Assessment of the calibration accuracy of IOMS for odour concentration estimation using model comparison methods

Santiago Marco

(work developed within CEN TC264/WG41)

Signal and Information Processing for Sensor Systems
Institute for Bioengineering of Catalonia (IBEC)
smarco@ibecbarcelona.eu

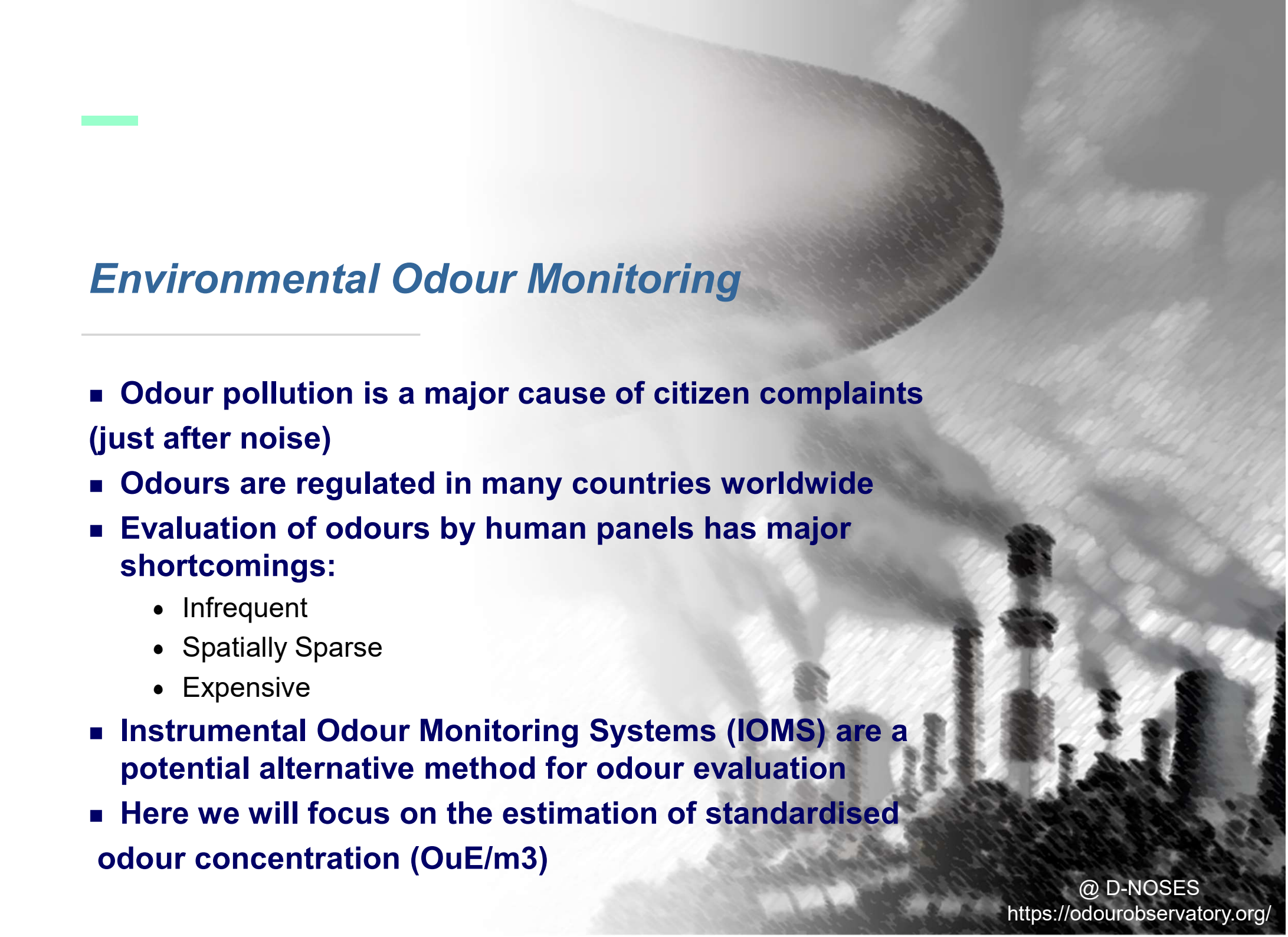
Department of Electronics and Biomedical Engineering
Universitat de Barcelona (UB)

Outline

- **Introduction & Motivation**
  - Dynamic Olfactometry vs IOMS
- **Method comparison**
  - Method comparison vs Regression
  - Pearson Coeficient
  - Regression methods:
    - Figures of merit
    - Errors in variables methods
  - Uncertainty Bands & Bland-Altman plots
    - Bland Altman plot
    - Chebyshev theorem
      - Sample size calculations
  - Dealing with Replicates / Bag Dilutions
- **Application examples**
- **Simulations studies**
- **Summary and conclusions**

Environmental Odour Monitoring

- **Odour pollution is a major cause of citizen complaints (just after noise)**
- **Odours are regulated in many countries worldwide**
- **Evaluation of odours by human panels has major shortcomings:**
  - Infrequent
  - Spatially Sparse
  - Expensive
- **Instrumental Odour Monitoring Systems (IOMS) are a potential alternative method for odour evaluation**
- **Here we will focus on the estimation of standardised odour concentration (OuE/m3)**



@ D-NOSES
https://odourobservatory.org/

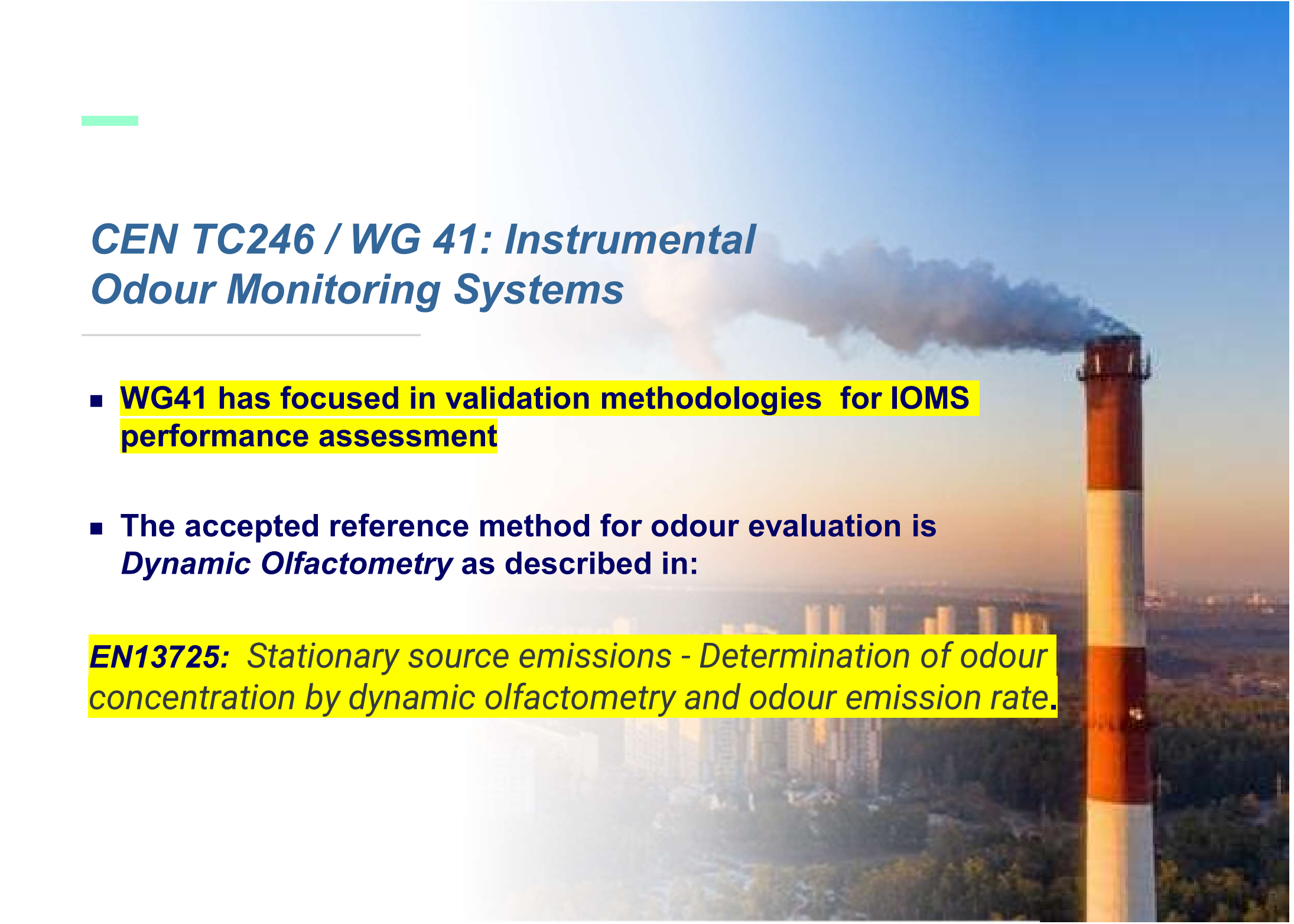
Standarization Needs & Difficulties

- **Odour is a human perception**

- **We need standard practices to verify the quality of the monitoring process: the instrument's performance for the task.**

- **Environmental odours are very complex mixtures with thousands of components**

- **Real environmental odours are not suitable reference materials for test**



CEN TC246 / WG 41: Instrumental Odour Monitoring Systems

- **WG41 has focused in validation methodologies for IOMS performance assessment**

- **The accepted reference method for odour evaluation is *Dynamic Olfactometry* as described in:**

EN13725: *Stationary source emissions - Determination of odour concentration by dynamic olfactometry and odour emission rate*.

Definition of equivalence

Would (someday) IOMS be considered as
an alternative method for Dynamic Olfactometry?

- **According to 'Terms of Reference for CEN/TC 264 Ambient-Air Standards':**

"An equivalent method to the reference method for the measurement of a specified air pollutant is a method meeting the data quality objectives for fixed measurements specified in the relevant air quality directive"

Odour concentration estimation

- **Dynamic Olfactometry for odour estimation features large uncertainties:**
  - EN13725: Intermediate precisión CI 95% <3 (stdev = 0.176)
  - EN13725: Accuracy (Bias) CI 95% < 1.64 (log10(Bias)<0.217)
  - A Factor 2 is often quoted as typical DO uncertainty at 95%

- **QUESTIONS:**
  - Can we safely replace the DO odour concentration estimation by IOMS?
  - What is the Acceptance Limit for the Differences between DO and Machine Olfaction readings?
  - How do we compare Machine Olfaction and Dynamic Olfactometry when the later features large uncertainties?

- **MODEL COMPARISON METHODS:**
  - Examples:
    - Regression Methods
    - Difference statistics(Chebyshev or Gaussian)

Introduction to Model comparison

- **The assessment of agreement between two methods of measurement is needed in diverse scientific and technical domains.**

- **Difference between Comparison (Agreement) and Calibration**

  - **Calibration:** Compare a new method **with a known accurate and precise method whose error is neglected.** The goal is to establish a mathematical relationship between their measurements so that the new method is an approximation of the 'true' measurement.

  - **Comparison**: a new method is evaluated by comparison with an established standard ('reference method'). **Both methods are not accurate or precise. If both methods sufficiently agree the alternative can replace the reference**.

- **Here we focus on 'Model Comparison': Agreement between Human Panels and Instrumental Solutions.**
- **We assume that IOMS has been previously calibrated and we just want to validate instrumental performance by comparing pairs of readings.**

Introduction to Model Comparison

- **The most common method to compare instrument readings is the correlation coefficient, which is considered insufficient.**

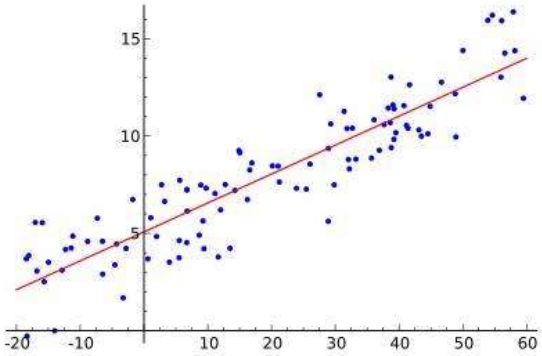
- **Innapropiate use of the correlation coefficient R (Bland-Altman, 1983)**

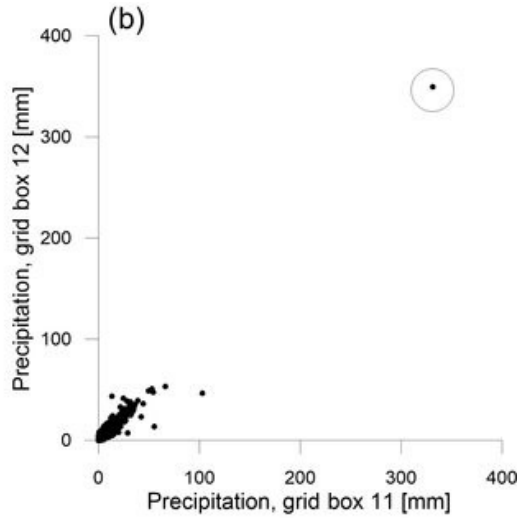
  - R measures the strength of a relation between two variables, not the agreement between them. We may have a perfect correlation in the presence of offsets and gain errors.

  - Correlation depends on the range of the true quantitity. The higher the range of the true quantity the greater the correlation.

  - R has no physical units, and hinders the interpretation of the expected Reading differences.

  - R can be sensitive to outliers.





Model Comparison based on Regression

Regression Methods in Model Comparison

- **Regression methods are often used in Model Comparison**

$$x = C_0 + C_1 z$$

$C_0=0$, $C_1=1$

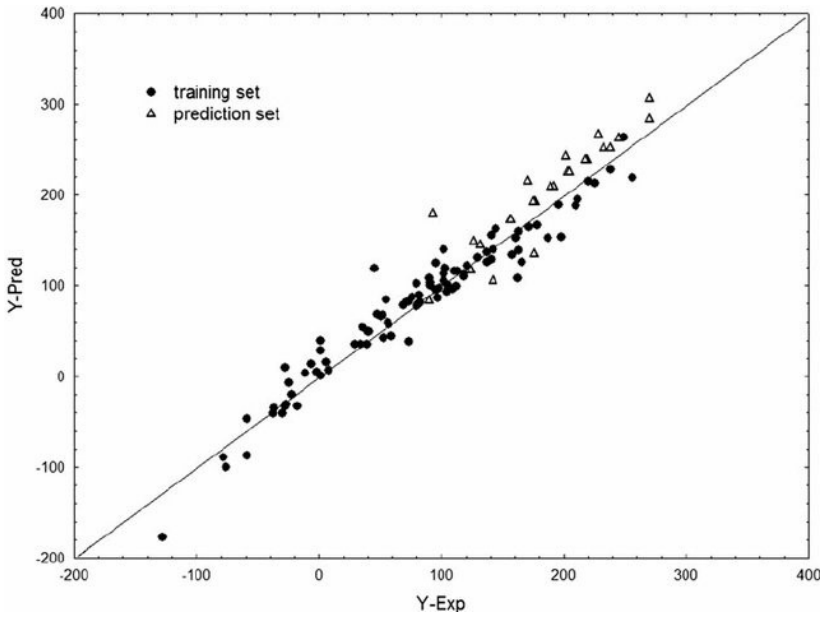
Alternative Method (AM)

Reference Method (RM)

EN14793: Stationary source emissions: Demonstration of the equivalence of an alternative method with a reference method

Table 4 — Compliance with the criteria

Verification tests	Value obtained		Criterion	Conclusion (result acceptable)
Systematic deviation				
correlation coefficient	r	value	$r \geq 0,97$	yes/no
slope	C_1	value	$1-\frac{s_R(\bar{\bar{z}})}{\bar{\bar{z}}} \leq C_1 \leq 1+\frac{s_R(\bar{\bar{z}})}{\bar{\bar{z}}}$	yes/no
intercept	C_0	value	$\lvert C_0 \rvert \leq s_R(\bar{\bar{z}})$	yes/no
Repeatability standard deviation				
	$s_r(\bar{\bar{z}})$ $s_r(\bar{\bar{x}})$	value value	$s_r(\bar{\bar{z}}) \leq s_{r,\text{limit}}(\bar{\bar{z}})$ $s_r(\bar{\bar{x}}) \leq s_{r,\text{limit}}(\bar{\bar{z}})$	yes/no



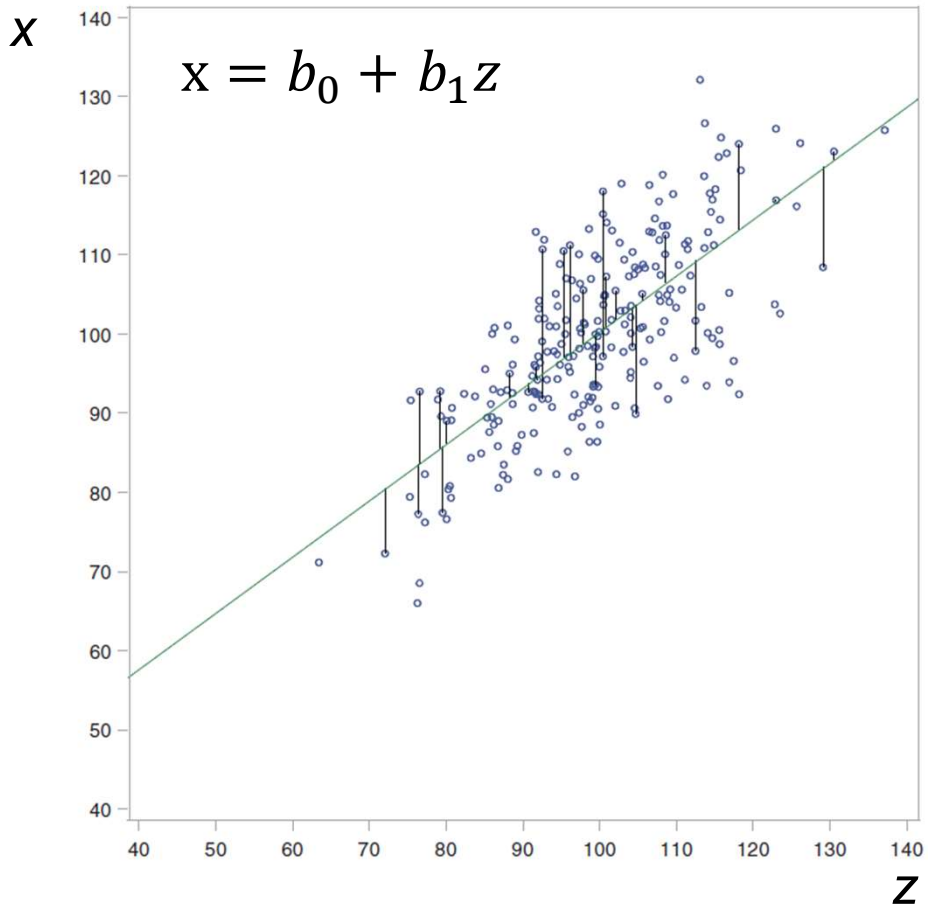
Ordinary Least Squares

- **Ordinary Least Squares assumes all the errors belong to the independent variable (alternative method) !!!.**
- **When both methods have uncertainties OLS is not recommended**

$$OLS\ (Y|X):\ b_1 = \frac{s_{xy}}{s_x^2}$$

$$b_0 = \bar{y} - b_1\bar{x}$$

where s_x^2 = sample variance of x
s_y^2 = sample variance of y
s_{xy} = sample covariance of x and y



Geometrical Mean Regression

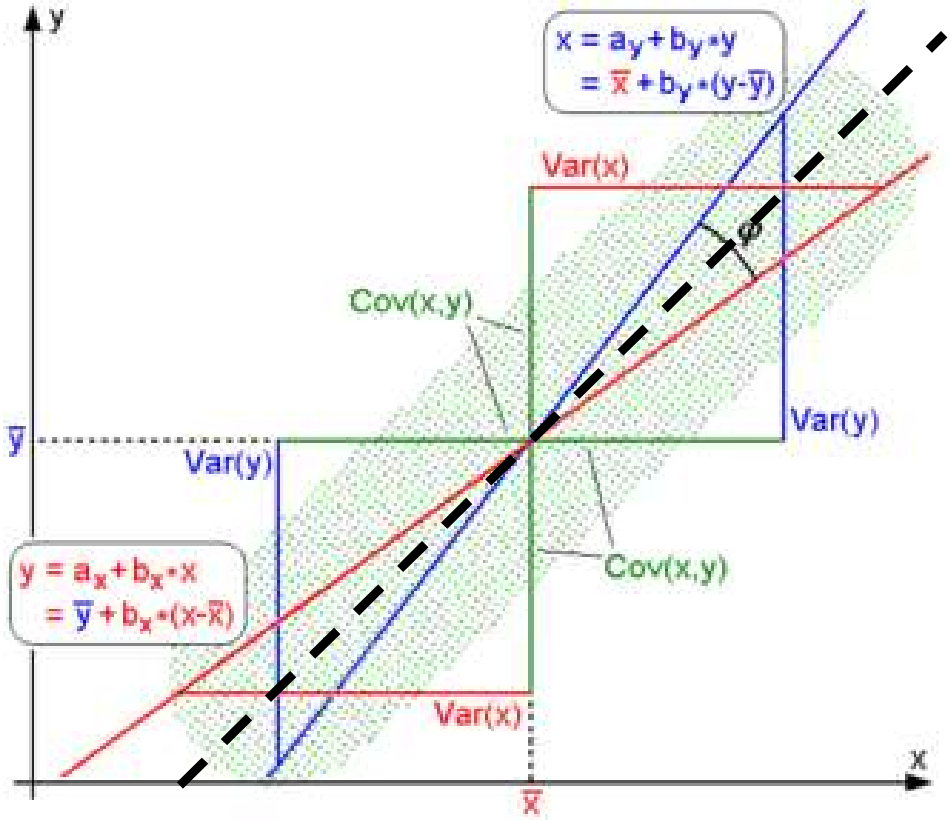
$$y = b_0 + b_1 x \qquad b_1 = \frac{s_{xy}}{s_x^2}$$

$$x = b_0 + b_1 y \qquad \widetilde{b_1} = \frac{s_{xy}}{s_y^2}$$

Geometrical Mean Regression

$$\widetilde{b_1} = \sqrt{\frac{b_1}{\widetilde{b_1}}} = \frac{s_y}{s_x}$$

Geometrical Mean Regression assumes errors in variables are proportional to their sample variances



Orthogonal Least Squares

- **Orthogonal Least Squares is the regression method recommended by EN14793**

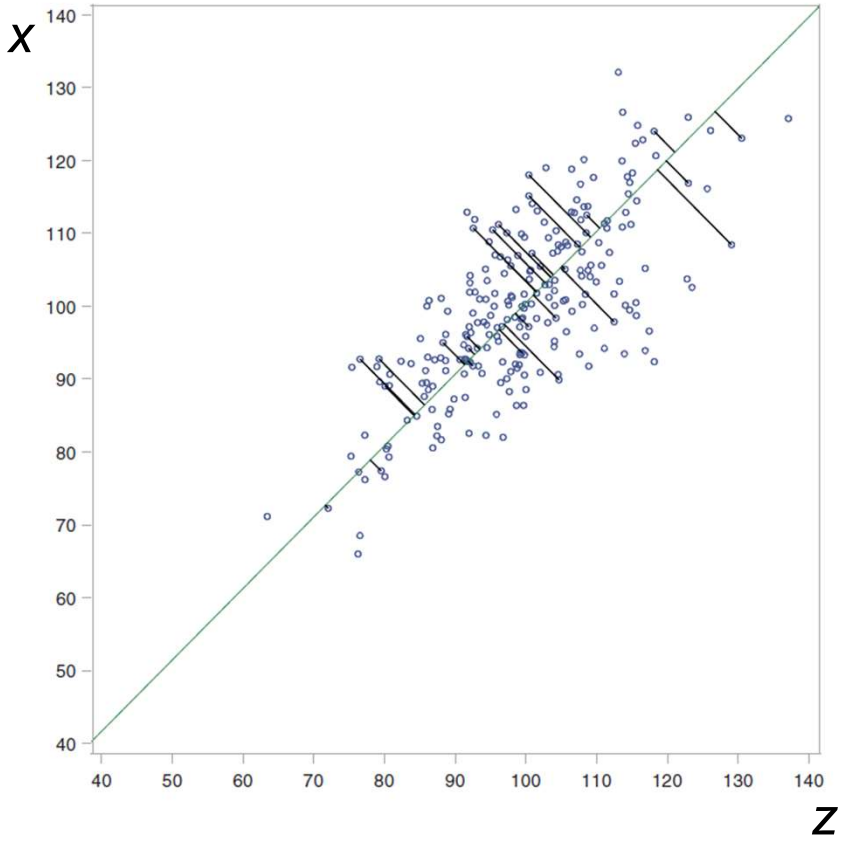
$$b_1 = \frac{s_y^2 - s_x^2 + \sqrt{\left(s_y^2 - s_x^2\right)^2 + 4s_{xy}^2}}{2s_{xy}}$$

$$b_0 = \bar{y} - b_1\bar{x}$$

where s_x^2 = sample variance of x
s_y^2 = sample variance of y
s_{xy} = sample covariance of x and y

It assumes uncertainties in both axis are equal !!!

$$x = b_0 + b_1 z$$



Deming regression

- **Considers a linear model with errors in both variables.**

$$y_i = \hat{y}_i + \varepsilon_i \qquad x = \hat{x}_i + \delta_i$$

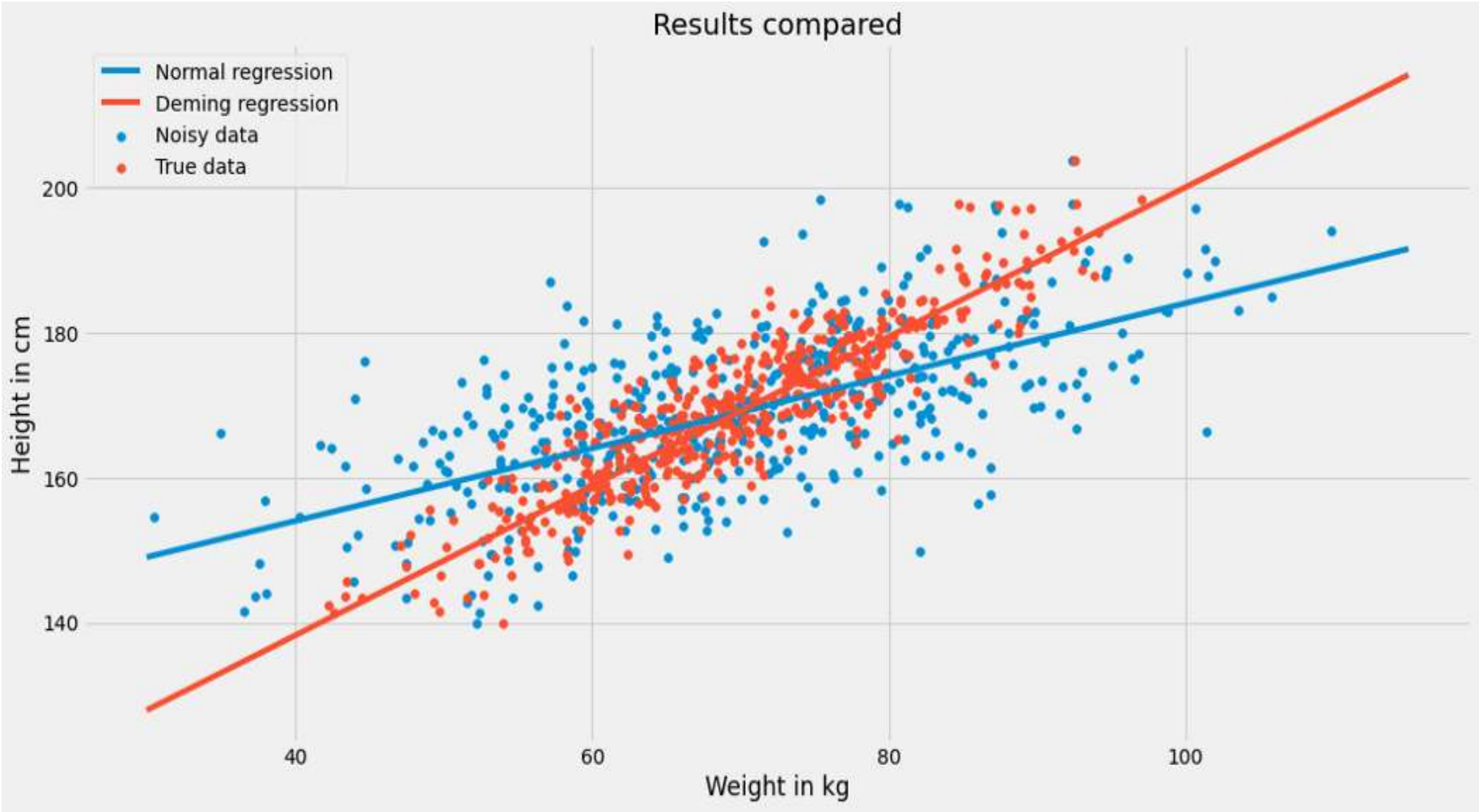
- **If both variables are normally distributed with constant variances and the ratio $\lambda = \sigma_\varepsilon^2/\sigma_\delta^2$ is known, then:**

$$DR:\ b_1 = \frac{s_y^2 - \lambda s_x^2 + \sqrt{\left(s_y^2 - \lambda s_x^2\right)^2 + 4\lambda s_{xy}^2}}{2s_{xy}}$$

- **Orthogonal Least Squares and GM can be considered as special cases of Deming Regression for different hypothesis on lambda**

Deming Regression example

- **Example when errors in X are double errors in Y**

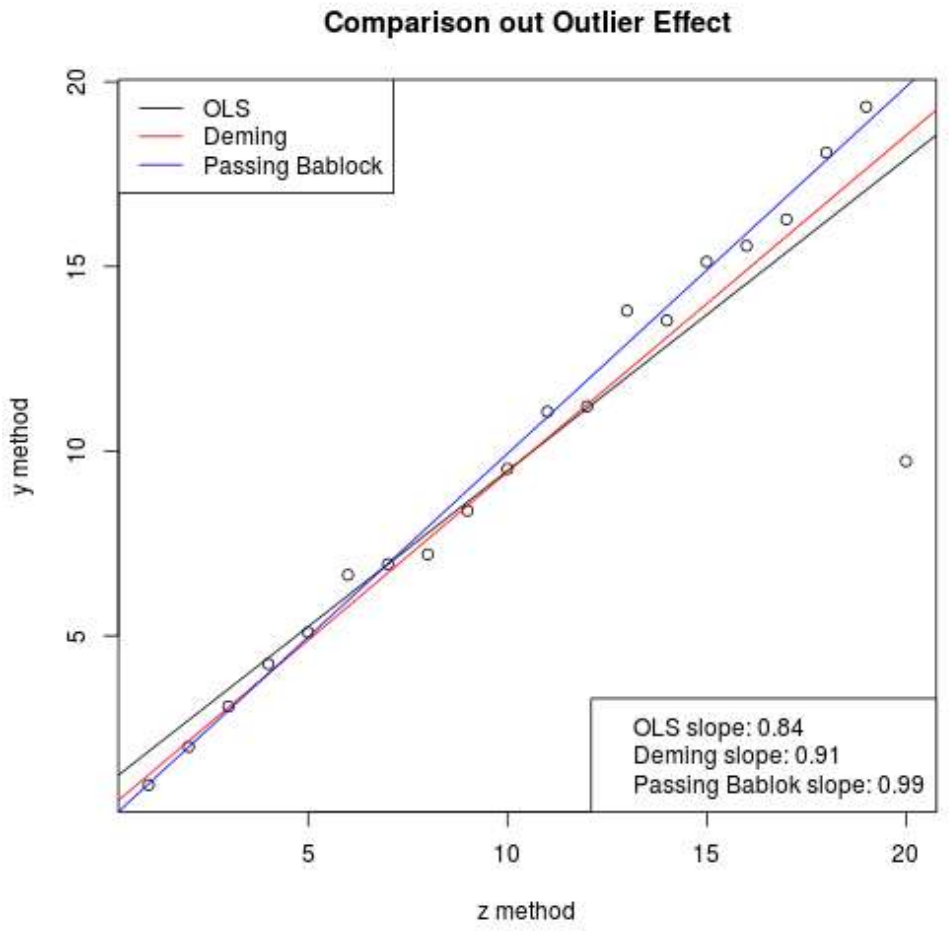


https://towardsdatascience.com/error-in-variables-models-deming-regression-11ca93b6e138

Passing-Bablok Regression

- **Passing-Bablok (1983) regression does not minimize any residuals**
- **PB regression calculates the slopes between all data point pairs and the final slope is the median of the slopes.**

- PB is a robust method and it does not require errors in variables to be Gaussian. There are not strong underlying hypothesis for the application of the method, but on the other hand it has not a figure of merit.
- PB inherently assumes that both errors are equally distributed for all data pairs.
- Only recently the statistics of the coefficients have been understood.



https://www.r-bloggers.com/2015/09/deming-and-passing-bablok-regression-in-r/

Tools for EiV Regressions

- **In R:**

  - **Package ‘mcreg’: Method Comparison Regression in CRAN implements, Deming, Passing-Bablok (among others)**
  - It does not consider the possibility to have replicates.
  - J. A. Budd et al.(2018, <https://clsi.org/standards/products/method-evaluation/documents/ep09/>)

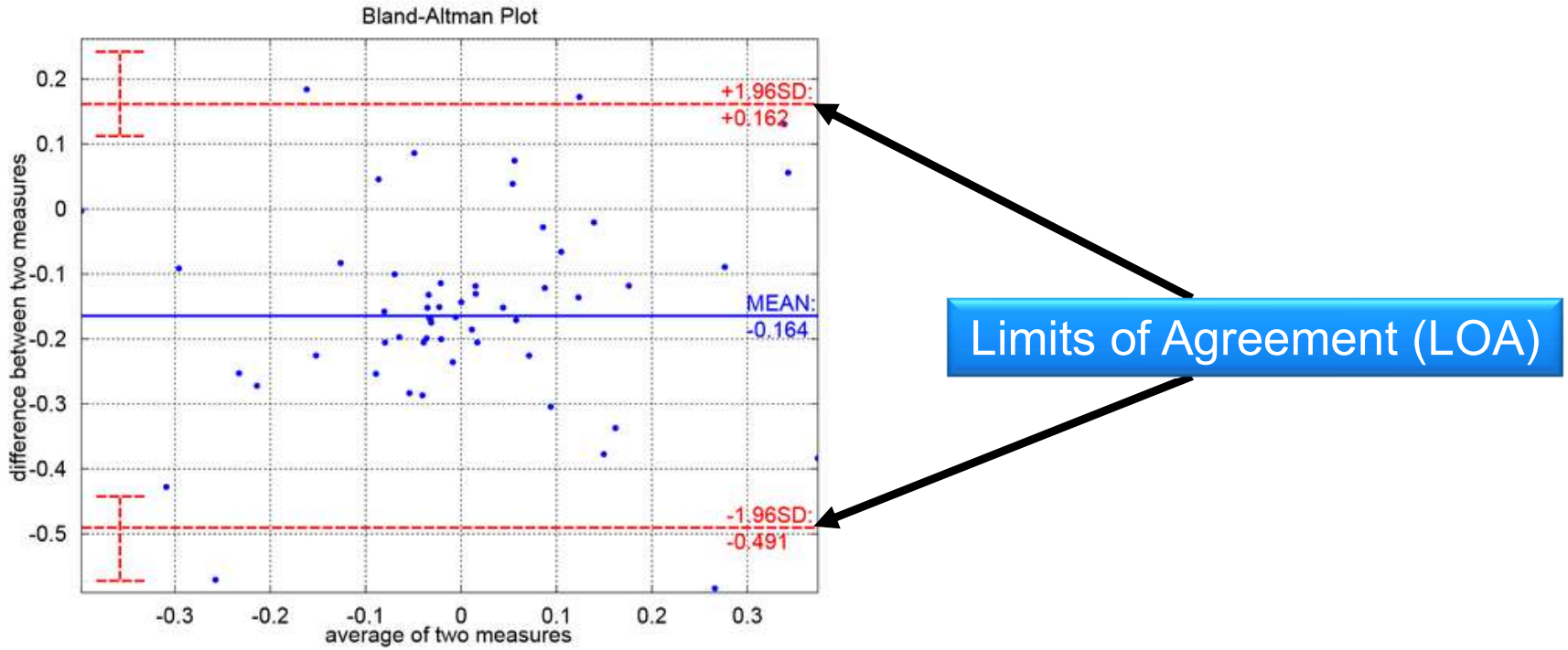
  - **Analysis of Agreement in Method Comparison Studies: Package ‘MethComp**’.
  - (See Carstensen B. (2010) ``Comparing Clinical Measurement Methods: A Practical Guide (Statistics in Practice)’)
  - It considers the possibility to have replicate measurements

- Deming WE. Statistical adjustment of data. New York: John Wiley & Sons, 1943, 1964:184.
- PASSING, H.; BABLOK, W. A new biometrical procedure for testing the equality of measurements from two different analytical methods. Application of linear regression procedures for method comparison studies in clinical chemistry, Part I. *Clinical Chemistry and Laboratory Medicine*, 1983, vol. 21, no 11, p. 709-720.

Method comparison based on statistics of differences

Bland Altman Plot (Gaussian Differences)

- In 1983 Bland and Altman proposed a way to compare two approximate methods that is independent of the regression method.
- The Bland Altman Plot represents the difference of readings with respect to the mean value of the both readings.
- Then statistics on the differences can be carried out, including means and confidence limits.



From wikipedia

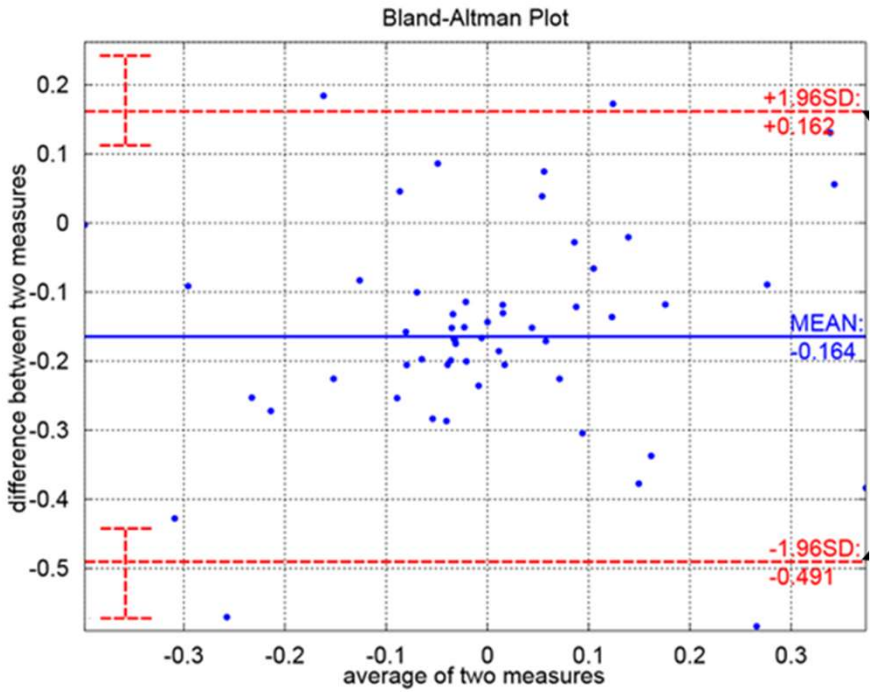
Bland Altman Plot

- **Criteria for sufficient agreement:**
  - 1) Limits of Agreement should include the '0', (Perfect Agreement)
  - 2) Sufficient agreement is declared if the differences between the readings are not practically important as determined by the application
  - 3) The limit to accept the differences should be setup a priori by the 'user': **Δ.**

- **If the differences are Gaussian the limits are easily caculated. For instance for a 5% risk.**

$$\left(\bar{d} - 1.96S_d, \bar{d} + 1.96S_d\right)$$

- **To decide if the differences are sufficiently normal the Saphiro-Wilk test is recommended**

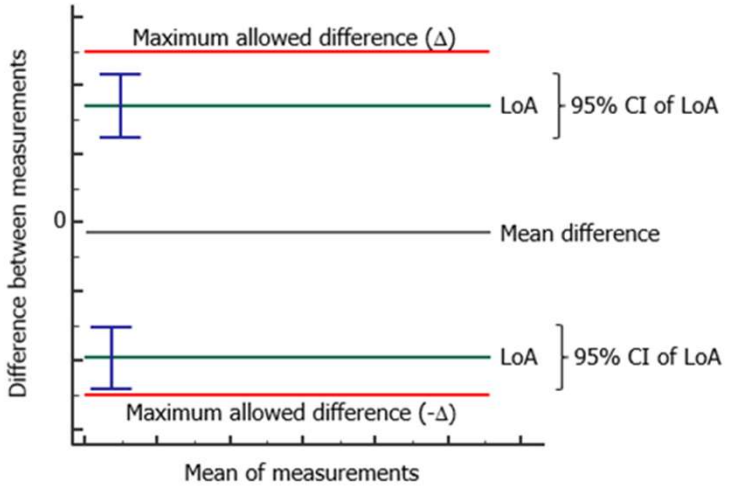


Variance of estimators

- **Variance:**

$$\widehat{Var}(\bar{d}) = \frac{S_d^2}{n}$$

$$\widehat{Var}(\bar{d} \pm 1.96 S_d) = \left(\frac{1}{n} + \frac{1.96^2}{2(n-1)}\right) S_d^2$$



- **Confidence Interval for Bias:**

$$\bar{d} \pm t_{1-\alpha/2,n-1}\sqrt{\widehat{Var}(\bar{d})}$$

$$(\bar{d} - 1.96\, S_d) - t_{1-\alpha/2,n-1}\sqrt{\widehat{Var}(\bar{d} \pm 1.96\, S_d)}$$

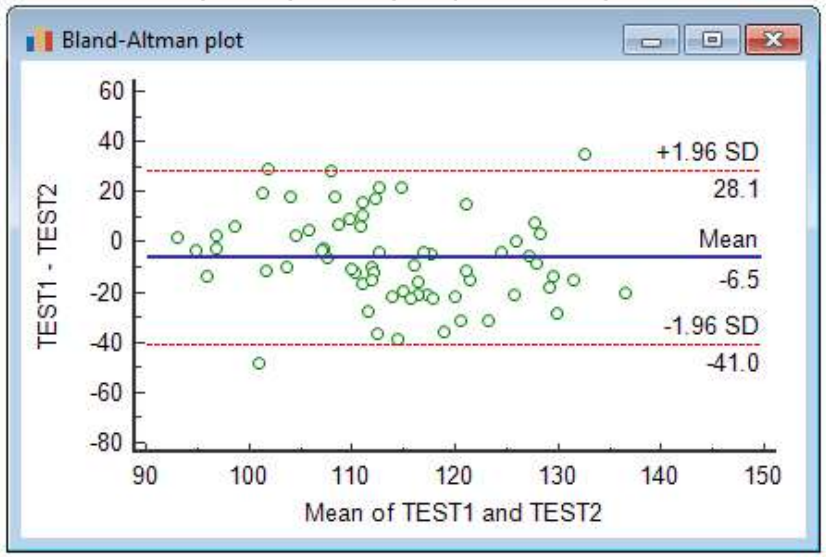
- **Confidence Interval for LoA:**

$$(\bar{d} + 1.96\, S_d) + t_{1-\alpha/2,n-1}\sqrt{\widehat{Var}(\bar{d} \pm 1.96\, S_d)}$$

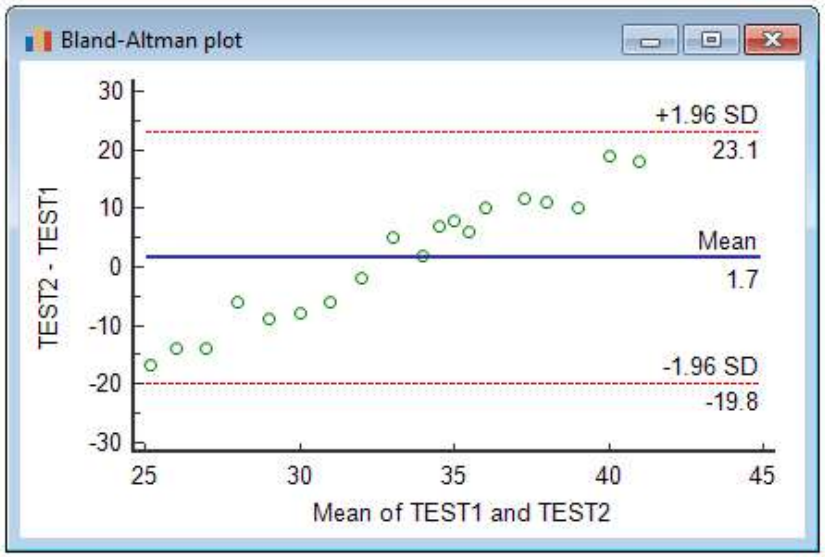
These expressions asume iid samples (no replicates)

Diagnostics from the Bland Altman Plot

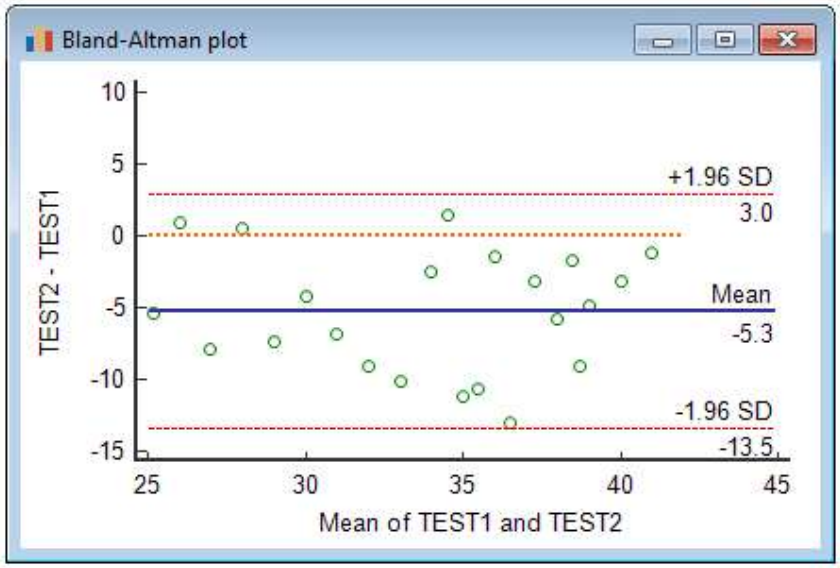
Normal Bland-Altman



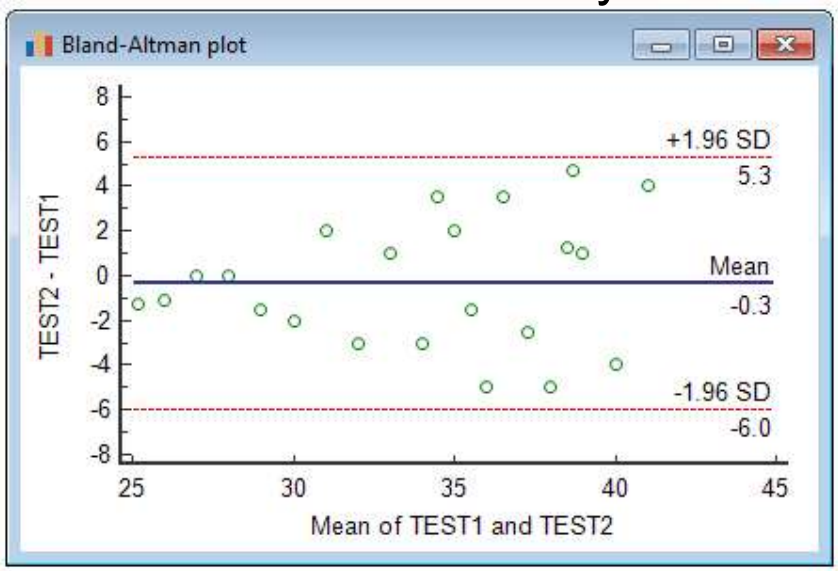
Proportional Error



Systematic Deviation



Heterocedasticity



Models using replicate measurements

- **Linear mixed effects model with:**
  - Bias, systematic factor (odor intensity)
  - Error terms:
    - i: Bag
    - m: Method
    - r: replicates

$$y_{imr} = \alpha_m + \mu_i + a_{ir} + c_{im} + e_{imr}$$

Readings

Bias

Odour intensity

Captures method independent Variability among replicates

Captures variability on ítems over methods

Residuals

$$a_{ir} \sim \mathcal{N}(0, \omega^2), \quad c_{mi} \sim \mathcal{N}(0, \tau_m^2), \quad e_{mir} \sim \mathcal{N}(0, \sigma_m^2)$$

These models can be estimated with the REML algorithm (Restricted Maximum Likelihood Algorithm)

CARSTENSEN, Bendix; SIMPSON, Julie; GURRIN, Lyle C. Statistical models for assessing agreement in method comparison studies with replicate measurements. The international journal of biostatistics, 2008, vol. 4, no 1

Harville, D. A. (1977). "Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems". *Journal of the American Statistical Association*. **72** (358): 320–338.

Chebyshev Inequality

- **Bland-Altman approach assumes Gaussian distribution of the errors**

- **The Chebyshev inequality puts distribution free bounds on the diffences to the mean,**

$$\Pr(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

- **The basic vesion of the Chebyshev inequality assumes perfect knowledge of Bias and Variance**

- **The confidence Interval at 5% risk is: X=bias ± 4.47 σ**
- **Chebyshev bands are way larger than equivalent bands assuming normal distribution.**

Chebyshev Inequality: Finite Sample size

- **The limits and the minimum number of samples to attain them have been proven by:**
- *SAW, John G.; YANG, Mark CK; MO, Tse Chin. Chebyshev inequality with estimated mean and variance. The American Statistician, 1984, vol. 38, no 2, p. 130-132.*

Max coverage	Min Number of samples	K (# Standard Deviations)
80%	4	2.5
85%	6	2.86
90%	9	3.33
95%	19	4.58
95%	Infinite samples	4.47

Formulas for minimum number of samples for coverage

$$\Pr(|X-\mu| \le k\sigma) \le \frac{N}{N+1} \qquad k = \frac{N+1}{\sqrt{N}}$$

Summary Limits of Agreement (LoA)

- Limits of Agreement can be calculated under the Bland Altman hypothesis (normal distribution) or the Chebyshev (distribution free).
- In both cases the coverage factor needs to be decided before hand and this controls the number of standard deviations.

- Basic Version LoA (neglecting variance of estimators):
  - Bland Altman: Bias ± K_{ba}*σ
  - Chebyshev: Bias ± K_c*σ

- Example: For 95% coverage (GUM) and large sample:
  - K_c=4.47
  - K_{ba}=1.96

- Chebyshev bands will be 2.3 times wider than B-A..

Odour Quantification by IOMS/DO

- As established in EN13725, DO uncertainty is better expressed as multiplicative errors: additive in the logarithmic domain
- The use of model comparison methods based on regression take the form

$$\log(IOMS_i) = b_o + b_1 \log(DO_i) = \log(Bias) + \log\left(DO_i{}^{b_1}\right)$$

$$IOMS = Bias \cdot DO^{b_1}$$

- Similarly when simply using the analysis of differences:

$$\log(IOMS_i) - \log(DO_i) = \log(Bias) + \log(\varepsilon_i) \qquad \frac{IOMS_i}{DO_i} = Bias \cdot \varepsilon_i$$

$$\log(Bias) = Mean(\log(IOMS_i) - \log(DO_i))$$

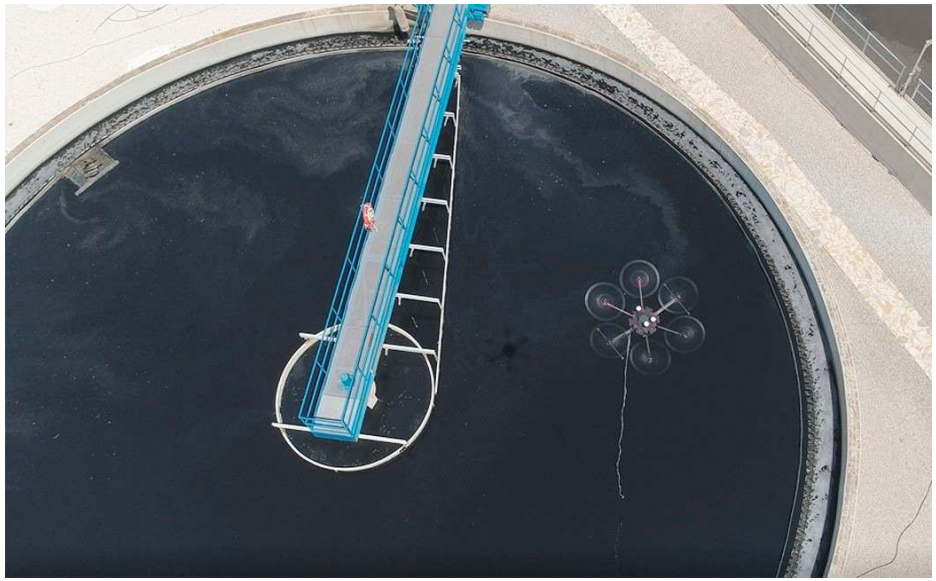
$$\log(sigma) = std(\log(IOMS_i) - \log(DO_i))$$

$$LoA = 10^{(Bias - kSigm \ , \ \ Bias + KSigma)}$$

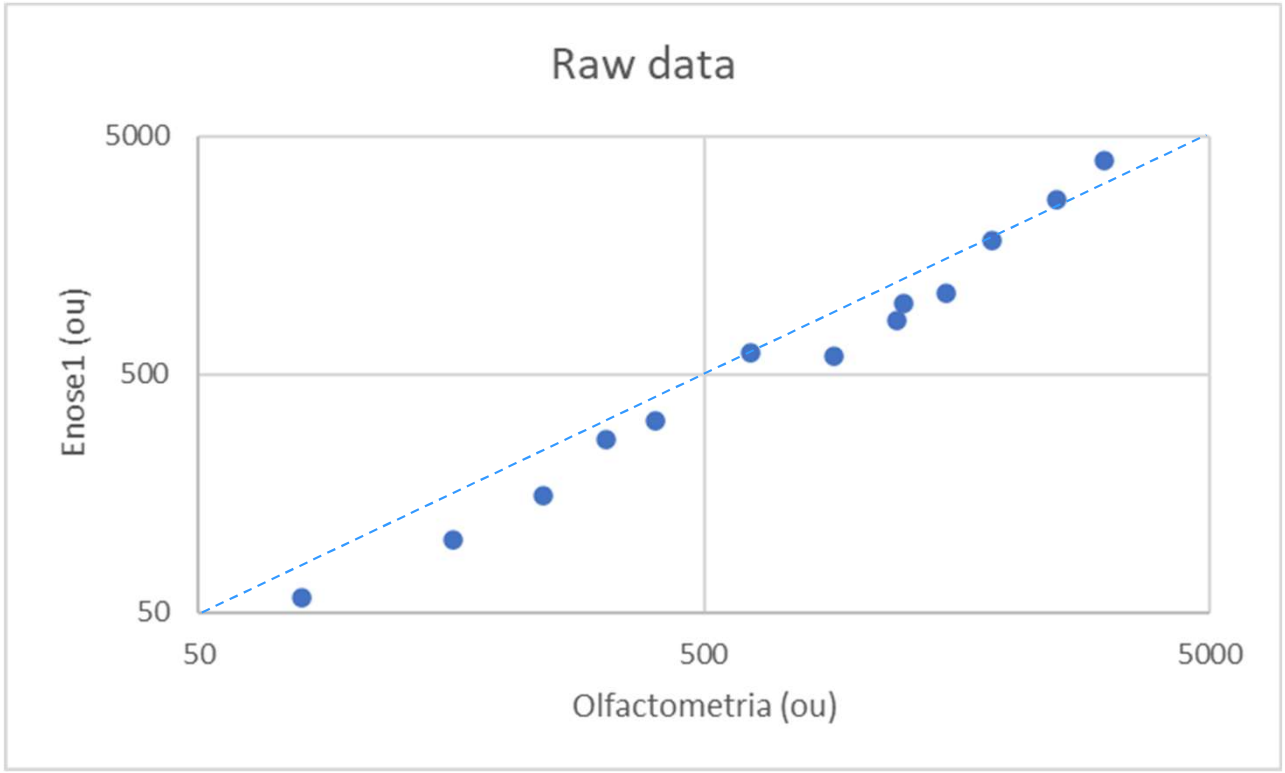
RESULTS:
BLIND COMPARISON DO/IOMS

Datasets

- Hospital Incinerator (Courtesy of Dra. Laura Capelli) N=13
- Landfill 1 (Courtesy of Dra. Laura Capelli) N=6, 10 dilutions
- WWTP (Courtesy of Dra. Anne Claude Romain) N=37
- WWTP (Courtesy of Dr. Santiago Marco) N=48
- Landfill 2 (Courtesy Dra. Laura Capelli) N=12



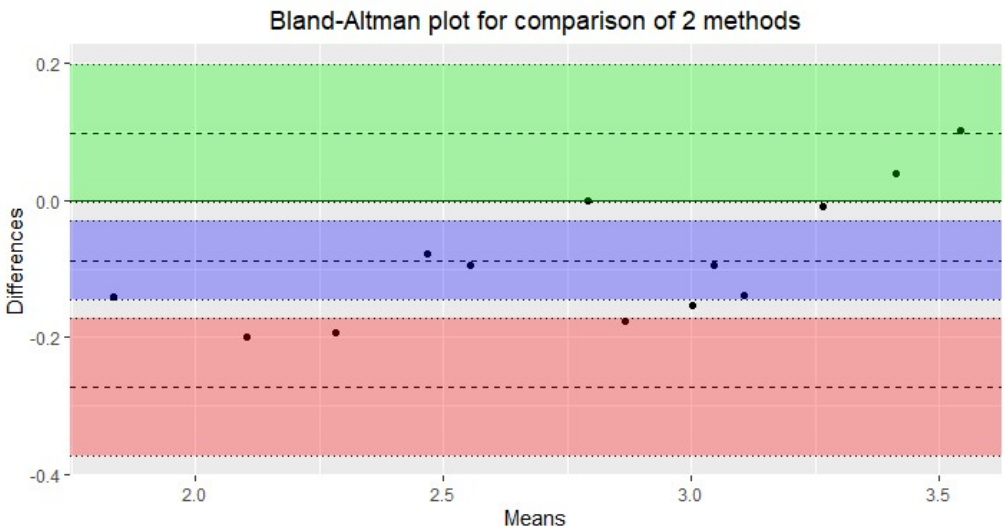
Hospital Incinerator data (N=13).



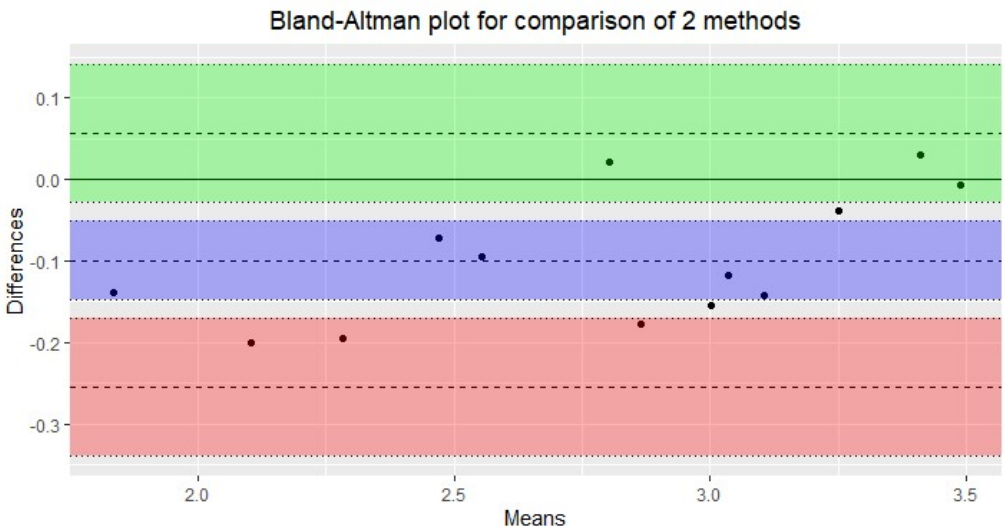
Bland-Altman plots on real enose data

Hospital Incinerator data (courtesy of Laura Capelli):
Olfactometry, enose1, enose2

Enose1



Enose2



Data is transformed to logarithmic scale (base 10).

Normality Test: Shapiro.Wilk test was done for both differences: H0 could not be rejected (Data is approximately normal)
Caution: with 13 samples the power of the test is very poor

Bland-Altman Bias and Limits of Acceptance

- **Bland-Altman (95% coverage, N=13, k=1.96)**
- **Chebyshev (93% coverage, N=13, k=3.88)**

Bland-Altman

Bland-Altman	Enose1 / Olfac	Enose2 /Olfac
Bias (Gain Error)	0.82 (0.72-0.93)	0.79 (0.71, 0.89)
Min (Gain Error) - BA	0.42	0.46
Max (Gain Error) -BA	1.58	1.38

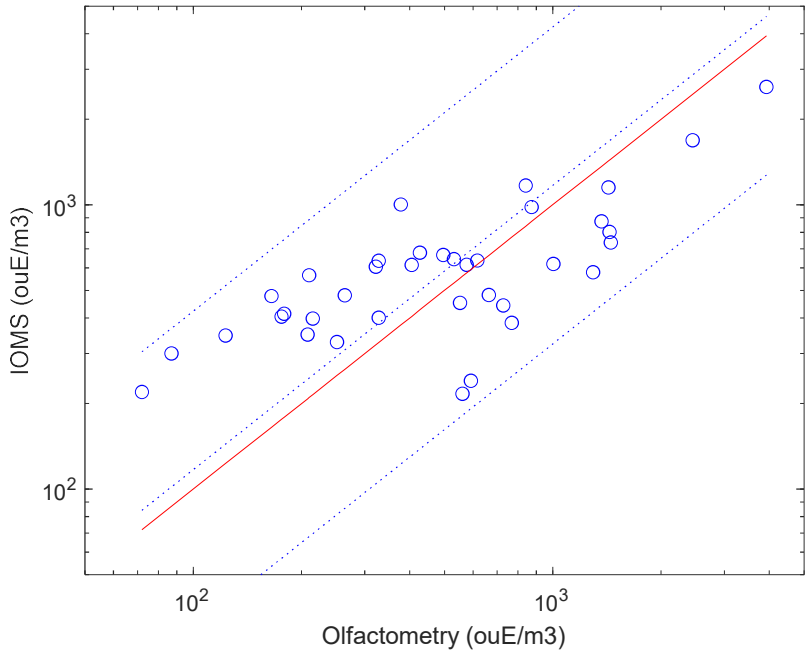
Sigma=1.39

Sigma=1.31

Chebyshev	Enose1 / Olfac	Enose2 /Olfac
Bias (Gain Error)	0.82 (0.72-0.93)	0.79 (0.71, 0.89)
Min (Gain Error) - Che	0.23	0.27
Max (Gain Error) -Che	2.93	2.33

Waste Water Treatment Plant

Courtesy Dra. A.C. Romain (N=37)



Bland Altman (95%) k=2

Upper LoA	4.21
Bias	1.17
Lower LoA	0.33

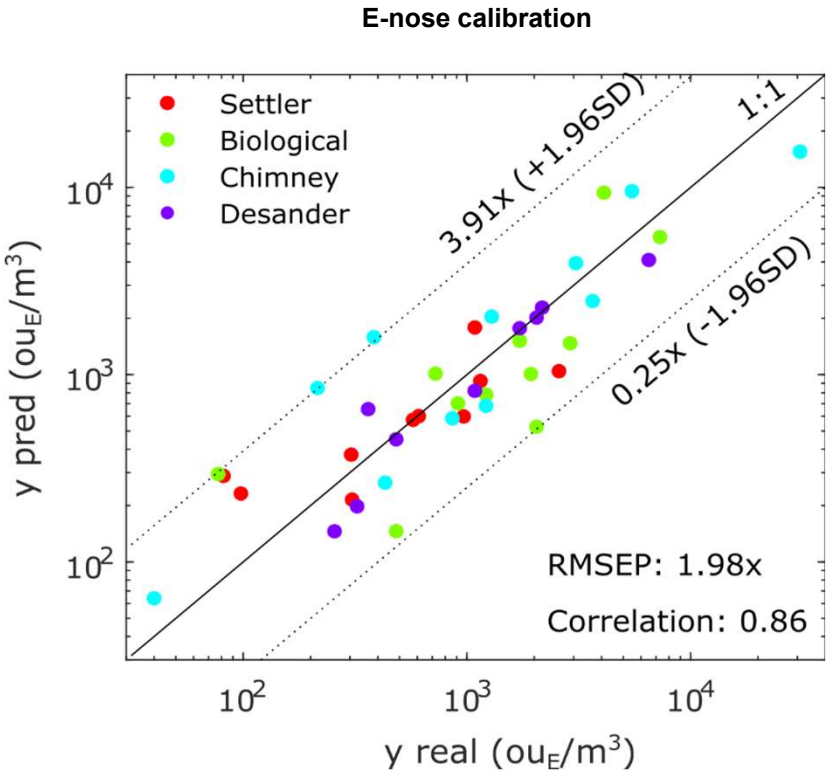
Chebyshev (95%) k=4.47

Upper LoA	20.89
Bias	1.17
Lower LoA	0.066

Sigma=1.90

Waste Water Treatment Plant

Courtesy of Dr. S. Marco (N=48 samples)



Bland Altman (95%) k=2

Upper LoA	3.91
Bias	0.99
Lower LoA	0.25

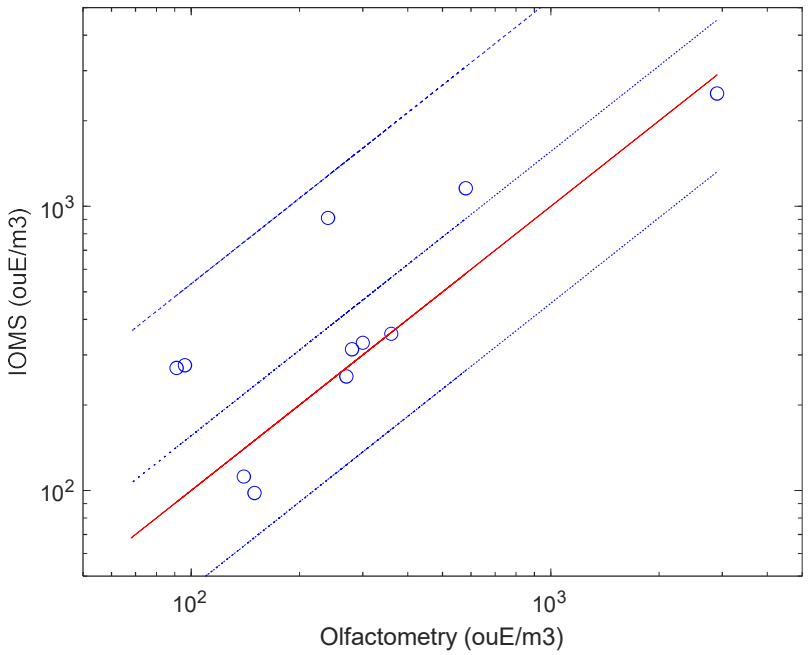
Chebyshev (95%) k=4.47

Upper LoA	20.98
Bias	0.99
Lower LoA	0.047

Sigma=1.98

Landfill 1

Courtesy (Dra. Laura Capelli) N=12



Bland Altman (95%) k=2

Upper LoA	5.34
Bias	1.56
Lower LoA	0.46

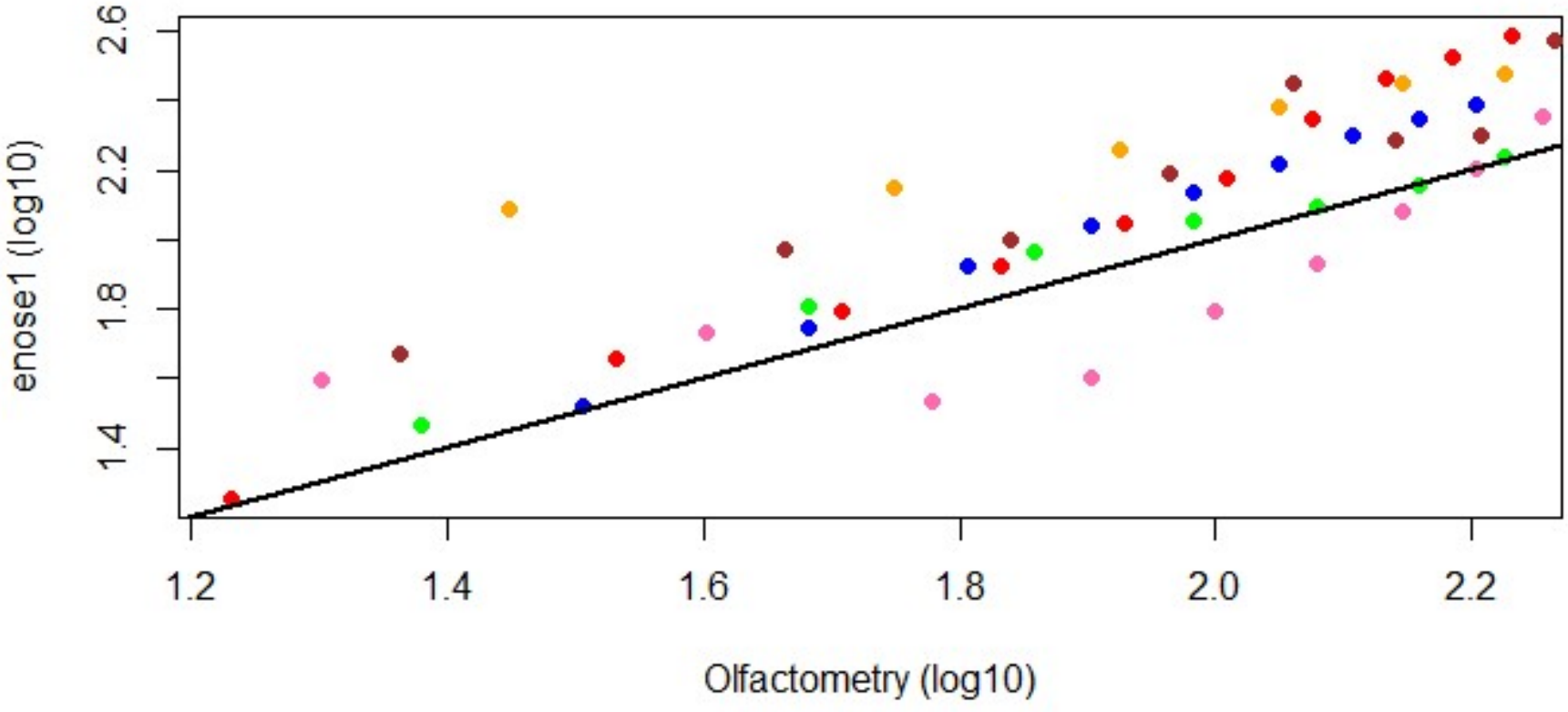
Chebyshev (92%) k=3.88

Upper LoA	16.97
Bias	1.56
Lower LoA	0.14

Sigma=1.85

Empirical data: landfill 2 (Dra. Capelli)

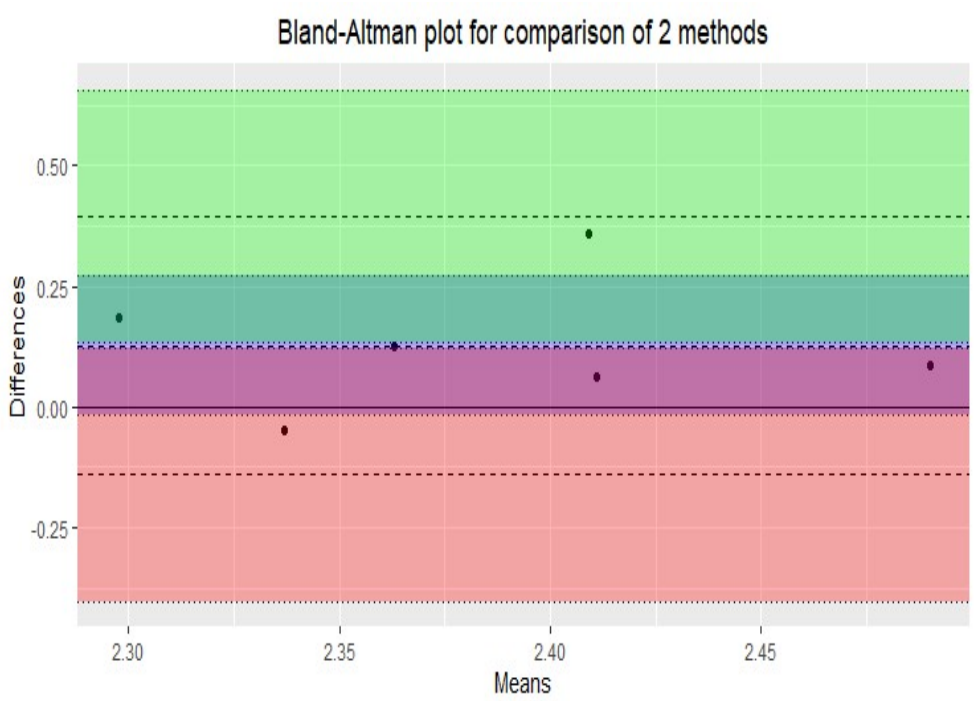
6 bags and 5 dilutions



Different colors represent different bags and their dilutions

Analysis using only 1 sample per bag

- **Analysis using only the most concentrated sample (N=6 bags)**



Bland Altman (95%)

Upper LoA	2.52
Bias	0.75
Lower LoA	0.22

Chebyshev (85%)

Upper LoA	4.26
Bias	0.75
Lower LoA	0.13

Sigma=1.83

Linear Mixed Effects Models for replicates

	Independent data (N=6) (log/Factor)		LME_Model (N=30) (log/Factor)	
Bias	-0.127	0.75	-0.136	0.73
LoA (upper limit) 5% risk	+0.402	2.52	+0.188	1.54
LoA (lower limit) 5 % risk	-0.657	0.22	-0.46	0.34
Range: max/min	0.3	2	1.47	30
Resolution	0.256	1.8	0.162	1.4
Dynamic Range (max/min) / resolution		1.1		21

Sigma=1.83

Sigma=1.45

- **Independent samples: N=6**
  - Concentration range explored : Factor 2
  - Resolution: Factor 1.8
- **Bags considered as 1 sample per Dynamic olfactometry and 10 dilutions per electronic nose**
  - Concentration range explored: Factor 30
  - Resolution: Factor 1.3
- **Conclusion: the use of dilutions decreases the statistical uncertainty but it also increases the Dynamic range considered in concentrations**

Summary of Results (Bland – Atlman)

Case	N	Bias	2*σ
Hospital Incinerator 1	13	0.82 (1.22)	1.93
Hospital Incinerator 2	13	0.79 (1.27)	1.71
WWTP (Liege)	37	1.17 (1.33)	3.61
Landfill POLIMI	6	0.75 (1.33)	3,35
Landfill POLIMI - Dilutions	30	0.73 (1.37)	2.1
Landfill 1 (POLIMI)	12	1.56	3.42
WWTP (IBEC)	48	0.99 (1.01)	3.93
Median	**13**	**0.82 (1.22)**	**3.34**

Limits EN13725: Bias

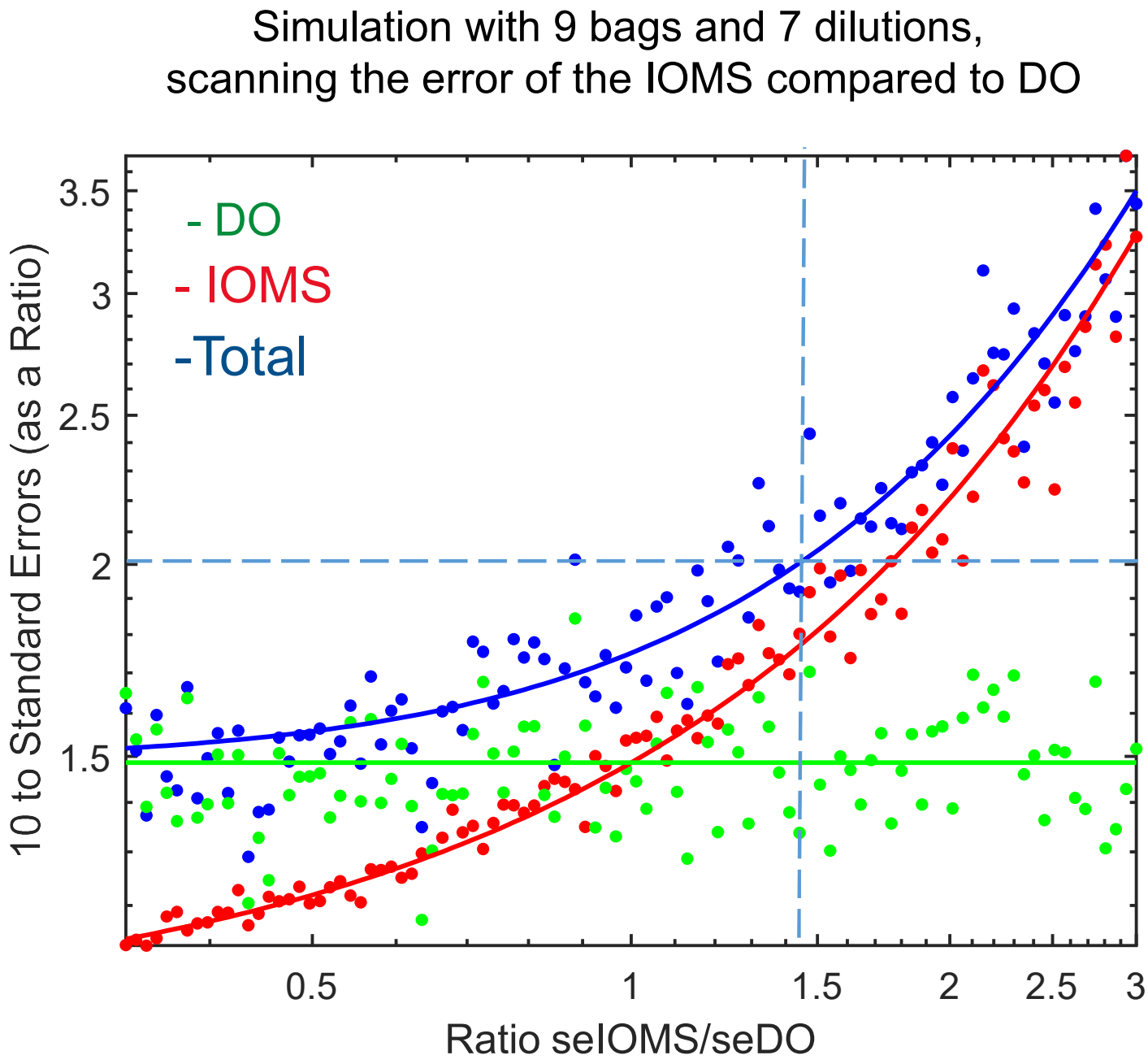
$10^{Abs(Bias)} < 1.64$

Limits EN13725: Intermediate Precision

3

SIMULATION STUDIES

Monte Carlo



In log10 scale

Sigma total = 0.301

Sigma DO= 0.173

Sigma IOMS= 0.245

Factor

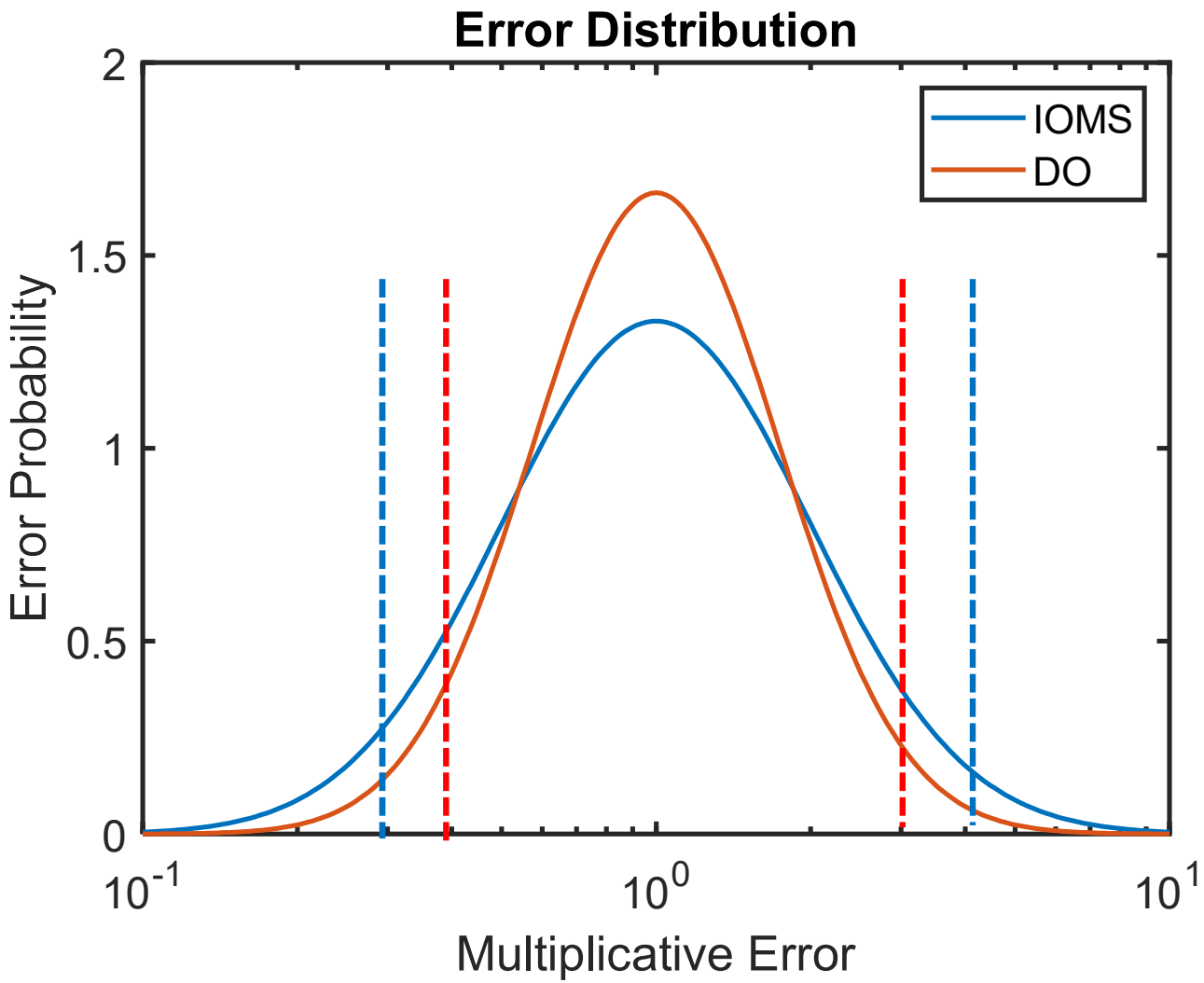
Sigma total = 2

Sigma DO= 1.49

Sigma IOMS= 1.76

Error Distributions: DO vs IOMS

For typical performance, multiplicative errors in IOMS have more tails tan DO



CI 95%

F=2 in DO

F=3 in IOMS

Summary ; Methods

- Model comparison is conceptually different from calibration.
- Correlation coefficient is not recommended as the single figure of merit to compare methods.
- Methods based on regression look for models with slope 1 and intercept 0 as acceptance criterio (EN14793)
- Methods based on EIV regressions have different underlying statistical hypothesis
- Statistical analysis of reading differences was proposed by Bland & Altman assuming Gaussian distribution.
- The extension of B&A to distribution-free (Chebyshev) bounds leads to wider uncertainty bands.

Summary: Results

- **Explored datasets do not suffer from Bias problems.**
- **In the worst-case LoA under normality are (0.25 – 4).**
- **This bands go to (1/20 to 20) if we prefer a distribution free statistic**
- **In all the analyzed datasets the differences look gaussian**
- **Chebyshev leads to very conservative LoA bands**
- **With few samples estimated LoA suffer from large variability**
- **Using a simple model we can estimate the variance of the IOMS.**
- **In the worst-case the CI 95% for IOMS readings are a factor of 3 compared to a factor of 2 for DO.**

- **Limits of the study:**
  - The lifetime of these uncertainties has not been explored.
  - All statistics assume that samples are representative of the population.
  - In case of seasonal effects, multiple sources, etc., all the sources of variance have to be properly sampled leading to higher sample size.

References

- ALTMAN, Douglas G.; BLAND, J. Martin. Measurement in medicine: the analysis of method comparison studies. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 1983, vol. 32, no 3, p. 307-317.
- BLAND, J. Martin; ALTMAN, Douglas G. Statistical methods for assessing agreement between two methods of clinical measurement. *International Journal of Nursing Studies*, 2010, vol. 47, no 8, p. 931-936.
- BLAND, Martin. *An introduction to medical statistics*. Oxford University Press (UK), 2015.
- CHOUDHARY, Pankaj K.; NAGARAJA, H. N. Measuring agreement in method comparison studies—a review. En *Advances in ranking and selection, multiple comparisons, and reliability*. Birkhäuser Boston, 2005. p. 215-244.
- BILIC-ZULLE, Lidija. Comparison of methods: Passing and Bablok regression. *Biochemia medica: Biochemia medica*, 2011, vol. 21, no 1, p. 49-52.