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?

<table>
<thead>
<tr>
<th>Quality aspect</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relevance</td>
<td>Statistics meet the needs of the users</td>
</tr>
<tr>
<td>Accuracy and reliability</td>
<td>Statistics accurately and reliably portray reality</td>
</tr>
<tr>
<td>Timeliness and punctuality</td>
<td>Statistics are released in a timely and punctual manner</td>
</tr>
<tr>
<td>Accessibility and clarity</td>
<td>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.</td>
</tr>
<tr>
<td>Coherence and comparability</td>
<td>Statistics are consistent internally, over time and comparable between regions and countries.</td>
</tr>
</tbody>
</table>

For more information please see: [Fundamental Principles of Official Statistics](https://www.oecd.org) and [International Recommendation for Energy Statistics (IRES)](https://www.iea.org)
Balance checks

- Internal consistency
- Statistical 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.

<table>
<thead>
<tr>
<th>Production</th>
<th>Use in transformation processes and Final consumption</th>
</tr>
</thead>
<tbody>
<tr>
<td>+ Imports</td>
<td></td>
</tr>
<tr>
<td>- Exports</td>
<td></td>
</tr>
<tr>
<td>- International marine and aviation bunkers</td>
<td>How much was consumed in the market?</td>
</tr>
<tr>
<td>- Stock changes</td>
<td></td>
</tr>
<tr>
<td><strong>= Supply</strong></td>
<td></td>
</tr>
</tbody>
</table>

How much is available to the internal 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...)

© OECD/IEA 2017
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.

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

**Expected GCV (kJ/m³) 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.

![Diagram of Natural Gas Total final consumption (mcm) with shaded areas and trend lines for 2011-2016 range, 2016, 2016, and 2017 years.]

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.

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

*Example*: Fuel-electricity transformation

\[
\frac{C}{A + B}
\]
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.

<table>
<thead>
<tr>
<th>Type of plant</th>
<th>Expected Efficiency range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Electricity only plants</td>
<td>10 - 40%</td>
</tr>
<tr>
<td>CHP plants</td>
<td>30-80%</td>
</tr>
<tr>
<td>Heat only plants</td>
<td>40-100%</td>
</tr>
<tr>
<td>Refineries</td>
<td>95-100%</td>
</tr>
<tr>
<td>Blast furnaces</td>
<td>35-45%</td>
</tr>
<tr>
<td>Coke ovens</td>
<td>67-100%*</td>
</tr>
<tr>
<td></td>
<td>*(coke oven coke + coke oven gas)</td>
</tr>
<tr>
<td>Patent fuel plants</td>
<td>90-100%</td>
</tr>
<tr>
<td>BKB</td>
<td>85-100%</td>
</tr>
<tr>
<td>Gas works</td>
<td>67-100%*</td>
</tr>
<tr>
<td></td>
<td>*(gas works gas + gas coke)</td>
</tr>
<tr>
<td>Charcoal</td>
<td>25-55%</td>
</tr>
</tbody>
</table>
## Transformation efficiencies – Dealing with unusual efficiencies

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

Data are received → Processing → Time Series Check → Further checking and peer review → Database is frozen → Publishing

- Inconsistencies such as breaks in stocks, negative values, sum errors, inter-table errors...
- Consistency, Balance Check, Seasonality and Outliers, Calorific Values...
- Outliers and Seasonality

Iterations with data providers
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?