Data Analysis & Checking Methods

Bülent TUNGUL

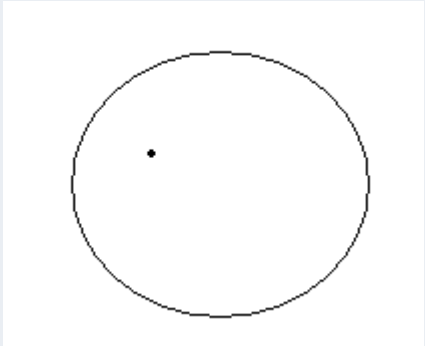
SESRIC Statistical Cooperation Programme
Workshop on External Trade Statistics

6-8 January 2013
Kuwait

Data correctness

Control of single data elements:

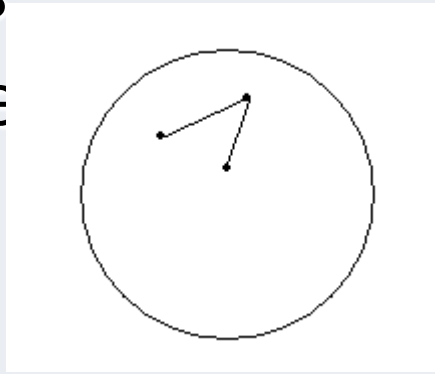
- existence and timeframe
- validity of codes,
- absence of data,
- negative values,
- extremely high values,
- zero values



Transaction coherence

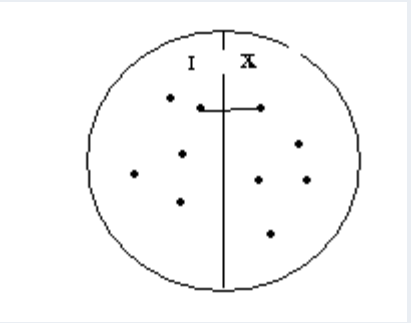
Coherence of different data elements:

- Variables within a transaction should be coherent:
  - customs procedure codes with flow,
  - commodity codes with country codes
  - commodity codes with transport code



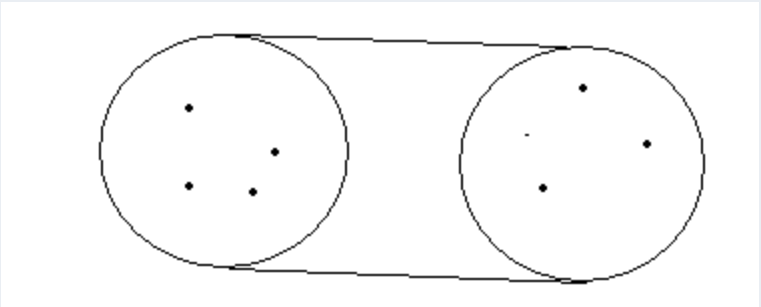
Declarant coherence

- Coherence of data elements over transactions: stability of products imported and exported
- Coherence of flows declared by the same declarant: monitor the trader



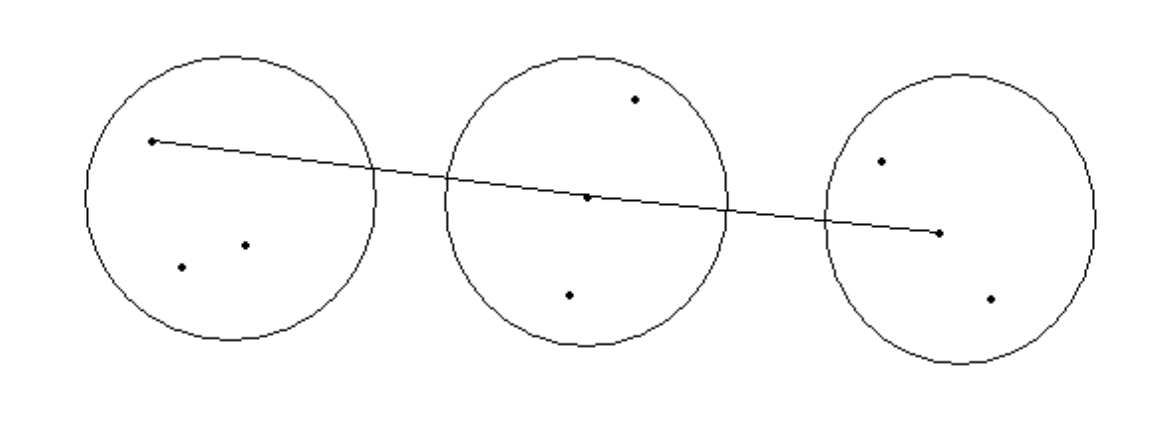
Declarant coherence

- Coherence with external data of the same declarant: comparison with VAT, prodcom (production for export), business surveys, turnover data from balance sheets



Declarant coherence

- Coherence of data elements over time: stability of products traded over time.



Population coherence

- Coherence of company prices to average prices for all other companies: distance to average market prices and average weight per unit (value:kilo, value:units, kilo:units)

(Unit value analysis)

Aggregates coherence

- Time series analysis, outlier detection:
- An outlier in a given month in a time series on HS2 level for example will initiate an identical outlier detection procedure for all the HS4 headings that are part of the failed HS2 heading, etc. Drilling down this way will allow the identification of the companies/transactions responsible for the outlier in the given month/HS2 combination.

Aggregates coherence

- Price jumps in indices
- Mirror statistics
- Asymmetry analysis
- Reconciliation analysis

Quality Choices

given:

- mountain of information: high number errors
- wide variety of possible checks
- limited staff
- limited patience of declarants

Quality Choices

- what to check, which checks/tools to use?
- which errors are important? (impact totals)
- which are urgent? (upcoming publication)
- limit the no. of checks, or correct automatically?
- which errors = absolute/probable? need correction or not?
- which errors require contact with declarant?
- optimal planning of checks? (in sync with publications)

VALIDATION ERRORS

- Data is certainly Incorrect
- May cause processing Problems
- Cannot be included in statistical analysis
- Must be removed/corrected
- Examples
  - text characters in value field
  - invalid commodity or country code

CREDIBILITY ERRORS

- Data is valid but possibly incorrect
- Can be processed
- Will distort statistical analysis if included
- May be expensive to correct
- Examples
  - inflated value or quantity (*000)
  - valid but wrong commodity code

RATIOS

- The traditional ratios that are used:
  - V/Q1
  - V/Q2
  - Q1/Q2
- These ratios should be within:
  - Min < V/Q1 < Max
  - Min < V/Q2 < Max
  - Min < Q1/Q2 < Max

What we do in TurkStat

- Validity of codes
- Imputation of missing values
- Time series analysis
- Specific analysis
- Unit value analysis

Validity of codes

- All code lists are available in Oracle database,
- All are checked automatically while importing Customs data by SAS,
- Autocorrection for some code errors (commodity, mode of transport, partner)
- Errors in codes are reported by SAS
- Code errors are corrected (commodity code, customs office code, currency code etc)

Imputation of missing values

- SAS programme sets default values for missing codes,
- Partner country code,

Time series analysis

- SAS generates automatically time series for last 2 years;
  - by chapters
  - by partner country
  - by customs office
  - By type of payment
  - By nature of transaction
  - By currency

Time series analysis

- SAS generates automatically time series for last 2 years;
  - by chapters
  - by partner country
  - by customs office
  - By type of payment
  - By nature of transaction
  - By currency

Specific analysis

- SAS generates automatically;
  - Ships transaction analysis
  - Aircraft transaction analysis
  - Repair
  - Goods for processing
  - New enterprises

Unit value analysis

- SAS programme enables;
  - Automatic flagging for implausible transaction
  - Selection of all transaction control or flagged transaction control
  - Generates excel table according to selection for each staff

Unit value analysis

- In current system;
  - Credibility intervals
    - Using last two years data,
    - For exports and imports
    - For each commodity
    - For unit value per kg, per supp.unit, per ratio (kg/s.u.)
  - Unit, Clossness, Min, Mean, Median, Max, Number of transaction, flags for indicators
  - Unit values for enterprises (by flow, enterprise and commodity)
  - Suggested quantity

Be careful!

If you find outlier;

- Quantity may be wrong (net weight or supp.quantity)
- Commodity code may be wrong

Outlier method

- In new system;
  - Z-score method will be used
    - Using last two-three years data,
    - For exports and imports
    - For each commodity
    - For unit value per kg, per supp.unit, per ratio (kg/s.u.)

Outlier method

Z-score method;

- N >= 30
- Z-score >3 outlier
- Median z-score is also applied when mean is lower than standard deviation
- Med-zscore > 5.2 outlier

Outlier method

Z-score method;

$$z\text{-}score = (y_t - \bar{y}) / s_y$$

$$median\ z\text{-}score = (y_t - median) / MAD$$

Conclusion

❖The z-score method is one of the most effective methods in the detection of outliers

- Detect important outliers
- Reduces number of false outliers

❖While implementing you may apply;

- Automatic correction
- Manual correction
  - Asking customs
  - Asking enterprises
  - Experience

Thank you.

Any questions?