Beijing | 18 May 2016

JODI Oil Data Quality Assessment

Takuya Miyagawa
Asia Pacific Energy Research Centre (APERC)

www.jodidata.org

Outline

- Data quality
- Data validation techniques
- Data quality assessment
  - Color codes assessment
  - Participation assessment (Smiley faces)
- Availability of metadata
- Resources for data

Elements of Data Quality

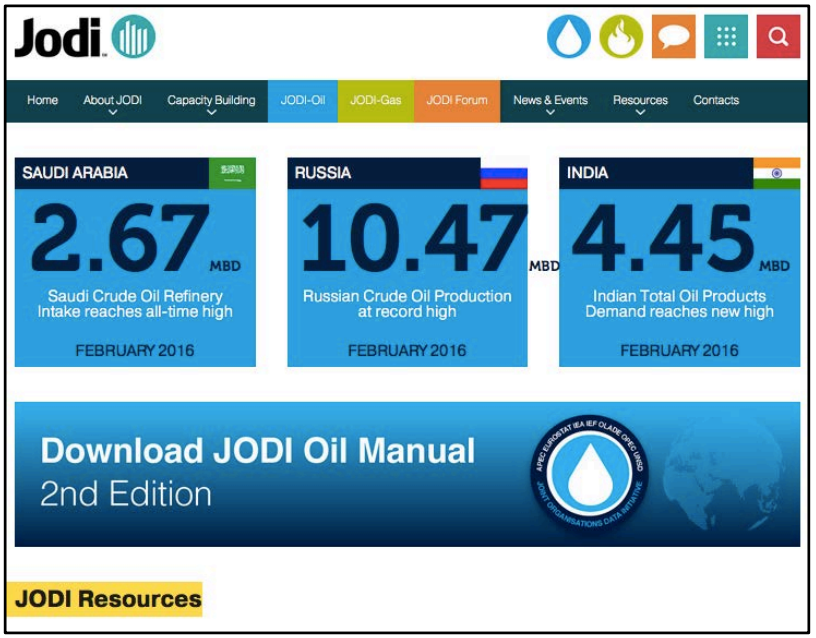
- Timeliness
- Relevance (of statistical concepts)
- Accessibility and clarity
- Coherence
- Accuracy
- Completeness/coverage
- Sustainability

Relevance

- Statistics should meet the current and potential user's needs
- Identification of users and their expectations is necessary
- Consult users
- Example: Consumer-Producer dialogue

Accessibility and Clarity

- Easily accessible to users
  - Available in the form users desire
  - Adequately documented metadata
  - User support



Coherence

- **Coherence** is the measure of the extent to which one set of statistical characteristics agrees with another and can be used together (with each other) or as an alternative (to each other)
- To assess the **coherence** of the statistics, comparisons with other statistics relating to the JODI data could be made, e.g. comparisons with monthly, quarterly and yearly oil statistics of international organisations

Data Accuracy

- Data Accuracy is an essential quality element of any database
- Closely related to usefulness of the database
- Usually negatively correlated to timeliness and completeness

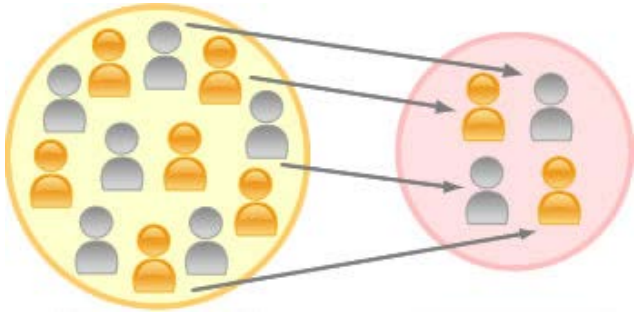
- Accuracy should be checked
  - At national level (before submitting the JODI questionnaire) and
  - At international level (OLADE, APEC, OPEC, IEA, etc)

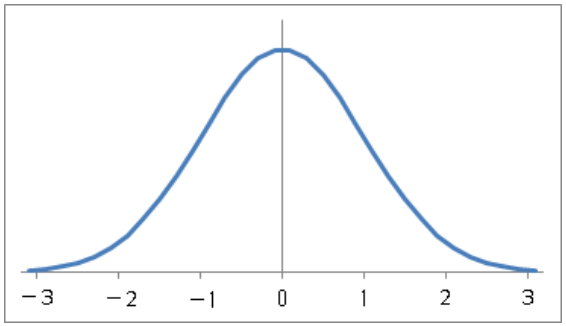
Data Accuracy

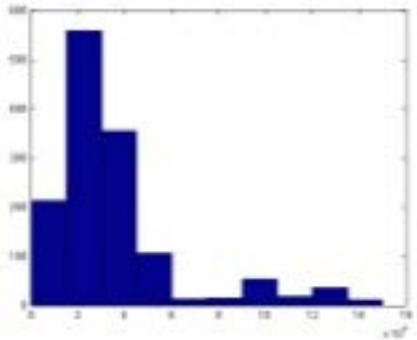
- Accuracy is defined as the proximity between the computations or estimates and the true (unknown) value
  - Sampling survey / Non-sampling survey
  - Sampling errors / non-sampling errors

Accuracy

- Non-sampling errors
  - Poor sampling method
  - Measurement errors
  - Processing errors
  - Non-response/behavioral errors
  - Model assumptions errors







Data Validation Techniques

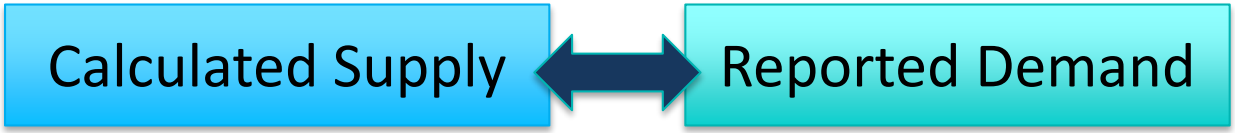
1. Balance Check: Supply vs Consumption
2. Refinery Input vs Output Check
3. Trend Check
4. Consistency with Other Statistics

Data Validation Techniques

1. Balance Check

Calculated Supply =
Production + From Other Sources + Imports
– Exports – Direct Use – Stock changes

- Large deviation means incorrect data in some or all flows



- This check is applicable only if data for all the flows are complete and reliable.

Data Validation Techniques

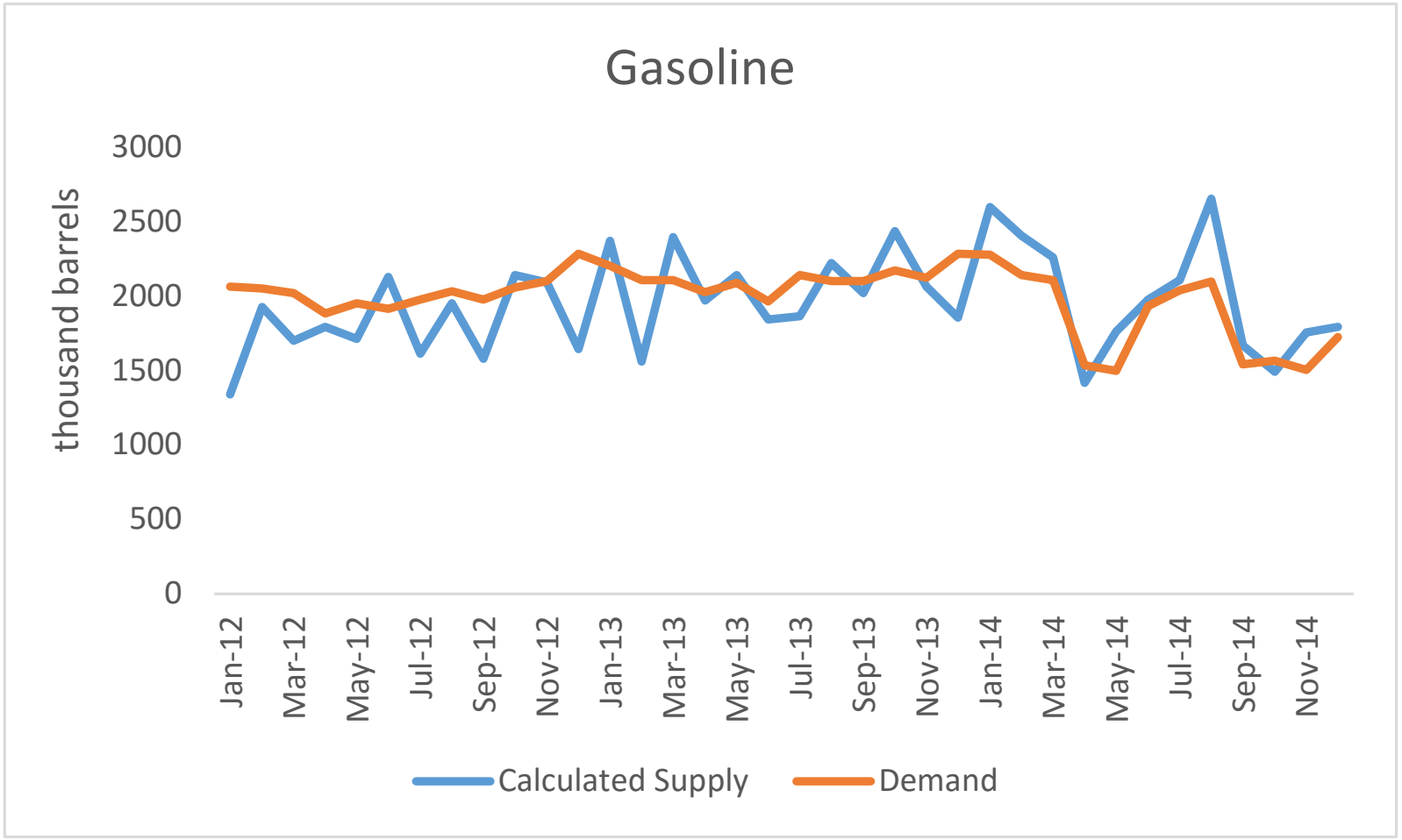
1. Balance Check

- Statistical Difference = Calculated Supply – Refinery Intake
- The absolute value of the deviation of “Statistical Diffrence” should not be higher than 10% of domestic supply of primary products
- and should not be higher that 10% of Final Consumption

	Crude Oil
	(1)
+ Production	314
+ From Other sources	
+ Imports	6,052
- Exports	314
+ Products Transferred /Backflows	
- Direct Use	-
- Stock Change	(1,006)
- Statistical Difference	(13)
= Refinery Intake	7,071
Closing stocks	4,584

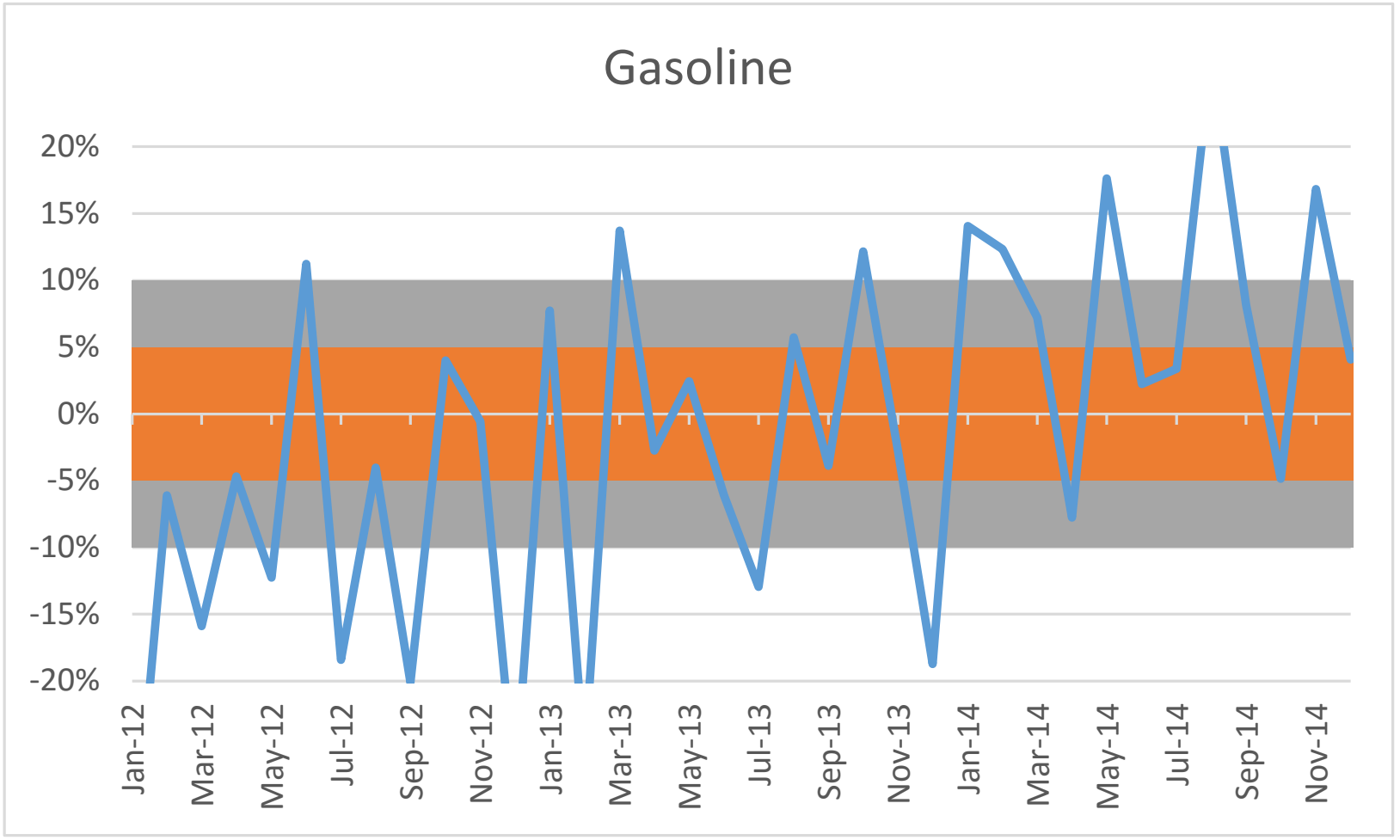
Data Validation Techniques

1. Balance Check



Data Validation Techniques

1. Balance Check



Data Validation Techniques

Other Consistency Checks

JOINT OIL DATA INITIATIVE

Closing minus opening level
Positive number corresponds to stock build, negative number corresponds to stock draw

Country **Country**

Month **Month** **Year** ***Unit :*** ***thousand tons***

	Crude Oil	NGL	Other	Total (1)+(2)+(3)		Petroleum Products									Checks
						LPG	Naphtha	Gasoline	Total Kerosene	*Of which: Jet Kerosene*	Gas/ Diesel Oil	Fuel Oil	Other Products	Total Products (5)+(6)+(7) +(8)+(10) +(11)+(12)	
	(1)	(2)	(3)	(4)		(5)	(6)	(7)	(8)	*(9)*	(10)	(11)	(12)	(13)	(14)
+ Production	12622	1883	3954	18,459	+ Refinery Output	125	274	2559	517	455	2536	397	1147	7,555	
+ From Other sources			0	0	+ Receipts	0	108	622	13	10	125	36	1487	2,391	
+ Imports	2453	59	0	2,512	+ Imports	6	0	229	156	127	86	90	393	960	
- Exports	9066	969	2310	12,345	- Exports	53	54	605	43	43	695	243	202	1,895	
+ Products Transferred /Backflows			536	536	- Products Transferred	0	25	0	0	0	0	2	509	536	
- Direct Use	0	602	0	602	+ Interproduct Transfers	216	-18	169	-23	-10	105	-26	-423	0	
- Stock Change	1012	315	0	1,327	- Stock Change	28	-50	-63	-33	-44	16	39	-87	-150	
- Statistical Difference	-911	-43	0	-954	- Statistical Difference	-46	30	49	11	12	-76	53	87	108	
= Refinery Intake	5908	99	2180	8,187	= Demand	312	305	2988	642	571	2217	160	1893	8,517	
Closing stocks	9246	1973	0	11,219	Closing stocks	258	100	1712	338	306	1757	315	1253	5,733	

Automatic Checks

Total sum		OK
Statistical Difference		OK
Stat. Diff./Refinery Intake		Statistical Difference above 10% of Refinery Intake, please investigate
Products Transferred		OK
Negative Products Transferred		OK
Blocked out cells		OK
Negative Stock Values		OK
Refinery Losses	632	OK

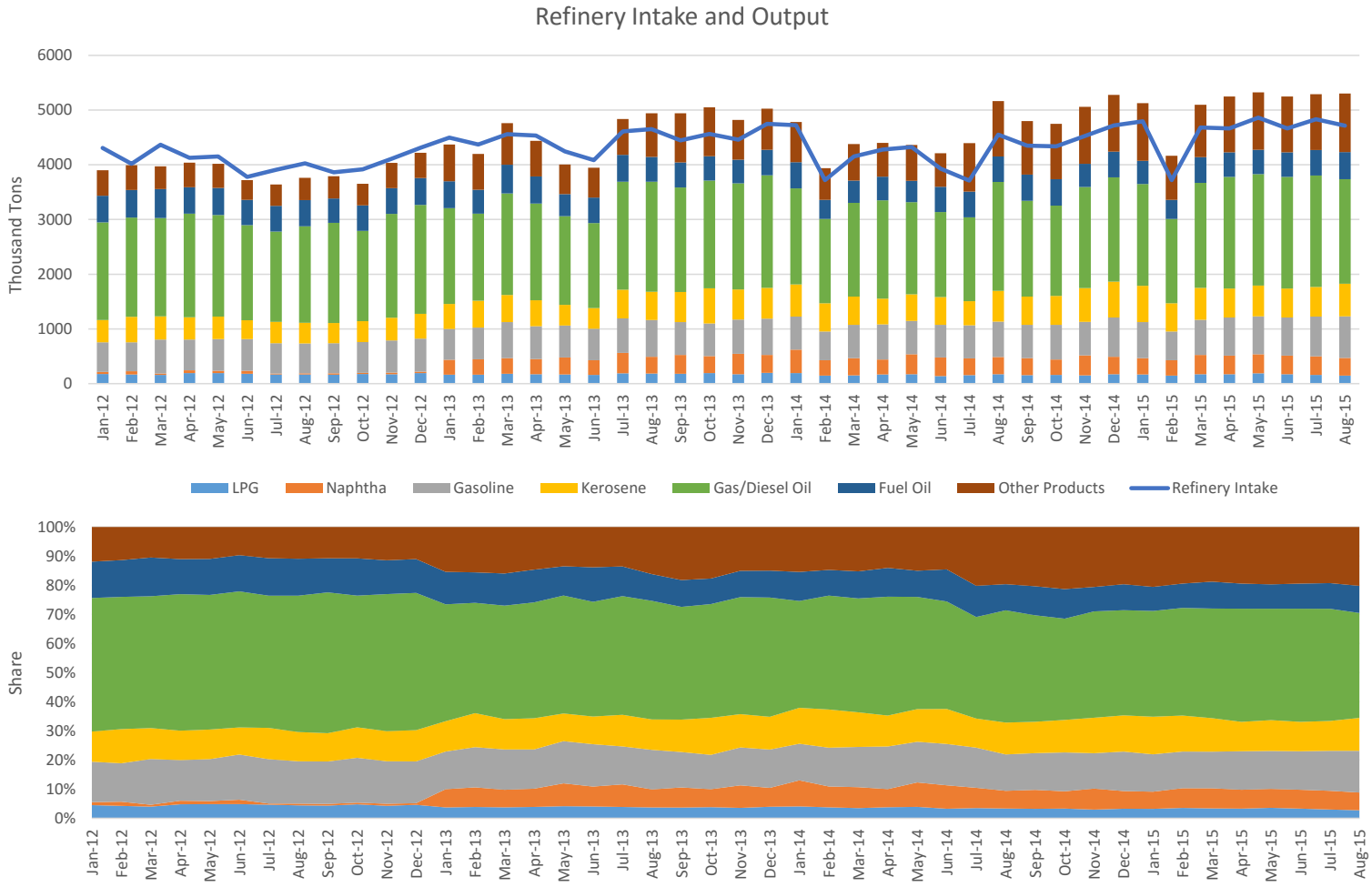
Automatic Checks Petroleum Products

Total Products sum	OK
Statistical Difference	OK
Stat. Diff. /Demand	Statistical Difference above 10% of Demand, please investigate
Negative Products Transferred	OK
Interproduct transfers	OK
Jet Kerosene	OK
Negative Stock Values	OK

embedded in the JODI Oil Questionnaire

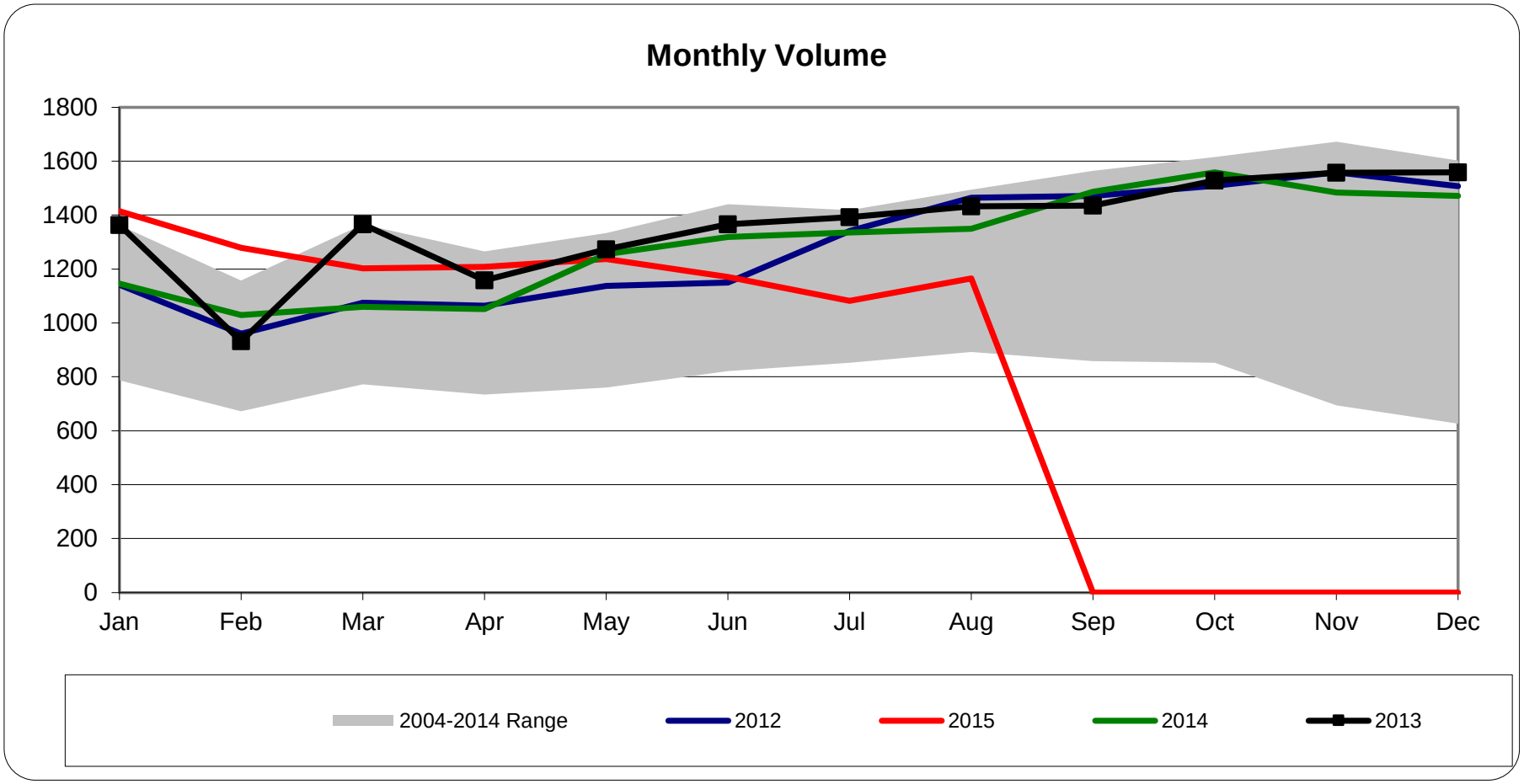
Data Validation Techniques

2. Refinery Data Check



Data Validation Techniques

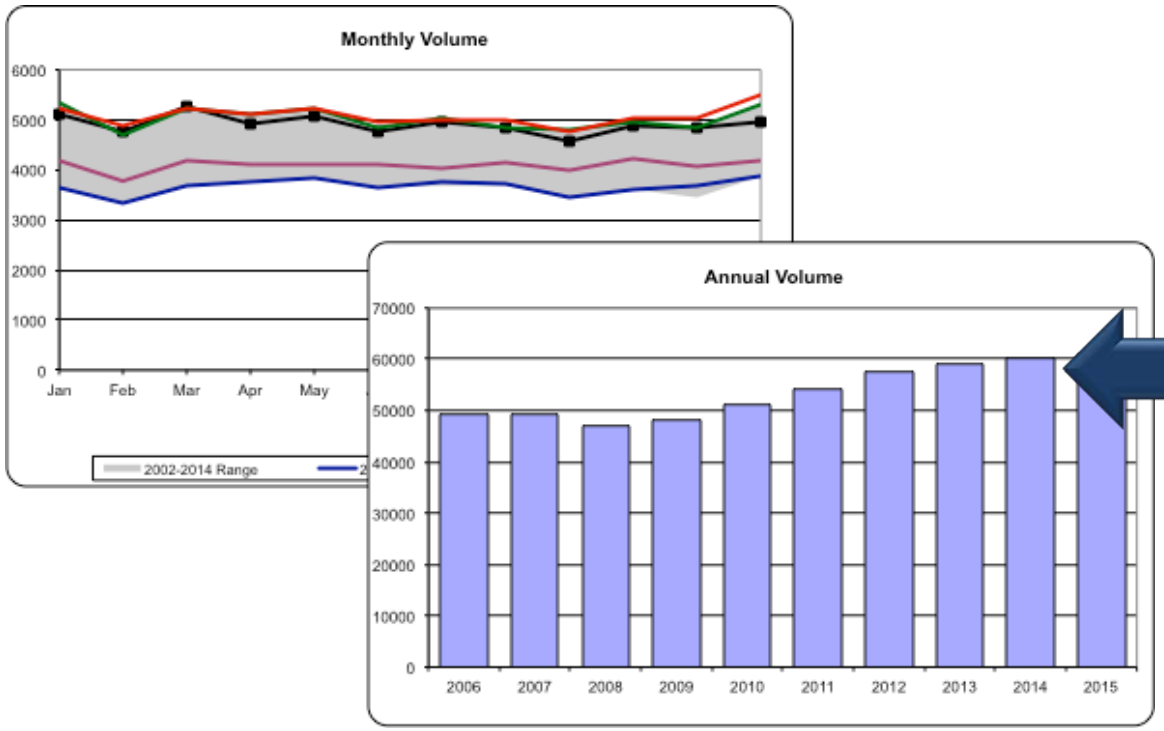
3. Trend Check



Data Validation Techniques

4. Consistency with other energy statistics

- Comparison with annual statistics (“APEC Energy Database”, etc.)
  - Sum of 12 months data is compared with annual data
  - JODI and annual data definitions are carefully considered



Data Quality Assessment

Color Codes (JODI-Oil World Database)

TIME / Country	Dec2014	Jan2015	Feb2015	Mar2015	Apr2015	May2015	Jun2015	Jul2015	Aug2015	Sep2015	Oct2015	Nov2015	Dec2015	Jan2016	Feb2016
Australia	356	338	319	251	297	265	336	369	370	347	338	357	343	314	305
Brunei Darussalam	118	107	117	114	128	119	126	127	84	119	101	116	127	130	123
Canada	2,916	2,917	2,923	2,835	2,786	2,638	2,675	2,898	2,925	2,758	2,879	2,927	3,020	2,846	2,885
Chile	5	5	5	5	5	5	5	7	8	8	8	7	7	8	0
China	4,327	4,237	4,237	4,266	4,270	4,282	4,420	4,275	4,290	4,329	4,271	4,309	4,287	4,161	4,161
Chinese Taipei	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Hong Kong China	0	0	0	0	0										
India	770	765	764	774	749	7									
Indonesia	770	743	847	739	814	7									
Japan	5	5	5	5	4										
Korea	0	0	1	0	0										
Malaysia	601	628	665	506	635	6									
Mexico	2,357	2,252	2,335	2,323	2,208	2,2									
Myanmar	15	15	0	0	13										
Papua New Guinea	0	0	0	0	0										
Philippines	11	11	15	1	9										
Russian Federation	10,086	10,110	10,167	10,110	10,102	10,0									
Singapore	0	0	0	0	0										
Thailand	237	234	236	254	249	2									
United States of America	9,422	9,345	9,456	9,653	9,694	9,4									
Vietnam	0	0	0	0	0										

Blue: indicates that the organization that assessed the data sees the data as reliable

Yellow: data might not be reliable, consult the metadata

White: data cannot be assessed

Purple: data is still under verification

Data Quality Assessment

Color Codes

- IEA Methodology
  - M-1 data vs. MOS data
  - MOS is the final monthly data (M-2)
  - Data with absolute value of deviation of at least 5% is colored blue
  - Higher than 5% is colored yellow
  - Data that cannot be assessed is colored white

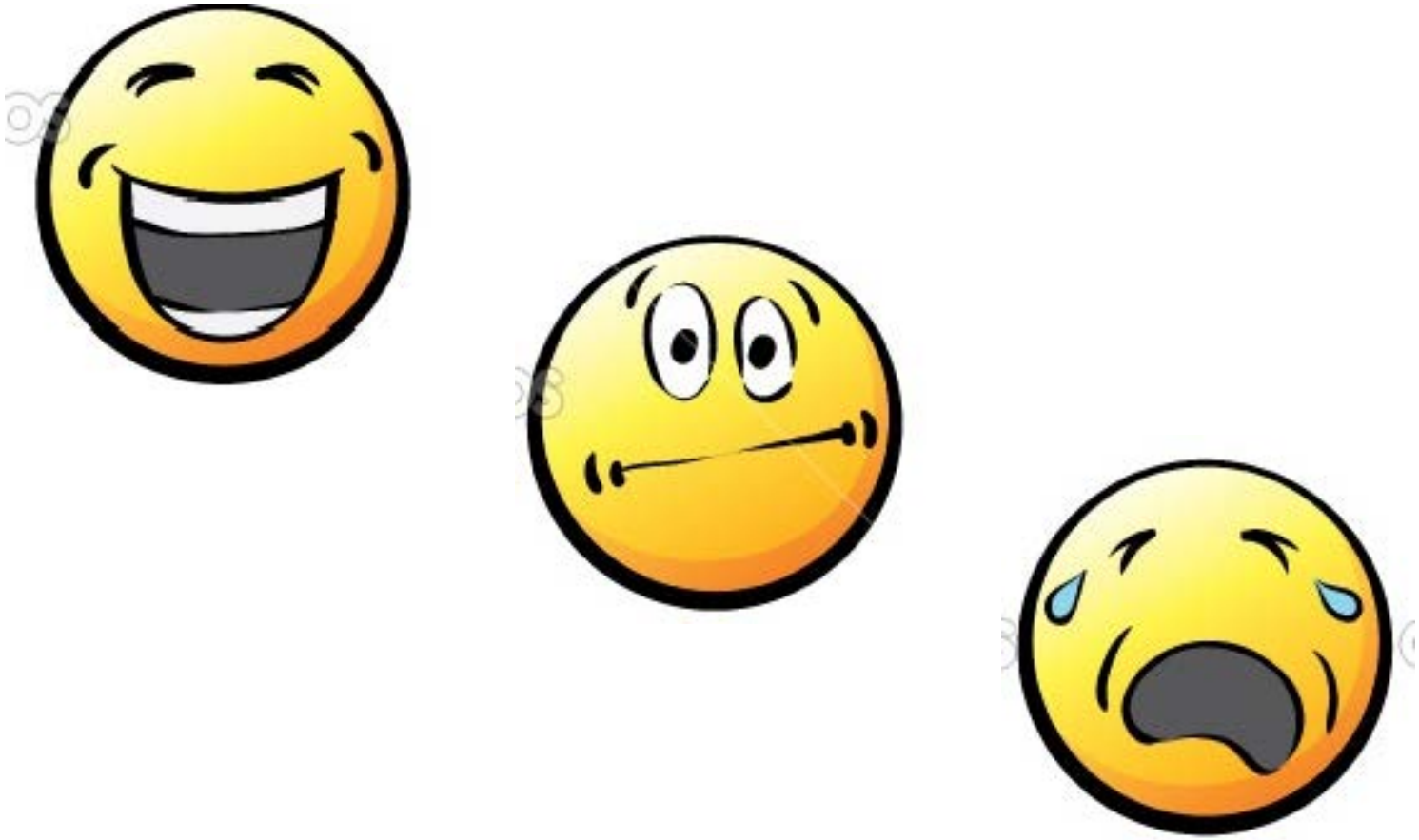
Data Quality Assessment

Color Codes

- APEC Methodology
  - Compared with data from other sources
    - Production and demand of large economies
  - Compared with quarterly data
    - Production and trade data
  - Compared with annual data
    - All other data
  - Data with absolute value of deviation of at least 5% is colored blue
  - Higher than 5% is colored yellow
  - Data that cannot be assessed is colored white

Data Quality Assessment

Participation Assessment Approach



Source: http://www.rovish.myewebsite.com/photos/cool-pictures/depositphotos_7272052-set-of-smiley-faces.html

Participation Assessment Approach

Smiley Faces

Brunei Darussalam	☺	☺	☺	Italy	☺	😐	☺
Bulgaria	☺	☺	☺	Jamaica	☺	☹	😐
Canada	☺	☺	☺	Japan	☺	☺	☺
Chile	☹	☹	☹	Kazakhstan	😐	😐	☹
China	☺	☺	😐	Korea	☺	☺	☺
Colombia	n.a.	n.a.	n.a.	Kuwait	😐	😐	☺
Costa Rica	☹	☹	☹	Latvia	☺	☺	☺
Croatia	☺	☺	☺	Libya	n.a.	n.a.	n.a.
Cuba	n.a.	n.a.	n.a.	Lithuania	☺	☺	☺
Cyprus	☺	☺	☺	Luxembourg	☺	☺	☺
Czech Republic	😐	😐	☺	FYR of Macedonia	😐	😐	😐
Denmark	☺	☺	☺	Malaysia	☺	☺	☹

Red: Sustainability
Blue: Timeliness
Green: Completeness

Participation Assessment Approach

Smiley Faces

- IEA Methodology
- Timeliness: Number of M-1 submissions within the 6-month period under review
  -  6 M-1 submissions
  -  4-5 M-1 submissions
  -  less than 4 submissions

Participation Assessment Approach

Smiley Faces

- IEA Methodology
- Completeness: Number of data points submitted based on the original JODI format
  -  above 90% of all data points
  -  60-90% of all data points
  -  less than 60% submissions

Participation Assessment Approach

Smiley Faces

- IEA Methodology
- Sustainability: M-1 and M-2 submissions within the 6-month period under review
  -  6 months of data
  -  4-5 months of data
  -  less than 4 months of data

Participation Assessment Approach

Smiley Faces

- APEC Methodology
- Timeliness: Number of M-1 & M-2 submissions within the 6-month period under review
  -  6 M-1 & M-2 submissions
  -  4-5 M-1 & M-2 submissions
  -  less than 4 M-1 & M-2 submissions

Participation Assessment Approach

Smiley Faces

- APEC Methodology
- Completeness: Number of data points submitted based on the original JODI format
  -  above 90% of all data points
  -  60-90% of all data points
  -  less than 60% submissions

Participation Assessment Approach

Smiley Faces

- APEC Methodology
- Sustainability: M-1 and M-2 submissions within the 6-month period under review
  -  6 months of data
  -  4-5 months of data
  -  less than 4 months of data

Availability of Metadata

- JODI Data should have metadata
- The simplest definition of metadata is that it is **data about data**. More specifically information (data) about a particular content (data)
- Metadata describes **how and when and by whom** a particular set of data was collected; how the data is **formatted**
- Metadata **must be updated** when there is a change in the resource it describes
- It can be useful to **keep** metadata even when the resource no longer exists
- Metadata **enhances data transparency** and is essential for understanding information stored in a database

Resources for Data (Quality vs Cost)

- The quality of the data will be affected by available resources to collect, analyze and store energy statistics
- Although not measures of quality, they are positively correlated with quality
- Costs: Office space, utility bills, staff-hours involved, software, etc.
- Cost is not only on the collector but also on the respondent
- Response burden: Simplest way to measure is the time spent by the respondent to provide information
- A compromise between quality and cost and burden must be achieved

Resources for Data (Quality vs Cost)

- Functions of cost/burden
  - Collection of data
  - Level of disaggregation
  - Time lags, frequencies of data
  - Applied methodologies
- Fortunately, administrative data are available; they are just to be found and collected

www.jodidata.org

APEC Asia-Pacific Economic Cooperation
eurostat EUROPEAN COMMISSION
GECF
iea
IEF
olade
OPEC
UN