Quantitative Data Quality Assessment for Normalization Reference

Sangwon Suh, K. Lee

Contents

- Normalization Reference(NR) calculation project In Korea (1998-1999)

  - NRs are calculated for 8 impact categories

  - 1 year for data collection and calculation

  - 4 years for revision and update

Roll of Data Quality Issue in NR

In developing NR

- Individual data source level :
- NR for an impact category level :

In using NR

-Whole NRs set level :

- (How much will it vary the result ?/record keeping) :

Matter of choice /requires a tool to assist the selection procedure

Data Quality(DQ)

-Degree of confidence(SETAC, 1994)

- Ability to satisfy stated requirements*(ISO, 1998)

- (Additional uncertainty or Variability on the final results)

- *(Separation between data quality and variability)*

---

* Data quality requirements :
time-related coverage, geographical coverage, technology coverage, precision, completeness, representativeness, consistency, reproducibility

DQA discussions

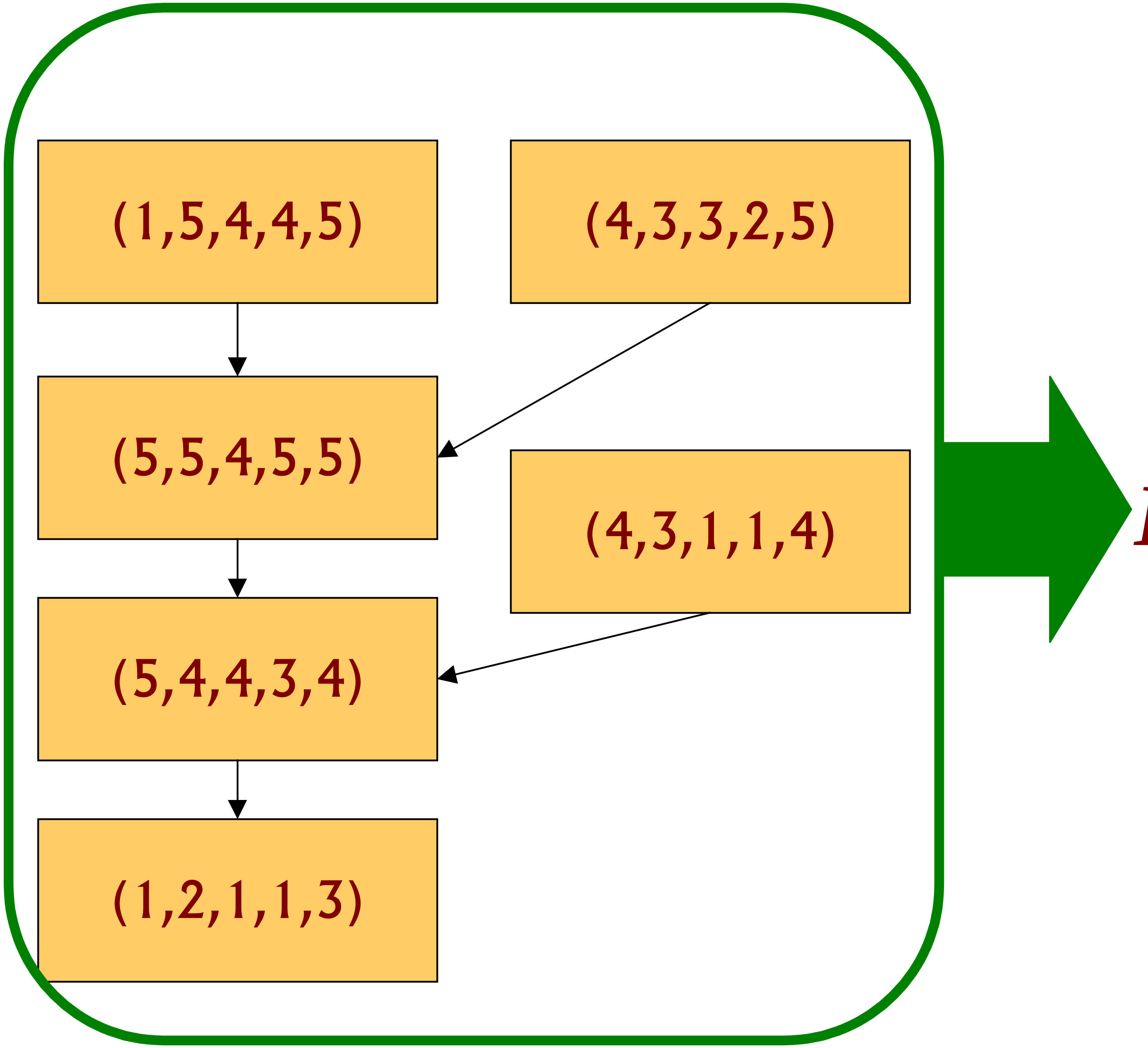
-Pedigree matrix (Weidema, 1994; Weidema, 1998)

- Semi-quantitative approach (Wrisberg, et al., 1996; Lindeijer & van den Berg,1997)

- Stochastic model (Kennedy, et al., 1997)

Introduction – Background [5]

	Pros	Cons
Pedigree Matrix	Widely used/Easy to apply Keeps important meta- data	No condensed overall information for the choice of a dataset in systems level
Semi-quantitative	Convey condensed informati on for users in systems level	Arbitrariness in scales between indicators and within a indicator (score 4 vs. score 2)
Stochastic model	Has scientific basis Looks objective	Doubt on the relevance of the basic assumption, etc. Arbitrariness in the absence of statistical info.



Data Quality?

Levels	Data acquisition method	Data age		
				...
1				...
2	2			...
3				...
4	4	4		...
5				...

CML

Scope & Objectives

- Provide a DQA method to assist selection procedure

  (for the purpose of internal use)

- Demonstrate it with actual NR data

- Survey on NR DQ

Definition

Data Quality

Utility of data determined by their attributes that are required to satisfy the goal of a study

(Variability or discernability using statistical data and simulation e.g. Monte Carlo Simulation is separated from the scope of DQ in this study)

Data source pool & Priority list

- Exhaustive list of data source selection priority

Priority List

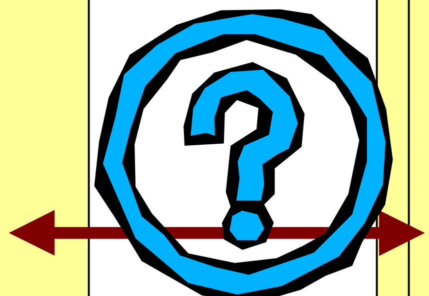
- Verified data published by competent international organization
- Verified data published by governmental organization
- Data reported by concerned parties(eg. companies) and aggregated by independent 3rd party

Which one should be chosen ?

Data from an concerned party

Based on on-site measurement

Exactly matches the reference year



Data from an Independent international organization

Based on estimation

2 years of deviation

Procedure of DQ calculation

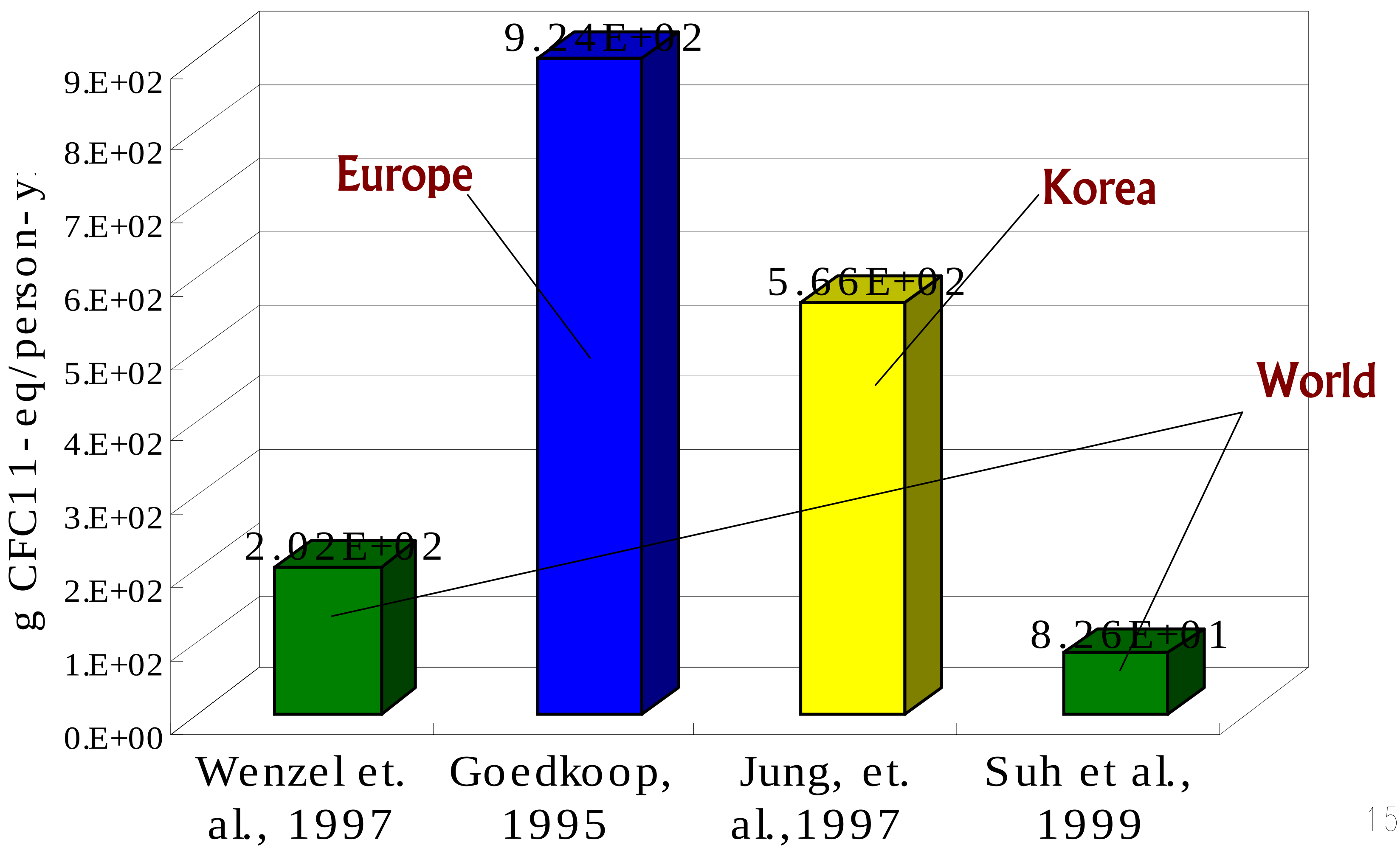
- DQ of a data source
  1. Indicator setting
  2. Selecting and grading quality levels of an indicator
  3. Weighting each indicator
  4. Summing up weighted indicators

- DQ of a NR

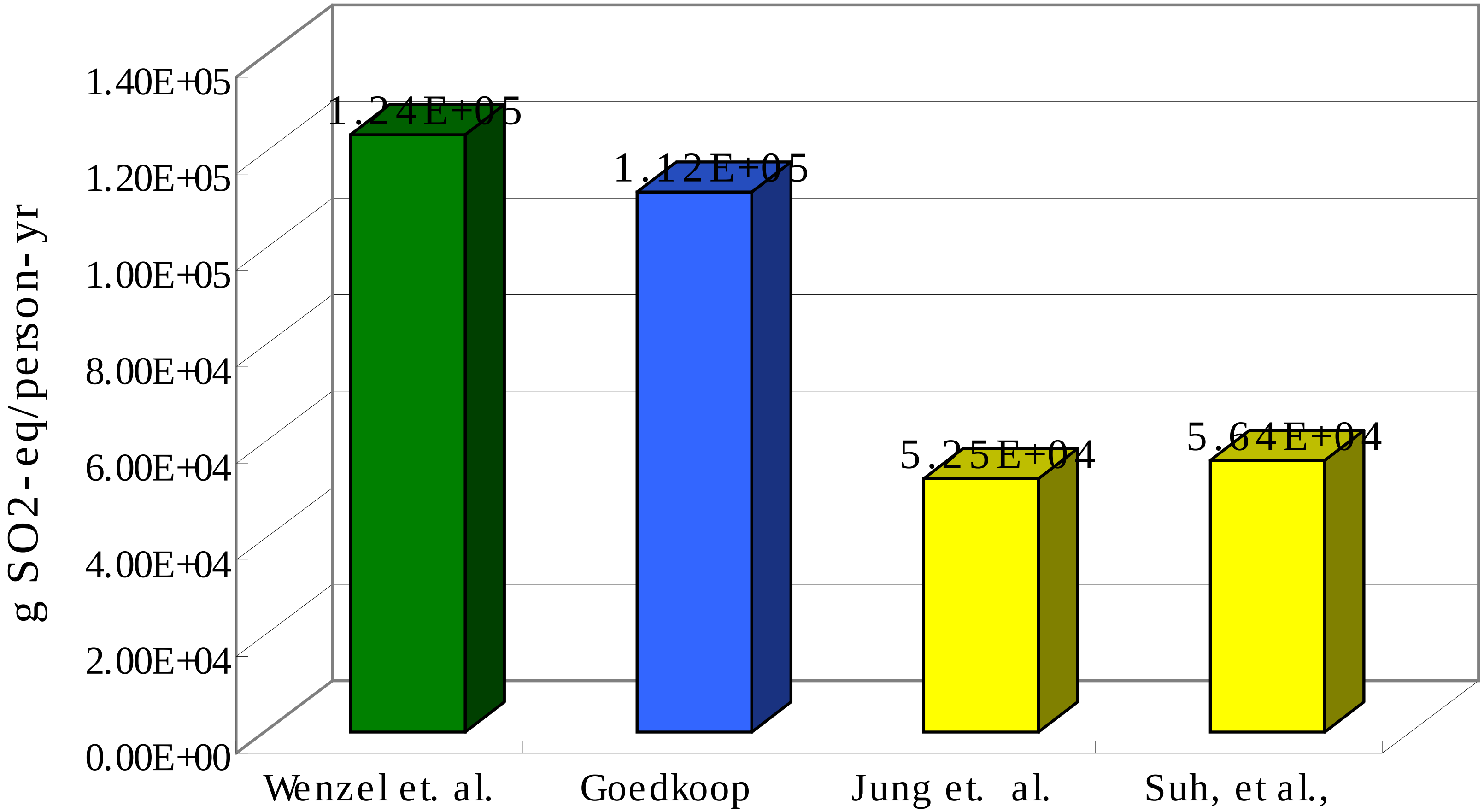
  Sum up according to the fractional contribution (characterized impact) of each data source

Indicator setting

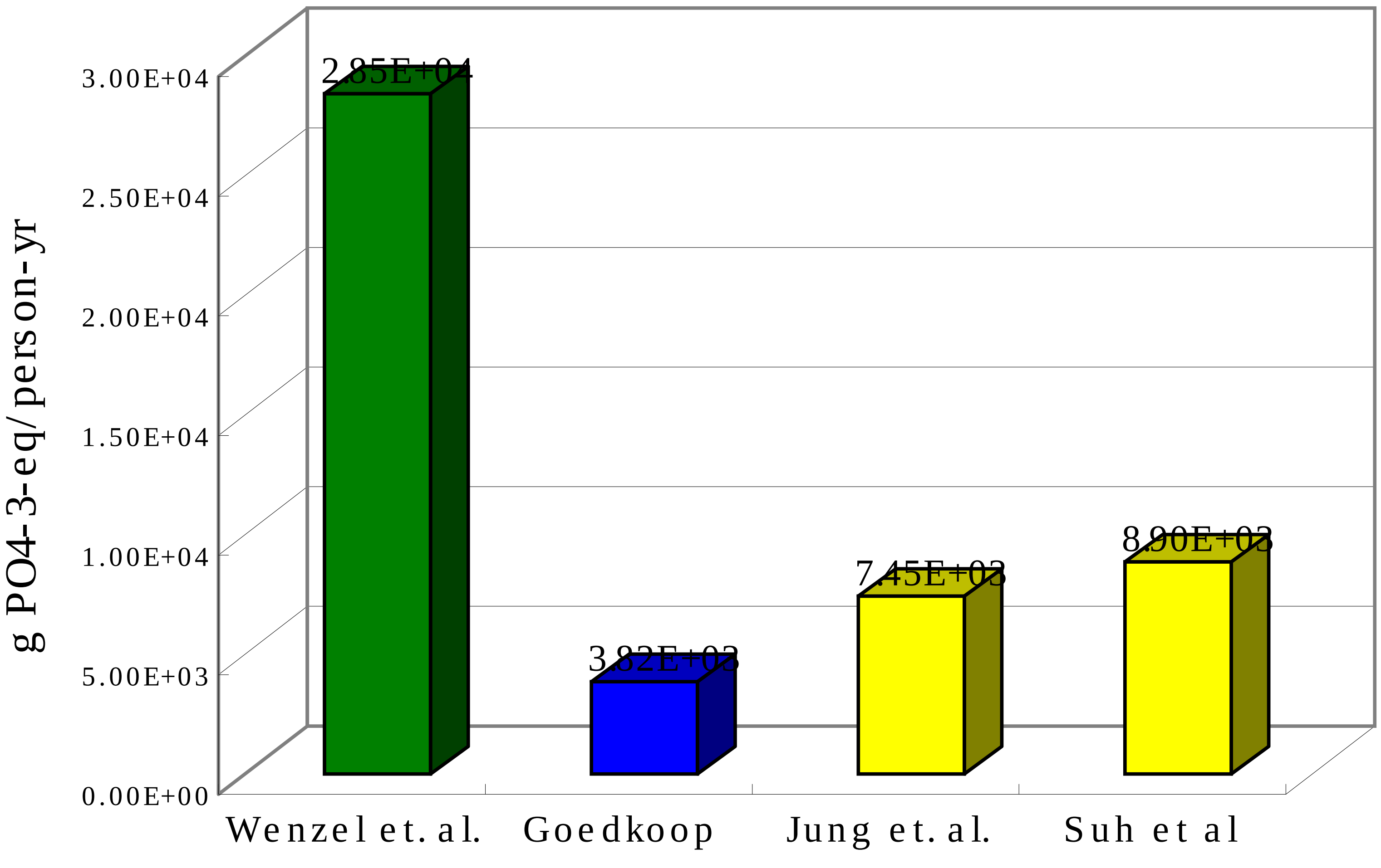
Based on the main origins cause deviation between studies



< Acidification >



< Eutrophication >



Selected Indicators

1. Data acquisition method
2. Independence of the data supplier
3. Completeness of the pollutant composition
4. Completeness of the pollution source composition
5. Data age

* Geographical system boundary was excluded since data sources of which geographical boundary were different from this study were eliminated from the data pool

CML

Selecting and grading quality level/weighting

Different importance-weighting ↔

Arbitrary scale-grading ↕

Levels	Data acquisition method	Data age	...
1	On-site measurement	0 years of interval	...
2	Emission factor based on on-site measurement	1 year	...
3	Calculation based on scientific assumption	2 years	...
4	Based on both scientific assumption and estimation	3 years	...
5	Assumption	More than 4 years	...

CML

Data quality score of *i*th data source

$$DQ_i = \sum_j w_j \cdot g_j$$

Where,

w_j : weighting factor for *j*th indicator

g_j : grade of a data source in a *j*th indicator

CML

Data quality score for NR of *k*th impact category

$$DQ_{NR_k} = \sum_i f_i \cdot DQ_i$$

Where,

f_i : fractional contribution of data source i for NR of kth impact category

DQ_i : Weighted data quality of ith data source

Survey design

Questionnaire distributed
: A total of 53 ISO delegates, LCIA practitioners and researchers from 9 countries

Questionnaire replied
: A total of 24 questionnaires were replied from 7 countries

Weights between indicators
: 10 questions of pair-wise comparison

Normalizing arbitrary scale of quality levels
: 25 questions on grading each levels

3. Compare A and B

A : methodology of data derivation
B : completeness in pollution source composition

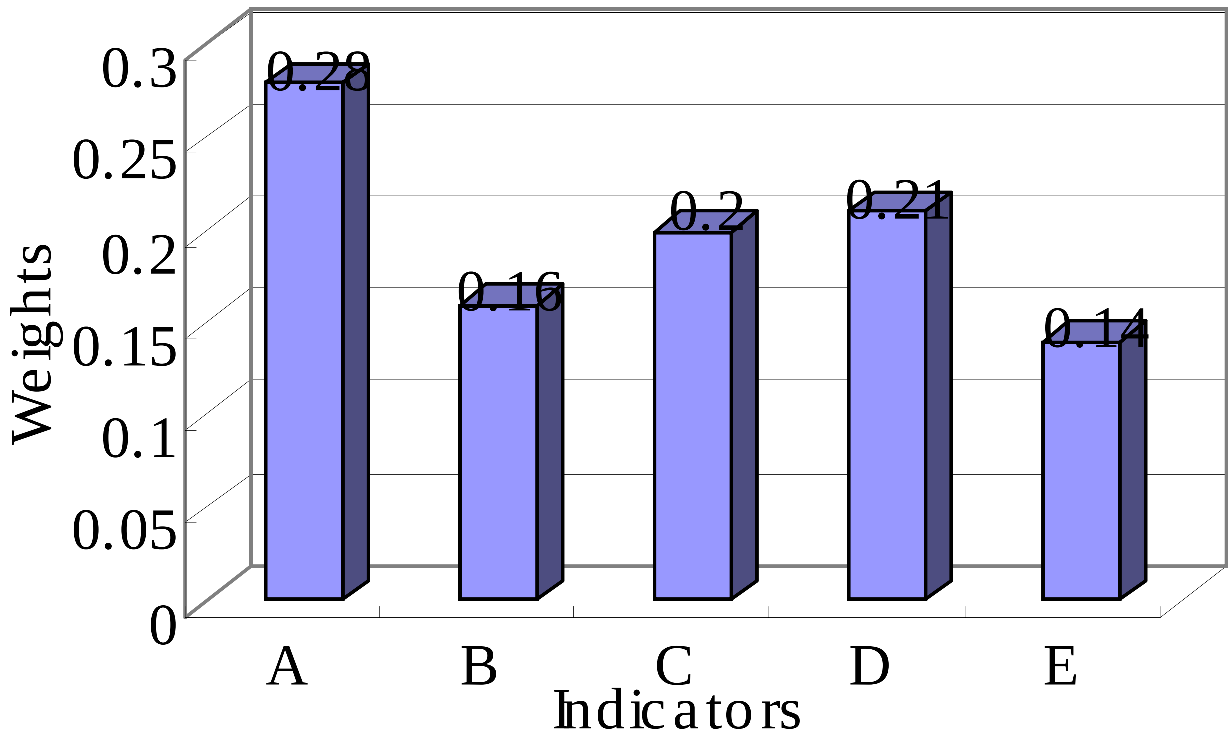
* When I compare A (methodology of data derivation) and the B (completeness in pollution source composition), I think (A B) is **times** more important compared to the other in relation to the quality of normalisation data.

1. Methodology of data derivation

Levels (*the data derived from*;)	Score
Using on-site measurement	100
Using emission factor based on the on-site measurement	______
Using calculation based on scientific assumption	______
Using calculation based on both scientific assumption and estimation	______
Using calculation based on estimation	______

Survey results

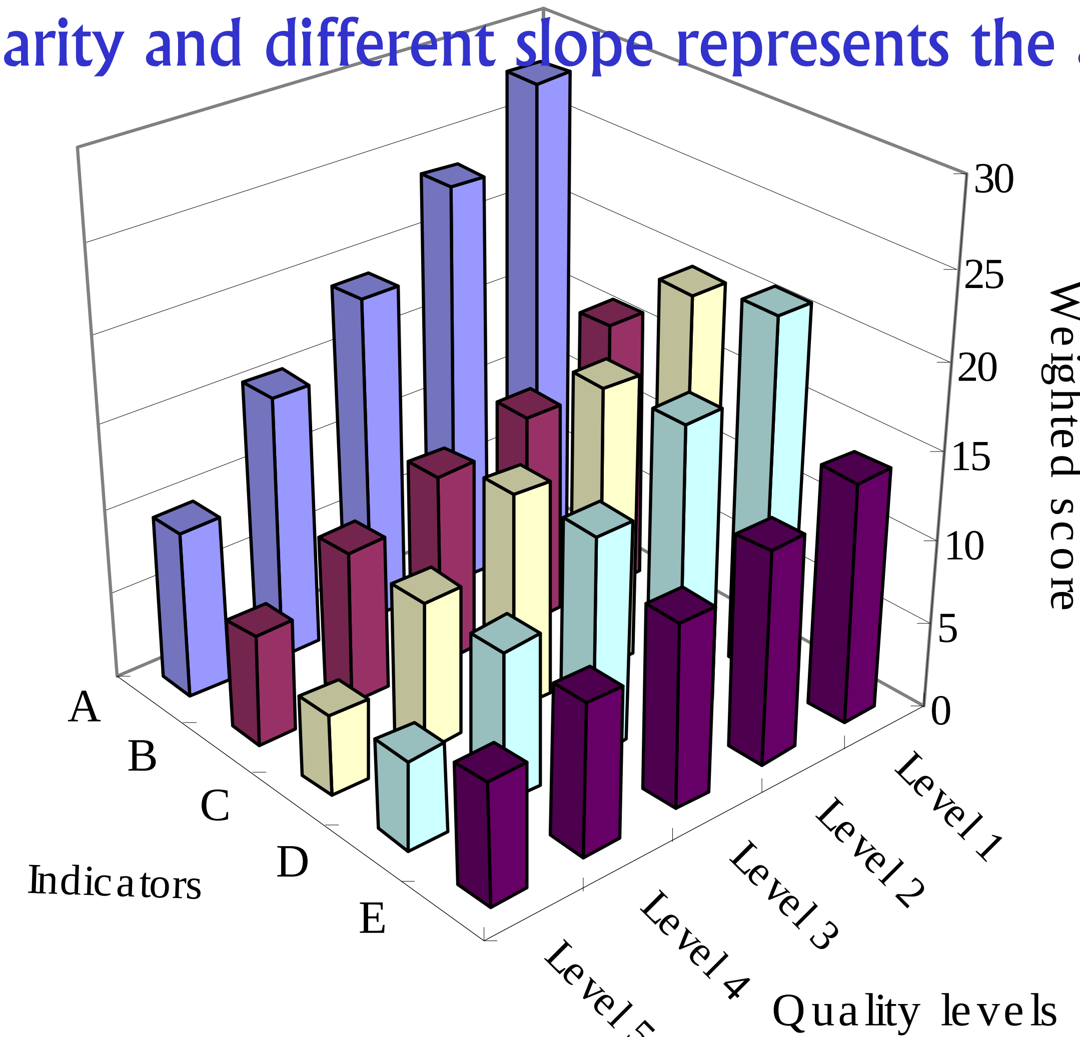
-Two replies under the Consistency Rate(CR) of 0.1 were excluded from further data processing



A: Data acquisition method, B: Independence of the data supplier, C: Completeness of the pollutant composition, D: Completeness of the pollution source composition, E Data age.

Weighted data quality score

(non-linearity and different slope represents the arbitrary scale)



Assessment of the priority for further revision of current NR

Procedure

1. Calculate data quality score for each indicator(ex. Data age)

E(Data Age)

	Description	Grade
EL1	0 year devi.	100
EL2	1 years	88.81
EL3	2 years	76.7
EL4	3 years	62.6
EL5	4 years	50.24

0 year devi. 20% of NR

1 years devi. 70% of NR

2 years devi. 10% of NR

g_{ij} f_{ij}

$100 \times 0.2 = 20$

$88.8 \times 0.7 = 62.16$

$76.7 \times 0.1 = 7.67$

$$\sum_i g_{ij} \times f_{ij} = 89.83$$

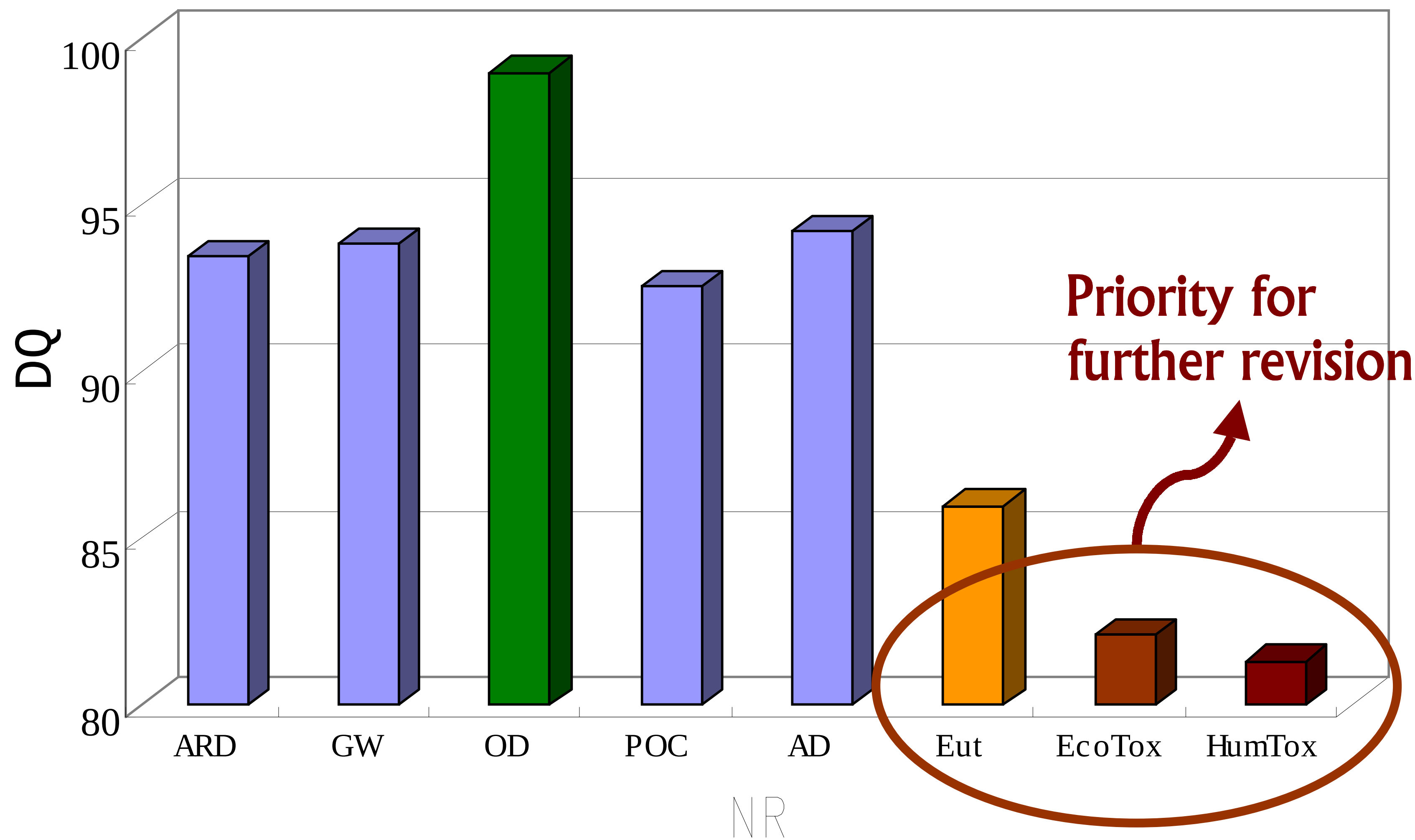
2. Calculate weighted data quality score for a NR(ex. NR for Ecotoxicit)

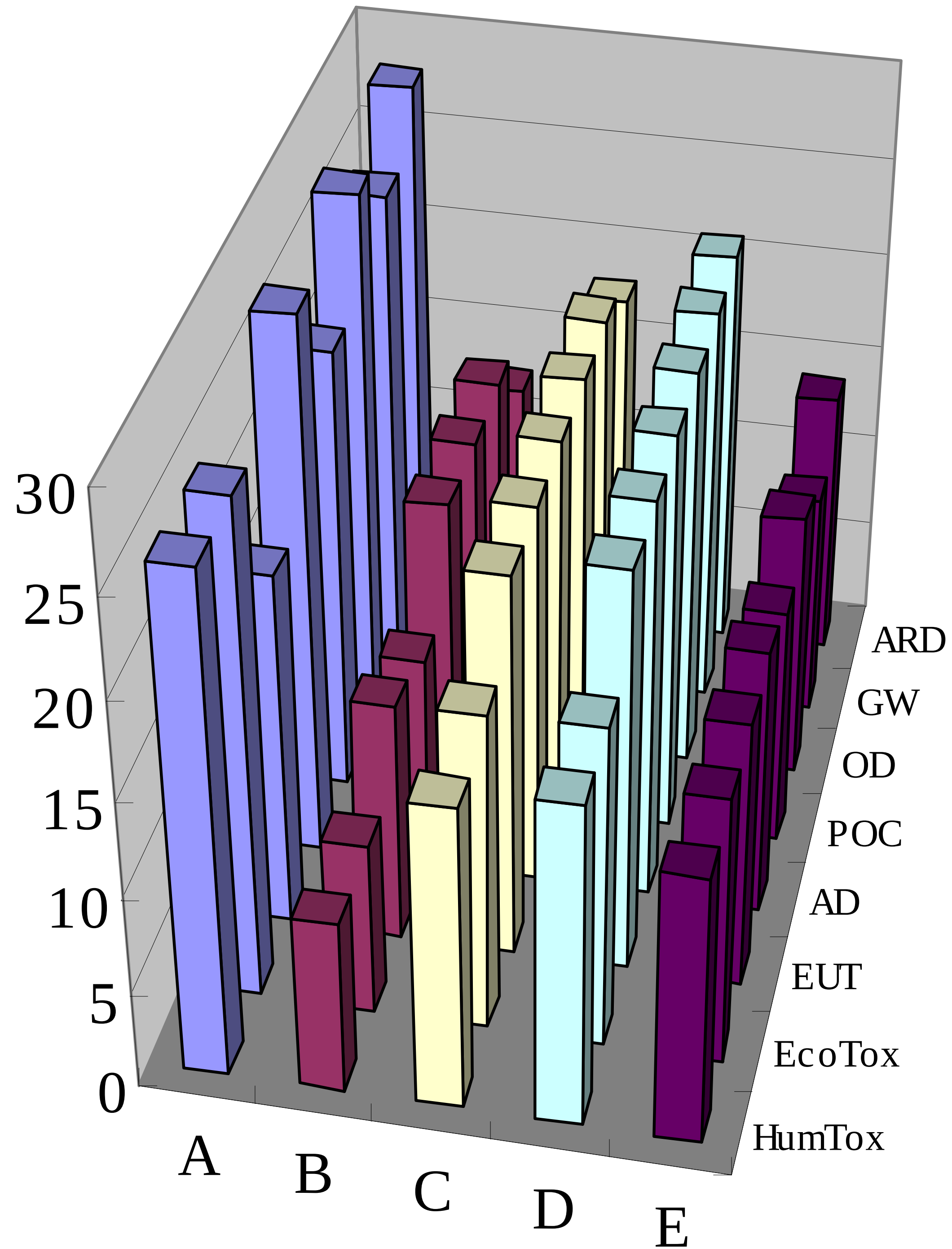
	DQ_j		w_j		$DQ_j \times w_j$
A: Data acquisition method	62.7	×	0.28	=	17.6
B: Independence of the data supplier	75.8	×	0.16	=	12.1
C: Completeness of the pollutant composition	40.5	×	0.20	=	8.1
D: Completeness of the pollution source composition	70.6	×	0.21	=	14.8
E Data age	89.83	×	0.14	=	12.6

$$DQ_{ET} = \sum_j DQ_j \times w_j$$

$$= 65.19$$

Application-Priority assessment [3]





30
25
20
15
10
5
0
A
B
C
D
E
ARD
GW
OD
POC
AD
EUT
EcoTox
HumTox

Summary & Conclusions [1]

- Semi-quantitative DQA method was proposed for the purpose of internal data evaluation.

- Survey on NR data quality shows that LCA practitioners give more importance on data acquisition method and less on data age.

- By applying grades and weights to each data quality level and DQI, respectively, the problem of 'arbitrary scale' can be alleviated.

-Resulting single DQ score is expected to be used as a tool to assist data evaluation e.g., selection between individual data sources, NRs for further revision or NRs sets as a whole.

-Proposed DQ calculation process was illustrated and applied for the Korean NRs to determine revision priority.

-Proposed DQA method was able to generate condensed information for any level of aggregation while keeping up meta data.

Limitations

-Arbitrariness in determining DQIs and scalin g DQ levels is still exist.

-Further collaboration to refine DQIs and DQ levels is needed.

Contact

Sangwon Suh

suh@cml.leidenuniv.nl

CML, Leiden University
the Netherlands