Input quality of administrative data

BLUE-ETS WP4

Piet Daas, Saskia Ossen, Martijn Tennekes, Joep Burger and Fannie Cobben

EUROPEAN COMMISSION
European Research Area

SEVENTH FRAMEWORK PROGRAMME

Funded under Socio-economic Sciences & Humanities

In this presentation…

- Introduction
- Dimensions of input data quality
- Examples of input data quality indicators
- What to expect from WP4?
- Questions

Introduction (1)

- More and more statistical institutes are using administrative sources for statistical purposes
- They become more *dependent* on data sources collected and maintained by *others*
- Need to monitor the *quality* of those data sources when they *enter* the office

Introduction (2)

- The main goal of WP4 is to improve the use of administrative sources
- By developing a **standardized** way to determine the **quality** of administrative sources for **statistical purposes**:
  - Dimensions of quality
  - Indicators for each dimension
  - Quality Report Card (QRC)

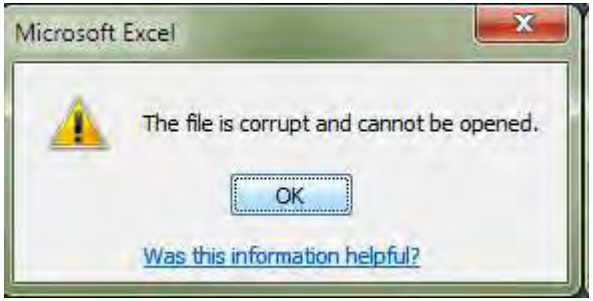
Dimensions of input data quality

- Essential quality dimensions for input data of administrative sources

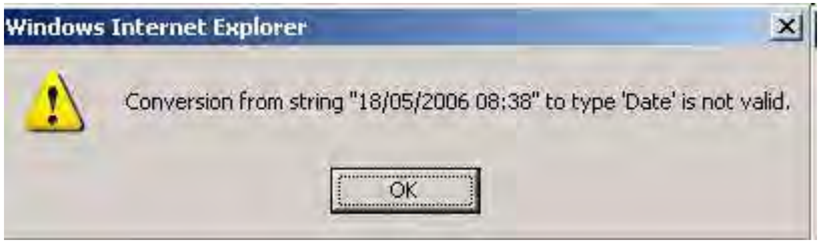
  1. Technical checks
     - Technical usability of the file and data in the file
  2. Accuracy
     - Extent to which data are correct, reliable and certified
  3. Completeness
     - Degree to which a data source includes data describing the corresponding set of real-world objects and variables
  4. Time-related dimension
     - Indicators that are time and/or stability related
  5. Integrability
     - Extent to which the data source is capable of undergoing integration or of being integrated

Examples of input data quality indicators:
Technical checks

- Very important for ***new*** sources, becomes somewhat less essential later on
  - Corrupt files
  - Encoded files of which decoding password is missing
  - Files of which the data is not compliant to the metadata description
  - Files with errors during/after conversion



The file is corrupt and cannot be opened.



Conversion from string "18/05/2006 08:38" to type 'Date' is not valid.

Examples of input data quality indicators:
Accuracy: Authenticity

- Objects with incorrect Identification numbers (ID’s)
- In the Netherlands all people have a Citizen’s Service Numbers
  - 9-digit number (e.g. 123456782)
  - Number has a feasibility check, last digit is a checking digit
  - Rule used: sum($9*n_1 + 8*n_2 + 7*n_3 + 6*n_4 + 5*n_5 + 4*n_6 + 3*n_7 + 2*n_8 – 1*n_9$)

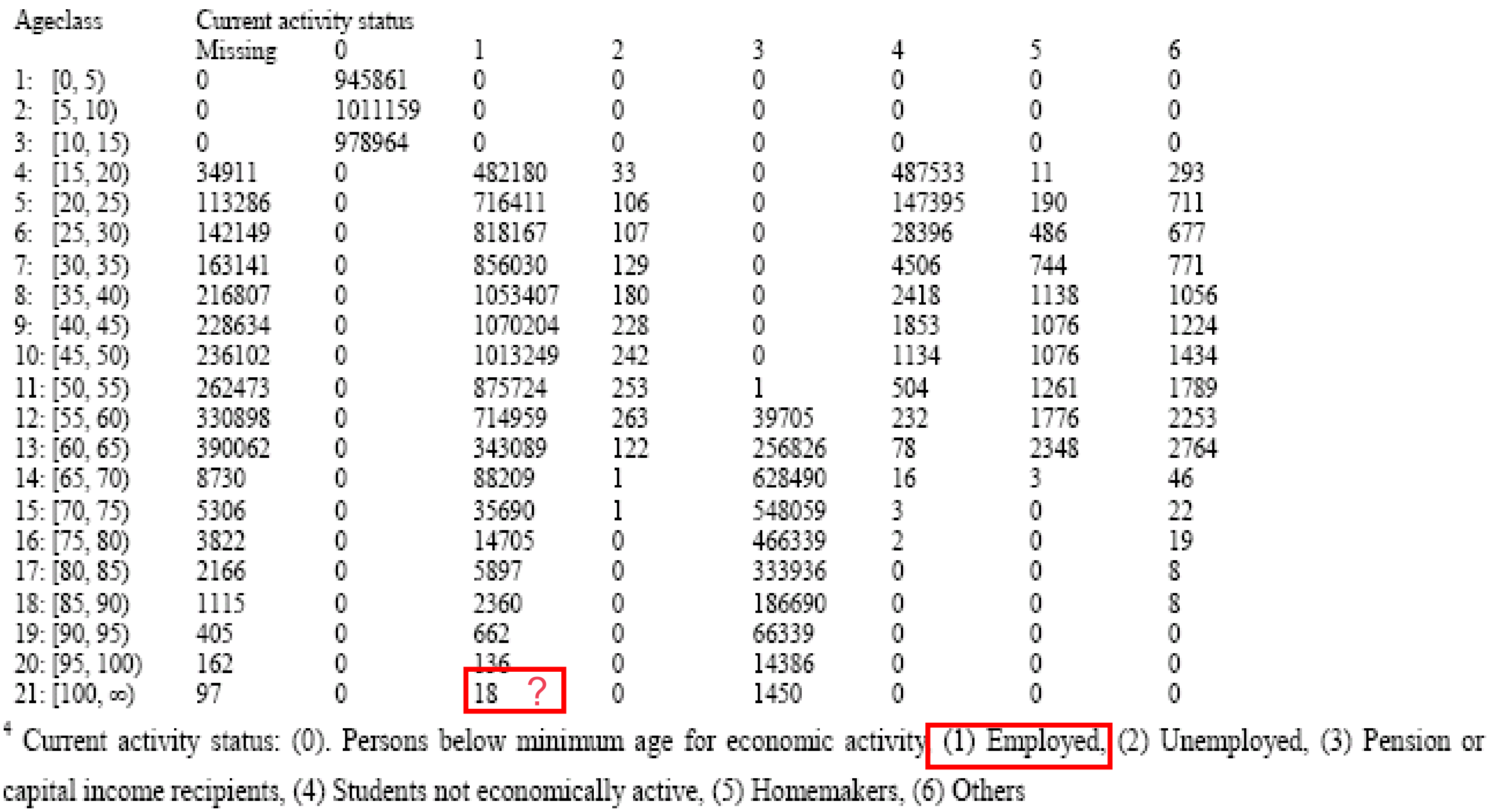
    Remainder of sum/11 should be 0
- In the Social Statistical Database* it was found (in 2000) that:
  - 0,3% of all persons in admin. data sources used had an invalid Citizen Service Number

*set of integrated admin. data sources and surveys (then ~100 million admin records)
Arts et al. (2000) *Netherlands Official Statistics* 15, pp. 16-22.

Examples of input data quality indicators:

Accuracy: Dubious values

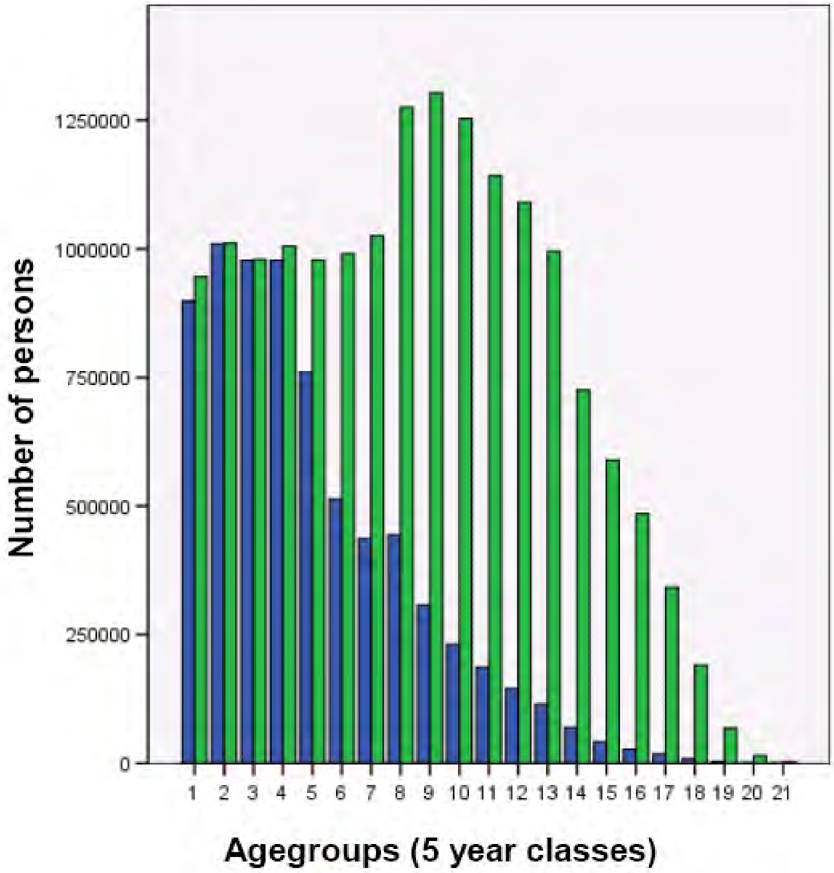
Cross tabulation of the variable “Current activity status” versus age group



Ageclass	Current activity status							
	Missing	0	1	2	3	4	5	6
1: [0, 5)	0	945861	0	0	0	0	0	0
2: [5, 10)	0	1011159	0	0	0	0	0	0
3: [10, 15)	0	978964	0	0	0	0	0	0
4: [15, 20)	34911	0	482180	33	0	487533	11	293
5: [20, 25)	113286	0	716411	106	0	147395	190	711
6: [25, 30)	142149	0	818167	107	0	28396	486	677
7: [30, 35)	163141	0	856030	129	0	4506	744	771
8: [35, 40)	216807	0	1053407	180	0	2418	1138	1056
9: [40, 45)	228634	0	1070204	228	0	1853	1076	1224
10: [45, 50)	236102	0	1013249	242	0	1134	1076	1434
11: [50, 55)	262473	0	875724	253	1	504	1261	1789
12: [55, 60)	330898	0	714959	263	39705	232	1776	2253
13: [60, 65)	390062	0	343089	122	256826	78	2348	2764
14: [65, 70)	8730	0	88209	1	628490	16	3	46
15: [70, 75)	5306	0	35690	1	548059	3	0	22
16: [75, 80)	3822	0	14705	0	466339	2	0	19
17: [80, 85)	2166	0	5897	0	333936	0	0	8
18: [85, 90)	1115	0	2360	0	186690	0	0	8
19: [90, 95)	405	0	662	0	66339	0	0	0
20: [95, 100)	162	0	136	0	14386	0	0	0
21: [100, ∞)	97	0	18 ?	0	1450	0	0	0

[4] Current activity status: (0). Persons below minimum age for economic activity (1) Employed, (2) Unemployed, (3) Pension or capital income recipients, (4) Students not economically active, (5) Homemakers, (6) Others

Examples of input data quality indicators:
Completeness: Selectivity

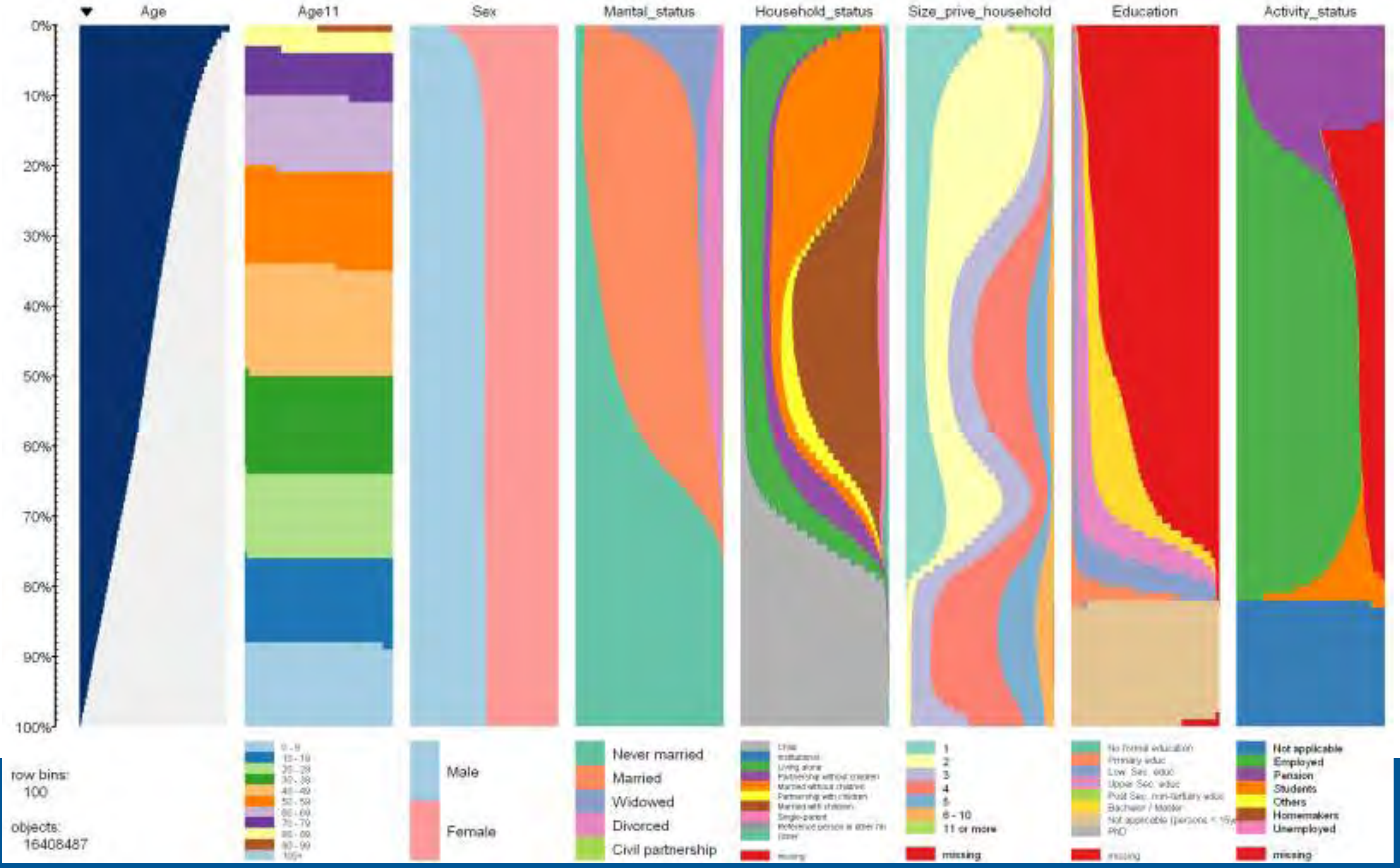


The education register has *age-related* undercoverage of educational attainment (56,3% is missing)

Explanation:
1) Children <15 age have a known level of education
2) Level of education of young adults is usually stored in recently created admin. data sources
3) Information from ‘middle-aged’ people is obtained from LFS-survey (small compared to admin. data info)
4) Information of ‘elderly’ people (≥65 year) almost completely missing (not surveyed and hardly registered)

Examples of input data quality indicators:
Completeness: Missing values

Tableplot of Dutch virtual census (Test version, ~16,5 million people)



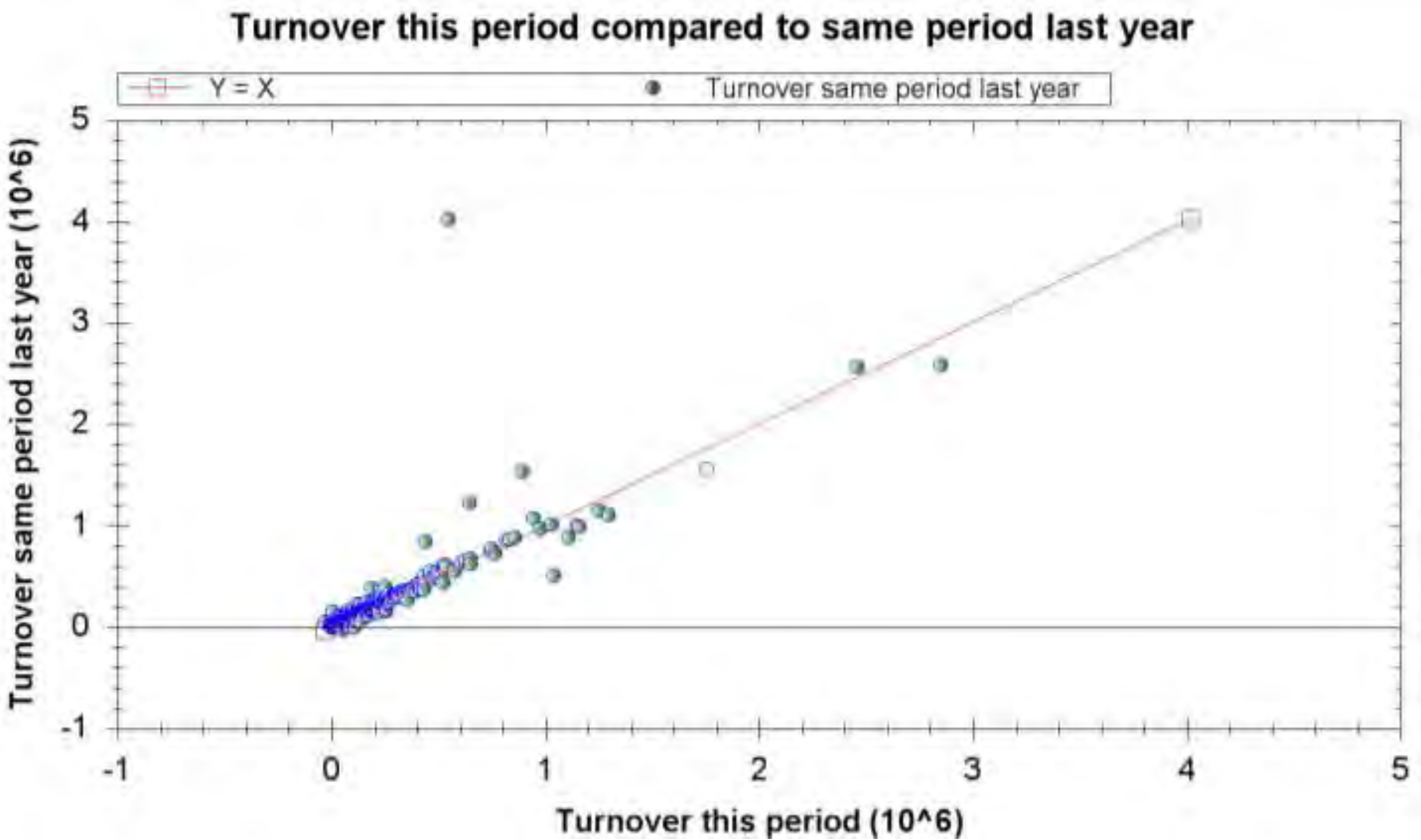
Examples of input data quality indicators:
Time-related: Delay

- Events recorded some time after they have occurred
  - Events are missing (or erroneously recorded)
  - Particularly important for sources used immediately

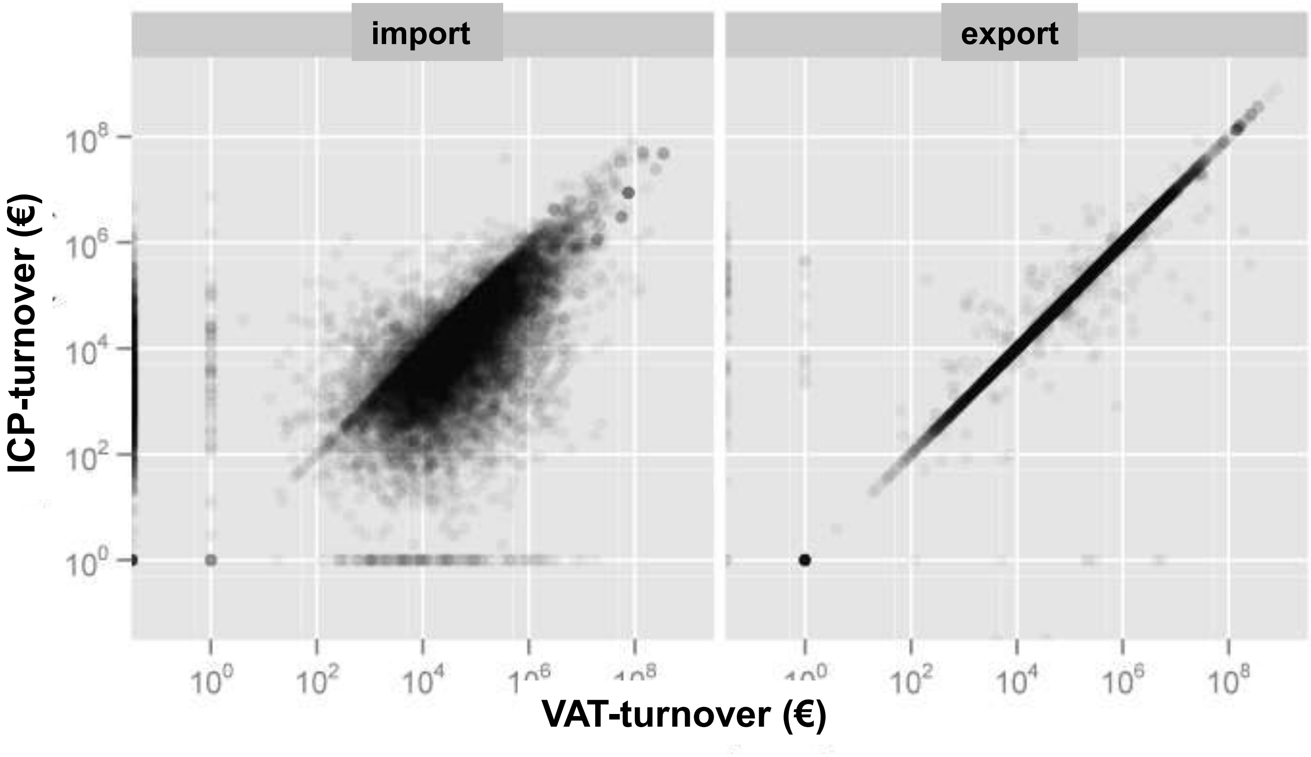
- Examples:
  - Netherlands: Marriages contracted in immigrants' country of origin are sometimes recorded two or three years after the event (Bakker et al. AIOS-paper 2008)
  - Norway: Corrections in Persons Register are received over a lengthy period. Even months after the event has taken place (Zhang, presentation in 2011)
  - ~ Netherlands and more: Part of VAT-data is reported later than is needed for monthly estimates (Vlag, ISI-paper 2011)

Examples of input data quality indicators:
Time-related: Stability

Type of comparison used in the Dutch Short term Statistics



Examples of input data quality indicators:
Integrability: Allignment



Differences between two admin. data sources (ICP and VAT) both used for International trade statistics

Export aligns good but import is much more problematic!

Explanation:
-*ICP import* units are difficult to identify and can therefore not always by linked correctly

-ICP export data can be integrated well.

VAT: Value Added Tax data, ICP: Intra-Community Product transactions (EU-countries)

What to expect from WP4?

- Scripts for measurement methods
  - In R
- Quality Report Card (QRC)
  - Scoring form
    - Score per dimension (+/o/- or smiley's ;-)
- Guidelines for QRC use
  - *Evaluation* sequence and *instructions* for use

What to expect from WP4?

- In June 2012
  - R-scripts, QRC and instructions will be available *within* the project (as a draft version)
- In 2012 case studies by each partner
  - Results will be combined
- Aim to finalize work at end of nov. 2012
  - To enable combined reporting in Jan. 2013

Thank you for your attention!
Questions?