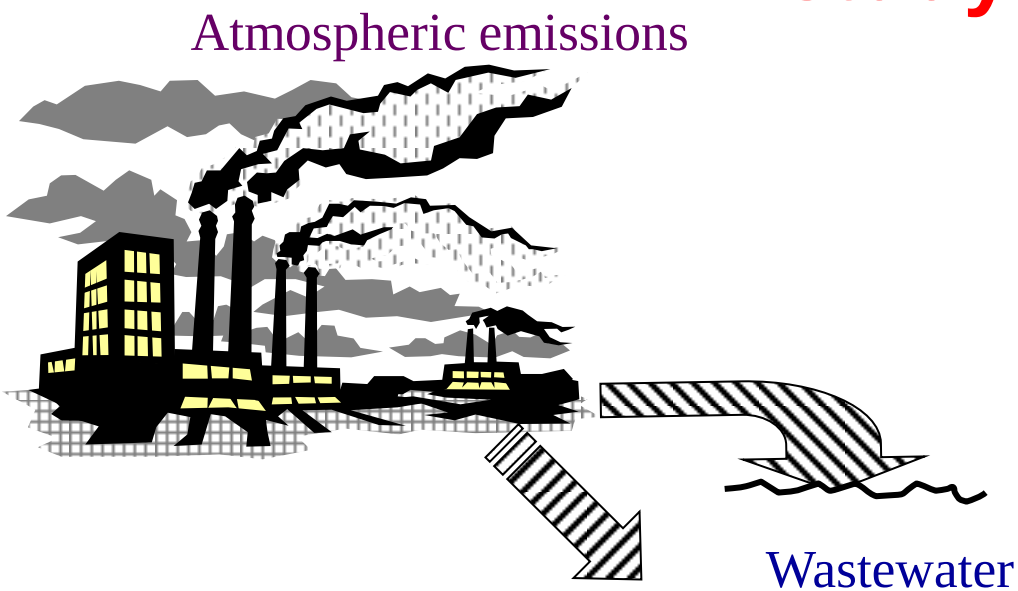
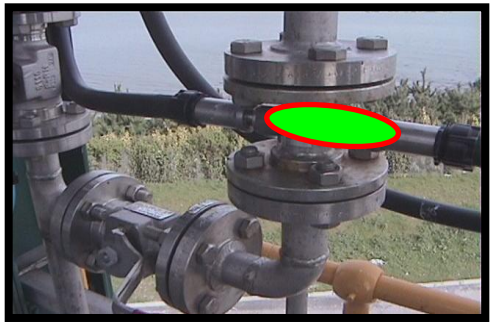
LAPPEENRANTA
UNIVERSITY OF TECHNOLOGY

ACABS
Research Group

Comparison of independent process analytical measurements — a variographic study





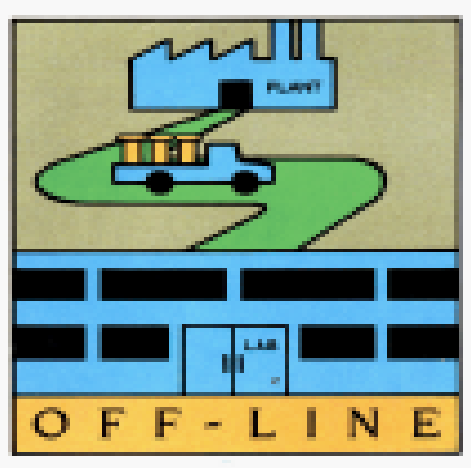
Pentti Minkkinen
1) Lappeenranta University of Technology
2) Aalborg University Campus Esbjerg
E-mail: Pentti.Minkkinen@lut.fi

Outline

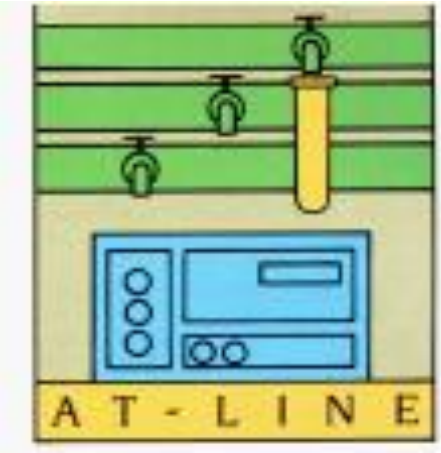
- Process analytical chemistry – measure-ment systems
- Why comparisons are needed
  - Calibration
  - Performance tests of measurement systems
  - Comparison of process means
- Data analysis methods
  - Incorrect
  - Correct

Process Analytical Chemistry

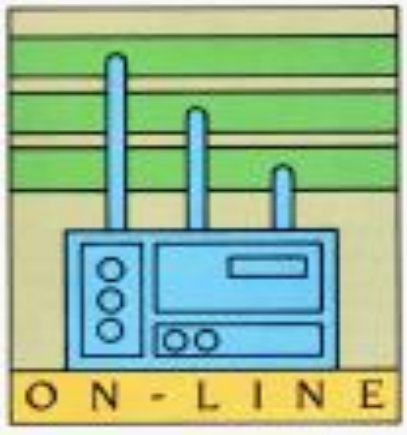
Off-line analysis



At-line analysis

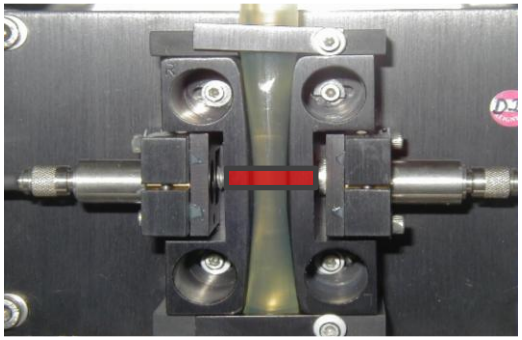


On-line analysis



In-line analysis

Non-invasive analysis



Student's t-tests

- Most widely used statistical test in testing the *equality* of the mean values of two measurement sets
- Basic assumptions:
  - Sets independent
  - Normal distribution (sensitive to non-normality)
- Assumptions seldom met in process analysis

t-tests for comparing 2 means

Two measurement sets:

$\mathbf{x}1, \quad \bar{x}_1, s_1, n_1, \nu_1$

$\mathbf{x}2, \quad \bar{x}_2, s_2, n_2, \nu_2$

n_i = number of measurements

$\nu_i = n_i - 1$ = degrees of freedom

$d = \bar{x}_2 - \bar{x}_1$ = difference of the two mean

QUESTION: Is d significantly different from zero?

A) Standard deviations of both sets can be assumed to be equal

$\sigma_1^2 = \sigma_2^2 \approx s^2$, confirmed with F-test

$$F = \frac{s_2^2}{s_1^2} \qquad (1)$$

If this is not significant the standard deviations can be pooled.

$$s = \sqrt{\frac{\nu_1 s_1^2 + \nu_2 s_2^2}{\nu_1 + \nu_2}}, \quad \nu = \nu_1 + \nu_2 \qquad (2)$$

The standard deviation of the difference with ν degrees of freedom, is

$$s_d = \sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}} \qquad (3)$$

Two ways to detect if *d* is different from zero

a) If the confidence interval, *ci*, does not include zero

$$ci = d \pm t_{2\alpha,\nu} \cdot s_d \qquad (4)$$

b) t-test

$$t = \frac{|d - 0|}{s_d} \qquad (5)$$

B) Standard deviations not equal

$$s_1^2 \neq s_2^2$$ (F-test significant)

The standard deviation of the difference *d* is calculated as

$$s_d^* = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \qquad (6)$$

The degrees of freedom have to be estimated by using Satterthwaite's formula

$$\nu^* = \frac{s_d^4}{\frac{s_1^4}{n_1^2 \nu_1} + \frac{s_2^4}{n_2^2 \nu_2}} \qquad (7)$$

Significance estimated either by using Eqs. 4 or 5

C) Parallel determinations on same samples

$$\mathbf{d} = \mathbf{x1} - \mathbf{x2}, \quad \bar{d}, s_d, n, \nu = n - 1$$

s_d → SD of the differences

Conclusion on significance can be based on confidence interval

$$ci = \bar{d} \pm t_{2\alpha,\nu} \cdot \frac{s_d}{\sqrt{n}} \qquad (8)$$

or on t-test

$$t = \frac{\left|\bar{d} - 0\right|}{s_d} \sqrt{n} \qquad (9)$$

Pitfalls

- Tests A) and B) cannot be used to test differences of analytical methods, if tests are carried out on different samples (between-samples variance will mask the analytical variance)
- Test C) eliminates the between-samples variance. However, if the analytical variance is dependent on concentration d_i's are not normally distributed
- All tests fail, or are inefficient in multivariate case, if correlated variables are tested one-at-time
- Autocorrelation is a problem in all

Estimation of the variance (uncertainty) of the mean of a data set from a dynamic (elongated) 1-D data set

Data sets are *autocorrelated*, i.e., samples taken within short intervals have values which differ less than samples taken far apart ➔ assumption on normality does not hold

Error made in estimating the mean of a continuous object from discrete samples is called Point Selection Error, PSE.

PSE is the error of the **mean** of a continuous lot estimated by using discrete samples.

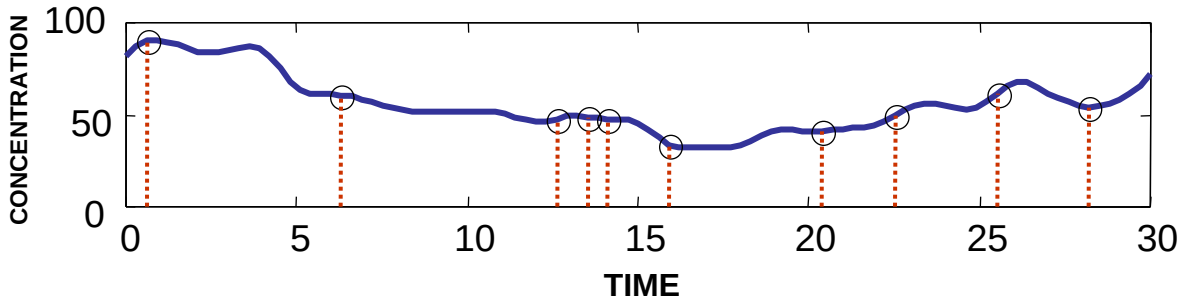
PSE depends on sample selection strategy, if consecutive values are *autocorrelated*. Selection options:

- *random*
- *stratified* random
- stratified *systematic.*

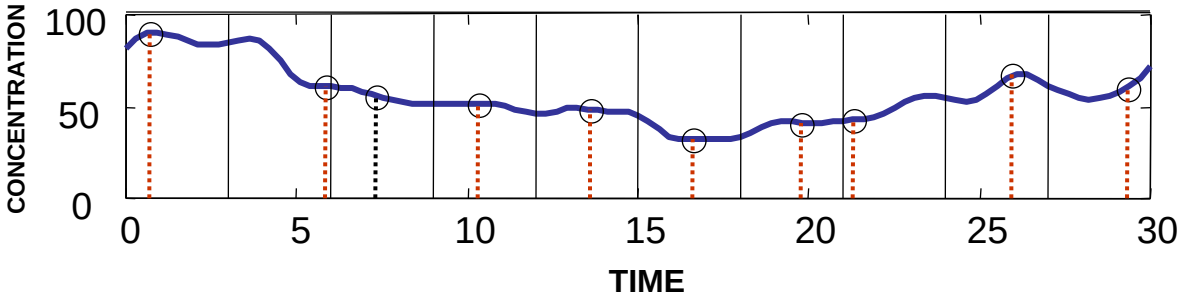
Point selection error has two components: $\boldsymbol{PSE = PSE_1 + PSE_2}$

- $\boldsymbol{PSE_1}$ ... error component caused by random drift
- $\boldsymbol{PSE_2}$ ... error component caused by cyclic drift

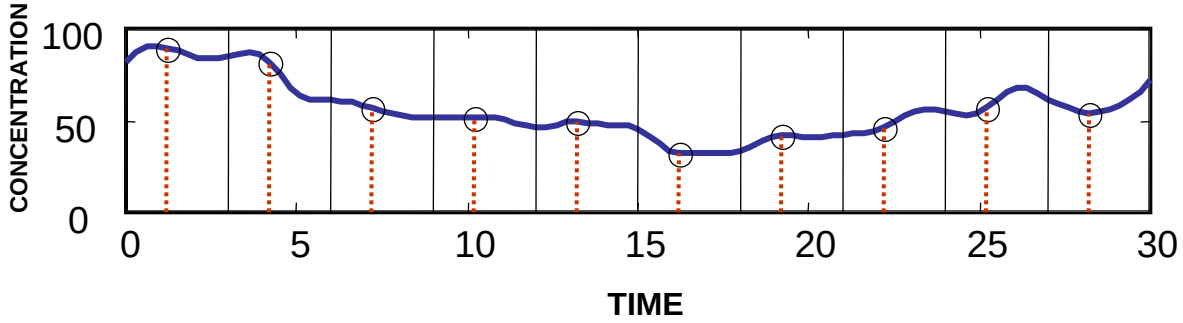
Statistics of ***correlated series*** is needed to evaluate the sampling variance of ***mean*** of the results .



Random selection



Stratified selection



Systematic selection

Sample selection modes

When sampling autocorrelated series the same number of samples gives different uncertainties for the mean depending on selection strategy

Random sampling: $s_{\bar{x}} = \frac{s_p}{\sqrt{n}}$

Stratified sampling: $s_{\bar{x}} = \frac{s_{str}}{\sqrt{n}}$

Systematic sampling: $s_{\bar{x}} = \frac{s_{sys}}{\sqrt{n}}$

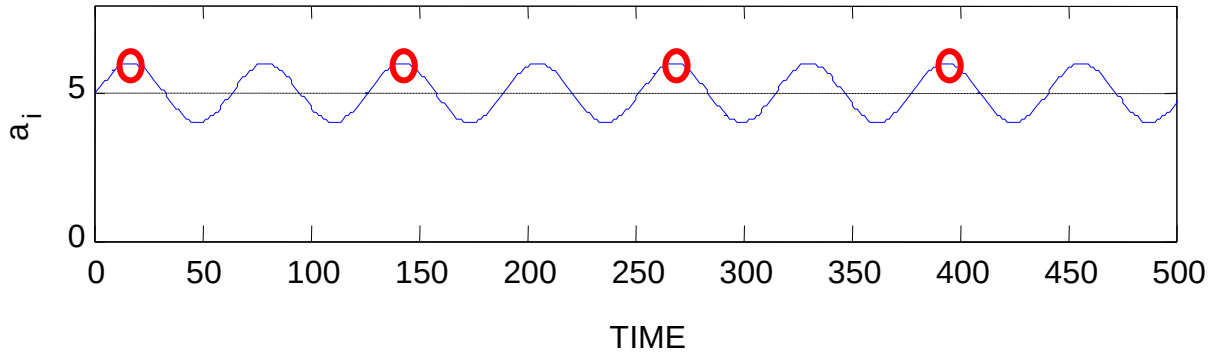
s_p is the process standard deviation

s_{str} and s_{sys} are standard deviation estimates where the autocorrelation has been taken into account.

Normally $s_p > s_{str} > s_{sys}$,

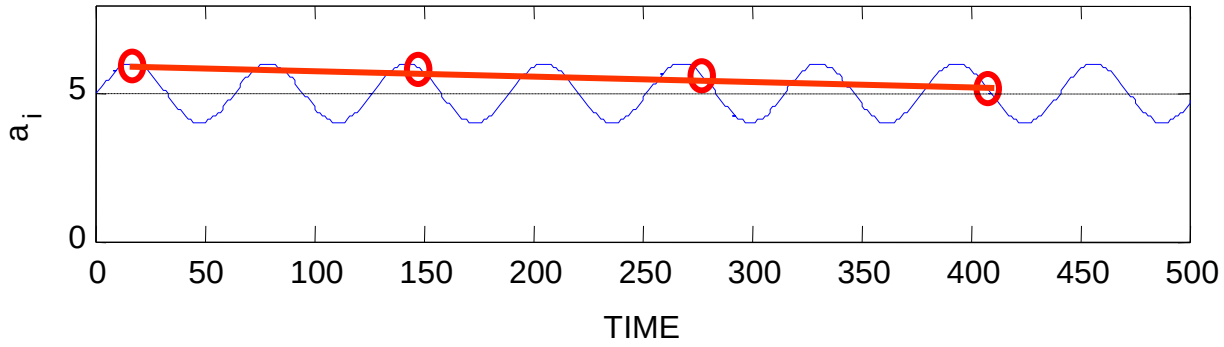
except in periodic processes, where s_{sys} may be the largest

Systematic sampling from periodic process



$a_L = 0$

$a_{sample} = 0.996$



$a_L = 0$

$a_{sample} = 0.689$

If too low sampling frequency is used in sampling periodic processes there is always a danger that the mean is *biased*

Estimation of *PSE* by variography

Variogaphic experiment: N samples collected at equal distances and analyzed, $a_i, M_{s_i}, \overline{M}$ are analytical results, sample sizes and mean sample size, respectively.

Mean of the process:

$$a_L = \frac{\sum M_{s_i} a_i}{\sum M_{s_i}} \qquad (10)$$

Relative heterogeneity of the process:

$$h_i = \frac{a_i - a_L}{a_L} \frac{M_{s_i}}{\overline{M}}, \qquad i = 1, 2, \ldots, N \qquad (11)$$

Absolute heterogeneity of the process:

$$h_i = (a_i - a_L) \frac{M_{s_i}}{\overline{M}}, \qquad i = 1, 2, \ldots, N \qquad (12)$$

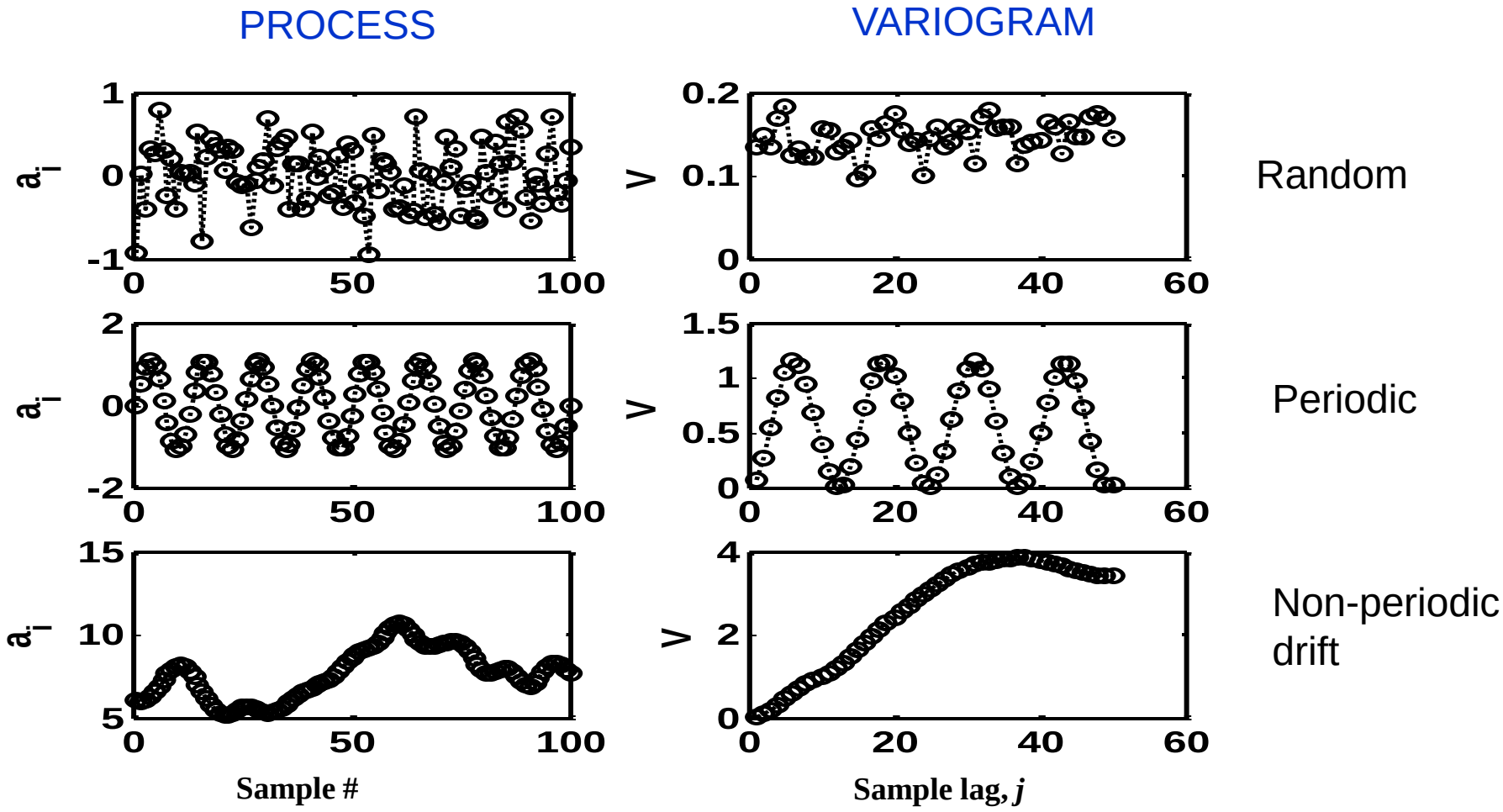
Variogram of heterogeneity as function of sampling interval j :

$$V_j = \frac{1}{2(N-j)} \sum_{i=1}^{N-j} (h_{i+j} - h_i)^2 , \quad j = 1,2,\ldots,\frac{N}{2} \qquad (13)$$

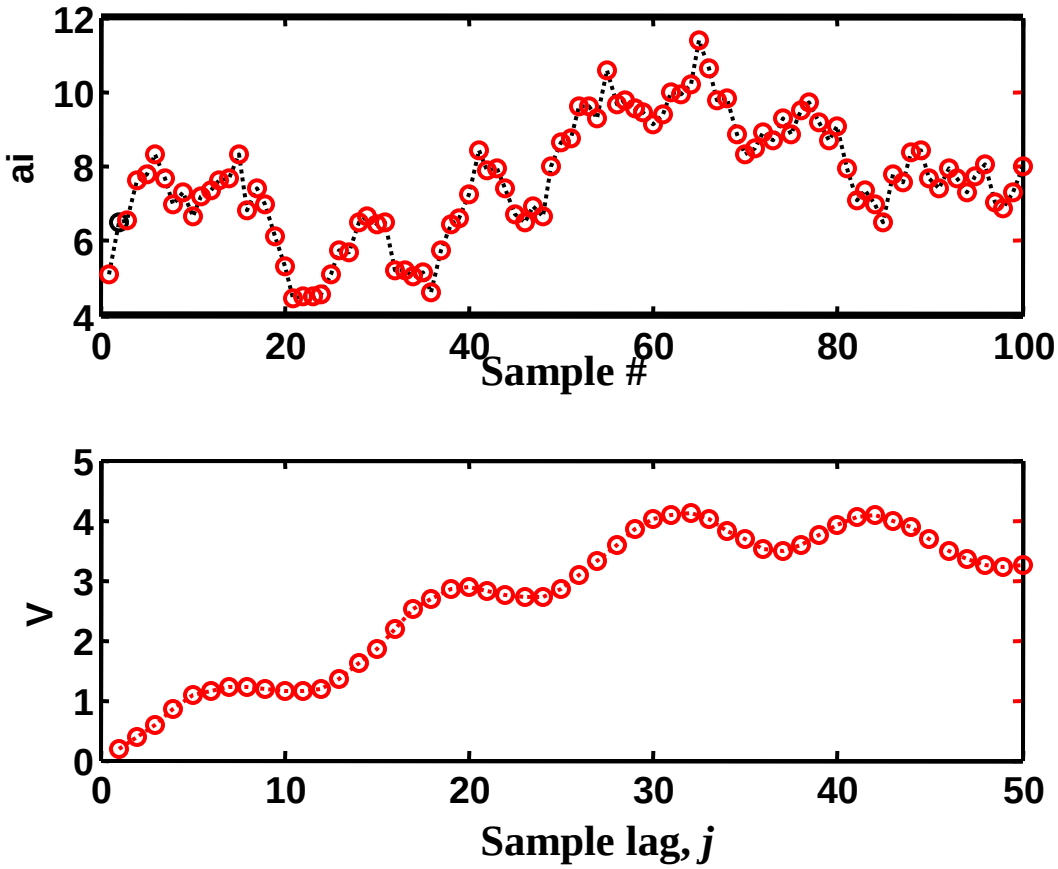
To estimate variances the variogram has to be integrated (numerically in Gy's method).

Analysis of variogram provides variance estimates for estimating the mean of the data set obtained by random, stratified or systematic sample selection mode

$$\mathrm{var}(a_L) = \frac{s^2_{ra,st,sy}(j)}{n} \qquad (14)$$

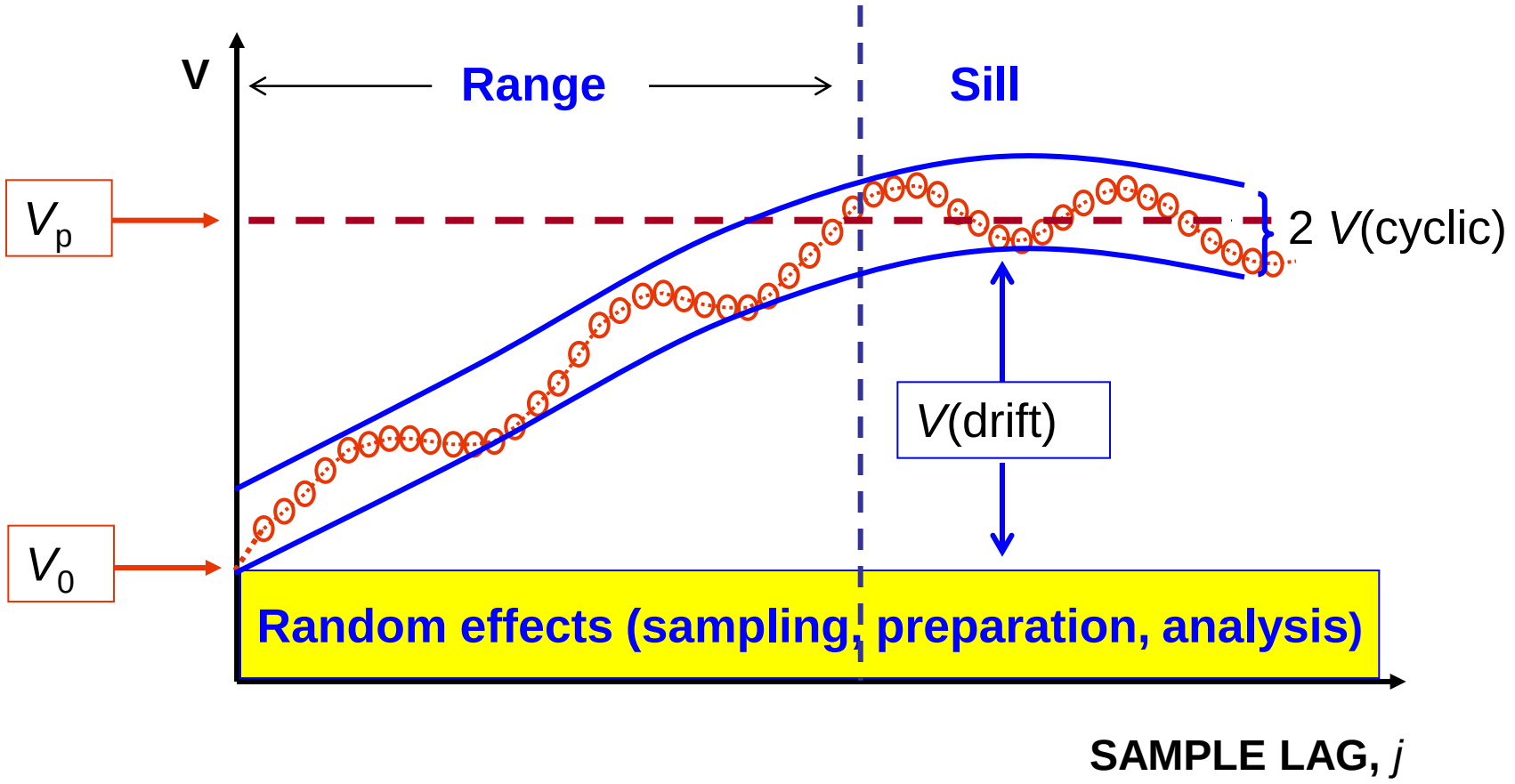


VARIOGRAMS FOR THREE DIFFERENT BASIC PROCESS TYPES



Data and variogram of a complex process

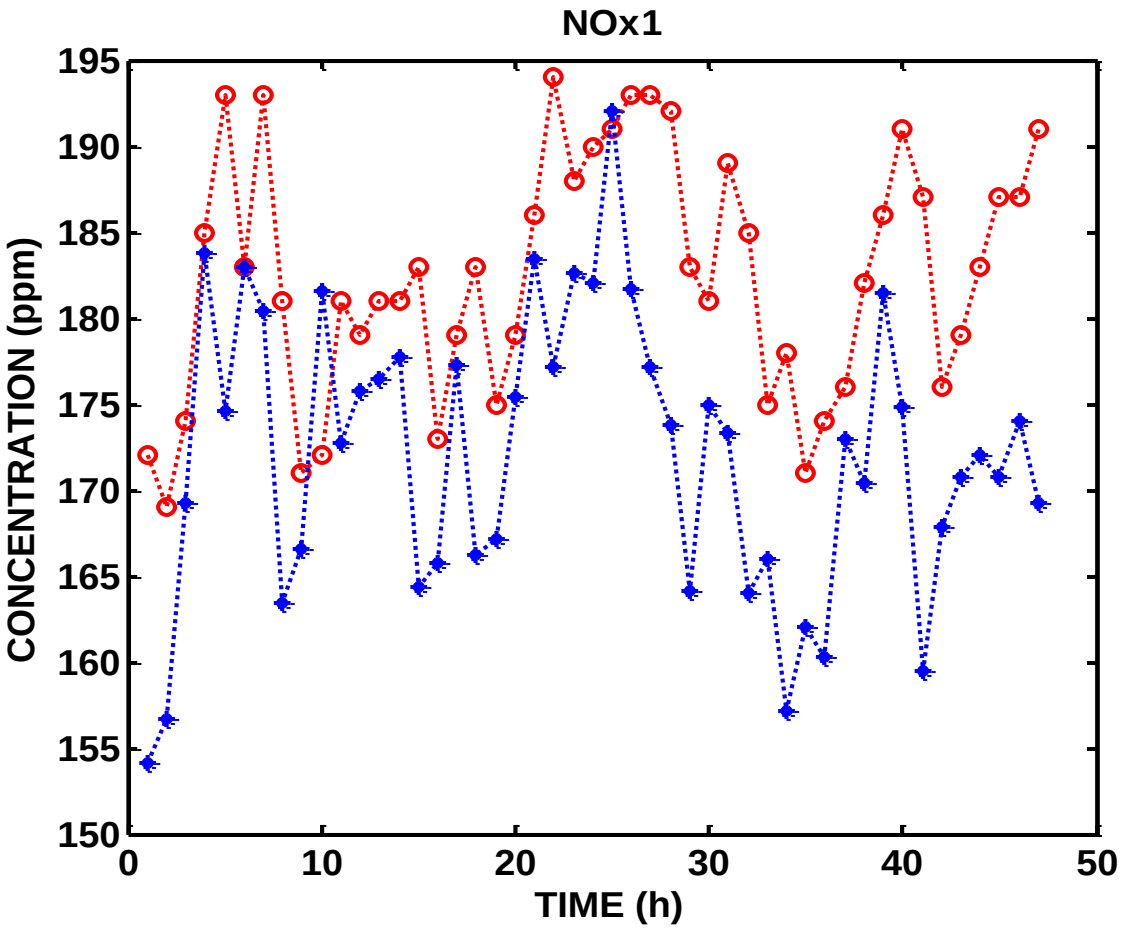
Interpretation of the variogram



Examples

- Simultaneous emission measurements by two different process analyzers/teams from a power plant
  - NOx emission
  - O2 in stack

NOx emission: Control measurements



Control

$\bar{x}_1$ = 182.4

s_1 = 7.1

Routine

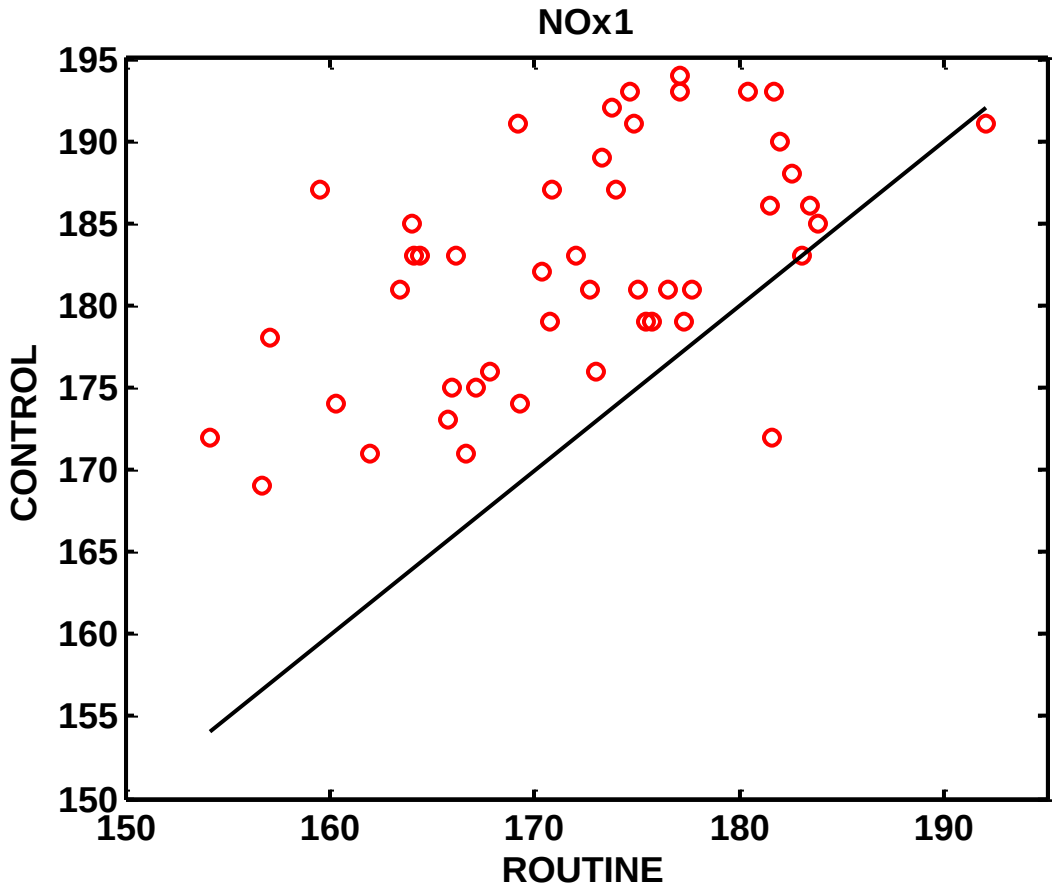
$\bar{x}_2$ = 172.1

s_2 = 8.3

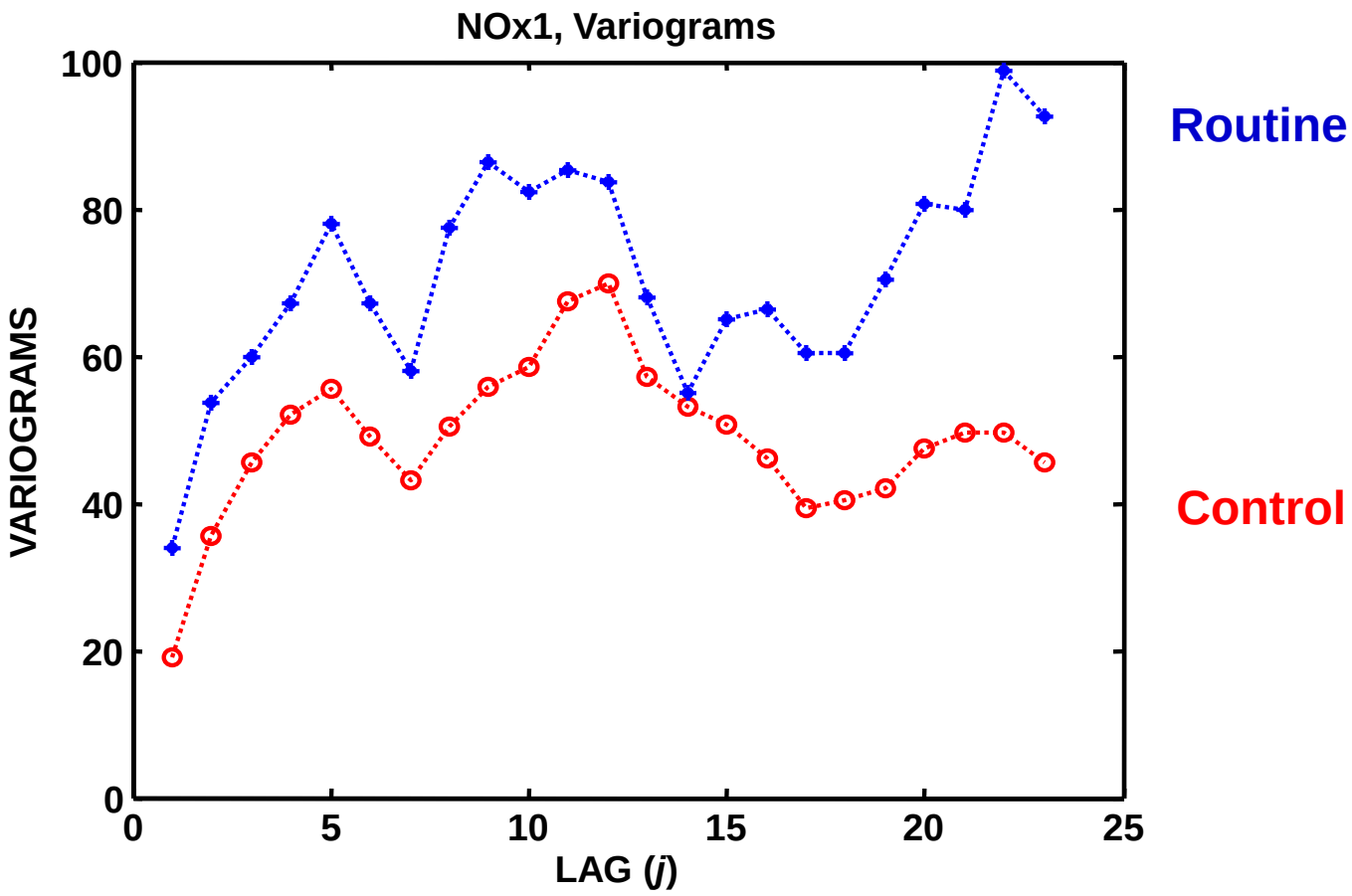
d = 10.3

t-test significant

Results of control vs. routine measurements, two process analyzers

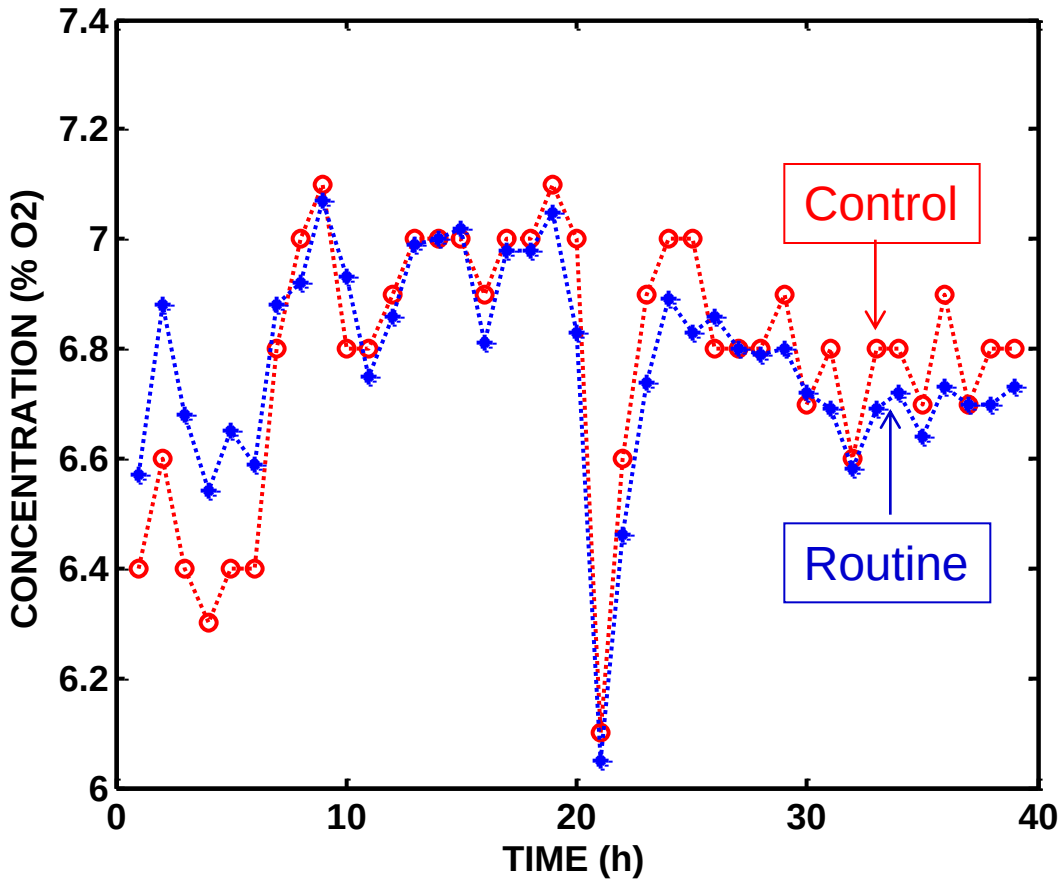


Routine vs. control measurement of NOx emissions



Variograms of the routine and control measurement sets

Oxygen



Control

$\bar{x}_1$ = 6.77

s_1 = 0.191

Routine

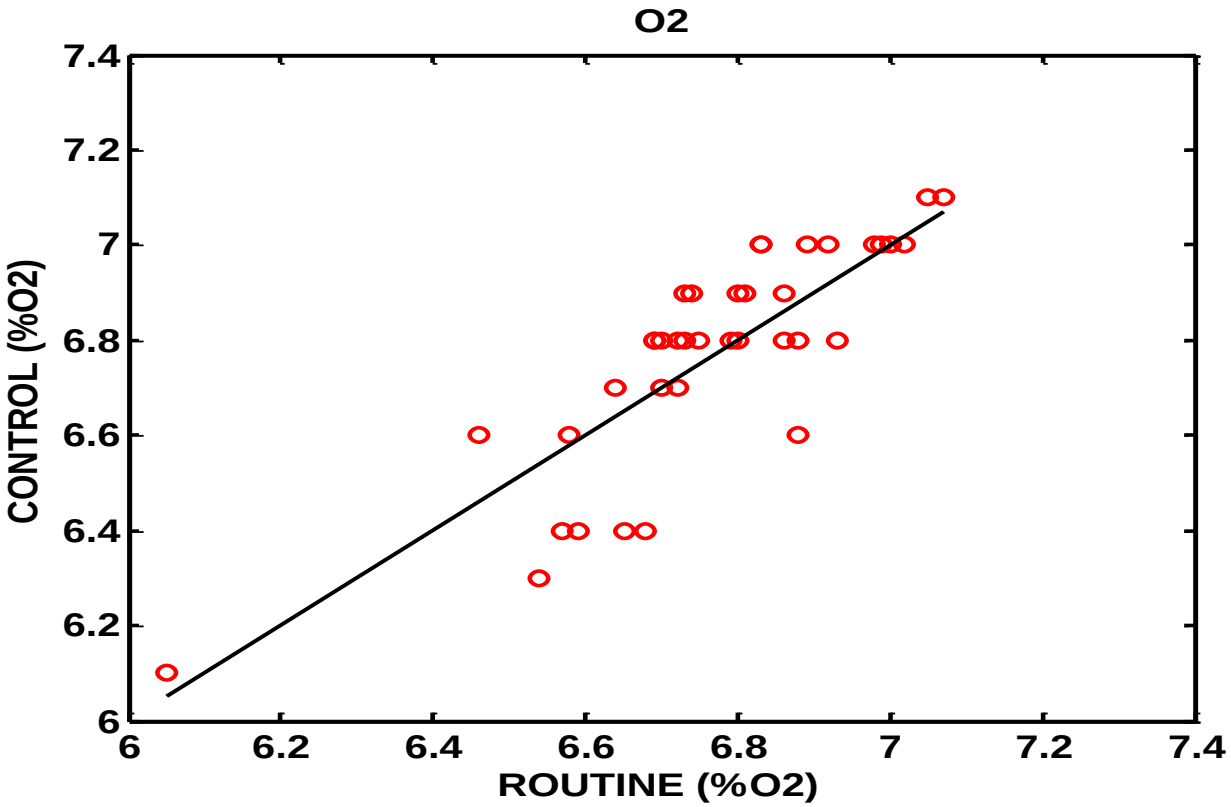
$\bar{x}_2$ = 6.78

s_2 = 0.253

t=0.224

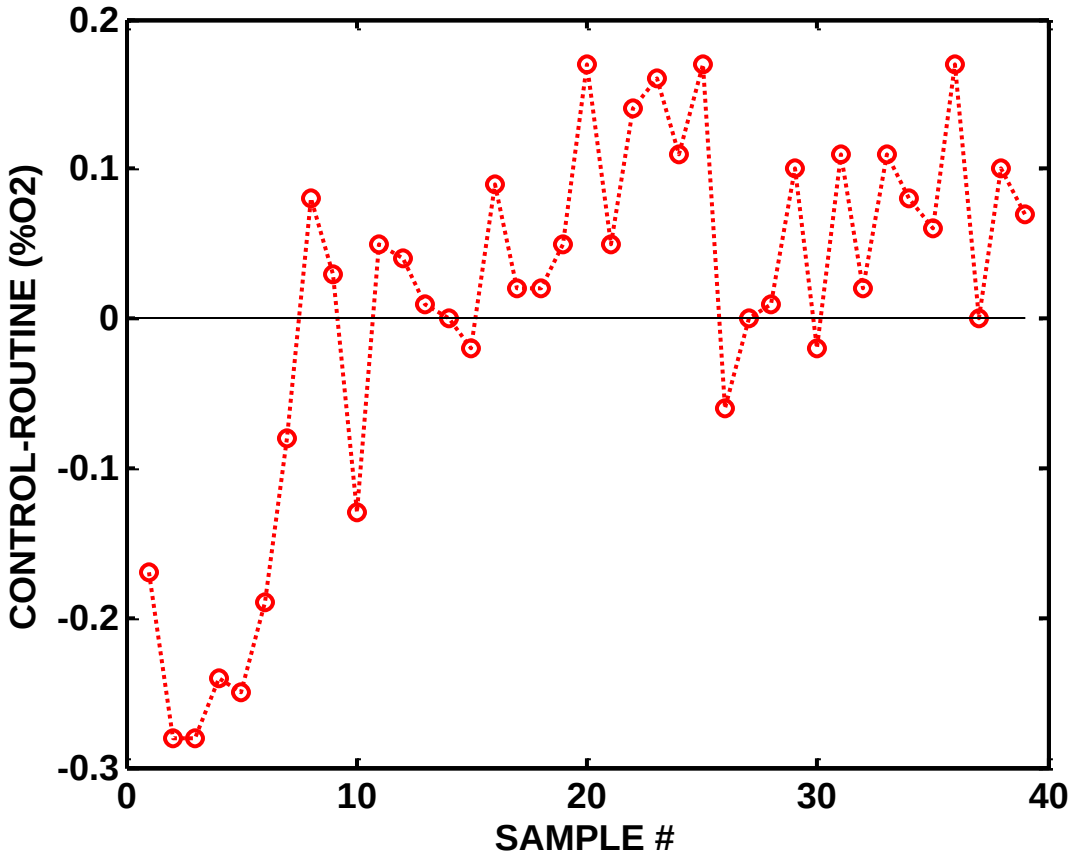
not significant

Results of control vs. routine measurements:
parallel O_2 measurements using two different process analyzers



Control vs. routine measurements, results of linear regression analysis:

Intercept = -0.3004 (95 % ci = -1.76 ... 1.16)
Slope = 1.046 (95 % ci = 0.83 1.26)



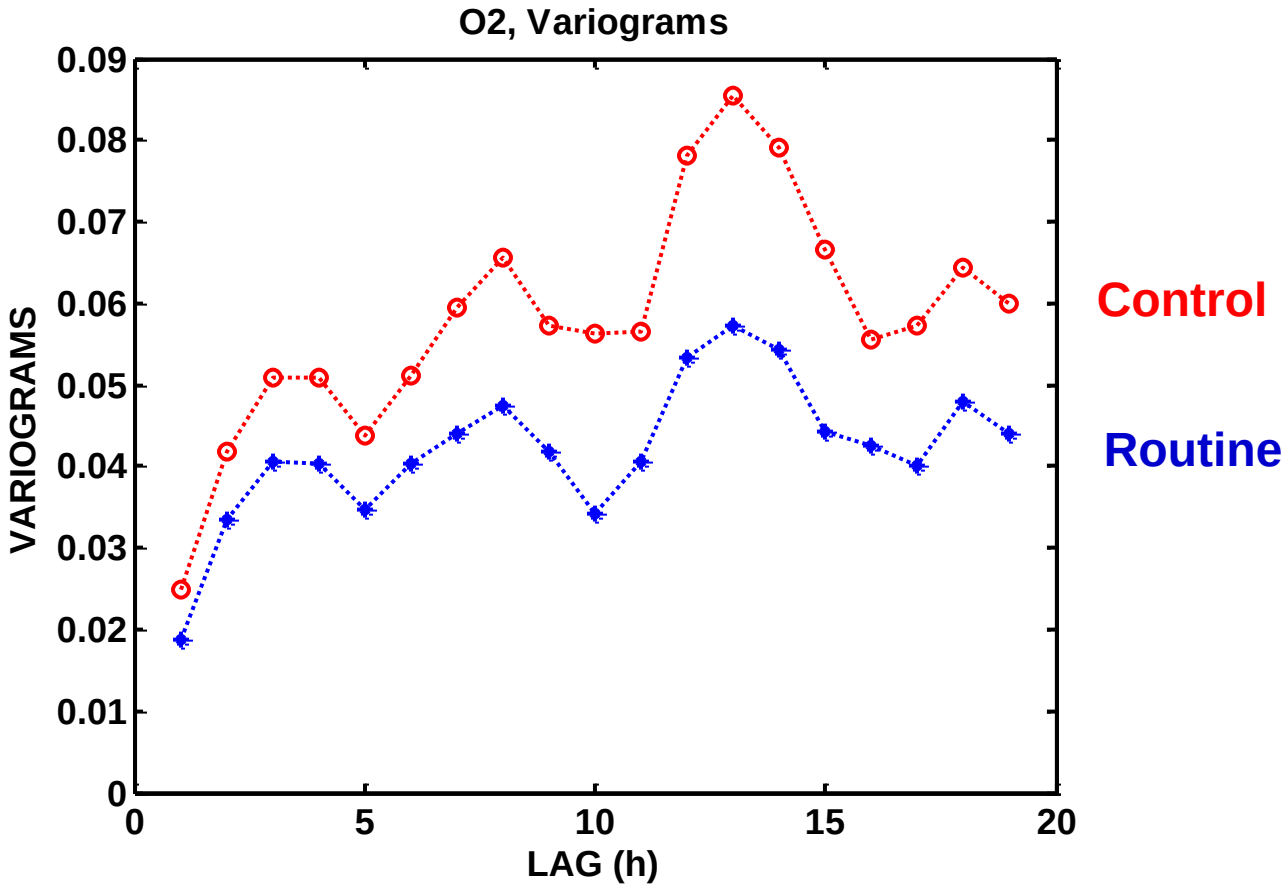
$\bar{d}$ = 0.0077

s_d = 0.125

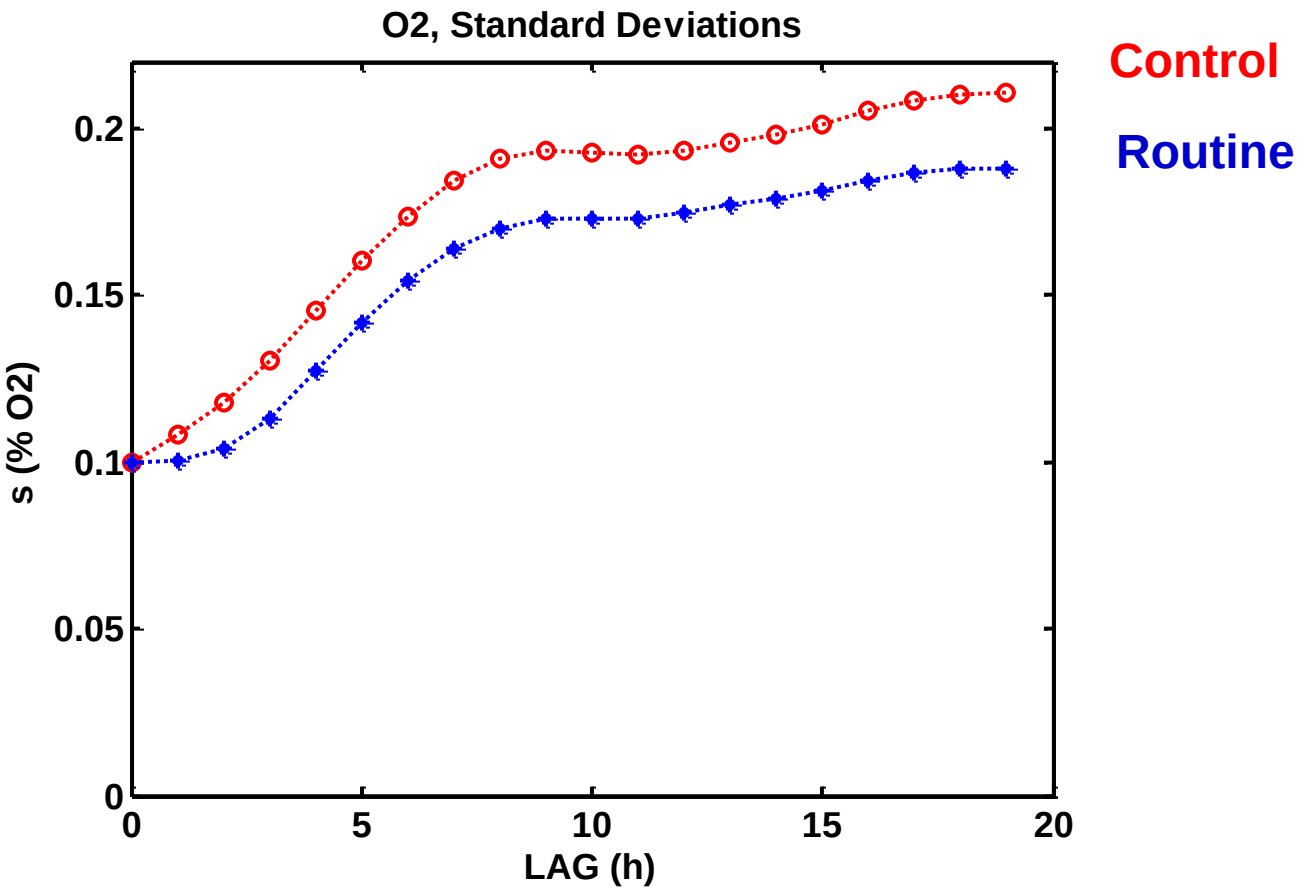
t=0.224

not significant

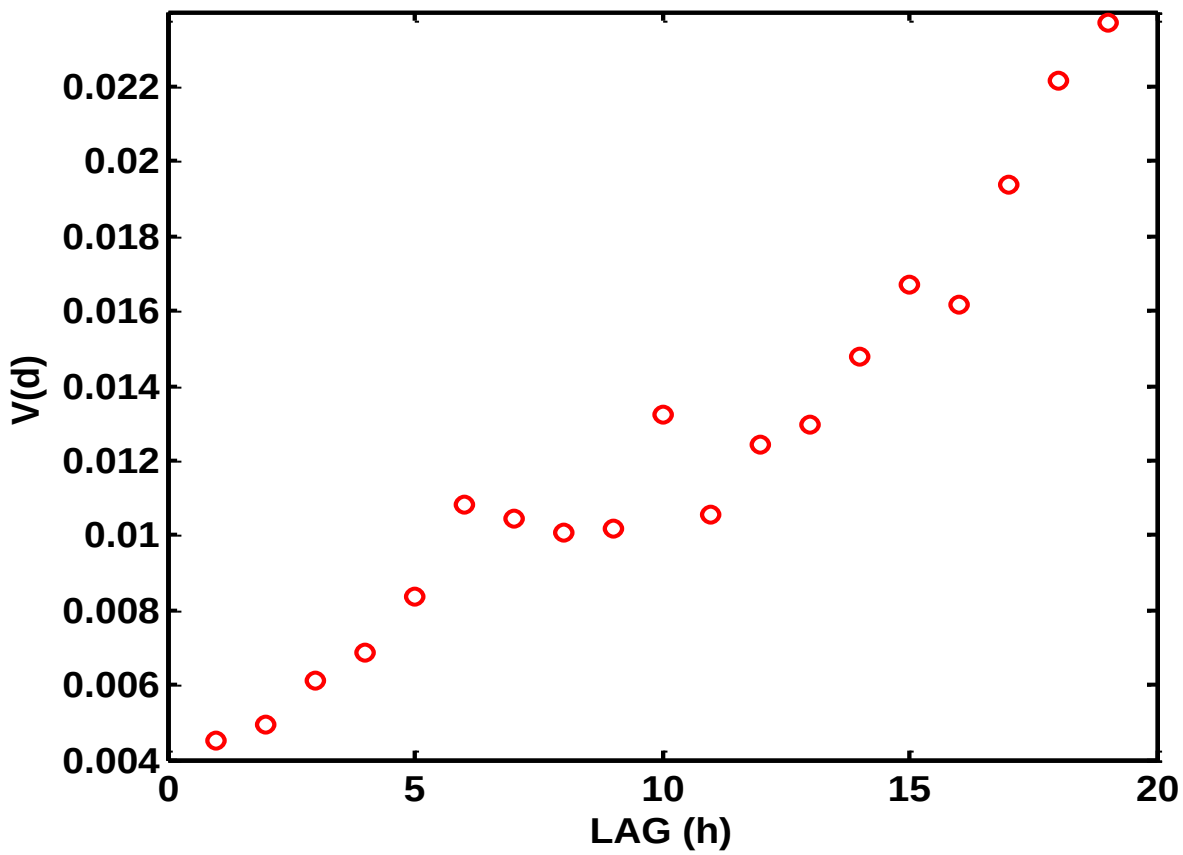
Differences of the control - routine measurements



Variograms of the routine and control measurement sets



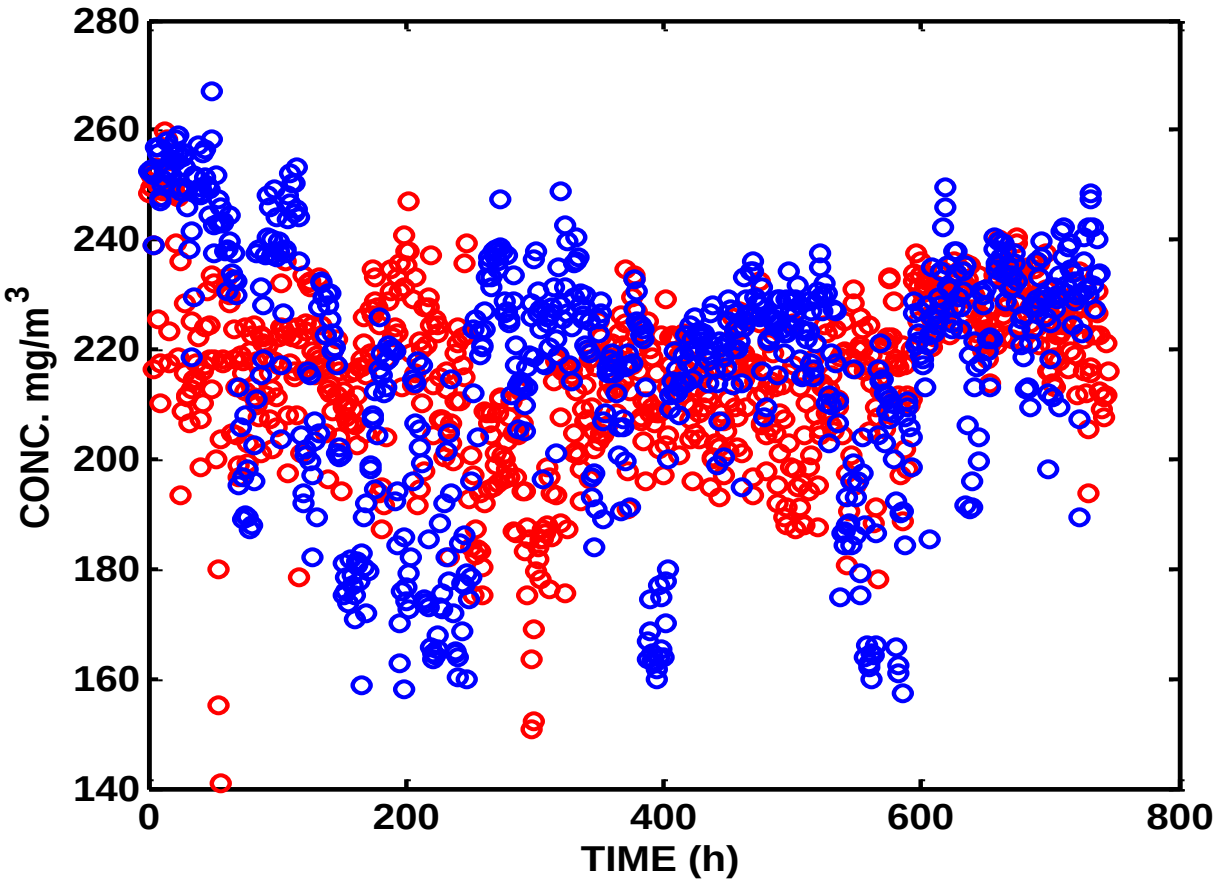
Standard deviations of systematic sampling mode estimated from the variograms of the routine and control measurement sets



Variogram of the difference d, control-routine measurement of O2

Comparison of two different process means

Does process change affect the result (or behavior of the process)?



Period 1

$\bar{x}_1$ = 214.8

s_1 = 15.4

Period 2

$\bar{x}_2$ = 216.5

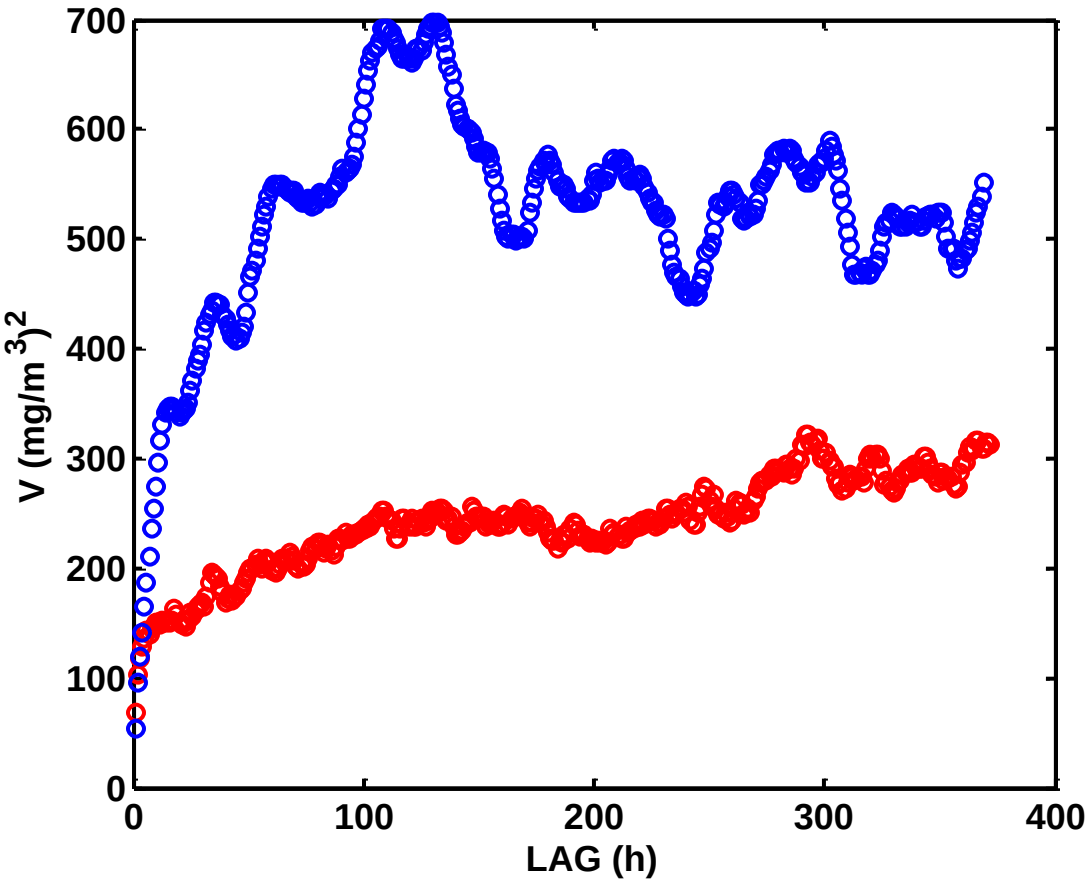
s_2 = 23.2

d =1.7

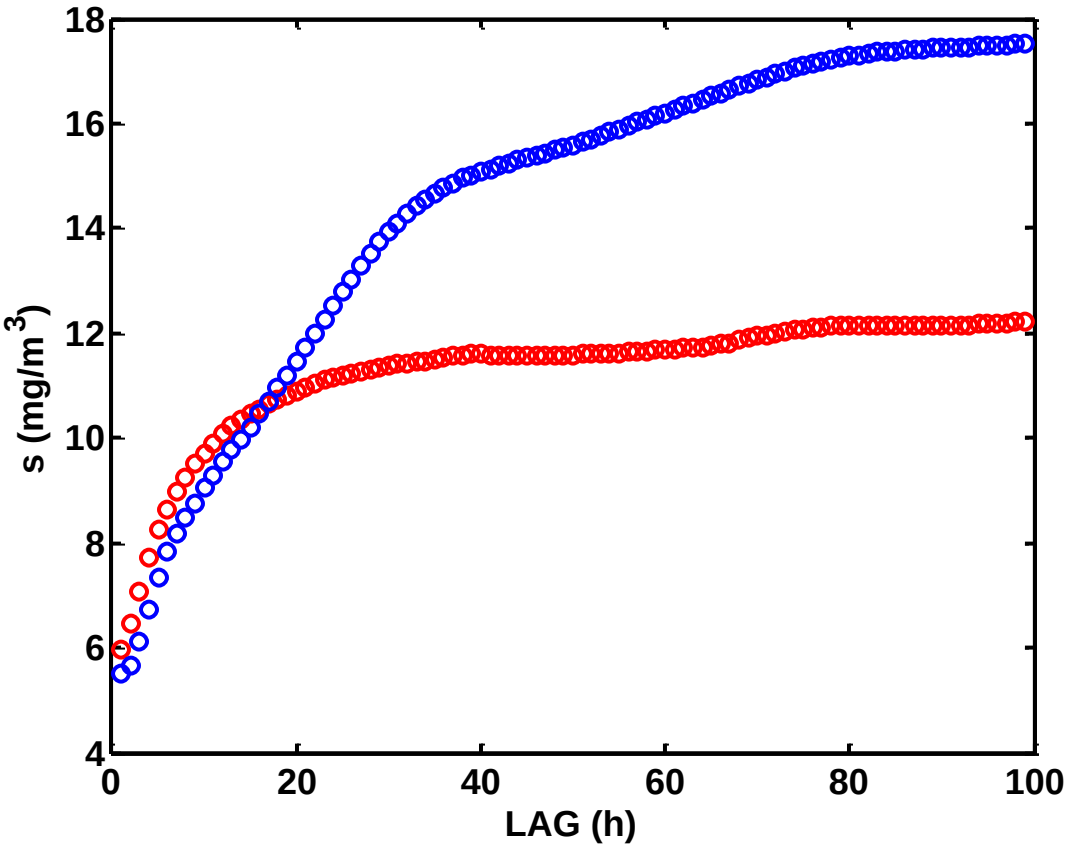
t = 1.68

Not significant

NOx emissions from a power plant within two different time periods, 745 and 738 data points



Variograms of the of the data sets from two different time spans



Standard deviations of systematic sampling mode estimated from the variograms of the two time spans

t-test taking the autocorrelation into account

$$\bar{x}_1 = 214.8, s_1 = 5.95, n_1 = 745$$

$$\bar{x}_2 = 216.5, s_2 = 5.49, n_2 = 738$$

t3 = 5.78 SIGNIFICANT

CONCLUSIONS

- Before selecting the statistical tools and drawing inferences try to plot the data so that it shows the desired phenomenon
- Variographic analysis of time series is a powerful tool. It separates random effects, non-periodic and periodic drift
- In multivariate case variographic analysis can be carried out, e.g., on PCA scores

THANK YOU, tack, kiitos, danke, merci, obrigado, gracias, grazie, tesekkür ederim, sukran, Спасибо

Graduate courses at Lappeenranta University of Technology (in English)

- Experimental Design, 25-26 March, 2010
- Sampling for Chemical Analysis,7-9 April, 2010