Best Practices for Integration of Alternative Data Sources into Petroleum and Biofuels Statistics

For

FedCASIC Workshop

May 4, 2016 | Suitland, Maryland

By

Shawna Waugh, Mathematical Statistician, Office of Petroleum and Biofuels Statistics

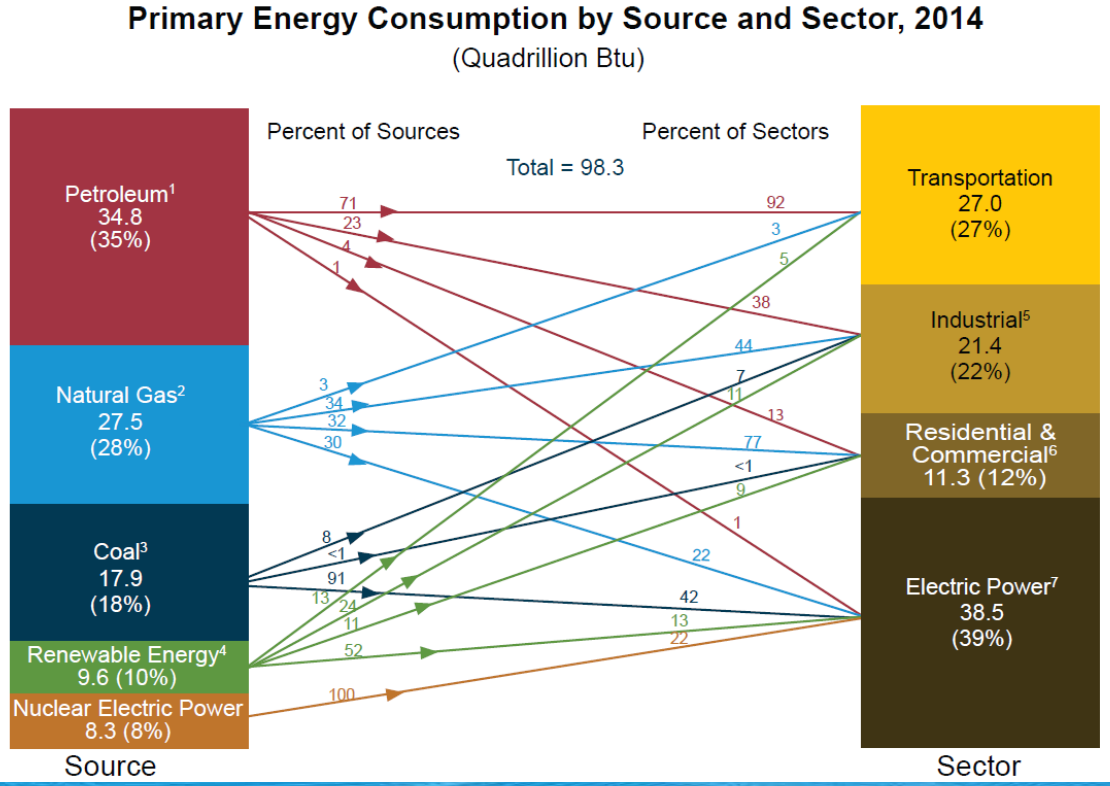
Roadmap

1. U.S. Energy Information Administration and the Office of Petroleum and Biofuels Statistics
   - U.S. petroleum supply chain
   - Petroleum supply data collection programs
   - A table displaying “shades of estimation” from EIA’s weekly petroleum supply publication
2. Motivation for a study on the integration of survey and non-survey data
3. Phases of the Generic Statistical Business Process Model (GSBPM)
4. Focus on two examples of uses of administrative and third-party data
   - Crude oil movements by rail
   - Monthly and weekly crude oil imports
5. Next steps

U.S. Energy Information Administration

Motto: We help people understand energy

Mission: The U.S. Energy Information Administration (EIA) collects, analyzes, and disseminates independent and impartial energy information to promote sound policymaking, efficient markets, and public understanding of energy and its interaction with the economy and the environment.

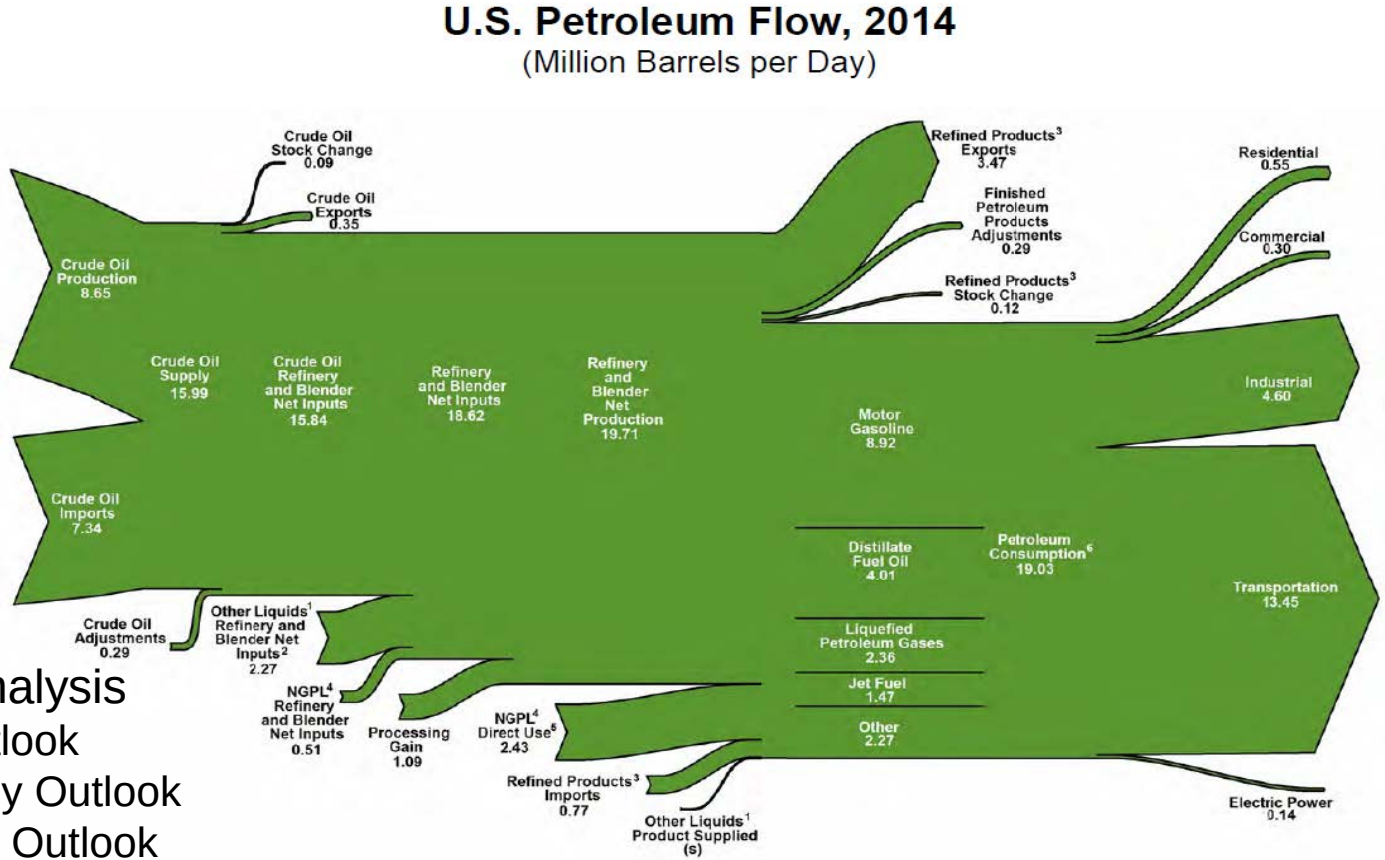


Office of Petroleum and Biofuels Statistics

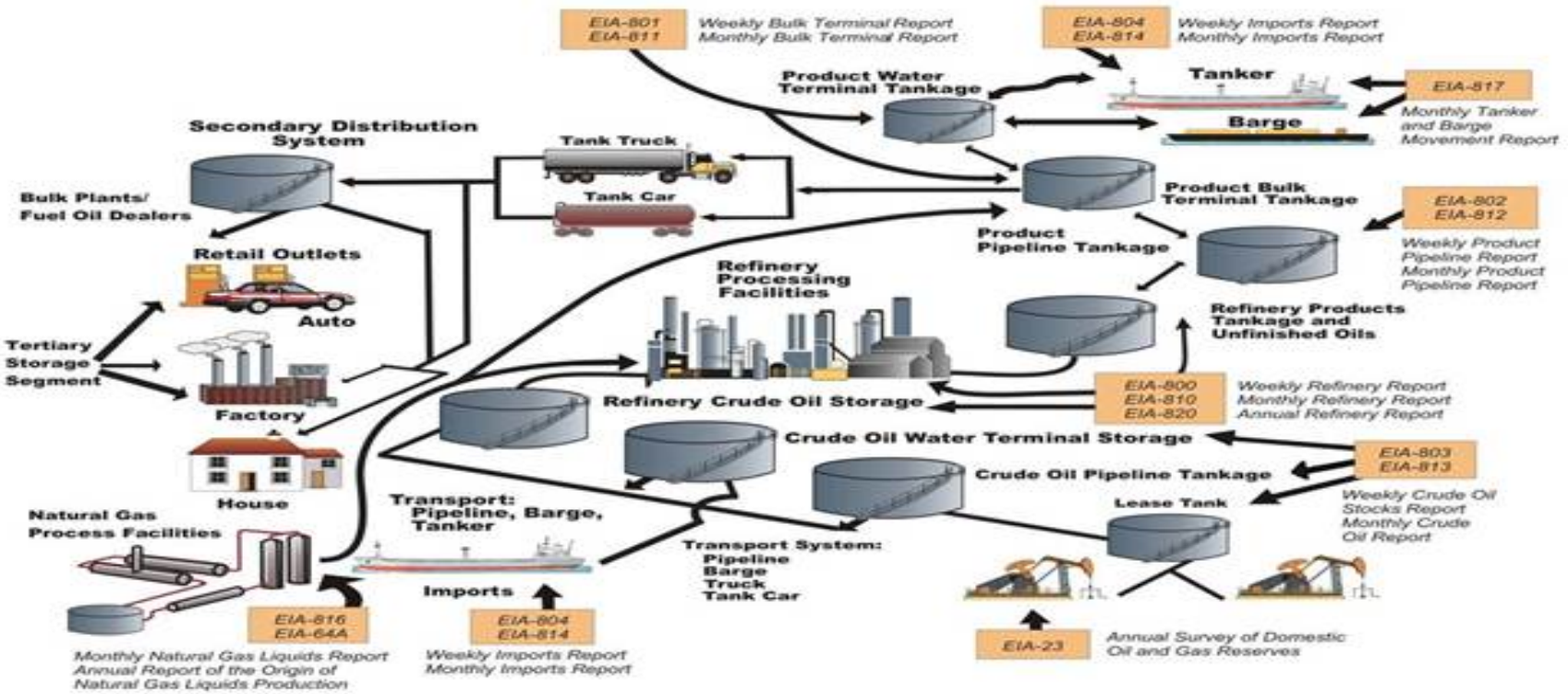
Family of surveys
- 9 weekly
- 13 monthly
- 2 annual

PBS publications
- 3 weekly
- 5 monthly
- 4 annually

Inputs for energy analysis
- Annual Energy Outlook
- International Energy Outlook
- Short-Term Energy Outlook



Petroleum Supply Data Collection



Many Shades of Estimation

Table 1. U.S. Petroleum Balance Sheet, Week Ending 10/12/2012

Petroleum Stocks (Million Barrels)	Current Week	Week Ago			Year Ago		
	10/12/2012	10/5/2012	Difference	Percent Change	10/14/2011	Difference	Percent Change
Crude Oil	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Commercial (Excluding SPR)[1]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Strategic Petroleum Reserve (SPR)[2]	694.952	694.952	0	0	695.951	-0.999	-0.1
Total Motor Gasoline	197.128	195.408	1.72	0.9	206.271	-9.143	-4.4
Reformulated	0.161	0.069	0.092	133.3	0.138	0.023	16.7
Conventional	47.451	47.521	-0.07	-0.1	55.174	-7.723	-14
Blending Components	149.516	147.818	1.698	1.1	150.959	-1.443	-1
Fuel Ethanol	18.989	19.256	-0.267	-1.4	17.053	1.936	11.4
Kerosene-Type Jet Fuel	42.974	44.126	-1.152	-2.6	46.452	-3.478	-7.5
Distillate Fuel Oil[3]	118.664	120.882	-2.218	-1.8	149.739	-31.075	-20.8
15 ppm sulfur and Under	85.831	89.297	-3.466	-3.9	98.18	-12.349	-12.6
> 15 ppm to 500 ppm sulfur	6.182	5.778	0.404	7	13.243	-7.061	-53.3
> 500 ppm sulfur[3]	26.651	25.807	0.844	3.3	38.317	-11.666	-30.4
Residual Fuel Oil	34.107	34.044	0.063	0.2	33.095	1.012	3.1
Propane/Propylene	74.628	75.896	-1.268	-1.7	58.929	15.699	26.6
Other Oils[4]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Unfinished Oils	83.928	84.097	-0.169	-0.2	84.454	-0.526	-0.6
Total Stocks (Including SPR)[1,2]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Total Stocks (Excluding SPR)[1]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Petroleum Supply (Thousand Barrels per Day)	Current Week	Week Ago		Year Ago		Four Week Averages Week Ending			Cumulative Daily Average		
	10/12/2012	10/5/2012	Difference	10/14/2011	Difference	10/12/2012	10/14/2011	Percent Change	10/12/2012	10/14/2011	Percent Change
Crude Oil Supply											
(1) Domestic Production[5]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
(2) Alaska	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
(3) Lower 48	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
(4) Net Imports (Including SPR)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
(5) Imports	8,347	8,221	126	7,921	426	8,067	8,852	-8.9	8,801	8,896	-1.1
(6) Commercial Crude Oil	8,347	8,221	126	7,921	426	8,067	8,852	-8.9	8,801	8,896	-1.1
(7) Imports by SPR	0	0	0	0	0	0	0	–	0	0	–
(8) Imports into SPR by Others	0	0	0	0	0	0	0	–	0	0	–
(9) Exports	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
(10) Stock Change (+/build; -/draw)	409	239	170	-676	1085	58	-219	--	130	-114	--
(11) Commercial Stock Change	409	239	170	-676	1085	58	-220	--	134	-7	--
(12) SPR Stock Change	0	0	0	-1	1	0	0	--	-4	-107	--
(13) Adjustment[6]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
(14) Crude Oil Input to Refineries	14,819	14,749	70	14,407	412	14,760	14,807	-0.3	14,948	14,714	1.6
Other Supply											
(15) Production	4,270	4,268	2	4,208	62	4,278	4,187	2.2	4,388	4,107	6.8
(16) Natural Gas Plant Liquids[7]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
(17) Renewable Fuels/Oxygenate Plant	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
(18) Fuel Ethanol	797	800	-3	908	-110	798	868	-8.1	870	890	-2.3
(19) Other[8]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
(20) Refinery Processing Gain	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
(21) Net Imports[9]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
(22) Imports[9]	2,197	1,532	665	1,413	784	1,942	1,851	4.9	2,004	2,373	-15.6
(23) Exports[9]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
(24) Stock Change (+/build; -/draw)[3,10]	-791	-863	72	-885	94	-491	-346	--	32	2	--
(25) Adjustment[11]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Products Supplied											
(26) Total[12]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
(27) Finished Motor Gasoline[13]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
(28) Kerosene-Type Jet Fuel	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
(29) Distillate Fuel Oil	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
(30) Residual Fuel Oil	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
(31) Propane/Propylene	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
(32) Other Oils[14]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Net Imports of Crude and Petroleum Products											
(33) Total	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Independent Statistics & Analysis
U.S. Energy Information Administration

Table 1: U.S. Petroleum Balance Sheet: Week Ending 10/12/2012 from the Weekly Petroleum Status Report (WPSR)

Estimates from weekly sample surveys

Estimates from the *Petroleum Supply Monthly*

– Natural Gas Plant Liquids Production

– Other oils (partially estimated)

Data from other EIA sources

– Domestic Production (*model from survey & other data*)

– Product Supplied (derived)

Modeled data from Census Bureau data

– Exports

Source: American Statistical Association Committee on Energy Statistics Committee Meeting, Washington, D.C., November 2, 2012

Motivation for study on the integration of survey and non-survey data

People

- Customers
- Survey respondents

Process

- Generic Statistical Business Process Model (GSBPM)
- Upcoming OMB clearance of Petroleum Marketing and Petroleum Supply Programs

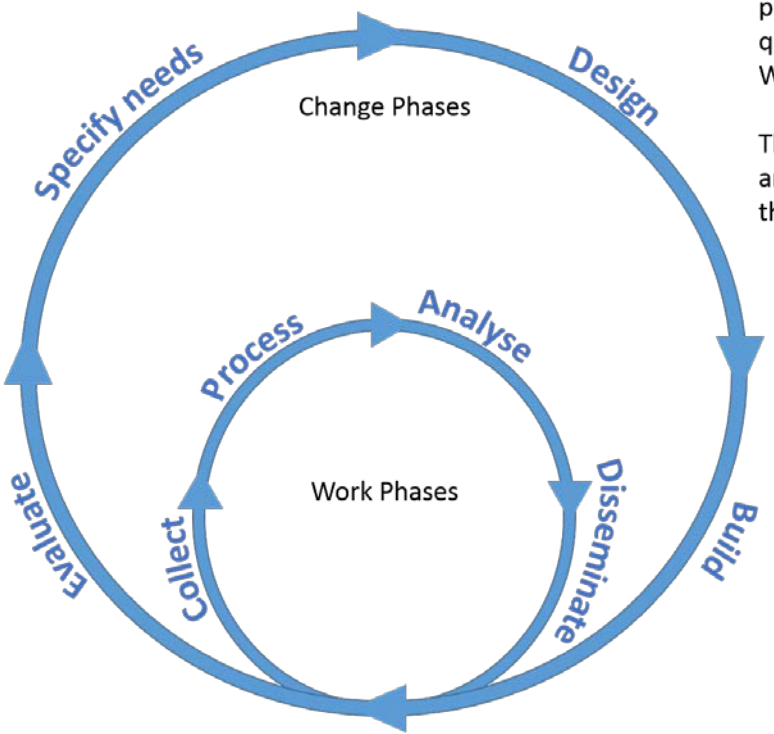
Standards

- OMB's Circular M-14-06
- OMB's Statistical Guidelines

Technology

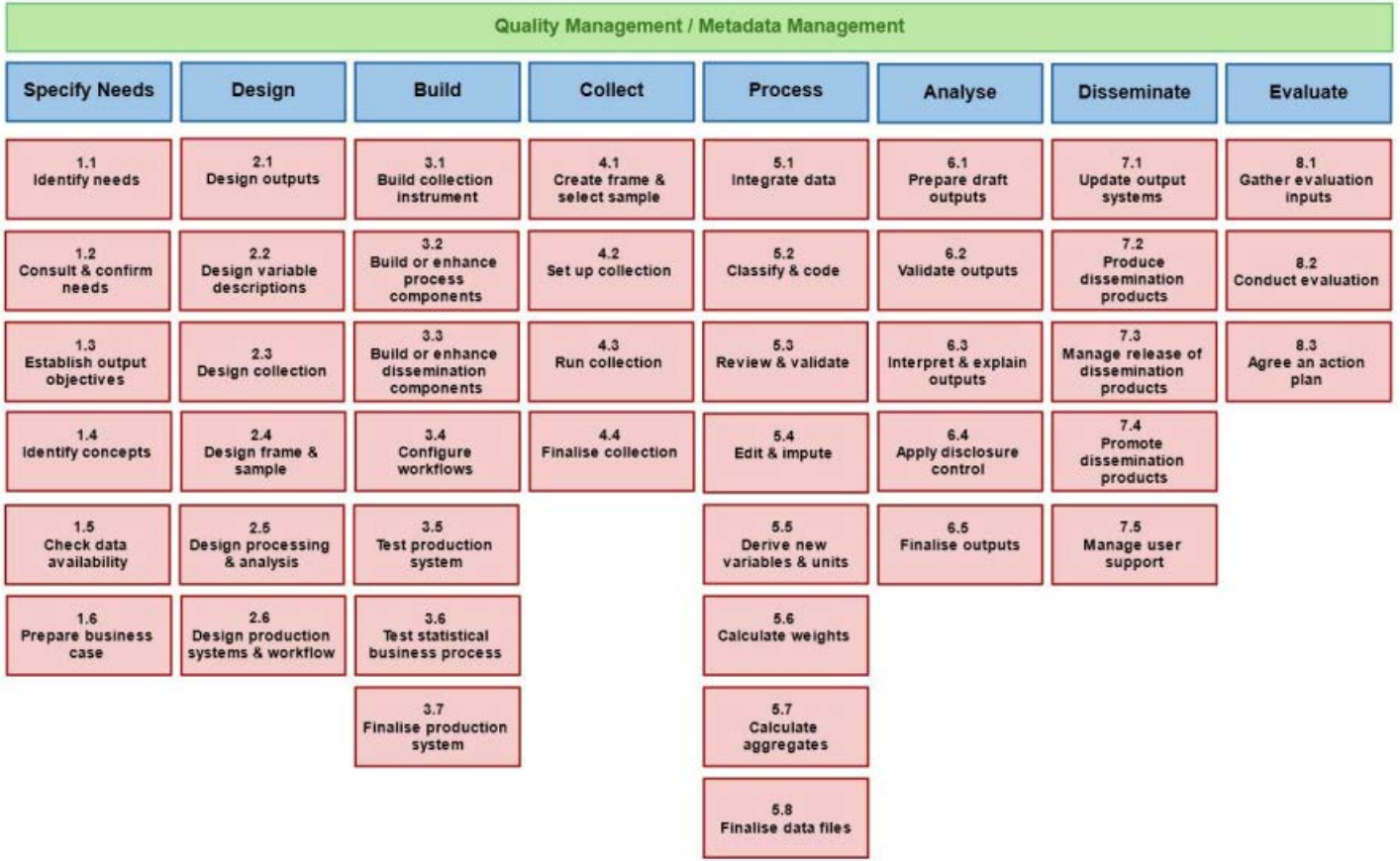
- Cost-effective
- Standardization - consistency within and across PBS programs
- Transformation of systems

GSBPM change vs. work (survey operations) cycles

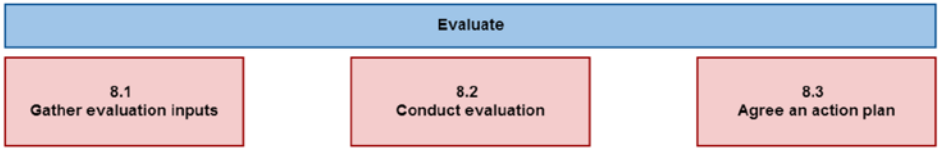


In GSBPM, there are some phases which are undertaken quickly and frequently – the Work Phases.

There are other phases which are undertaken less often - the Change Phases.



Source: United Nations Economic Commission for Europe, GSBPM (version 5.0), December, 2013



Assess quality of estimates: Crude oil movements by rail

Data Sources	Example	Accuracy	Complete	Cost	Timeliness
Survey (gov't) - census	EIA-817 Monthly	High	100%	$$$	Very timely
Administrative data (gov't)	Surface Transportation Board (Waybill)	Inconsistencies over time and across companies	80%	$	Timely, some delays in reporting

Best Practices

√ Use OMB tool to assess information quality of administrative and third-party data sources

√ Use other data sources to assess quality of estimates

√ Assess pros and cons of all options – survey, non-survey, or both

Standards

OMB Standards and Guidelines for Statistical Surveys, Standard 3.5

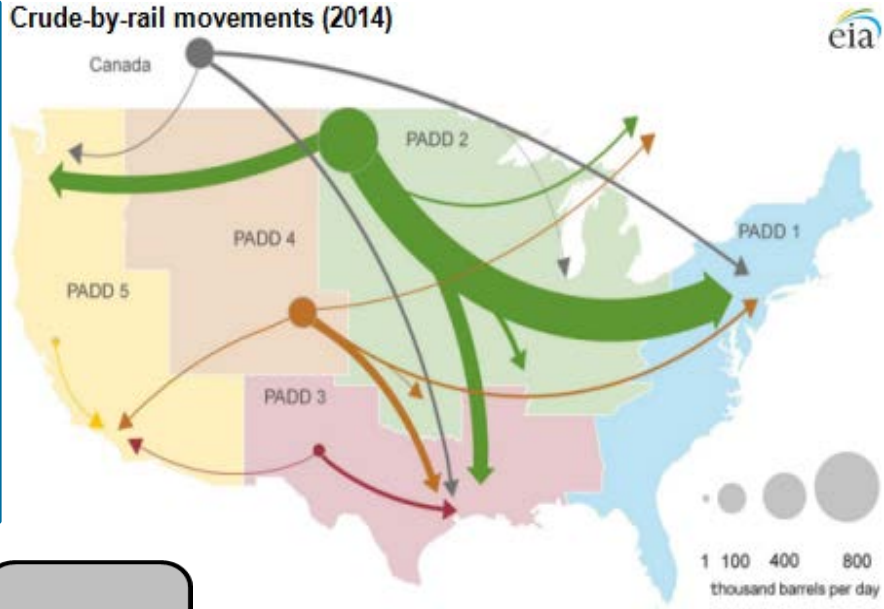
Specify Needs					
1.1 Identify needs	1.2 Consult & confirm needs	1.3 Establish output objectives	1.4 Identify concepts	1.5 Check data availability	1.6 Prepare business case

Crude oil movements by rail

Best practices

√ Identify needs, including changes in the industry
√ Identify "similar" data sources
√ Prepare business case (template) of options
√ Prioritize among potential improvements

OMB Clearance Process (Paperwork Reduction Act)



Codes and definitions

Geographic codes

- Country
- Census regions and districts
- PADDs, Sub-PADDs, and Refining Districts
- Port of origin
- State

Other codes

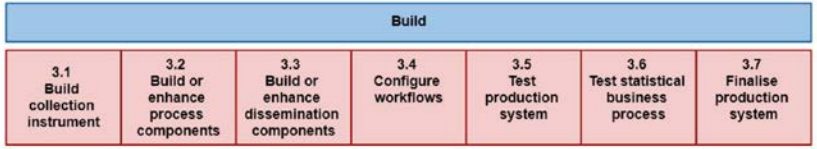
- PBS Crude streams
- PBS Product codes
- IRS Terminal Control Number (TCN)

Best practices

√ Design or modify codebook so codes are used consistently across surveys in agency programs
√ Develop crosswalk between EIA and other U.S. and International codes – i.e., Harmonized System
√ Develop and routinely update codebook and crosswalks

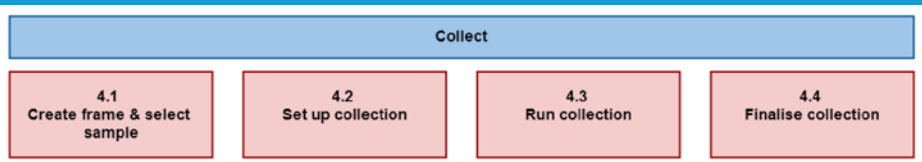
Standards

OMB Standards and Guidelines for Statistical Surveys, Standards 1.1 – 1.4
EIA Business Standards



Data integration and standardization

Today	Future Best Practices
Manual interventions and processes	Automated to the extent possible
Over 39 separate applications	Integrated collection/analysis system
Disparate survey-specific processes	Standardized survey processes
Insufficient integration and coordination across survey programs	Increased collaboration and integration of survey planning
Limited documentation on processing and analysis systems	Transparency on process & analysis systems
Separate operations for surveys	Coordinated, streamlined applications
Limitations responding to needs	Proactive customer outreach



Frame maintenance and sample selection

Data Sources	Example	Accuracy	Completeness	Cost	Timeliness
Survey (gov't) - census	EIA-815 Terminals EIA-816-NG plants	High	100%	$$$$$	Timely (annually)
Administrative data (gov't)	IRS' Terminal Control Number	High	100%	$	Very timely
Third-party data (private)	Natural Gas Plant Almanac	High	100%	$	Timely (annually)

Best practices

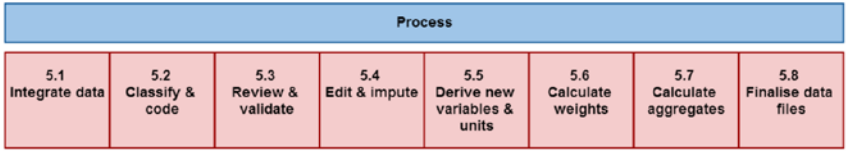
√ Record linkage for building frame – integrate data from multiple sources

√ Use IDs to monitor change in ownership – mergers, acquisitions, and divestures

√ Regularly update frame with other (industry) sources prior to sample selection

Standards

OMB Standards and Guidelines for Statistical Surveys, Standards 2.1 – 2.3



Micro-level editing and imputation

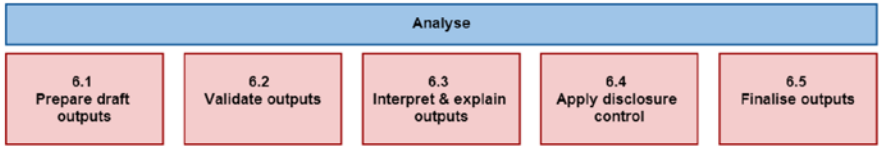
Data Sources	Example	Accuracy	Completeness	Cost	Timeliness
Survey (gov't) - sample	EIA-804 (Weekly)	High	90% cut-off sample	$$$	Very timely, some delays
Survey (gov't) - census	EIA-814 (Monthly)	Very high	Census	$$$$$	Timely
Administrative (gov't)	Census (CBP)	High	Census	$	Very timely, some delays

Best practices

√ Compare survey and administrative data from Census
√ Develop concordance between EIA codes and Harmonized System codes
√ Use port of entry codes maintained by Customs and Border Patrol

Standards

OMB Standards and Guidelines for Statistical Surveys, Standards 3.1 – 4.1



Macro-level validation

Data Sources	Example	Accuracy	Completeness	Cost	Timeliness
Survey (gov't) - sample	EIA-804 (Weekly)	High	90% cut-off sample	$$$	Very timely, some delays
Survey (gov't) - census	EIA-814 (Monthly)	Very high	Census	$$$$$	Timely
Administrative (gov't)	Census (CBP)	High	Census	$	Timely, some delays

Best practices

√ Conduct evaluation of models periodically

√ Develop quality control charts to monitor differences between data sources

√ Understand how delays in processing either survey or other data impacts results

Standards

OMB Standards and Guidelines for Statistical Surveys, Standards 5.1 – 6.1

Integrate estimates from other data sources

Data Sources	Example	Accuracy	Complete	Cost	Timeliness
Survey (gov't) - census	EIA-817	Very high	100%	$$$$	Timely (monthly)
Administrative (gov't)	Surface Transportation Board	Inconsistency over time & across companies	80%	$	Timely (monthly, some delays)

Best practices

√ Use other data sources to publish estimates

√ Conduct testing to evaluate estimates periodically

√ Document methods and post with data to ensure transparency

Standards

OMB Standards and Guidelines for Statistical Surveys, Standards 7.1 to 7.4

Best practices

DARE to improve

√ Be Discerning and cautious

√ Be Adaptive and innovative

√ Be Relevant and timely

√ Be Efficient and resourceful

Standards (Circular M-14)

√ Foster collaboration between program and statistical offices

√ Develop data stewardship policies and practices for administrative data

√ Require the documentation of quality control measures

Codebook

√ Create crosswalk between EIA and other codes

√ Routinely update codebooks, including new options

Other

√ Apply OMB guidelines and standards to use of administrative records and third-party data

√ Document critical decisions and best practices

√ Select database tools to integrate data from other sources for data harmonization and normalization

√ Select appropriate statistical tools (record linkage, regression, etc.) to asses use other data sources

√ Use administrative and third-party data appropriately

What's next?

1. Catalogue PBS uses of A3P data sources
2. Report on PBS best practices by phases of the GSBPM
3. Present findings and recommendations
4. Adopt and implement best practices
5. Update OMB clearance packages with information on uses of A3P data

For more information

Petroleum supply publications

- Crude Oil Imports | www.eia.gov/petroleum/imports/companylevel/
- Petroleum Supply Annual | www.eia.gov/petroleum/supply/annual/volume1/
- Petroleum Supply Monthly | www.eia.gov/petroleum/supply/monthly
- Weekly Petroleum Supply Report | www.eia.gov/petroleum/supply/weekly

Analytical publications

- Annual Energy Outlook | www.eia.gov/aeo
- Short-Term Energy Outlook | www.eia.gov/steo
- International Energy Outlook | www.eia.gov/ieo

EIA's home page

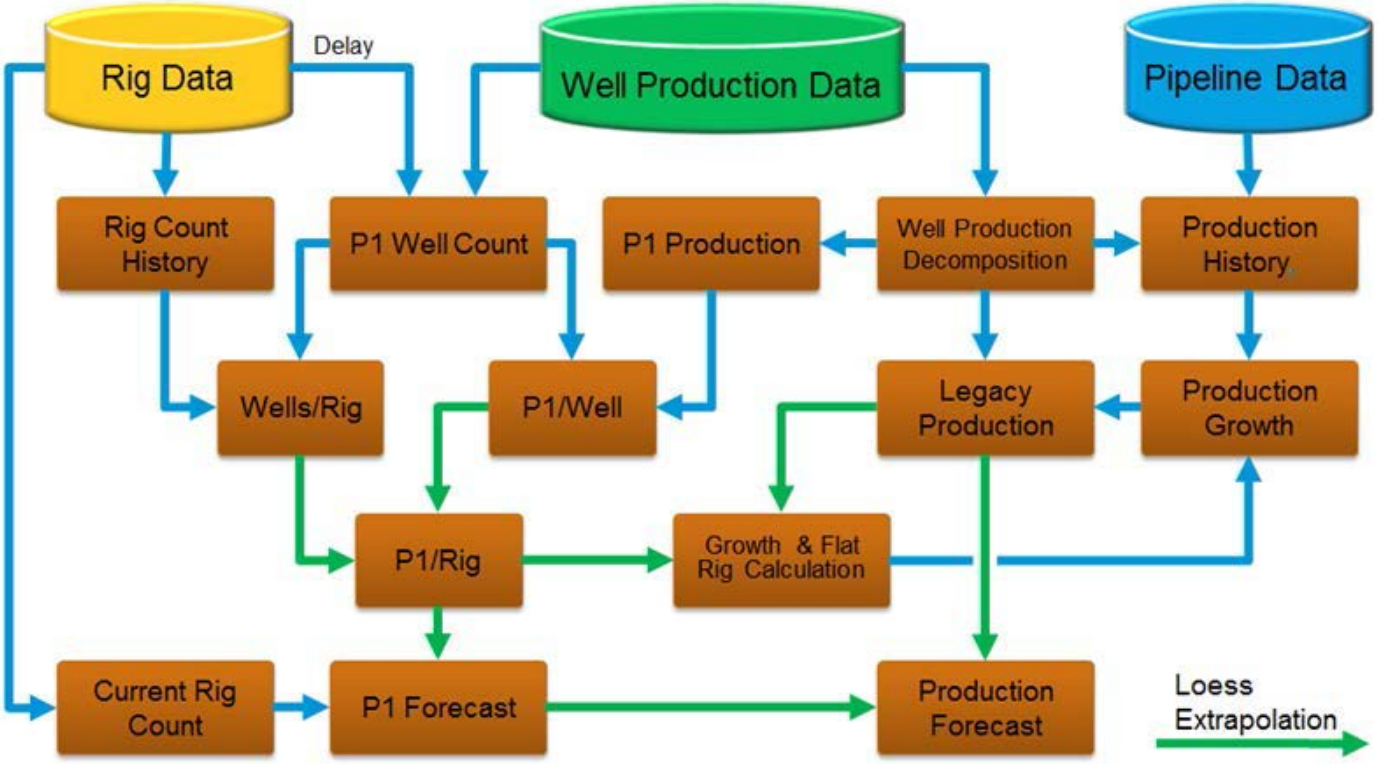
U.S. Energy Information Administration home page | www.eia.gov

Optional slides

Desired results

1. *Prepare systematic inventory of uses of other data by survey and GSBPM phases*
   - *Current uses*
   - *Potential uses*
2. *Research and understand issues pertaining to uses of A3P data*
   - *Administrative*
   - *Legal*
   - *Technical*
3. *Identify and adopt best practices to PBS uses of A3P data to enhance data quality and reduce program costs*
4. *Document use of other data sources in future Information Collection Request (ICR)*
   - *Petroleum Marketing*
   - *Petroleum Supply*

Field production estimates (if time permits)



Source: EIA's Drilling Productivity Report Documentation, August, 2014

Data sources

- **Rig count data** from Baker Hughes, Inc. North America Rig Count.
- **Oil and natural gas well production data** from DrillingInfo, Inc. (DI) Production Database.
- **Interstate natural gas pipeline flow data from** Ventyx Velocity Suite Daily Natural Gas Receipts Database.

Source: EIA's Drilling Productivity Report Documentation, August, 2014

ACTIONAL INFORMATION * **BIG DATA** *BLENDED DATA

DATA INTEGRATION AND NORMALIZATION * **DATA VISUALIZATION**

DISCOVERABLE DATA * INFORMATION QUALITY * LINGUISTICS

LEVERAGE ALTERNATIVE DATA * METADATA

METHODOLOGY MOBILE DEVICES AND INTERFACES ***PREDICTIVE ANALYTICS**

RECORD LINKAGE *RELEVANCE * RELIABILITY * **SOCIAL MEDIA**

SOCIAL NETWORK DATA * STATISTICS * SURVEY DATA *TAXONOMY

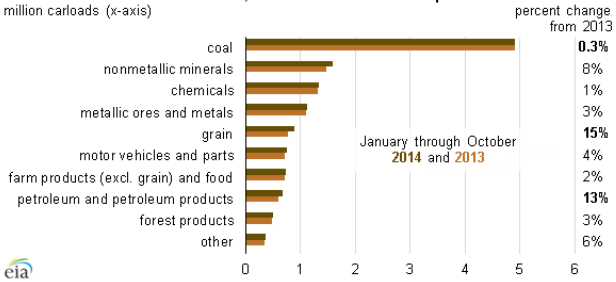
TEXT ANALYTICS * TIMELINESS * TRANSPARENCY * TYPOLOGY * VARIETY * VEROCITY * WEBSCRAPPING

Growth in crude oil and petroleum product movements by rail

NOVEMBER 13, 2014

Rail shipments of oil and petroleum products through October up 13% over year-ago period

Rail carloads of select commodities, Jan-Oct 2014 versus same period of 2013



Source: U.S. Energy Information Administration, based on Association of American Railroads
Note: These carloadings do not include intermodal traffic.

U.S. rail traffic, including carloadings of all commodity types, has increased 4.5% through October 2014 compared to the same period in 2013. Crude oil and petroleum products had the second-biggest increase in carloadings through the first 10 months of this year, with these shipments occurring in parts of the country where there is also strong demand to move coal and grain by rail. In response to shipper concerns over the slow movement of crude oil, coal, grain, ethanol, and propane, federal regulators are closely tracking service among the major U.S. freight railroad companies.

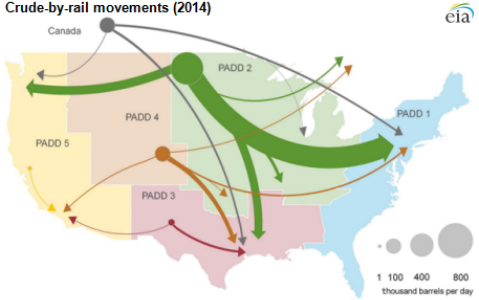
Rail carloadings of oil and petroleum products totaled 672,118 tank cars during January-October 2014, 13.4% higher compared to the same period last year, according to the Association of American Railroads (AAR). Rising U.S. crude oil production, particularly in North Dakota's Bakken Shale formation, where pipeline takeaway capacity is limited in moving the state's growing oil volumes to market, is one of the main reasons for this increase in rail shipments of petroleum and petroleum products.

Rail shipments of coal were up a relatively small 0.3% during the same period, but coal is still by far the largest commodity volume moved by rail, with 4.9 million carloadings. Power plant operators are seeking more coal deliveries by rail to rebuild their coal stockpiles, which were drawn down during last winter's colder-than-normal weather. Rail also moves U.S. coal to various points for export. At the national level, coal exports were down nearly 16% during the first half of this year, but coal exports from the Seattle Customs District (mostly sourced from Wyoming's Powder River Basin) were up 2.4% during the first half of 2014.

MARCH 31, 2015

New EIA monthly data track crude oil movements by rail

Crude-by-rail movements (2014)



Source: U.S. Energy Information Administration based on data from the Surface Transportation Board and other information
Note: Crude-by-rail movements greater than 1,000 barrels per day are represented on the map; short-distance movements between rail yards within a region are excluded. PADD denotes Petroleum Administration for Defense District.

For the first time, EIA is providing monthly data on rail movements of crude oil, which have significantly increased over the past five years. The new data on crude-by-rail (CBR) movements are integrated with EIA's existing monthly petroleum supply statistics, which already include movements by pipeline, tanker, and barge. The new monthly time series of crude oil rail movements includes shipments to and from Canada and dramatically reduces the absolute level of *unaccounted for* volumes in EIA's monthly balances for each region.

EIA is initiating the new series with monthly data from January 2010 through the current reporting month, January 2015. CBR activity is tracked between pairs of Petroleum Administration for Defense District (PADD) regions (inter-PADD), within each region (intra-PADD), and across the U.S.-Canada border. EIA developed the new series using information provided by the U.S. Surface Transportation Board (STB) along with data from Canada's National Energy Board, and EIA survey data.

Total CBR movements in the United States and between the United States and Canada were more than 1 million barrels per day (bbl/d) in 2014, up from 55,000 bbl/d in 2010. The regional distribution of these movements has also changed over this period.

The maps below provide general flows of CBR movements annually from 2010 through 2014.

2010 2011 2012 2013 2014