Data quality and verification

Amani Al-Saidi

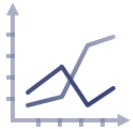
Training Workshops on Energy Statistics: Energy Balances

December 13-14, 2018, Beirut

IEA

Outline

Balance checks



Time series analysis

Efficiencies

Data validation in practice

What is data quality?

Data quality is most commonly assessed in terms of its **"fitness for use"**.

How well do the statistical outputs meet user needs?

Relevance	Statistics meet the needs of the users
Accuracy and reliability	Statistics accurately and reliably portray reality
Timeliness and punctuality	Statistics are released in a timely and punctual manner
Accessibility and clarity	Statistics are presented in a clear and understandable form, released in a suitable and convenient manner, available and accessible on an impartial basis with supporting metadata and guidance.
Coherence and comparability	Statistics are consistent internally, over time and comparable between regions and countries.

For more information please see:
Fundamental Principles of Official Statistics and International Recommendation for Energy Statistics (IRES)

Balance checks

Internal consistency

Statisical differences

Own use and losses

Calorific values

Balance checks - Internal consistency

- Some things to look out for…
  - If one flow or product is supposed to be the sum of two others , is it?
  - Does the detailed breakdown match the total (e.g. for trade or demand data)?
  - Are stock changes correctly calculated?
  - Is the data set complete?
  - Decrease in use of one fuel coupled with increase in use of another one?
  - Is the story logical?

- It is good practice to analyze the data processes to identify possible points of failure and introduce checks.

Balance checks - What are statistical differences?

- When collecting energy data we try to get a picture of supply and demand in a market.
  - Economic theory tells us that when a market is in equilibrium supply = demand.
- In practice, measuring supply and demand comes with many challenges.
- Statistical differences measure the level of mismatch between the collected supply and demand data.

Production
+ Imports
- Exports
- International marine and aviation bunkers
- Stock changes
= Supply

How much is available to the internal market?

Use in transformation processes and Final consumption
= Demand

How much was consumed in the market?

Balances checks - When are statistical differences a problem?

- High statistical differences can be a red flag (often when more than 5% of deliveries to the market).
- But… a statistical difference of 0 is also suspicious!
- Calculating the **ratio of statistical difference to total primary energy supply (TPES)** can help us evaluate the size of the problem.

What can we do?

Check the completeness and coverage of the data

Review definitions and methods of data sources

Check calorific values for supply and demand

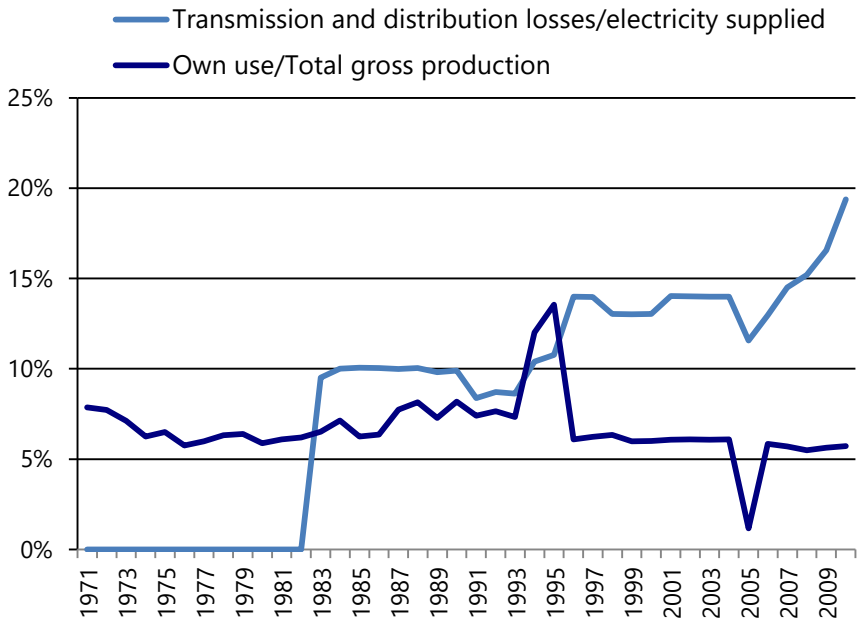
Are there any new trends missing? (renewables, electric vehicles…)

Balance checks - Own use and losses

Case study: Electricity and heat

- **Own use** of electricity by power plants (*main electricity plants only*) as a proportion is not expected to vary significantly year-on-year (**≤ 3%**)

- **Transmission and distribution losses** should technically stay **below 10%**.
  - Bad maintenance and theft can lead to higher transmission and distribution losses (~ 30%)



Balance checks - Calorific values

- Calorific values are essential to building an energy balance and for converting to/from energy units.

- For every energy product there is an expected range for the calorific values.
  - The collected Net Calorific Values (NCVs) should be compared against default NCV
  - And also with historical NCVs for the product as calorific values can be region specific (except for natural gas as it is mostly methane)
  - Check if NCVs used for consumption is in line with the one used for supply.

Fuel	Expected calorific value (kJ/kg, MJ/ton)			GCV estimation
Coking coal	25000	-	33000	≈ NCV + 5%
Anthracite	22000	-	29000	≈ NCV + 5%
Other bituminous coal	22000	-	29000	≈ NCV + 5%
Sub-bituminous coal	16000	-	24000	≈ NCV + 5%
Lignite	5000	-	18000	≈ NCV + 5%
Peat	7000	-	13000	≈ NCV + 5%
Oil Shale	2500	-	12000	≈ NCV + 5%

Expected GCV (kJ/m^3) range for natural gas

30 000 – 45 000

Different qualities of bituminous coal around the world

28 201 kJ/kg

18 581 kJ/kg

Time series analysis

Growth rates

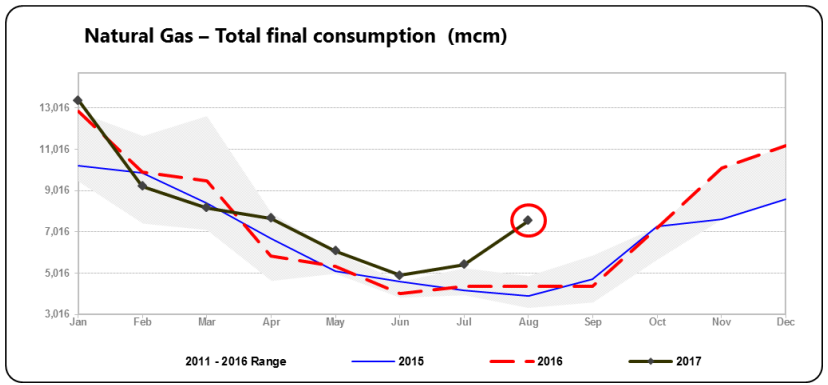
Time series breaks

Outliers

Time series analysis - Growth rates

$$\text{Growth Rate (\%)} = \frac{\text{Current period} - \text{Previous period}}{\text{Previous period}} \times 100$$

- When seasonality is strong or we are dealing with monthly data
  - It is often more useful to compare the current month with the same month of the previous year.



Production
Consumption

Less useful for

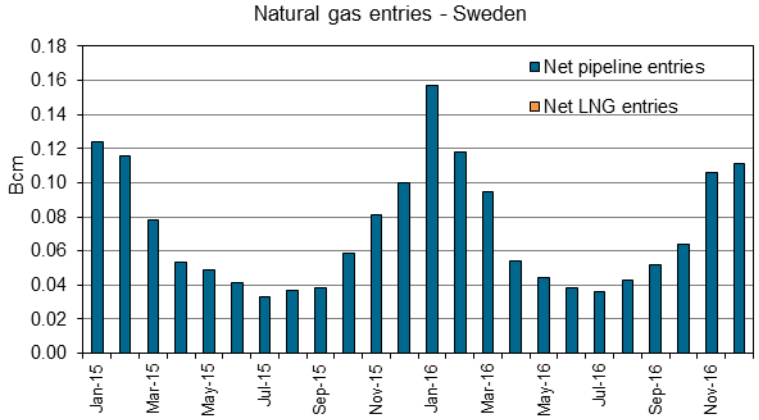
Stocks
Trade

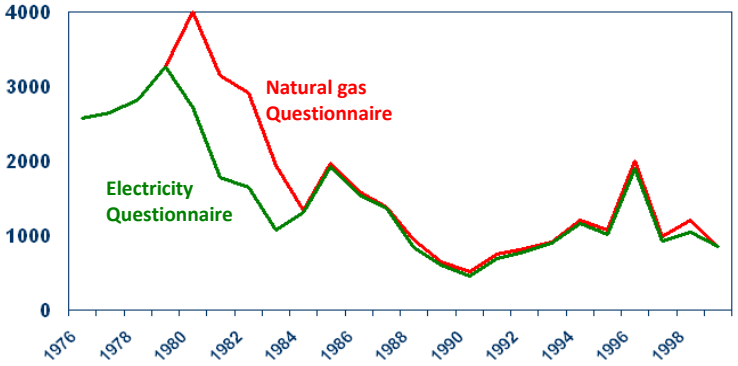
Time series analysis - Is the data set coherent?

- Does the data set match expectations for trends? (e.g. known maintenance, weather disruptions, seasonality patterns, changes in policy…).
  - Certain patterns may be fuel specific:
    - Natural gas demand and storage is linked to weather patterns.
    - Oil and gas fields follow cyclical maintenance patterns.

- Does the data match similar data reported in other questionnaires?
  - E.g. does inputs for electricity and heat match in the fuel and electricity data sets?

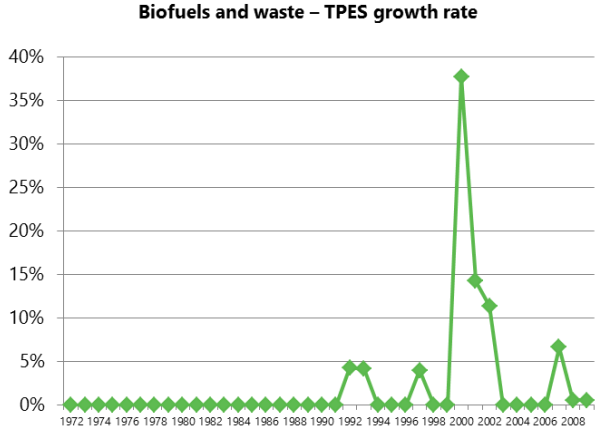
Example. *Sweden imports more natural gas in the winter months to cope with increasing demand that comes from colder temperatures.*

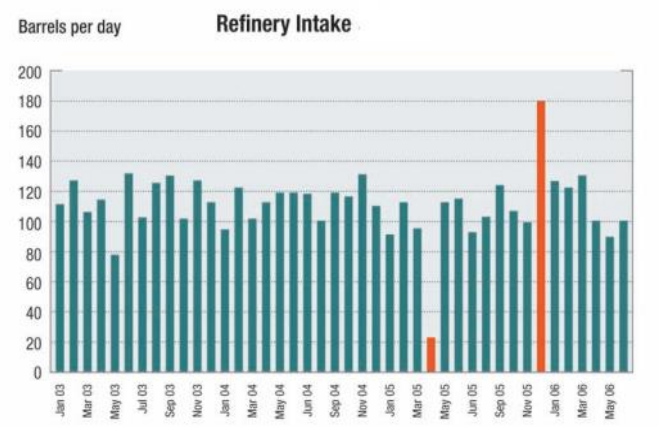




Time series analysis - Breaks in time series & unexplained outliers

- Although breaks can happen when coverage improves or methodology changes in general one should aim for a **consistent time series**.
- **Unexplained** breaks or big changes in trends should be avoided.
- **Breaks in stocks** are rare and should be investigated.
- Certain flows are generally consistent over time, thus sudden increases or decreases should raise a red flag.





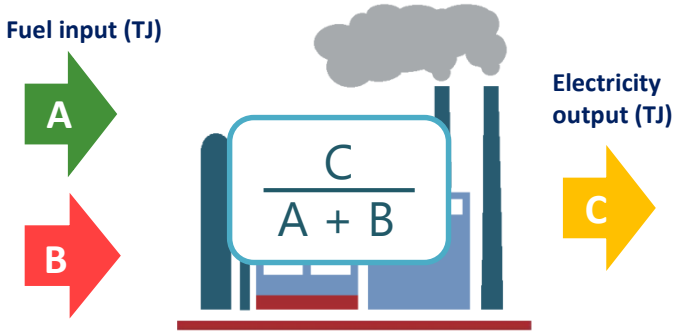
Efficiencies

Transformation efficiencies - Overview

- For every energy transformation process there is an **input** and an **output**.
  - There is a relationship between the inputs and outputs which will depend on the technology used.
- Both inputs and outputs need to be in:
  - The same energy unit.
  - The same energy form (net or gross basis)
- This efficiency then needs to be checked against the expected ranges for that process and technology.
  - An efficiency larger than 100% is implausible.

$$Efficiency = \frac{Total\ "useful"\ output\ (net\ energy\ units)}{Total\ input\ (net\ energy\ units)}$$

***Example**. Fuel-electricity transformation*



Transformation efficiencies - Expected efficiency ranges

- Calculating and checking efficiencies allows us to check the **consistency** of our data set.
- Efficiencies are expected to be within a certain range depending on the **type of plant** and the **fuel used**.
- The ranges are wide because efficiency can vary widely depending on the **state of technology** (old v. new).
- Efficiencies **do not vary** significantly from year to year as technology upgrades are time and capital intensive.

Type of plant	Expected Efficiency range
Electricity only plants	10 - 40%
CHP plants	30-80%
Heat only plants	40-100%
Refineries	95-100%
Blast furnaces	35-45%
Coke ovens	67-100%* *(coke oven coke + coke oven gas)
Patent fuel plants	90-100%
BKB	85-100%
Gas works	67-100%* *(gas works gas + gas coke)
Charcoal	25-55%

Transformation efficiencies – Dealing with unusual efficiencies

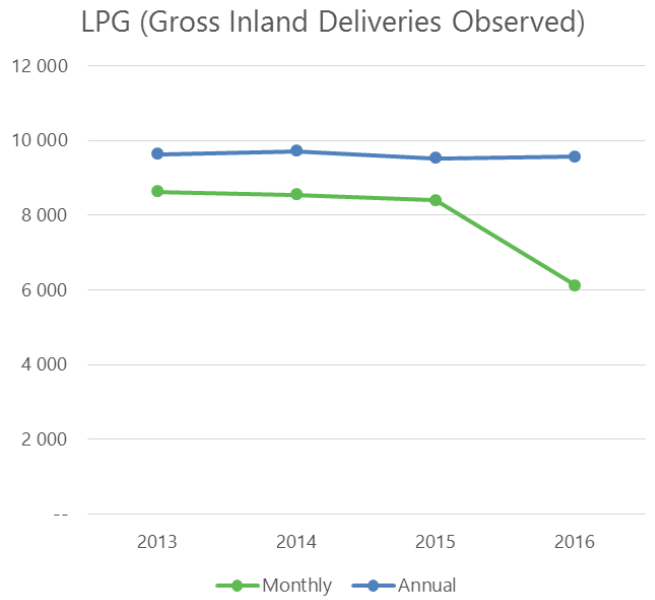
Potential issue	What can be done…
Incomplete **input** or **output** data (measurement errors, different source institution…)	Check the correct **definitions** are being used. Missing cells in energy balance, check **logical flows** for the fuel.
Heat sold output for CHP and heat only plants is not being correctly accounted for.	Review **methodology for heat output** estimation.
Conversion to energy units of either the input our output.	Are all the flows being converted on a **net basis** (NCV)? Is the **calorific value** data within the expected ranges?
Incorrect accounting of industry **own use** and **output sold**.	Check **boundaries** of statistical reporting and that the correct definitions are being used.

Checking for coherence – Monthly/quarterly data

- Good quality monthly data can be a tool for data validation as well as key for energy security purposes.
  - Does the aggregate of the monthly/quarterly data make sense with the annual data?

Somethings to look out for…

- Revisions
- Different definitions or coverage due to timing constraints.

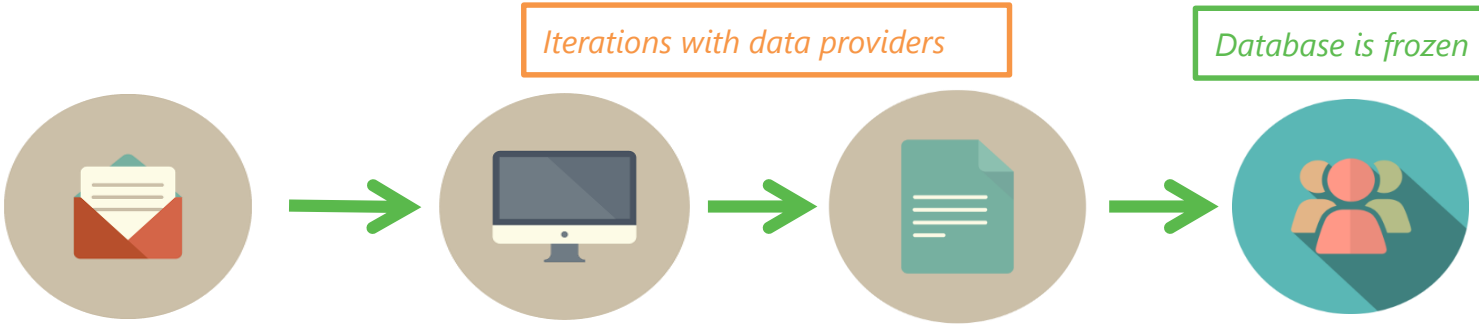


Source: IEA, Monthly oil data service, Oil Information 2017

Data validation in practice

Data validation processes - Approaches

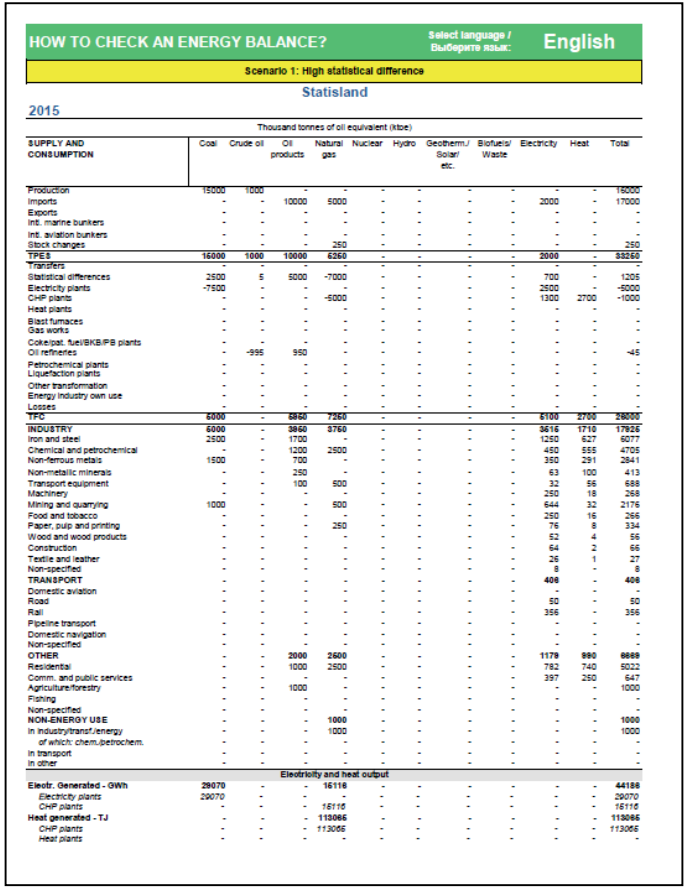
An overview of how data is processed and validated at the IEA



Practical exercise – Checking the energy balance

- In groups discuss the different scenarios for checking the energy balance sheet.

- As a group, go through the questions raised in the files for each scenario, thinking about:
  - Checks that could be applied for checking the data when received?
  - What are some of the systems that are in place in your countries?



HOW TO CHECK AN ENERGY BALANCE? Select language / Выберите язык: English

Scenario 1: High statistical difference

Statisland

2015

Thousand tonnes of oil equivalent (ktoe)

SUPPLY AND CONSUMPTION	Coal	Crude oil	Oil products	Natural gas	Nuclear	Hydro	Geotherm./ Solar/ etc.	Biofuels/ Waste	Electricity	Heat	Total
Production	15000	1000	-	-	-	-	-	-	-	-	16000
Imports	-	-	10000	5000	-	-	-	-	2000	-	17000
Exports	-	-	-	-	-	-	-	-	-	-	-
Int. marine bunkers	-	-	-	-	-	-	-	-	-	-	-
Int. aviation bunkers	-	-	-	-	-	-	-	-	-	-	-
Stock changes	-	-	-	250	-	-	-	-	-	-	250
TPES	**16000**	**1000**	**10000**	**5250**	-	-	-	-	**2000**	-	**33250**
Transfers	-	-	-	-	-	-	-	-	-	-	-
Statistical differences	2500	5	5000	-7000	-	-	-	-	700	-	1205
Electricity plants	-7500	-	-	-	-	-	-	-	2500	-	-5000
CHP plants	-	-	-	-5000	-	-	-	-	1300	2700	-1000
Heat plants	-	-	-	-	-	-	-	-	-	-	-
Blast furnaces	-	-	-	-	-	-	-	-	-	-	-
Gas works	-	-	-	-	-	-	-	-	-	-	-
Coke/pat. fuel/BKB/PB plants	-	-	-	-	-	-	-	-	-	-	-
Oil refineries	-	-995	950	-	-	-	-	-	-	-	-45
Petrochemical plants	-	-	-	-	-	-	-	-	-	-	-
Liquefaction plants	-	-	-	-	-	-	-	-	-	-	-
Other transformation	-	-	-	-	-	-	-	-	-	-	-
Energy industry own use	-	-	-	-	-	-	-	-	-	-	-
Losses	-	-	-	-	-	-	-	-	-	-	-
TFC	**5000**	-	**5950**	**7250**	-	-	-	-	**5100**	**2700**	**26000**
INDUSTRY	**5000**	-	**3950**	**3750**	-	-	-	-	**3615**	**1710**	**17925**
Iron and steel	2500	-	1700	-	-	-	-	-	1250	627	6077
Chemical and petrochemical	-	-	1200	2500	-	-	-	-	450	555	4705
Non-ferrous metals	1500	-	700	-	-	-	-	-	360	291	2841
Non-metallic minerals	-	-	250	-	-	-	-	-	63	100	413
Transport equipment	-	-	100	500	-	-	-	-	32	56	688
Machinery	-	-	-	-	-	-	-	-	250	18	268
Mining and quarrying	1000	-	-	500	-	-	-	-	644	32	2176
Food and tobacco	-	-	-	-	-	-	-	-	250	16	266
Paper, pulp and printing	-	-	-	250	-	-	-	-	76	8	334
Wood and wood products	-	-	-	-	-	-	-	-	52	4	56
Construction	-	-	-	-	-	-	-	-	64	2	66
Textile and leather	-	-	-	-	-	-	-	-	26	1	27
Non-specified	-	-	-	-	-	-	-	-	8	-	8
TRANSPORT	-	-	-	-	-	-	-	-	**408**	-	**408**
Domestic aviation	-	-	-	-	-	-	-	-	-	-	-
Road	-	-	-	-	-	-	-	-	50	-	50
Rail	-	-	-	-	-	-	-	-	356	-	356
Pipeline transport	-	-	-	-	-	-	-	-	-	-	-
Domestic navigation	-	-	-	-	-	-	-	-	-	-	-
Non-specified	-	-	-	-	-	-	-	-	-	-	-
OTHER	-	-	**2000**	**2500**	-	-	-	-	**1179**	**990**	**6669**
Residential	-	-	1000	2500	-	-	-	-	782	740	5022
Comm. and public services	-	-	-	-	-	-	-	-	397	250	647
Agriculture/forestry	-	-	1000	-	-	-	-	-	-	-	1000
Fishing	-	-	-	-	-	-	-	-	-	-	-
Non-specified	-	-	-	-	-	-	-	-	-	-	-
NON-ENERGY USE	-	-	-	**1000**	-	-	-	-	-	-	**1000**
In industry/transf./energy	-	-	-	1000	-	-	-	-	-	-	1000
of which: chem./petrochem.	-	-	-	-	-	-	-	-	-	-	-
In transport	-	-	-	-	-	-	-	-	-	-	-
In other	-	-	-	-	-	-	-	-	-	-	-
Electricity and heat output											
Electr. Generated - GWh	**29070**	-	-	**16116**	-	-	-	-	-	-	**44188**
Electricity plants	*29070*	-	-	-	-	-	-	-	-	-	*29070*
CHP plants	-	-	-	*16116*	-	-	-	-	-	-	*16116*
Heat generated - TJ	-	-	-	**113066**	-	-	-	-	-	-	**113066**
CHP plants	-	-	-	*113066*	-	-	-	-	-	-	*113066*
Heat plants	-	-	-	-	-	-	-	-	-	-	-

www.iea.org/statistics

IEA