Port of Spain| 20 October 2015

JODI Oil Data Quality Assessment

Edito Barcelona
Senior Coordinator, Asia Pacific Energy Research Centre, Tokyo

www.jodidata.org

Outline

- Data Quality
- Data accuracy
- Data validation techniques
- Color codes in the JODI database
- Smiley faces
- Availability of metadata
- Quality vs. Cost of data collection

Elements of Data Quality

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

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
- There is adequately documented metadata
- User support should be available from providers

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 errors and non-sampling errors
  - Sampling errors: due to problems in the design of the sample survey

Accuracy

- Non-sampling errors
  - Poor sampling method
  - Measurement errors
  - Processing errors
  - Non-response/behavioral errors
  - Model assumptions errors
- Country level: report collected information – revisions
- International level: revision of the time series

Data Validation Techniques

1. Balance Check ~ Supply vs Consumption
2. Refinery Input vs Output Check
3. Reported Total vs Calculated Total Check
4. Negative number check
5. Reported Stock Change vs Calculated Stock Change Check
6. Trend Check

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

Activity	Oil kbbl	Natural Gas Mm3	NGL kbbl	Liquefied Natural Gas kbbl	Natural Gasoline kbbl	Virgin Naphtha kbbl	LPG kbbl	Gasolines kbbl	Kerosene kbbl	Jet fuel kbbl
Production	120.00	120.00	120.00	120.00	10.00	20.00	20.00	20.00	20.00	20.00
Imports										
Exports				100.00						
Bunker										
Initial stocks										
Closing stocks										
Domestic Supply	120.00	120.00	120.00	20.00	10.00	20.00	20.00	20.00	20.00	20.00
Transformation	100.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
- Refinery load	100.00		100.00							
- Refinery output						20.00	20.00	20.00	20.00	20.00
- Recovery plants inputs										
- Recovery plantas outputs					10.00					
- Power plants										
- Self - producers										
- Transfers inputs										
- Transfer outputs										
- Recycles										
Losses										
Own Consumption										
Adjustments	**20.00**	120.00	**20.00**	**20.00**	**2.00**	**2.00**	**-3.00**	**3.00**	0.00	-1.00
Final Consumption					8.00	18.00	23.00	17.00	20.00	21.00

Data Validation Techniques

1. Balance Check

- Adjustments = Domestic Supply - (Transformation + Losses + Own use + Final Consumption)
- The absolute value of the deviation of “Adjustments” should not be higher than 10% of domestic supply of primary products
- and should not be higher that 10% of Final Consumption

ITEM	Activity	Oil kbbl	Natural Gas Mm3	NGL kbbl
1	Production	120.00	120.00	120.00
2	Imports			
3	Exports			
4	Bunker			
5	Initial stocks			
6	Closing stocks			
7	Domestic Supply	120.00	120.00	120.00
8	Transformation	100.00	80.00	100.00
9	- Refinery load	100.00		100.00
10	- Refinery output			
11	- Recovery plants inputs			
12	- Recovery plantas outputs			
13	- Power plants		80.00	
14	- Self - producers			
15	- Transfers inputs			
16	- Transfer outputs			
17	- Recycles			
18	Losses			
19	Own Consumption			
20	Adjustments	20.00	40.00	20.00
21	Final Consumption			

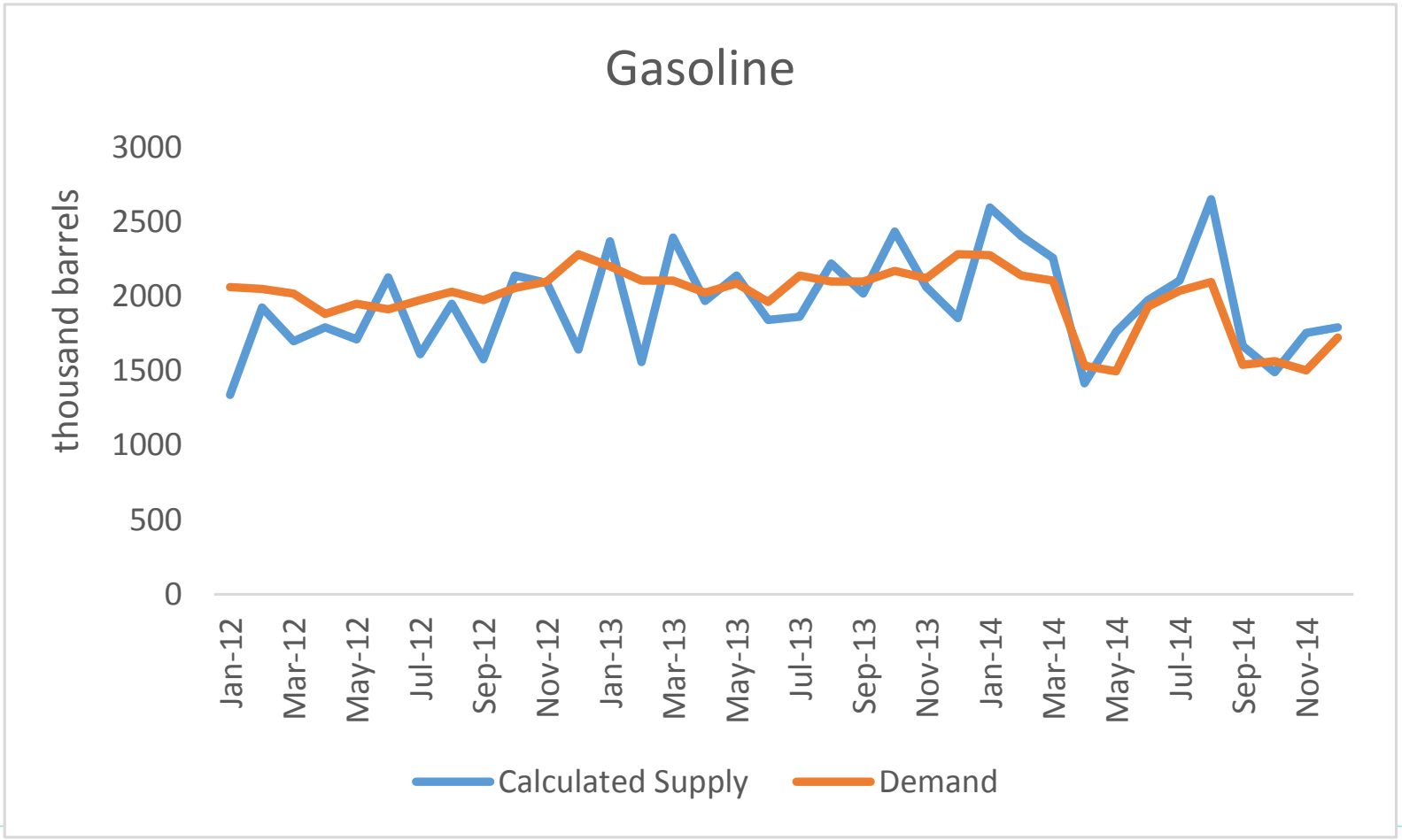
Data Validation Techniques

1. Balance Check

Activity	Natural Gasoline kbbl	Virgin Naphtha kbbl	LPG kbbl	Gasolines kbbl	Kerosene kbbl	Jet fuel kbbl	Diesel oil kbbl	Fuel oil kbbl	Refinery Gas kbbl	Others kbbl	Non Energy kbbl
Production	10.00	20.00	20.00	20.00	20.00	20.00	20.00	20.00	20.00	0.00	20.00
Imports											
Exports											
Bunker											
Initial stocks											
Closing stocks											
Domestic Supply	10.00	20.00	20.00	20.00	20.00	20.00	20.00	20.00	20.00	0.00	20.00
Transformation	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
- Refinery load											
- Refinery output		20.00	20.00	20.00	20.00	20.00	20.00	20.00	20.00		20.00
- Recovery plants inputs											
- Recovery plantas outputs	10.00										
- Power plants											
- Self - producers											
- Transfers inputs											
- Transfer outputs											
- Recycles											
Losses											
Own Consumption											
Adjustments	2.00	2.00	-3.00	3.00	0.00	-1.00	-2.00	-3.00	1.00	-18.00	3.00
Final Consumption	8.00	18.00	23.00	17.00	20.00	21.00	22.00	23.00	19.00	18.00	17.00

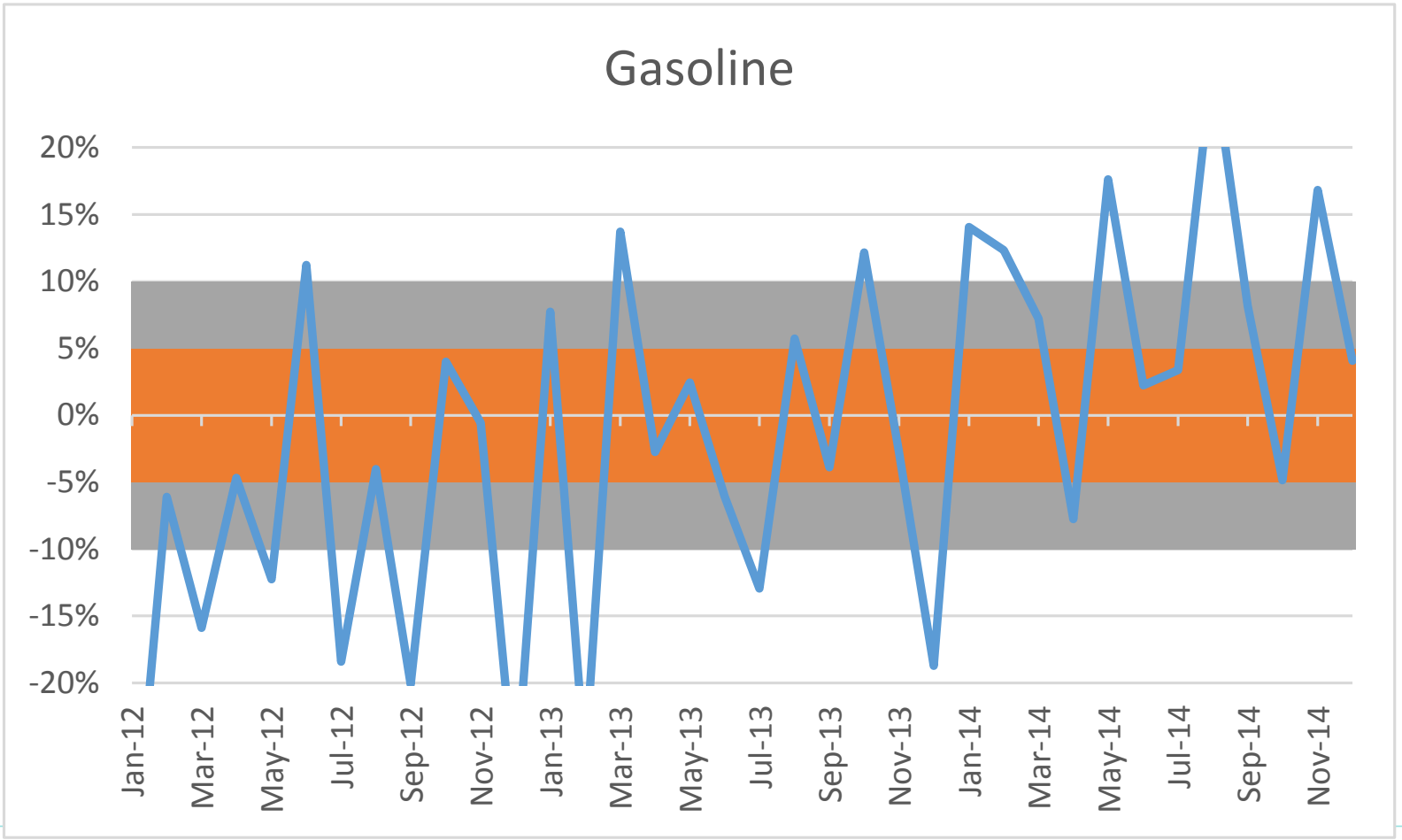
Data Validation Techniques

1. Balance Check



Data Validation Techniques

1. Balance Check



Data Validation Techniques

2. Other Consistency Checks – embedded in the JODI Oil Questionnaire

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

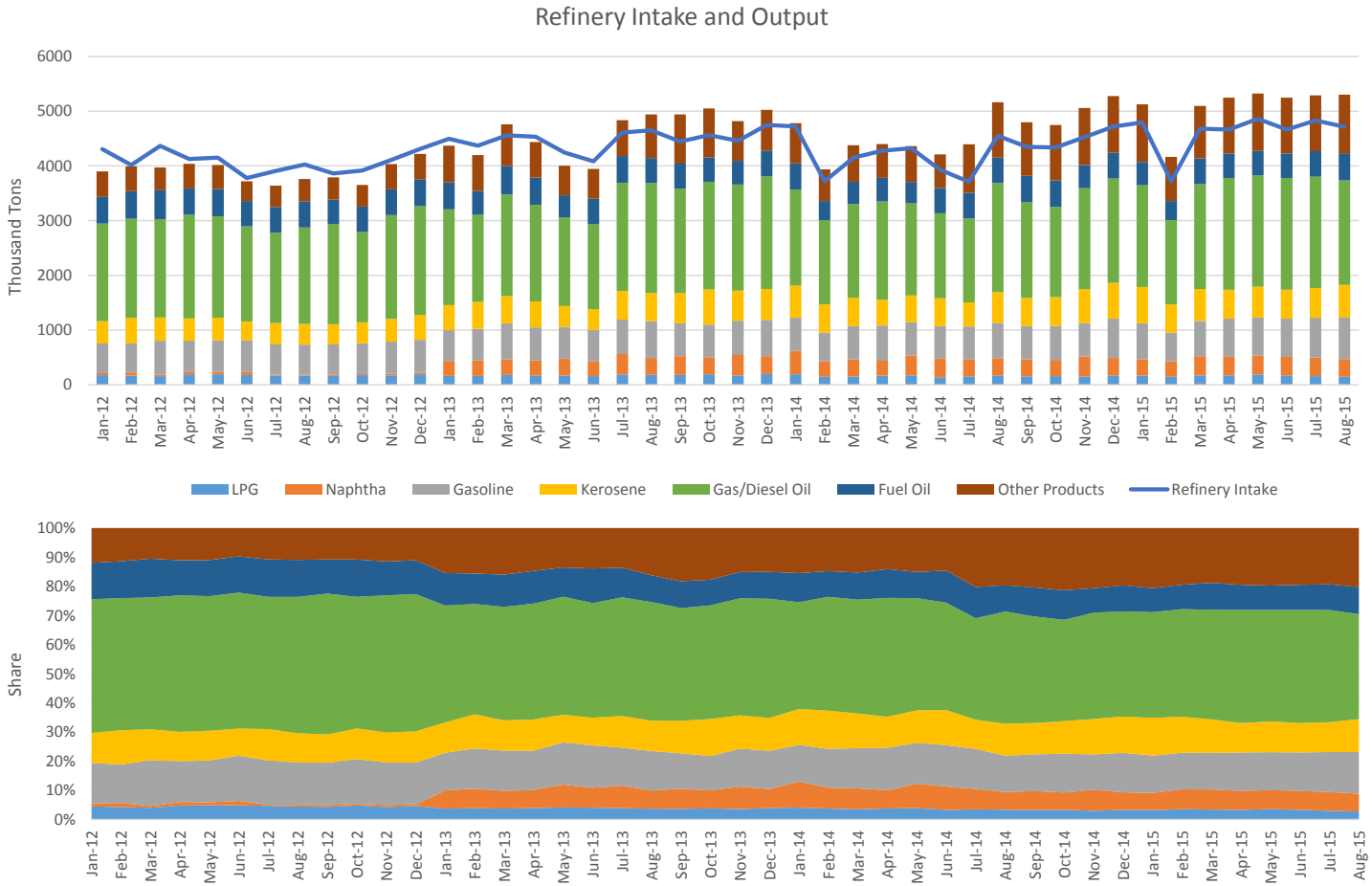
Total Products sum	OK
Statistical Difference	OK
Stat. Diff. /Demand	Statistical Difference above 10% of Demand, please investigate
Negative Products Transferred	OK
Interproduct transfers	OK
Jet Kerosene	OK
Negative Stock Values	OK

Jodi

BETTER DATA
BETTER DECISIONS

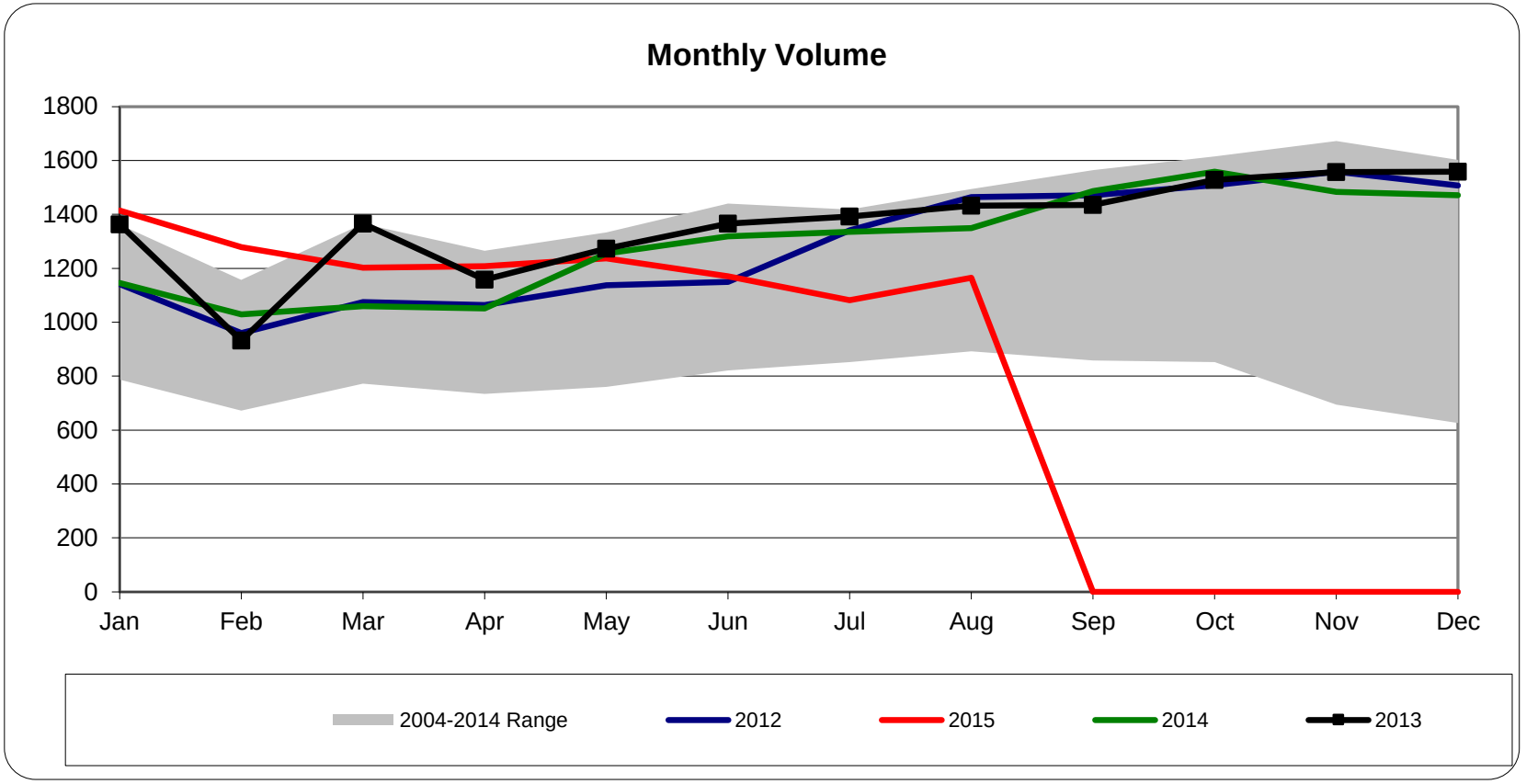
Data Validation Techniques

3. Refinery Data Check



Data Validation Techniques

4. Trend Check



Data Quality Assessment

Color Codes

Reports | Table | Chart

JODI-Oil World Database - Simplified Version (Last 15 Months)

Other: | Unit ▼ Thousand Barrels per day (kb/d) | Product ▼ Crude Oil | BALANCE ▼ Production/Refinery output

TIME / Country	May2014	Jun2014	Jul2014	Aug2014	Sep2014	Oct2014	Nov2014	Dec2014	Jan2015	Feb2015	Mar2015	Apr2015	May2015	Jun2015	Jul2015
Argentina	501	532	522	525	529	535	530	521	537	525	528	530	533	535	538
Barbados	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Belize	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0
Bermuda	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Bolivia	65	65	65	64	63	63	58	58	61	62	62	0	0	0	0
Brazil	2,195	2,252	2,272	2,332	2,364	2,398	2,363	2,503	0	0	2,419	2,400	2,418	2,401	2,471
Chile	7	6	7	7	7	6	6	5	5	5	5	5	5	5	0
Colombia	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Costa Rica	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Dominican Republic	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Ecuador	554	555	558	561	551	557	563	558	558	553	553	548	543	541	538
El Salvador	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Grenada	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Guatemala	10	10	10	10	10	9	11	10	0	0	0	0	0	0	0
Guyana	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Haiti	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Honduras	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Jamaica	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Mexico	2,496	2,440	2,392	2,418	2,393	2,366	2,366	2,357	2,252	2,335	2,323	2,208	2,233	2,252	2,275
Nicaragua	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Panama	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Paraguay	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Peru	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Suriname	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Trinidad/Tobago	79	82	82	81	84	81	79	83	83	85	81	81	78	80	0
Uruguay	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Venezuela	2,690	2,676	2,696	2,664	2,644	2,675	2,695	2,733	2,705	2,735	2,727	2,717	2,759	2,669	2,659

Data Quality Assessment

Color Codes

- **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 is compared with MOS data (12 month rolling average)
  - MOS is the monthly oil statistics which is submitted at 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

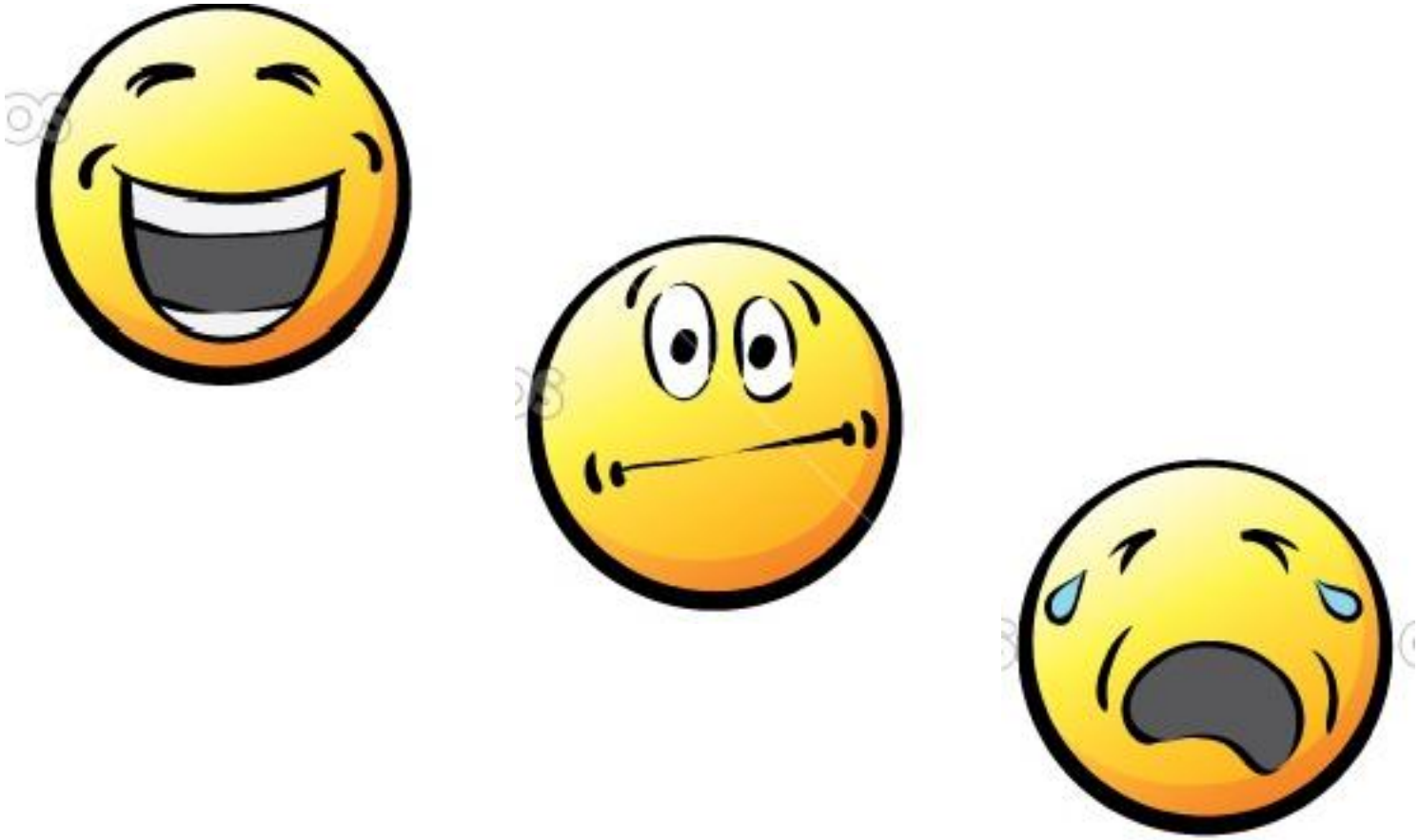
Data Quality Assessment

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

Assessment of Participation to JODI

Smiley Faces



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

Assessment of Participation to JODI

Smiley Faces

Algeria	☺	☺	☺	Germany	☺	☺	☺	Nigeria	☺	☺	☺
Angola	☺	☺	☹	Greece	☺	☺	☺	Norway	☺	☺	☺
Argentina	☺	☹	☺	Grenada	☹	☹	☹	Oman	☺	☺	☺
Armenia	☺	☺	☹	Guatemala	☹	☹	☹	Panama	😐	☹	☹
Australia	☺	😐	☺	Guyana	☹	☹	☹	Papua New Guinea	☹	☹	☹
Austria	☺	☺	☺	Haiti	n.a.	n.a.	n.a.	Paraguay	☺	☺	☹
Azerbaijan	☺	☺	☺	Honduras	n.a.	n.a.	n.a.	Peru	n.a.	n.a.	n.a.
Bahrain	☺	☺	☺	Hong Kong, China	☺	☺	☺	Philippines	☺	☺	☺
Barbados	☹	☹	☹	Hungary	☺	☺	☺	Poland	☺	☺	☺
Belarus	☺	☺	😐	Iceland	☺	☺	☺	Portugal	☺	☺	☺
Belgium	☺	☺	☺	India	☺	☺	☺	Qatar	☺	☺	☺
Belize	☺	☹	☹	Indonesia	☹	☹	☹	Romania	☺	☺	☺
Bermuda	☺	☹	☹	Iran	☺	😐	😐	Russian Federation	☺	☺	😐
Bolivia	☹	☹	☹	Iraq	☺	☺	☺	Saudi Arabia	☺	☺	☺
Brazil	☺	☹	☺	Ireland	☺	☹	☺	Singapore	☺	☺	☹
Brunei Darussalam	☺	☺	☺	Italy	☺	😐	☺	Slovak Republic	☺	😐	☺
Bulgaria	☺	☺	☺	Jamaica	☺	☹	😐	Slovenia	☺	☺	☺
Canada	☺	☺	☺	Japan	☺	☺	☺	South Africa	☺	😐	☺
Chile	☹	☹	☹	Kazakhstan	😐	😐	☹	Spain	☺	☺	☺
China	☺	☺	😐	Korea	☺	☺	☺	Suriname	n.a.	n.a.	n.a.
Colombia	n.a.	n.a.	n.a.	Kuwait	😐	😐	☺	Sweden	☺	☺	☺
Costa Rica	☹	☹	☹	Latvia	☺	☺	☺	Switzerland	☺	☺	☺
Croatia	☺	☺	☺	Libya	n.a.	n.a.	n.a.	Syria	n.a.	n.a.	n.a.
Cuba	n.a.	n.a.	n.a.	Lithuania	☺	☺	☺	Taiwan, China	☺	☺	☺
Cyprus	☺	☺	☺	Luxembourg	☺	☺	☺	Thailand	☺	😐	☺

Assessment of Participation to JODI

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

Assessment of Participation to JODI

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

Assessment of Participation to JODI

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

Assessment of Participation to JODI

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

Assessment of Participation to JODI

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

Assessment of Participation to JODI

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

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

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

Cost of Data Collection

- 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