Comparison of integration rules in the case of very narrow chromatographic peaks

Yuri Kalambet[a]*, Yuri Kozmin[b], Andrey Samokhin[c]

[a] Ampersand Ltd., Kurchatov sq.2 bld. 2, Moscow 123182, Russian Federation
e-mail: kalambet@ampersand.ru

[b] Shemyakin–Ovchinnikov Institute of bioorganic chemistry RAS, Ulitsa Miklukho-Maklaya, 16/10, GSP-7, 117997, Moscow, Russian Federation

[c] Chemistry Department, Lomonosov Moscow State University, 1-3 Leninskiye Gory, GSP-1, 119991, Moscow, Russian Federation

* Corresponding author

Abstract

Theory of peak integration is revised for very narrow peaks. It is shown, that Trapezoidal rule area is efficient estimate of full peak area with extraordinary low error. Simpson's rule is less efficient in full area integration. Theoretical conclusions are illustrated by digital simulation and processing of experimental data. It was shown that for Gaussian peak Trapezoidal rule requires 0.62 points per standard deviation (2.5 points per peak width at baseline) to achieve integration error of only 0.1%, while Simpson's rule requires 1.8 times higher data rates. Asymmetric peaks require higher data rates as well. Reasons of poor behavior of Simpson's rule are discussed; averaged Simpson's rules are constructed, these rules coincide with those based on Euler-Maclaurin formula. Euler-Maclaurin rules can reduce error in the case of partial peak integration. Higher peak moments (average retention time, dispersion, skewness, etc.) also exhibit extraordinary low errors and can potentially be used for evaluation of peak shape.

Keywords: narrow chromatographic peak; data sampling error; Simpson's rule; Trapezoidal rule; integration; Euler-Maclaurin formula.

1. Introduction

Extension of data rate range in the direction of low data rates is an important capability that can be very useful in the case of fast chromatography, hyphenated techniques, chromatography–mass spectrometry data processing. These techniques sometimes produce data with quite little number of points per peak, and capability to extract useful information from these data can significantly help researchers.

In this article we focused on the theory of data processing in the case of very low data rates, typically considered as unacceptable due to insufficient number of points per peak [1]. Main attention is paid to peak area, which is the major metrological characteristic of the peak.

The task of evaluation of sufficient data rate in chromatographic analysis started to be discussed in early 1970's [2–9] after appearance of computer data processing in chromatography. Authors paid great attention to influence of noise level, in most studies proper determination of peak height, width and asymmetry factor was required. Approaches to the problem of area integration were usually based on:

1) theoretical conclusions made using Fourier transform and information theory: according to [6] 0.9 pts/σ is needed to achieve <0.1% of integration error;
2) digital modeling experiments: according to [7] 0.5 pts/σ is needed to achieve <1% of integration error, lower error limits were not achieved due to noise simulation. Seeley [9] studied dependence of peak parameters on duty cycle using rectangle rule and confirmed value of 0.5 pts/σ for small duty cycles;
3) theoretical conclusions made using textbook error formulas for integration rules: Trapezoidal integration rule requires 14 pts/σ to achieve <0.1% of integration error [8]; Simpson's rule for the same accuracy requires 1.7 pts/σ [2] or 2.5 pts/σ [8].

All data rate requirements correspond to Gaussian peak. In papers that used textbook error formulas [2,8] no digital modelling was made; estimates were based entirely on theoretical considerations. Requirement of such a large number of points per peak in third approach contradicts results of first two approaches [3,4,6], and our own estimates. We decided to revise argumentation used in textbooks, especially for the case of peak-like function. This is especially important, as some of papers insist on using Simpson's rule in chromatographic integration software [8].

2. Theory

2.1. Notations and formulas

Variable x stays for elution time, volume, distance or other independent retention parameter

Peak is a real analytic (in the math sense) function *f*(*x*) of one real variable *x*, such that the function itself and all its derivatives can be considered equal to zero outside of finite interval $x \in (a, b)$. Mathematical term “Analytic” means, that we can use Taylor series for analysis of the function.

Our definition of peak is too simple from the rigorous mathematical point of view, but perfectly fits the case of experimental data processing. It allows us to avoid using o(), O() and ∑, presenting just the ideas of proofs in the simplest way. Exact zero outside the interval (a,b) is not a must, but in practice of data processing all experimental data are produced by some analog-to-digital converters (ADCs), that output integers as a result of conversion. Signal in the region, where peak function |*f*(*x*)| (with subtracted baseline) becomes smaller, than ADC conversion unit (or baseline noise), can be considered as zero together with all statistically significant derivatives. Experimental data processing should be arranged so, that derivatives, which cannot be measured with sufficient accuracy, can be neglected.

Data grid (frame). We assume, that function *f*(*x*) is measured or calculated at discrete set of *N*+1 points {x_0, x_1, x_2, … x_N} with equidistant *sampling period* (*step*) *h*:

$$x_i = a + i \cdot h + \varepsilon;\ i = 0 \ldots N; h = \frac{b-a}{N}; -h/2 < \varepsilon < h/2$$

where ε is a *digitization grid shift*, which is a random real number uniformly distributed on the interval –h/2 < ε < h/2 (probability density equals to *P*(ε)=*1/h* inside this interval and *P*(ε)=0 outside it). The reason of introducing term ε is in the lack of advance knowledge about the position of the peak apex with respect to grid (e.g. due to variability of chromatographic retention time from run to run).

Exponentially Modified Gaussian (EMG) function [10–13]

$$f(x) = h_G \cdot e^{\frac{-(\mu_G - x)^2}{2\sigma_G^2}} \cdot \frac{\sigma_G}{\tau} \cdot \sqrt{\frac{\pi}{2}} \cdot \operatorname{erfcx}\left(\frac{1}{\sqrt{2}}\left(\frac{\mu_G - x}{\sigma_G} + \frac{\sigma_G}{\tau}\right)\right) \qquad 1$$

where h_G is height, μ_G – position of the apex, σ_G – standard deviation of unmodified Gaussian; τ – time constant of modifying exponent; erfcx() – scaled complementary error function [14]. Dispersion of EMG σ^2 can be calculated [10–13] as $\sigma^2 = \sigma_G^2 + \tau^2$

Euler-MacLaurin formula [15]

$$\int_{x_0}^{x_N} f(x)dx = h\left(\sum_{i=0}^{N} f(x_i) - \frac{f(x_0)+f(x_N)}{2}\right) + \sum_{k=1}^{m} h^{2k} \frac{B_{2k}}{(2k)!}\left(f^{(2k-1)}(x_0) - f^{(2k-1)}(x_N)\right) + R_m \quad (2)$$

where B_{2k} is a Bernoulli number (B_2=1/6; B_4=-1/30; …), 2*m* is maximal derivative order used in calculation and R_m is a remainder term, evaluating contribution of derivatives, higher than 2m. We present Euler-Maclaurin formula not exactly like in textbook, but solved for integral part.

Peak moments

Zeroth peak moment is peak area

$$M0 = \int_{-\infty}^{\infty} f(x)\,dx \approx h \cdot \sum_{i=0}^{N} f(x_i) \quad (3)$$

First moment is average retention time (unfortunately, it is rarely used in chromatography)

$$M1 = \frac{1}{M0}\int_{-\infty}^{\infty} x \cdot f(x)\,dx \approx \frac{1}{M0}\sum_{i=0}^{N}(x_i \cdot f(x_i)) \quad (4)$$

Second central moment is a dispersion of the peak (standard deviation σ is a square root of dispersion):

$$M2 = \sigma^2 = \frac{1}{M0}\int_{-\infty}^{\infty}(x - M1)^2 \cdot f(x)\,dx \approx \frac{1}{M0}\sum_{i=0}^{N}((x_i - M1)^2 \cdot f(x_i)) \quad (5)$$

Other moments usually are presented not only central, but also normalized to σ^n.

$$Mn = \frac{1}{M0 \cdot \sigma^n}\int_{-\infty}^{\infty}(x - M1)^n \cdot f(x)\,dx \approx \frac{1}{M0 \cdot \sigma^n}\sum_{i=0}^{N}((x_i - M1)^3 \cdot f(x_i)) \quad (6)$$

Instead of third moment, it is convenient to use estimate of parameter *τ* of EMG peak function with the same third moment [10–13]:

$$\tau = \sigma(M3/2)^{1/3} \quad (7)$$

For simplicity of presentation digital formulas of moments correspond to Midpoint Rectangle integration rule.

Phase shift φ=ε/h

Duty cycle – fraction of sampling period where function (signal) is averaged during measurement. Can be expressed as a fraction of one or in percent. All considerations of this paper correspond to instantaneous measurements with duty cycle of 0.0. For integrating ADC with duty cycle 1.0, peak area is defined by the sum of measurements by default and even one-point peak in the absence of noise will have exactly measured area. Duty cycle for the first dimension in the 2-D chromatography is usually close to 1.0, while duty cycles for fast scanning UV detectors or single quadrupole GC-MS are close to 0.0.

Data rate v=σ/h

2.2. Integration rules

The task of integration is to estimate area – definite integral of function $f(x)$ on (a, b).

All composite integration rules can be represented by a single formula:

$$A = h \cdot \sum_{i=0}^{N} w(x_i) f(x_i) \qquad 8$$

Where A is area, $w(x_i)$ – weight coefficients. Rectangle, Trapezoidal, Simpson's and other composite integration rules differ from each other by the set of coefficients $w(x_i)$.

2.2.1. Rectangle and Trapezoidal rules give identical peak areas

Let us set limits of summation in formula 8 so, that $f(x_1)= f(x_N)=0$ according to our definition of peak. For the Rectangle rule, all coefficients are ones $w()=\{1,1,1,\ldots,1,1,1\}$. Weight coefficients of Trapezoidal rule are $w()=\{1,2,2,\ldots,2,2,1\}/2$. As peak function is equal to zero on the boundaries of the integration interval, areas of the peak calculated using Rectangle and Trapezoidal rules are exactly equal.

In general case we should note that integration limits for composite Midpoint Rectangle rule are from position x_0-h/2 to position x_N+h/2, and integration limits for Trapezoidal rule are from x_0 to x_N. If we adjust integration limits for Midpoint Rectangle rule to those of Trapezoidal rule by throwing away half of first and last rectangles, weight coefficients of two rules will exactly coincide; their common weight formula is that of Trapezoidal rule.

2.2.2. Simpson's composite integration rule provides two estimates

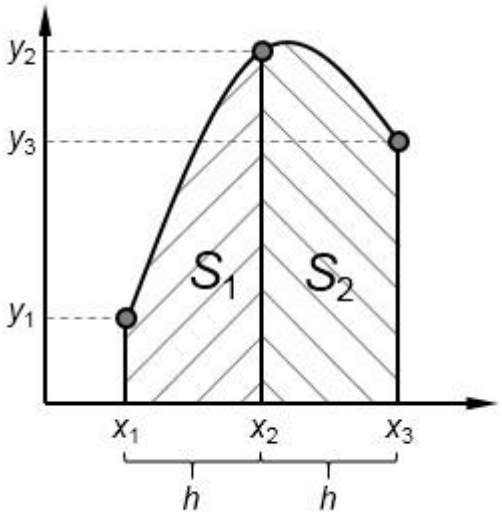
Simpson's 1/3 (further named just Simson's) rule utilizes parabola built for three successive points (Figure 1). It has coefficients $w()=\{1, 4, 1\}/3$ for three successive

nodes (elementary rule) and *w*()={1,4,2,4,…,4,2,4,1}/3 for odd number of nodes (composite rule) [15]. In the case of peak, we can get two different estimates of area, shifting first point of integration frame by one point. These two implementations of Simpson's rule give different results. In most cases, averaging two independent estimates improves accuracy of the answer. Averaging two implementations of Simpson's rule gives the following set of weight coefficients:

w() = ({0,1,4,2,…,4,2,4,1}/3 + {1,4,2,4,…,2,4,1,0}/3)/2 = {1,5,6,6,…,6,6,5,1}/6.

When responses at boundary points (2 points on each side) are equal to zero, integration result, provided by "averaged" Simpson's rule is exactly equal to that of Rectangle/Trapezoidal rule. Besides, this way of averaging is good to illustrate the problem but is not suitable in the software implementation. Below we discuss correct ways of composite Simpson's rule averaging.

2.2.3. Partial integration by Simpson's and Euler-Maclaurin integration rules



$$S_1 = \frac{1}{12}\cdot(5y_1+8y_2-y_3)\cdot h$$

$$S_2 = \frac{1}{12}\cdot(-y_1+8y_2+5y_3)\cdot h$$

$$S_1 + S_2 = \frac{1}{3}\cdot(y_1+4y_2+y_3)\cdot h$$

Figure 1. Elementary Simpson's figure split in two parts

It is possible to split 3-point elementary Simpson's region at the position of the middle point and calculate left and right half-areas separately as shown on Figure 1. Splitted areas have coefficients *w*()={5,8,-1}/12 for left part and *w*()={-1,8,5}/12 for right part. As expected, their sum equals {4, 16, 4} /12 = {1, 4, 1}/3. Left and right parts can be used to extend conventional Simpson's composite integration rule by one point at the ends of integration interval, allowing even number of points in the integration region and shifting frame start by one point.

Now we construct averaged Simpson's integration rules suitable for partial peak integration. Let's assume that we have 7-point region and we want to integrate it using Simpson's approach. No assumptions about zeros of function at the end of region are made. It is easy to apply one Simpson's frame to this region, as number of points is odd; another frame needs to be extended at the ends. Constructed frames are averaged, and we get following integration rules:

Table 1. Simpson's averaged Rule 1

Node index	-1	0	1	2	3	4	5	6	7	Divisor
Simpson 1		4	16	8	16	8	16	4		12
Simpson 2			4	16	8	16	4			12
Simpson 2 add-on	-1	8	5				5	8	-1	12
Average(Rule1)	-1	12	25	24	24	24	25	12	-1	24
Trapezoidal	0	12	24	24	24	24	24	12	0	24
Difference	-1	0	1	0	0	0	1	0	-1	24

Table 2. Simpson's averaged Rule 2

Node index	-1	0	1	2	3	4	5	6	7	Divisor
Simpson 1		4	16	8	16	8	16	4		12
Simpson 2			4	16	8	16	4			12
Simpson 2 add-on		5	8	-1		-1	8	5		12
Average(Rule2)		9	28	23	24	23	28	9		24
Trapezoidal		12	24	24	24	24	24	12		24
Difference		-3	4	-1	0	-1	4	-3		24

Rules 1 and 2 differ from each other by the way, how area of the region x_0 to x_1 is calculated. In Rule 1, we use right part S_2 of (x_{-1}, x_0, x_1) Simson's elementary figure; in

Rule 2 left part S_1 of (x_0, x_1, x_2) figure. Selection of the rule depends on availability of function estimate $f(x_{-1})$ at point x_{-1} outside the region being integrated. It is easy to see, that composite Rules 1, 2 differ from Trapezoidal rule at boundaries only. Note, that both rules give exact answer in the case of parabola integration. Integration region can be extended to any number of points, even or odd.

Rules 1, 2 are implementations of Euler-MacLaurin formula 2 derived in 1738. First term of this formula is just a Trapezoidal rule sum. If we evaluate first derivative at point x_0 as

$f'(x_0) \approx (f(x_1)-f(x_{-1}))/2h$,

calculate add-on from the k=1 member of the second sum in formula 2:

$$h^2 \frac{B_2}{2!} f'(x_0) \approx \frac{h^2}{12}\left(\frac{f(x_1)-f(x_{-1})}{2h}\right) = \frac{h}{24}\left(f(x_1) - f(x_{-1})\right)$$

and use this term to modify Trapezoidal rule, we will get exactly coefficients of Rule 1. If the first derivative is evaluated at point x_0 using values $f(x_0)$, $f(x_1)$ and $f(x_2)$ we will get coefficients of Rule 2. Therefore, Rules 1 and 2 can be considered as rules, derived from Euler-Maclaurin formula, with first derivative term included, and named Euler-Maclaurin rules. Note, that it is not a must, that Rules 1, 2 or their average will work better than trapezoidal rule for very narrow peaks, as in this case derivative, calculated by finite differences, is very inaccurate.

In the case of peak integration we can extend grid and peak boundaries by one or two points so, that area, calculated by any of Euler-Maclaurin rules, equals that of Rectangle/Trapezoidal rule. We can conclude that Rectangle, Trapezoidal and Euler-Maclaurin's (or averaged Simpson's) integration rules provide identical results, when the same discrete stand-alone peak is considered. Textbook statement, that error of composite Simpson's rule is much smaller, than error of Trapezoidal rule, for the case of peaks has to be reconsidered.

2.2.4. Trapezoidal rule is the most efficient integration rule for the whole peak

We can represent composite Simpson's rule coefficients

$w() = \{1,4,2,\ldots,4,2,4,1\}/3$
as $w() = \{1,2,2,\ldots,2,2,2,1\}/3 + \{0,2,0,\ldots,2,0,2,0\}/3$ 9

Note, that first summand of formula 9 is 2/3 of "traditional" Trapezoidal rule with step *h* and second summand is 1/3 of Rectangle rule with the step 2*h*.

We can evaluate error of this rule, assuming that error E of composite Trapezoidal rule depends on step size as $E \infty O(h^2)$. Then error of the rule with double step is $E_{2h} \approx 2^2 E = 4E$. In the case of formula 9 two summands and their errors are strongly correlated, so we should sum up errors, not dispersions. Total error E_s of the estimate can be evaluated as

$$E_s \approx 2E/3 + E_{2h}/3 \approx 2E/3 + 4E/3 \approx 2E$$

That is, Simpson's rule in integration of peaks is approximately twice less accurate than Trapezoidal rule provided Trapezoidal rule error has $O(h^2)$ dependence. Real correlation of errors of h- and 2h-step estimates may be slightly below one, as 2h-estimate has only one of two measurements of the h-estimate, but error drop coefficients may be significantly higher than 4 if error drops down faster than $O(h^2)$. From the theory we should expect exponentially-small function of 1/h [15], so error will drop much faster than evaluated. In this case, major part of error comes from the second summand of formula 9, and to achieve the same accuracy as Trapezoidal rule, Simpson's rule should have approximately twice higher data rate.

Increase of error in peak integration exists for all other composite integration rules with periodically repeating coefficients. E.g. Simpson's 3/8 rule can be represented as weighted superposition of original Trapezoidal rule with step h and Rectangle rules with step 3h and thus its error should be even higher than error of 1/3 rule. Trapezoidal rule with smallest step has the lowest possible estimate error, and therefore this estimate is efficient.

We do not consider here errors, caused by noise. Besides, difference of the two Simpson's estimates in the presence of highly correlated noise can be quite big, that means, that Simpson's rule is not robust with respect to noise.

2.2.5. Average Rectangle/Trapezoidal rule area estimate equals true peak area

We can approximate the function in the neighborhood of every node by Taylor series,

$$f(x_i + \tau) = f(x_i) + f'(x_i) \cdot \tau + \frac{1}{2} f''(x_i) \cdot \tau^2 + \cdots + \frac{1}{n!} f^{(n)}(x_i) \cdot \tau^n + \cdots; \ -h/2 < \tau < h/2,$$

For the purpose of evaluation of full definite integral, each term of Taylor series should be definitely integrated by τ in the (-h/2, h/2) neighborhood of every node. Integrals of terms with odd degrees of τ are equal to zero, as function $g(\tau) = \tau^{(2k+1)}$ is odd, $g(-\tau) = -g(\tau)$,

and integral of odd function on the symmetric (-h/2, h/2) neighborhood equals zero. After integration of Taylor series peak area *A* can be evaluated by the sum

$$A = A_0 + \sum_{k=1}^{\infty} \Delta A_{2k} \qquad 10$$

where

$$A_0 = h \cdot \sum_{i=0}^{N} f(x_i)$$

is area, calculated by Rectangle rule, and

$$\Delta A_{2k} = \frac{2}{(2k+1)!}\left(\frac{h}{2}\right)^{2k+1} \cdot \sum_{i=0}^{N} f^{(2k)}(x_i) = K_{2k} \cdot \sum_{i=0}^{N} f^{(2k)}(x_i) \qquad 11$$

is add-on term from (2*k*)-derivative.

For k=1 this term equals

$$\Delta A_2 = \frac{2}{6}\left(\frac{h}{2}\right)^3 \sum_{i=0}^{N} f''(x_i) = \frac{h^3}{24}\sum_{i=0}^{N} f''(x_i) \qquad 12$$

We can compare this term with the estimate of integration error used in [8].

$$E = I_{true} - I_{meas} = \left(\frac{W_b^3}{12n^2}\right)|f''(x)| \qquad 13$$

where W_b is the peak base width; $|f''(x)|$ is the absolute value of the second derivative of the function with respect to retention parameter *x*. If we replace $W_b = N \cdot h$ and evaluate $|f''(x)|$ as maximum of the second derivative, formula 13 can be written as

$$E = \frac{h^3}{12} \cdot N \cdot \max_{x \in (a,b)}(|f''(x)|) \qquad 14$$

We should not pay too much attention to coefficients; the major difference is between sum of second derivatives in formula 12 and N times maximum derivative in formula 14. As we will see later, expected value of the sum of derivatives equals zero, and N times maximum derivative in formula 14 is comparatively quite a big number. Even if we would use sum of modules of the second derivative instead of N times maximum, we would get significantly overestimated integration error. Theory of errors states, that there exists a number t=ξ between *a* and *b*, such that formula 13 is valid. In the case when region contains both convex and concave parts, $|f''(\xi)|$ can be quite small. This

means, that formula 13 in the interpretation formula 14 as used in [8] just is not suitable for error evaluation in the case of peaks.

Now we should calculate average of A_0 and every term ΔA_{2k} on epsilon

$$\overline{\mathrm{A_0}} = \int_{-h/2}^{h/2} A_0(\varepsilon)P(\varepsilon)d\varepsilon = h \cdot \int_{-h/2}^{h/2} \sum_{i=0}^{N} f(x_i) \cdot \frac{1}{h} d\varepsilon$$

$$= \sum_{i=0}^{N} \int_{-h/2}^{h/2} f(i \cdot h + \varepsilon)\, d\varepsilon = \int_{a-h/2}^{b+h/2} f(\varepsilon)\, d\varepsilon = A$$

is true area of the peak, and

$$\overline{\Delta \mathrm{A_{2k}}} = \int_{-h/2}^{h/2} \Delta A_{2k}(\varepsilon)P(\varepsilon)d\varepsilon = \frac{1}{h} \cdot K_{2k} \cdot \int_{-h/2}^{h/2} \sum_{i=0}^{N} f^{(2k)}(x_i)d\varepsilon = \frac{1}{h} \cdot K_{2k} \cdot \sum_{i=0}^{N} \int_{-h/2}^{h/2} f^{(2k)}(x_i)d\varepsilon$$

is average add-on term from (2*k*)-derivative. We should note, that sum of integrals of (2*k*)-derivatives in neighborhoods of all nodes equals full integral of this derivative. This sum

$$\sum_{i=0}^{N} \int_{-h/2}^{h/2} f^{(2k)}(x_i)d\varepsilon = \sum_{i=0}^{N} \int_{-h/2}^{h/2} f^{(2k)}(i \cdot h + \varepsilon)d\varepsilon = \int_{a-h/2}^{b+h/2} f^{(2k)}(\varepsilon)d\varepsilon$$

$$= f^{(2k-1)}(a - h/2) - f^{(2k-1)}(b + h/2) = 0$$

equals zero for the peak function for any positive integer *k*. Hence, average on ε value of ΔA_{2k} equals zero. After averaging, formula 10 has only one non-zero term A_0 left. Hence, area A_0 measured by Rectangle rule being averaged on ε equals true area of the peak. Dispersion of estimate by Midpoint Rectangle rule depends on the distribution of the sum $\sum \Delta A_{2k}(\varepsilon)$ on ε. This sum obviously depends on the function being integrated.

3. Digital modelling

Theoretical considerations were verified by digital simulations and processing of experimental data presented below. We estimated maximum integration error of Trapezoidal and Simpson's rules for three model peak shapes, including Gaussian, and one experimental peak shape.

3.1. Simulation

Model peak shape was described by Exponentially Modified Gaussian (EMG) function [10–13]

Three model peak shapes were considered: symmetric (Gaussian, $\tau=0$), slightly asymmetric (EMG-1, $\tau=\sigma_G$) and strongly asymmetric (EMG-3, $\tau=3\sigma_G$). Data rate varied from 0.4 to 3.2 measurements per σ.

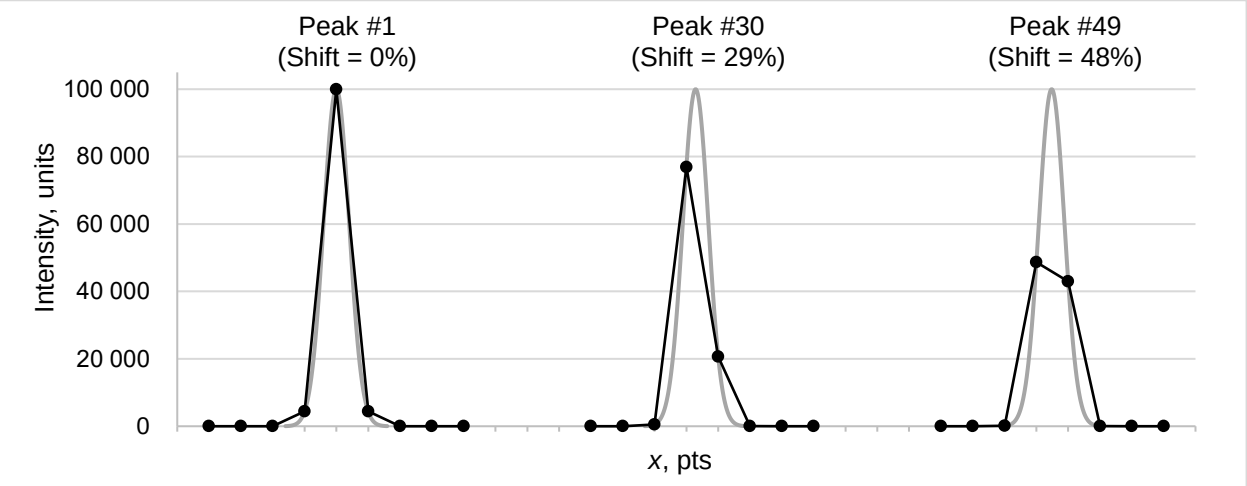


Figure 2. Examples of generated peaks. Abscissa – data point ordinal number; ordinate – modeled detector response. Parent peak is Gaussian, $\sigma_G=0.4$

All responses were given as integers. Heights of continuous peaks (h_G) were equal to 10^5 units, corresponding to full range of 17-bit ADC. For each particular peak shape and data rate, 100 different discrete peaks were generated. These peaks differed in phase shift of digitization grid nodes relative to apex of continuous peak (each successive discrete peak was shifted by 0.01 of the node-to-node distance). For convenience, each set of 100 peaks was located on the same chromatogram. Three discrete peaks are shown in Figure 2 as an example. All peaks were baseline-separated, intensity reached zero in the space between adjacent peaks. All modeled chromatograms were “noise-free”. Baseline level equals zero. To calculate area (zero-order moment M0) of chromatographic peak all non-zero responses and additionally two nearest zero responses on each side of the peak were integrated by the formula 8 with coefficients of Trapezoidal and Simpson’s rules. Model data and calculations for Gaussian peak are available in the file supplementary.xlsx

3.2. Experimental peak

Experimental peak of nitrate was extracted from calibration chromatogram (Figure 3) recorded by ICNet 2.0 software (Metrohm AG, Switzerland) [16].

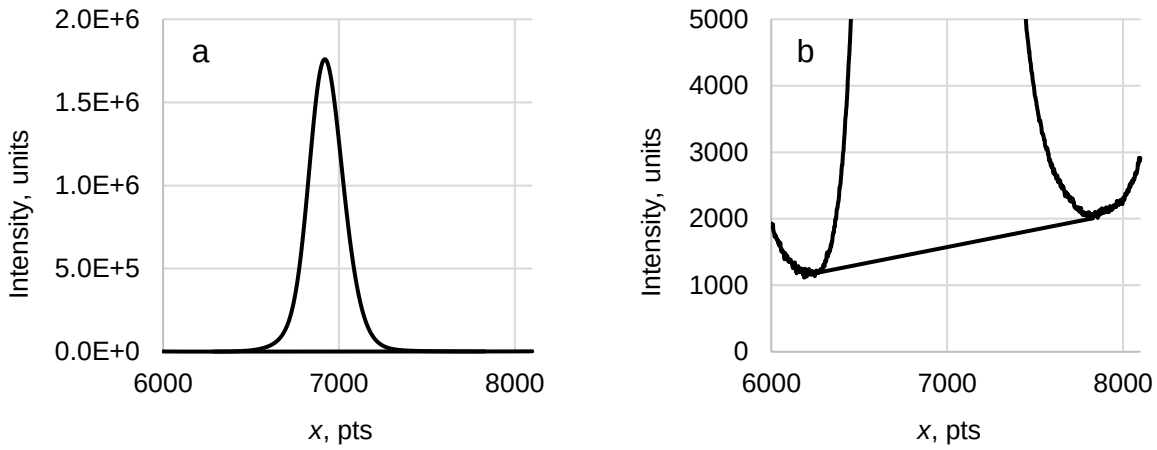


Figure 3. a) Fragment of calibration chromatogram with Nitrate peak. b) Baseline part of the peak amplified

Metrohm advanced IC chromatograph was used, consisting of 732 IC detector, 733 IC separation center, 709 IC pump. Column 4.6x75mm packed with Metrosep Anion Dual 2 6.1006.100, flow rate 1 ml/min, eluent 2 mmol/L $NaHCO_3$ / 1.3 mmol/L Na_2CO_3. Data rate was 10 pts/s.

Raw data were exported to text file and imported to Excel software. Peak consisted of 1543 points (Figure 3a), full width at half maximum 226 points, height $1.76*10^6$ conversion units; peak start, end and baseline were as they are shown on Figure 3b. Baseline was subtracted from initial raw data. Corrected peak was integrated by Trapezoidal rule and its moments were calculated. Second central moment was used to calculate peak standard deviation (σ), which was found to be 112.1 points. This σ was later used as σ_{ini} of the peak. Zeroth moment M0 (Area) was used as "true" area of the peak. Then for J=60 to J=230 step 10 every J'th point was picked from the corrected peak to form a new partial peak, and this partial peak was integrated having in mind J times higher time constant. J frames were used, differing by the first point of digitization grid, providing J partial peaks. Digitization error was calculated as maximum absolute difference between "true" area and area calculated from partial peaks. σ_{true} in points for the partial peak was calculated as initial sigma divided by J: $\sigma_{true} = \sigma_{ini}/J$.

4. Results and discussions

In accordance with literature data, we considered two thresholds of peak integration error: 1% and 0.1%. For target level of uncertainty 0.1%, minimum length of integration interval of Gaussian function equals ±4.0σ or ±3.3σ, depending on baseline position (Figure 4). Our calculations used definition from Figure 4b.

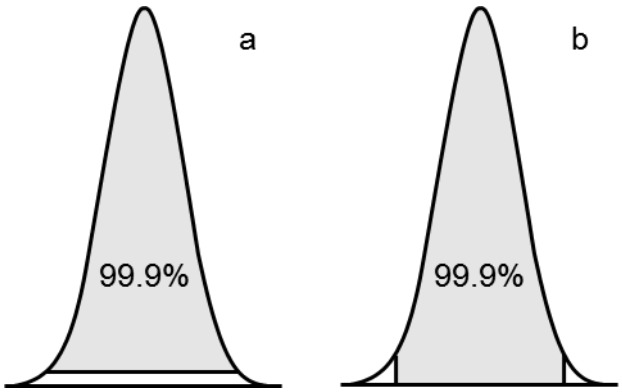


Figure 4. Examples of peak area distortion depending on integration model. a) baseline is drawn between peak points; b) "true" baseline is used

As expected, mean error of peak area (calculated over 100 discrete peaks using Rectangle/Trapezoidal rule) was equal to zero; example peaks from the series are presented in Figure 2. As can be seen in Figure 5, dependence of peak area error on shift of digitization grid resembles sine wave. Similar curves for Simpson's 1/3 and 3/8 rules together with raw data are available in supplementary.xlsx file.

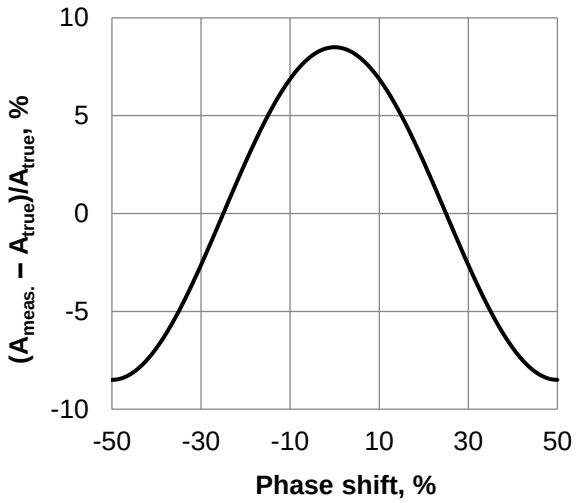


Figure 5. Dependence of peak area error, calculated by Trapezoidal rule on phase shift. Peaks had Gaussian shape, σ_G=0.4. Shape of generated peaks is shown in Figure 2

Figure 6 illustrates dependence of maximum peak area error on data rate. We are using peak σ as a measure of peak broadness, not Full Width at Half Maximum (FWHM) used in many papers. The reason for that is in the fact, that neither height, nor FWHM can be evaluated for narrow peaks like those from Figure 2, while σ can be evaluated from the available set of data points. Besides, data rate is normalized to true standard deviation of original peak, and not to standard deviation, calculated from the particular generated peak.

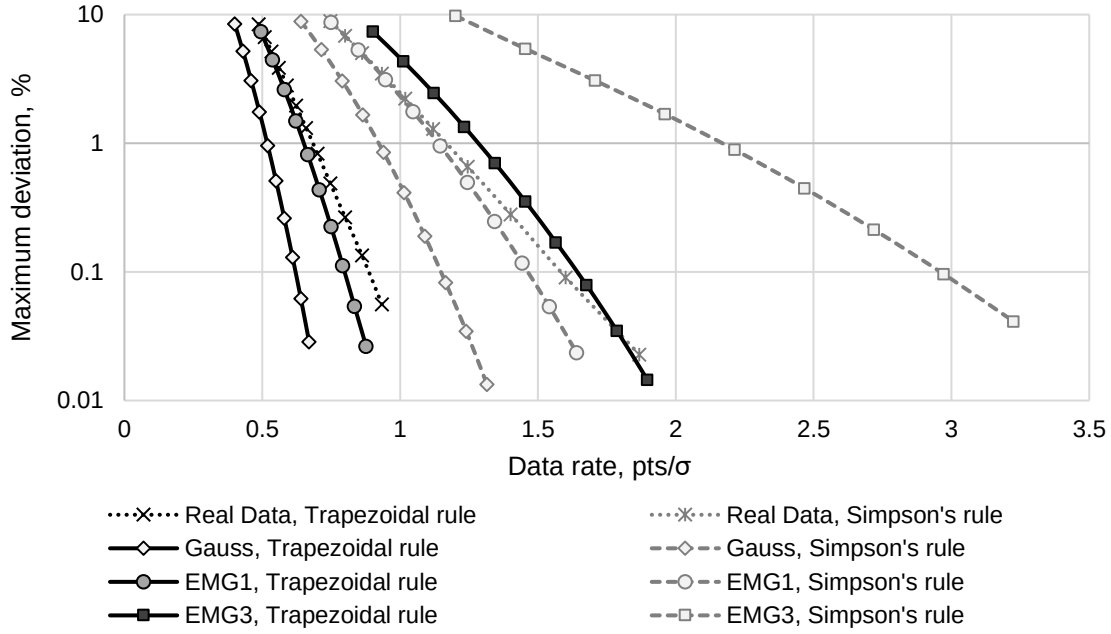


Figure 6. Dependence of maximum error of peak area (calculated by Trapezoidal and Simpson's rules) on data acquisition rate. Peaks of different shape were considered: Gaussian, EMG-1, EMG-3 and experimental (real) data.

It's easy to see from Figure 6, that maximum error drops down very fast as data rate increases. This conclusion is true for generated and experimental peaks. Drop down rate is even slightly faster, then exponential, as curves are slightly convex; this supports the conclusion made in Theory, that error of peak integration by Trapezoidal rule should be described as an exponentially-small function of 1/h rather as $O(h^2)$. Our results for peaks of different shapes are summarized in Table 3.

Table 3. Minimum data rates to measure peak area. All values in pts/σ_{true}

	Trapezoidal rule		Simpson's rule	
Threshold	1%	0.1%	1%	0.1%
Gaussian	0.52	0.62	0.92	1.15
EMG-1	0.65	0.80	1.14	1.46
EMG-3	1.28	1.64	2.17	2.96

Experimental peak integration by rectangle rule give errors slightly higher than those of EMG-1 (Figure 6). Attempt to approximate this peak with EMG as described in [13] gives τ/σ_G ratio of 0.64 and residual error of deconvolution (average RMS error divided by average response) of 2.3%. This residual error should be considered high compared

to target uncertainty level of 1% or 0.1%, so we should not expect that integration errors correspond to evaluated τ/σ_G value; peak shape significantly deviates from EMG.

It's clear from Figure 6, that Trapezoidal rule performs for peaks much better, than Simpson's rule, confirming our theoretical conclusions. Maximum error of Simpson's rule was calculated for both frames of the rule without averaging. Error of Simpson's rule can be compared to error of trapezoidal rule using curves for integration of Gaussian from Figure 6: lowest point for Trapezoidal rule is (0.67pts/σ, 0.029%) and second highest point for Simpson's rule error is (0.72pts/σ, 5.4%). Error of Simpson's rule at 0.67 pts/σ is higher, than at 0.72 pts/σ, therefore error of Simpson's rule at 0.67pts/σ is at least 5.4/0.029=186 times higher, than error of Trapezoidal rule. As can be expected from Theory, Simpson's 3/8 rule perform for peaks even worse, than 1/3 Rule (see supplementary.xlsx).

Proper estimate of Gaussian peak area by Simpson's rule requires 1.15 pts/σ for 0.1% precision or 1.8 times higher data rate, than for Trapezoidal rule. This fully supports the conclusion made in Theory that Simpson's rule requires approximately twice-higher data rate to achieve the same precision as Trapezoidal rule. Higher value (1.7 pts/σ) proposed by Kishimoto and Musha [2] overestimates required data rate, as it is obtained from textbook formulas and does not account exceptional integration properties of all rules applied to peaks. Goedert and Guiochon used criteria (Table 1 in [5]), required for reconstruction of peak height rather than area, and thus their estimates of required data rates are also much higher than ours.

The reason, why Rectangle/Trapezoidal rule performs for peaks so well is in the nature of the peak function: as full integral of every peak derivative is zero, errors, caused by this derivative in one peak region are (almost) compensated at other regions. Simpson's rule in the case of peak also has efficiency higher than "normal" $O(h^5)$, but it's less efficient than Trapezoidal rule (Figure 6).

Simpson's rule seems to be never good in processing of peaks, as it gives two different results, depending on index of integration start point (see Theory). After averaging, these two values will give exactly the same peak area estimate, as Rectangle and Trapezoidal rules. Without averaging, original Simpson's rule gives error that is more than hundred times bigger, than that made by Rectangle/Trapezoidal rules as seen from our simulation. Therefore, in the case of peaks, superior accuracy of Simpson's rule compared to Trapezoidal rule seems to be just a myth, created by textbooks by

manipulation of error estimates. Formula 14 exaggerates error of full peak area evaluated by composite Trapezoidal rule by several orders of magnitude.

Exceptional efficiency of Trapezoidal rule in the case of integration of periodic functions with integration interval equal to function period was noted long ago [15]. This efficiency is caused by the fact that all derivatives at the ends of integrated interval are equal, and all derivative terms disappear from Euler-Maclaurin formula 2. Similar situation is valid for peaks. According to our definition, peak function has zero derivatives on the edges. In this case, all derivative terms of Euler-Maclaurin formula are zero and peak area exactly equals Trapezoidal rule area. Our simulation experiments have shown that this statement is not true at very low data rates. Apparent disagreement indicates that our peak definition is not perfect, and some of high-order derivatives may become significant at very low data rates. This problem does not have easy solution, as at very low data rates all schemes of derivative calculation based on finite differences stop working. If we have access to formulas of the function and its derivatives, the problem can be resolved using these formulas.

We tested different integration rules applied to partial integration of Gaussian peak (see supplementary file Rules_Erf.xlsx). In the case, when we assume equation for calculation of derivative to be known, results of integration using “true” Euler-Maclaurin rule with calculated derivatives have extraordinarily high precision. Euler-Maclaurin formulas with derivatives, calculated by finite differences, are much less precise, as well as composite Simpson’s rule.

In the case of experimental data processing, in particular for peak integration, we have to work in the data rate region, where required part of the integral belongs to Trapezoidal rule sum. It is possible to understand, whether data rate is high enough, by digital modelling which we summarized in Figure 6. Required data rate depends on peak function (peak shape). In the case of peaks, dependence of Trapezoidal rule contribution on data rate is monotone (Figure 6), error drops down exponentially on 1/h, and thus at threshold and higher data rates (sufficient data rate) it’s safe to use Trapezoidal rule.

In the case of peak, digitized with sufficient data rate, the function itself and all derivatives are negligibly small on boundaries, thus formula 2 gives Trapezoidal rule area. For partial peak integration, estimates of function and its derivatives should be made by approximation of function and only statistically significant derivative terms

should be accounted for. Euler-Maclaurin rules as presented in Tables 1, 2 are suitable for partial integration of math functions at sufficient data rates only.

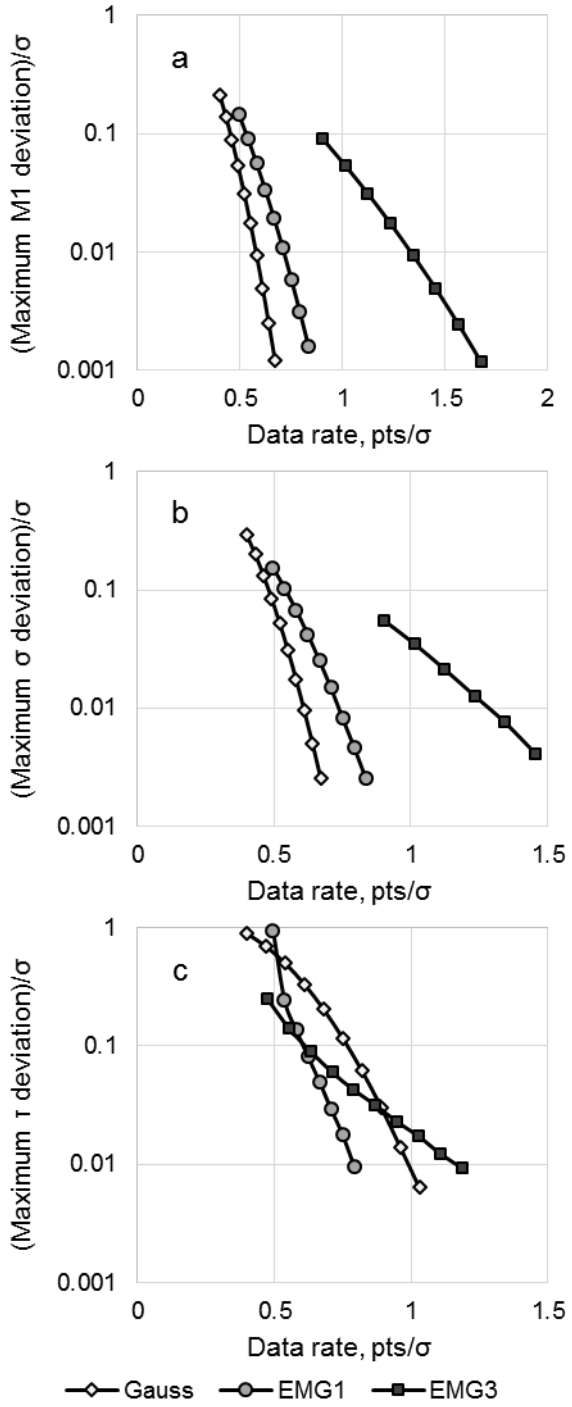


Figure 7. Dependences of maximum error of M1 (a), σ (b) and τ (c) calculated by Trapezoidal rule on data acquisition rate. Error values are normalized to true sigma

As can be seen from the formulas 3-7, each peak moment is the integral of the product of peak function $f(x)$ and some other function $g(x)$. As defined in 2.1, $f(x)$ equals zero outside the interval (a, b). Therefore, product of $f(x)$ and $g(x)$ equals zero outside this interval as well, and all functions located under integral sign (i.e. $f(x)\cdot g(x)$) can be

considered as peak functions. In reality for some type of functions, typically treated as peaks, such as Cauchy function, moments cannot be treated as peaks. Besides, in the case of exponentially decreasing functions, such as Gaussian or EMG, they can. As moments are evaluated by integration of peak-like function, all considerations, concerning area integration, are valid, and estimate of moments by Trapezoidal rules is very efficient and its average equals true value of the moment.

Dependences of maximum errors of Trapezoidal rule estimates of average retention ($M1$), standard deviation ($\sigma=M2^{1/2}$) and relaxation time ($\tau=\sigma\cdot(M3/2)^{1/3}$) on data acquisition rate are shown in Figure 7. For Gaussian peak we had to set $h_G=10^9$ to construct τ curve, as at $h_G=10^5$ response rounding error became significant at data rates higher than 0.75 pts/σ; so it is not clear, how much one can rely on τ estimate for narrow peaks in practice. As we are studying data rate (discretization) errors, we checked, that errors, produced by rounding of real number to integer (simulating Analog-to-Digital Conversion) for all peak moments are negligible. Rounding errors may become significant for smaller peak heights (lower ADC resolution). One should also have in mind, that higher-order moments are more sensitive to noise. Corresponding minimum data rates are summarized in Table 4. Target levels of uncertainty for calculated $M1$, σ, τ in Table 2 are selected having in mind, that these parameters are used in various validation criteria, rather than quantitative analysis. All values are normalized to true σ, and absolute errors are quite small: data rate, corresponding to the first moment $M1$ error of $0.1\sigma_{true}$, is 0.45 pts/σ; so, maximal absolute error equals $\Delta M1 = 0.1 \cdot 0.45 = 0.045$ pts, hence retention $M1$ measured in points has at least one valid decimal digit at this and higher data rates. At data rate above 0.58 pts/σ two significant decimal digits of $M1$ are guaranteed.

Table 4. Minimum data rates to measure M1, σ, τ by Rectangle or Trapezoidal rule. All values in pts/σ_{true}.

	Average retention $M1$		Standard deviation σ		Asymmetry τ	
Threshold	$0.1\cdot\sigma_{true}$	$0.01\cdot\sigma_{true}$	$0.1\cdot\sigma_{true}$	$0.01\cdot\sigma_{true}$	$0.1\cdot\sigma_{true}$	$0.01\cdot\sigma_{true}$
Gaussian	0.45	0.58	0.48	0.61	0.77	0.99
EMG-1	0.53	0.71	0.54	0.74	0.60	0.79
EMG-3	0.88	1.33	0.75	1.29	0.61	1.17

Every analysis is an implementation of the measurement process. Within our model, this process has random parameter ε – distance from (true) peak apex to nearest point of digitization grid. Every chromatographic peak has only one random value of this parameter implemented, adding random integration error, corresponding to this ε, to the result. Maximum on ε deviation is the worst-case error, caused by discretization. It should be noted, that in this work we evaluated maximum errors of peak parameters (*M*0, *M*1, σ and τ), rather than respective standard deviations. Standard deviation is lower than maximum error, so our results can be considered as an upper limit of the error; besides, probability distribution of the error value in this case has quite specific bimodal shape (Figure 8) and most probable error value equals maximum on ε deviation. Such probability distribution makes maximum error and standard deviation very close to each other.

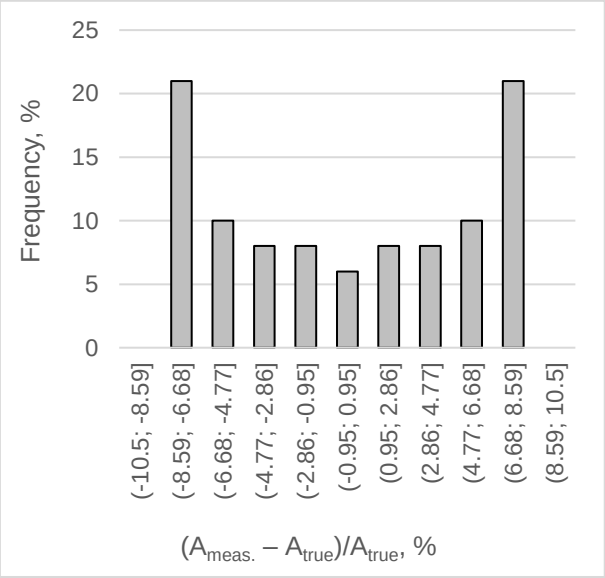


Figure 8. Histogram, reflecting probabilities of different deviations of area from the mean (see Figure 5)

Table 3 should be used in conjunction with error propagation law. Values in this table are just a basis for decision, whether data rate error is significant or not compared with errors from other sources, such as noise. If we assume that EMG-3 corresponds to the worst acceptable in chromatography peak shape, then peak can be considered to be too narrow, when it has 1.64 pts/σ (Table 1) or less. If data acquisition rate is higher than 1.64 pts/σ, integration error of area caused by signal discretization is not significant even for the most accurate analytical procedures. In the case of reasonably symmetric peaks much lower data rates are acceptable (Table 1).

In this paper main attention is given to measurement of chromatographic peak area and higher moments. Other parameters, such as retention time, height, full width at half-height, asymmetry factor were not considered. Estimates of data rates for these parameters made formerly [1–7] seem reasonable. We can propose that in the case of very low number of points per peak these parameters are replaced by other more robust ones (retention time – first moment or median time; width – standard deviation or distance between 25% and 75% quantiles; asymmetry – estimate of relaxation time τ or quantile spans ratio). Height has no definite robust analog; it can be evaluated as a function of area/σ ratio, or it can be calculated via peak reconstruction. Probably, for narrow peaks height should not be used at all.

If peak shape is known in advance, peak reconstruction may be done instead of integration by Rectangle/Trapezoidal rule. Theoretically, reconstruction of peak shape requires three data points for Gaussian and four points for EMG. If more points are available, respective equation system is overdetermined and can be solved by least square minimization [17]. Reconstruction by Fourier Transform Analysis seems also very efficient [18]. All proper reconstruction methods would provide very precise set of numbers for all shifted peaks, generated in our examples; differences between numbers of the set would be caused only by limitations of computer math used for modelling. Probably, peak reconstruction is the future of data processing. Besides, reconstruction methods in the case when model is wrong or in the presence of noise may be less robust than Trapezoidal rule integration. Initial guess of parameters for peak reconstruction procedures may also effectively utilize information about peak moments.

This paper is most valuable to researchers, working with narrow chromatographic peaks in GC/MS, 2D chromatography, fast chromatography. We are not sure, that we are the first, who found why Rectangle and Trapezoidal rules give extraordinarily efficient estimate of the whole peak area: mathematics of this kind can be met in all fields, and time span to investigate is nearly 300 years. We are not able to perform complete literature search in this case. Nevertheless, we are confident that considerations presented above are not available in the popular textbooks or chromatographic literature and it is important to draw the attention of the chromatographic community to the issue.

5. Conclusions

- Narrow peaks are suitable for quantitative analysis by Trapezoidal rule integration at very low data rates; minimal data rates for 0.1% accuracy vary from 0.62 to 1.64 pts/σ depending on peak shape;
- Composite Trapezoidal rule provides efficient estimate of the full peak area with average (on relative position of peak and digitization grid) value equal to true area of the peak;
- Composite Simpson's rule requires 1.8 times higher data rate to achieve efficiency, provided by Trapezoidal rule;
- Composite Simpson's rule averaged for two frames provides estimate equivalent to the rule based on Euler-Maclaurin formula;
- Peak moments for exponentially decaying peaks, such as Gaussian or Exponentially Modified Gaussian functions are also peak functions, and can be efficiently evaluated using Trapezoidal rule;
- Errors of peak area estimate by all rules drop down exponentially on data rate.

Funding: The work was partially supported by Russian Foundation for Basic Research, research project No. 16-33-60169 mol_a_dk

6. Supplementary files

Data set https://data.mendeley.com/datasets/xs7b5ckzsj/draft?a=b120e72f-e523-4ac2-af16-604a1e963a0b

Excel file Supplementary.xlsx

Demonstration of performance of different integration rules (Trapezoidal, Simpson's 1/3 and Simpsons's 3/8) applied to full integration of Gaussian. Graphs similar to Figure 5 for all three rules are constructed. Error of all rules drops down abnormally fast as data rate increases. Trapezoidal rule performs best and Simpson's 3/8 worst in full accordance with paragraph 2.2.4 of Theory section.

Data set https://data.mendeley.com/datasets/pvsyjfv5th/draft?a=d3fe6f80-80c4-4230-a56e-69e681e497cf

Excel file Rules_Erf.xlsx

This spreadsheet demonstrates errors of partial integration of Gaussian peak using different rules. Peak section is digitized using 3,5,7,9,11 or 15 points. Integration is performed using Trapezoidal, Simpson's, and rules based on Euler-Maclaurin formula. Column with the name Euler-Maclaurin contains rule with properly calculated first derivative term. Columns Average rule 1, Average rule 2 and (Av.rule 1 + Av.rule2)/2 correspond to Euler-Maclaurin formula with 1st derivative calculated using finite differences as described in Theory section of the paper. True Euler-Maclaurin rule is always the best. Trapezoidal rule is preferable at very low data rates (less than 0.7 pts/sigma) and full area integration. Errors of Simpson's 1/3 and Average rule1, 2 and (Av.1+Av.2)/2 rules are comparable, as all of them account for the second derivative term of Taylor series and use finite differences.

7. References

[1] M.F. Wahab, P.K. Dasgupta, A.F. Kadjo, D.W. Armstrong, Sampling frequency, response times and embedded signal filtration in fast, high efficiency liquid chromatography: A tutorial, Anal. Chim. Acta. 907 (2016) 31–44. doi:10.1016/j.aca.2015.11.043.

[2] K. Kishimoto, S. Musha, Investigation of Sampling Interval for GC Data Reduction, J. Chromatogr. Sci. 9 (1971) 608–611.

[3] S.N. Chesler, S.P. Cram, Effect of peak sensing and random noise on the precision and accuracy of statistical moment analyses from digital chromatographic data, Anal. Chem. 43 (1971) 1922–1933. doi:10.1021/ac60308a005.

[4] P.C. Kelly, G. Horlick, Practical Considerations for Digitizing Analog Signals, Anal. Chem. 45 (1973) 518–527.

[5] M. Goedert, G. Guiochon, Sources of error in chromatographic analysis. Effect of sampling parameters on the accuracy of numerical measurements, Chromatographia. 6 (1973) 76–83.

[6] P.J.H. Scheeren, Z. Klous, H.C. Smit, D.A. Doornbos, A software package for the orthogonal polynomial approximation of analytical signals, including a simulation program for chromatograms and spectra, Anal. Chim. Acta. 171 (1985) 45–60.

[7] D.T. Rossi, A Simplified Method for Evaluating Sampling Error in

Chromatographic Data Acquisition, 26 (1988) 101–105.

[8] N. Dyson, Peak distortion, data sampling errors and the integrator in the measurement of very narrow chromatographic peaks, J. Chromatogr. A. 842 (1999) 321–340. doi:10.1016/S0021-9673(99)00299-X.

[9] J. V. Seeley, Theoretical study of incomplete sampling of the first dimension in comprehensive two-dimensional chromatography, J. Chromatogr. A. 962 (2002) 21–27. doi:10.1016/S0021-9673(02)00461-2.

[10] E. Grushka, Characterization of exponentially modified Gaussian peaks in chromatography, Anal. Chem. 44 (1972) 1733–1738. doi:10.1021/ac60319a011.

[11] R. Delley, Series for the exponentially modified Gaussian peak shape, Anal. Chem. 57 (1985) 388–388. doi:10.1021/ac00279a094.

[12] J. Foley; J. Dorsey, A review of the exponentially modified gaussian (EMG) function: evaluation and subsequent calculation of universal data, J. Chromatogr. Sci. 22 (1984) 40–46.

[13] Y.A. Kalambet, Y.P. Kozmin, K.V. Mikhailova, I.Y. Nagaev, P.N. Tikhonov, Reconstruction of chromatographic peaks using the exponentially modified Gaussian function, J. Chemom. 25 (2011). doi:10.1002/cem.1343.

[14] W.J. Cody, Rational chebyshev approximations for the error function, Math. Comput. 23 (1969) 631–637. doi:10.1090/S0025-5718-1969-0247736-4.

[15] DLMF, F.W.J. Olver, A.B. Olde Daalhuis, D.W. Lozier, B.I. Schneider, R.F. Boisvert, C.W. Clark, B.R. Miller, B. V. Saunders, (eds.), DLMF: NIST Digital Library of Mathematical Functions, Release 1.0.13. (2016). http://dlmf.nist.gov/.

[16] IC Net ©1999-2011., (1999) Accessed April 02, 2018. https://www.metrohm.com/en/support-and-service/software-center/ic-net/.

[17] N. Hagen, E.L. Dereniak, Gaussian profile estimation in two dimensions., Appl. Opt. 47 (2008) 6842–6851. doi:10.1364/AO.47.006842.

[18] A. Felinger, A. Kilár, B. Boros, The myth of data acquisition rate, Anal. Chim. Acta. 854 (2015) 178–182. doi:10.1016/j.aca.2014.11.014.

8. Figure Captions

Figure 1. Elementary Simpson's figure split in two parts

Figure 2. Examples of generated peaks. Abscissa – data point ordinal number; ordinate – modeled detector response. Parent peak is Gaussian, σ_G=0.4

Figure 3. a) Fragment of calibration chromatogram with Nitrate peak. b) Baseline part of the peak amplified

Figure 4. Examples of peak area distortion depending on integration model. a) baseline is drawn between peak points; b) “true” baseline is used

Figure 5. Dependence of peak area error, calculated by Trapezoidal rule on phase shift. Peaks had Gaussian shape, σ_G=0.4. Shape of generated peaks is shown in Figure 2

Figure 6. Dependence of maximum error of peak area (calculated by Trapezoidal and Simpson’s rules) on data acquisition rate. Peaks of different shape were considered: Gaussian, EMG-1, EMG-3 and experimental (real) data.

Figure 7. Dependences of maximum error of M1 (a), σ (b) and τ (c) calculated by Trapezoidal rule on data acquisition rate. Error values are normalized to true sigma

Figure 8. Histogram, reflecting probabilities of different deviations of area from the mean (see Figure 5)