Available online at www.sciencedirect.com
ScienceDirect
Analytica Chimica Acta 595 (2007) 209–215
ANALYTICA CHIMICA ACTA
www.elsevier.com/locate/aca

Effect of missing values in estimation of mean of auto-correlated measurement series

Maaret Paakkunainen*, Jarmo Kilpeläinen, Satu-Pia Reinikainen, Pentti Minkkinen

Lappeenranta University of Technology, Department of Chemical Technology, P.O. Box 20, 53851 Lappeenranta, Finland

Received 27 October 2006; received in revised form 5 January 2007; accepted 10 January 2007
Available online 16 January 2007

Abstract

Sampling and uncertainty of sampling are important tasks, when industrial processes are monitored. Missing values and unequal sources can cause problems in almost all industrial fields. One major problem is that during weekends samples may not be collected. On the other hand a composite sample may be collected during weekend. These systematically occurring missing values (gaps) will have an effect on the uncertainties of the measurements. Another type of missing values is random missing values. These random gaps are caused, for example, by instrument failures.

Pierre Gy's sampling theory includes tools to evaluate all error components that are involved in sampling of heterogeneous materials. Variograms, introduced by Gy's sampling theory, have been developed to estimate the uncertainty of auto-correlated process measurements. Variographic experiments are utilized for estimating the variance for different sample selection strategies. The different sample selection strategies are random sampling, stratified random sampling and systematic sampling.

In this paper both systematic and random gaps were estimated by using simulations and real process data. These process data were taken from bark boilers of pulp and paper mills (combustion processes). When systematic gaps were examined a linear interpolation was utilized. Also cases introducing composite sampling were studied.

Aims of this paper are: (1) how reliable the variogram is to estimate the process variogram calculated from data with systematic gaps, (2) how the uncertainty of missing gap can be estimated in reporting time-averages of auto-correlated time series measurements.

The results show that when systematic gaps were filled by linear interpolation only minor changes in the values of variogram were observed. The differences between the variograms were constantly smallest with composite samples. While estimating the effect of random gaps, the results show that for the non-periodic processes the stratified random sampling strategy gives more reliable results than systematic sampling strategy. Therefore stratified random sampling should be used while estimating the uncertainty of random gaps in reporting time-averages of auto-correlated time series measurements.

Keywords: Sampling; Sampling uncertainty; Missing values; Variographic analysis; Pierre Gy's sampling theory; Auto-correlated data; Process data

1. Introduction

Process data used to estimate time-averages often include missing values and they may increase the uncertainty of the mean values. The handling of missing values is a significant task in process industry. In Finnish legislation, there is a decree, which regulates the emissions and measurement uncertainty of combustion processes [1]. This decree states, e.g., that the emissions should be measured as 1 h mean values with 95% uncertainty. If more than three of the 1 h mean values per day are rejected, the emission value of that day should also be rejected (service and malfunction are used as criteria for rejection). Furthermore, if 10 daily values per year had to be excluded, then the environmental authorities may require corrective actions, or even restrict the operating time of the combustion plant.

The purpose of this paper is firstly to describe, how reliable the process variogram is if data includes systematic gaps, and secondly to find out how the uncertainty generated by missing values can be estimated when time-averages of auto-correlated time series measurements are reported.

Variograms are estimated with variographic analysis that is included in the Pierre Gy's sampling theory. Pierre Gy's sampling theory, which is more than 50 years old gives tools to evaluate all error components that are involved in the sampling

* Corresponding author. Tel.: +358 5 621 2263.
E-mail address: maaret.paakkunainen@lut.fi (M. Paakkunainen).

0003-2670/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.aca.2007.01.020

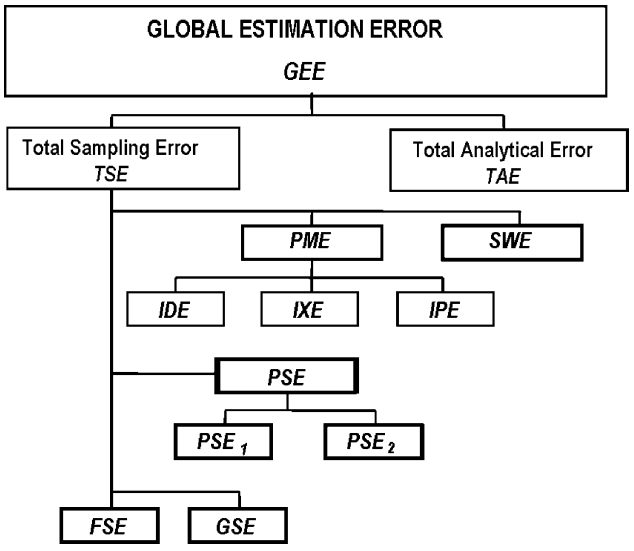

Fig. 1. Components of sampling errors according to Pierre Gy.

of heterogeneous materials [2–4]. Fig. 1 shows the different error sources of an analytical determination according to Gy. Global estimation error (GEE) is the sum of total analytical error (TAE) and total sampling error (TSE). Total sampling error consists of eight different error components: sample weighting error (SWE) which is caused if the fluctuation in the flow-rate of the process stream is ignored. Incorrect delimitation error (IDE), incorrect extraction error (IXE) and incorrect processing error (IPE) are due to the incorrect sampling and can be controlled by proper design of sampling equipment. Fundamental sampling error (FSE) is pure random error and the only error component that is left even if all other error components are eliminated. FSE can be estimated theoretically if all necessary material properties are known. Grouping and segregation error (GSE) occurs when the sampling increments are not ideal. It is a result of the material heterogeneity and the sampling process. Long-term point selection error (PSE_1) and periodic point selection error (PSE_2) are caused from the random and the cyclic drifts in the process. Point selection error ($PSE = PSE_1 + PSE_2$) is the error of the mean of a continuous lot estimated by using discrete samples. The size of this error component depends on the sample selection strategy and the degree of auto-correlation [4,5].

The optimal sampling frequency and the uncertainty of the sampling process often depend on the sampling strategy. The different sample selection strategies are shown in Fig. 2. When

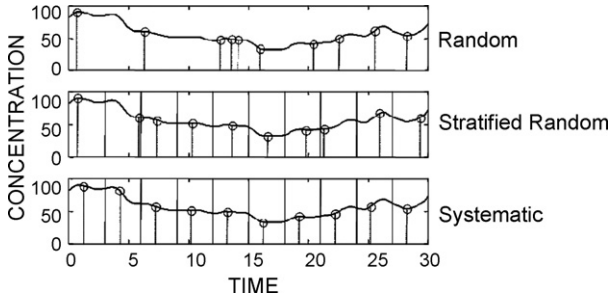

Fig. 2. Sample selection strategies.

using random sampling the samples are taken randomly from the lot. If stratified random sampling or systematic sampling is utilized, the lot is first divided into equal sizes sub-lots. In the stratified random sampling one sample is taken randomly from each sub-lot. In the systematic sampling one sample is taken systematically from each sub-lot with constant intervals. Usually the random sampling gives the highest standard deviation while the systematic sampling gives the lowest standard deviation. If systematic sampling is utilized, the process periodicity is critical. If sampling frequency is less than 2 samples per one period, and the process frequency is a multiple of the sampling frequency the mean may be biased. In those cases stratified random sampling should be used as it is the safest sample selection mode.

1.1. Variographic analysis

A variogram describes the variability of the process and is utilized for estimating the variances for different sample selection strategies. The experimental variogram is calculated from the heterogeneities. According to Pierre Gy the heterogeneity of the sampling target is defined as the relative fluctuation about the mean of the lot to be examined.

Let i be the index of a sample, a_i the analytical result and a_L the weighted mean of the lot. M_{s_i} is the size of sample i, if the sample is cut across the process stream so that it is proportional to the flow-rate, or the flow-rate, if constant sample volume is taken. $\bar{M}_s$ is the mean sample size (or flow-rate), and N the total number of samples. The heterogeneity is ([2], pp. 64–65):

$$h_i = \frac{a_i - a_L}{a_L} \frac{M_{s_i}}{\bar{M}_s}, \quad i = 1, 2, \ldots, N \tag{1}$$

The experimental variogram, V_j, is ([2], p. 91):

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

The experimental variogram is calculated up to sampling interval $N/2$ (rounded down). Petersen et al. [5] have shown that if higher lags are used, the center experiments will not be included in the calculations.

When the experimental variogram is calculated, the intercept, V_j ($j = 0$) is needed. Several techniques to obtain V_0 are presented by Heikka and Minkkinen [6]. They have found that the most reliable way to estimate V_0 is to carry out a separate test, where a series of pairs of increments are taken at the shortest possible sampling interval. A graphic approach was also found relatively reliable, and it was utilized in these case studies.

Several applications, including studies with missing values, to estimate sampling variance of process analytical measurements have been presented by Paakkunainen et al. [7].

2. Methods

In the experiments simulations and real process data set were studied. For the simulations several auto-correlated time series were generated. In the simulations to estimate the effect of systematic gaps, the amount of the random variance was 20 or