obs. Coded: 100%		Percent of Obs. Coded: 100%		Percent of Obs. Coded: 100%		
	Clear Error Points: 1		Clear Error Points: 3		Clear Error Points: 8		Clear Error Points: 9		Clear Error Points: 4		Clear Error Points: 9		Clear Error Points: 4		
Google Geocode Type	Count	Percent	Count	Percent	Count	Percent	Count	Percent	Count	Percent	Count	Percent	Count	Percent	Multiple Geocoding Results At Same Address Due To Sites With Multiple Tanks
Rooftop	227	63.4%	287	57.1%	166	44.99%	243	56.8%	230	58.2%	215	60.7%	227	59.6%	
Range Interpolated	110	30.7%	174	34.6%	131	35.50%	133	31.1%	137	34.7%	105	29.7%	130	34.1%	
Geometric-Center	7	2.0%	26	5.2%	25	6.78%	24	5.6%	12	3.0%	25	7.1%	10	2.6%	
Approximate	14	3.9%	16	3.2%	47	12.74%	28	6.5%	16	4.1%	9	2.5%	14	3.7%	
Total	358	100.0%	503	100.0%	369	100.00%	428	100.0%	0	100.0%	0	0.0%	0	0.0%	
Rooftop + Range Interpolated	337	94.1%	461	91.7%	297	80.5%	376	87.9%	242	92.9%	320	90.4%	237	93.7%	

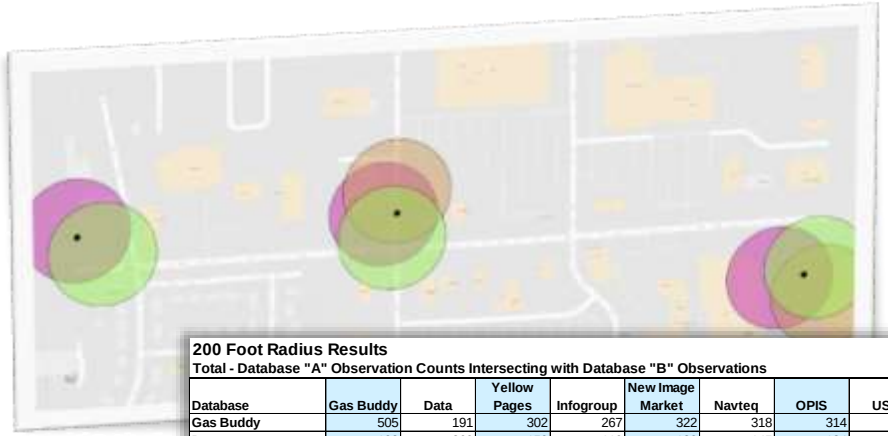
Coverage - Geospatial Analysis (cont.)

Wyoming Databases

Geocoded Addresses Summary Results By Database

	Yellow Pages		Gas Buddy		Data.Com		Infogroup		New Image Market		Navteq		OPIS		UST	
	Orig. # Obs: 360		Orig. # Obs: 505		Orig. # Obs: 370		Orig. # Obs: 428		Org.# Obs: 395		Orig. # Obs: 354		Org.# Obs: 381		Orig. # Obs: 1,797	
	Percent of Obs. Coded: 99.4%		Percent of Obs. Coded: 99.6%		Percent of Obs. Coded: 99.7%		Percent of Obs. Coded: 100%		Percent of Obs. Coded: 100%		Percent of Obs. Coded: 100%		Percent of Obs. Coded: 100%			
	Clear Error Points: 1		Clear Error Points: 3		Clear Error Points: 8		Clear Error Points: 9		Clear Error Points: 4		Clear Error Points: 9		Clear Error Points: 4		Multiple Geocoding Results At Same Address Due To Sites With Multiple Tanks	
Google Geocode Type	Count	Percent	Count	Percent	Count	Percent	Count	Percent	Count	Percent	Count	Percent	Count	Percent		
Rooftop	227	63.4%	287	57.1%	166	44.99%	243	56.8%	230	58.2%	215	60.7%	227	59.6%		
Range Interpolated	110	30.7%	174	34.6%	131	35.50%	133	31.1%	137	34.7%	105	29.7%	130	34.1%		
Geometric-Center	7	2.0%	26	5.2%	25	6.78%	24	5.6%	12	3.0%	25	7.1%	10	2.6%		
Approximate	14	3.9%	16	3.2%	47	12.74%	28	6.5%	16	4.1%	9	2.5%	14	3.7%		
Total	358	100.0%	503	100.0%	369	100.00%	428	100.0%	0	100.0%	0	0.0%	0	0.0%		
Rooftop + Range Interpolated	337	94.1%	461	91.7%	297	80.5%	376	87.9%	242	92.9%	320	90.4%	237	93.7%		

Coverage – Geospatial Analysis (cont.)



200 Foot Radius Results

Total - Database "A" Observation Counts Intersecting with Database "B" Observations

Database	Gas Buddy	Data	Yellow Pages	Infogroup	New Image Market	Navteq	OPIS	UST
Gas Buddy	505	191	302	267	322	318	314	1,231
Data	193	369	159	119	130	145	134	449
Yellow Pages	329	172	358	240	252	252	259	913
Infogroup	265	119	213	428	208	211	214	718
New Image Market	352	141	243	234	395	273	328	937
Navteq	359	148	256	237	279	354	269	962
OPIS	369	147	260	247	337	272	381	1,030
UST	395	125	275	240	280	263	303	1,789

Percent- Database "A" Observation Counts Intersecting with Database "B" Observations

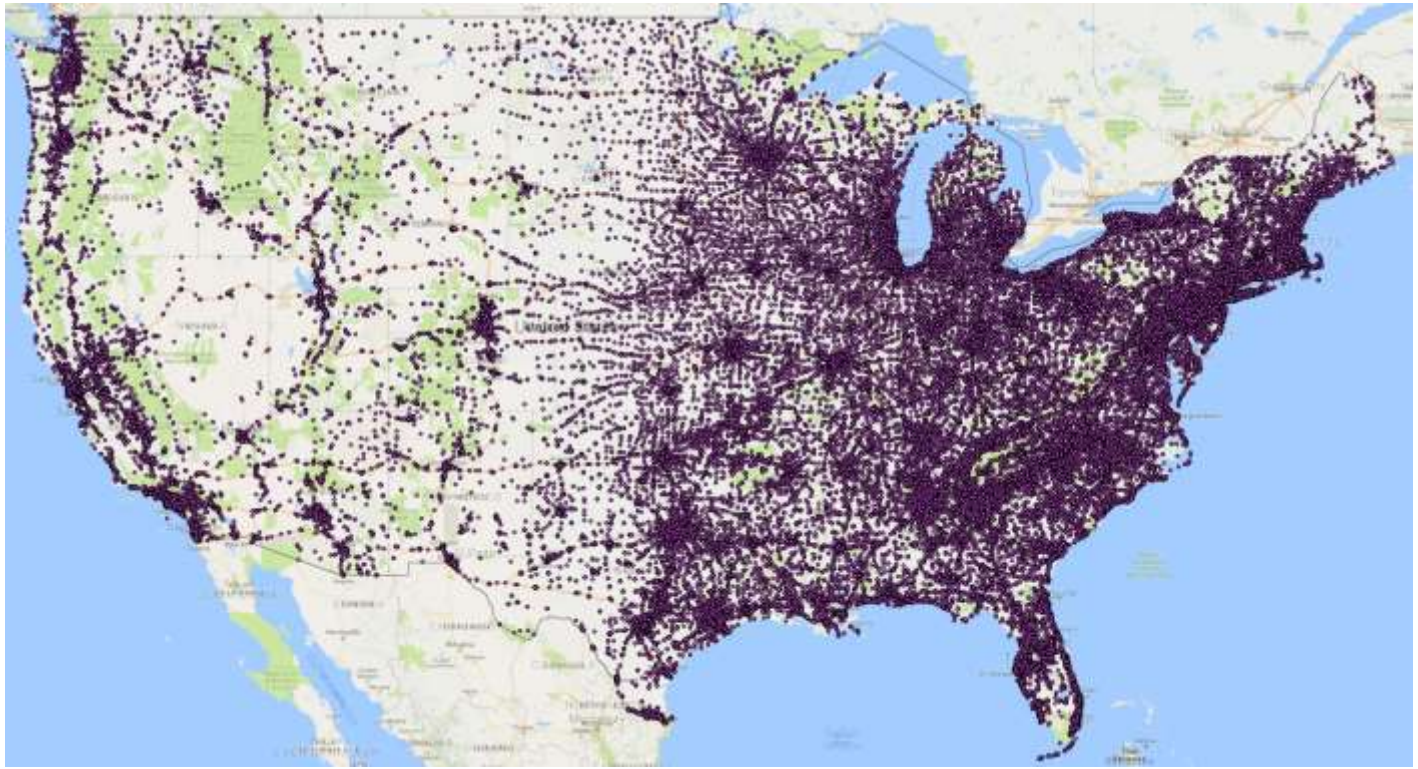
Database	Gas Buddy 505 Obs.	Data 369 Obs.	Yellow Pages 358 Obs.	Infogroup 428 Obs.	New Image Market 395 Obs.	Navteq 354 Obs.	OPIS 381 Obs.	UST 1,789 Obs.
Gas Buddy	100%	52%	84%	62%	82%	90%	82%	69%
Data	38%	100%	44%	28%	33%	41%	35%	25%
Yellow Pages	65%	47%	100%	56%	64%	71%	68%	51%
Infogroup	52%	32%	59%	100%	53%	60%	56%	40%
New Image Market	70%	38%	68%	55%	100%	77%	86%	52%
Navteq	71%	40%	72%	55%	71%	100%	71%	54%
OPIS	73%	40%	73%	58%	85%	77%	100%	58%
UST	78%	34%	77%	56%	71%	74%	80%	100%

- Buffer areas of 100 to 200 feet were created around provided GPS or geocoded coordinates
- Overlapping location points were compared each to each in terms of total matching overlaps and percentage of overlap
- Matching Rates Range:
  - 100 Foot Buffer: 13% to 84%
  - 200 Foot Buffer: 28% to 90%
- Overall process allowed for the elimination of 3 databases as primary sources

Expense and Age of Data

Source	Cost	Level of effort	Frequency of Updates
Census County Business Patterns	NA	NA	NA
Data.com	Medium	High	Few times a year
Gasbuddy	Low	Medium	Ad hoc
Infogroup	Medium	Low	Daily
NAVTEQ	Low	High	Unknown
New Image Marketing	Medium	High	2015 or 2016
Nielsen (NACS)	High	Low	Daily
Oil Price Information Service (OPIS)	Medium	Low	Daily
Underground Storage Tanks (UST)	Low	High	Unknown/varies
Yellow Pages	Low	High	Unknown

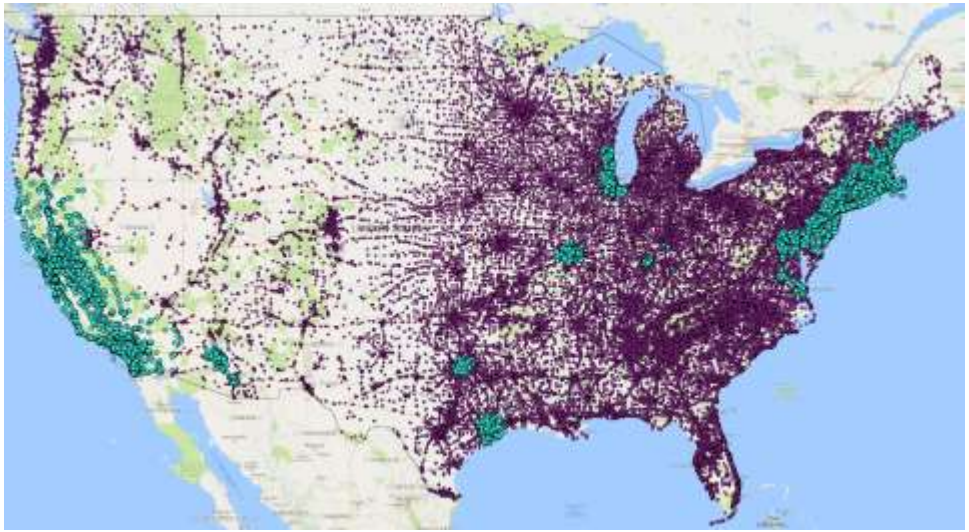
OPIS Database (133,616 Initial Observations)



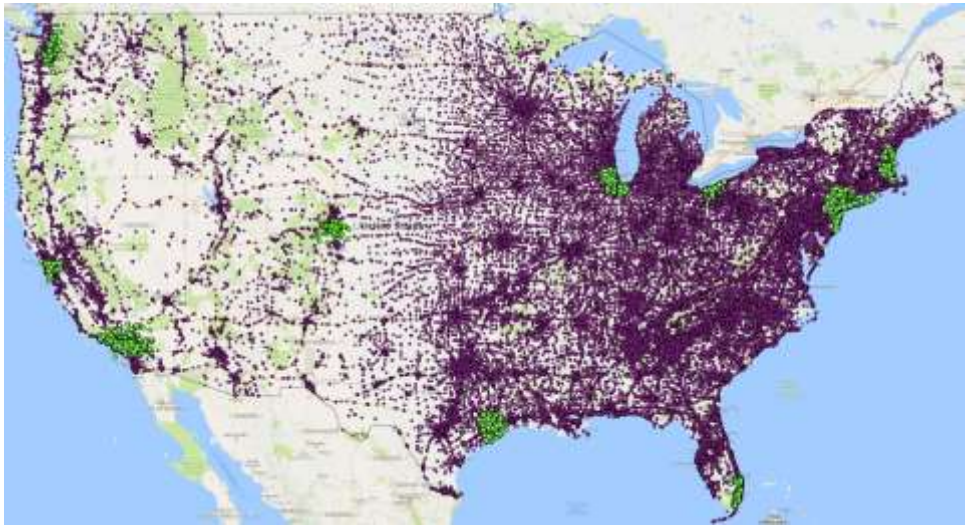
Auxiliary Files

- Yellow Pages – fill in missing phone numbers
- Underground Storage Tank – station owner information
- Geocoding
  - Used to identify stations within the reformulated gasoline areas as defined by the EPA
  - Used to tag stations for the 10 published cities

Geospatial Analysis – Reformulated Gas Area Definitions



Geospatial Analysis – City Area Definitions



Challenges

- Finding measure of size
- Difficulty getting enough information from private vendors
- Changing requirements
- Merging files

Ongoing and Future Work

- Merging information from Underground Storage Tank database
- Obtaining additional station owner information
- States' Weights and Measures Office

Acknowledgements

Nanda Srinivasan	Tammy Heppner
Nathan Agbemenyale	Andrew Thomson
Maura Bardos	Dan Walzer
Ruey-Pyng Lu	Jeramiah Yeksavich
Renee Miller	Bin Zhang
Marcela Rourk	Shala Brown
Cecile Sano	Julian Castillo
Lou Schloss	Eliza Goren
Michael Scott	Ethan Walker-Seim