# Statistical Quantification of LEIS Data

**Low Energy Ion Scattering (LEIS) has two key characteristics: It is extremely surface sensitive (to the outermost atomic layer) and it is quantitative. To quantify LEIS data it is necessary to determine the peak area as good as possible. Subsequently, the peak areas need to be converted to surface concentrations or surface fractions. Both steps require data fitting. To quantify LEIS data it is necessary to use the proper statistical procedures.**[1]

## Fitting and Least Squares

This note will start with simple examples and gradually work towards more complicated, and more LEIS related, situations. Though the complexity increases, the main principles remain. Figure 1 shows the simplest example of "fitting" of data. The diameter of a ball needs to be determined. It was measured six times. The bar graph shows the obtained measurement results.

It is important to realize what is going on here.

- We do not know the diameter of the ball.
- We try to guess, as good as possible, what the diameter of the ball is.

In many cases, the average (or mean) of the measurement values is the best guess. This average is indicated with the blue dashed line in the graph. It has a value of 87 mm.

There is a good statistical reason why the average value is the best guess in many cases. The reasoning will be explained here, since it is the basis for the more complex fitting.

The reasoning is as follows:

- We take a model value $x^*$ and assume it is the correct diameter of the ball. Now, we calculate the error in our model, $\varepsilon$:

$$\varepsilon = x^* - x \qquad (1)$$

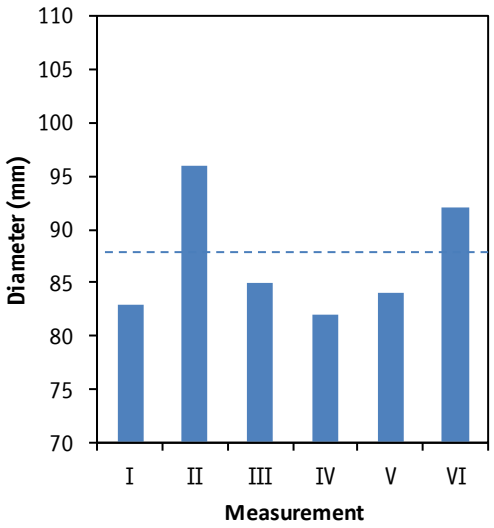


*Figure 1: Results of the measurement of the diameter of a ball*

- We make the model error as small as possible. The reason for this is that a larger discrepancy between measurement value and the "true" value would make it "stranger" to obtain that measurement result. For a single measurement, this leads to a trivial result: $x^* = x$. If the ball's diameter is measured once as 83 mm then the best guess is that it is 83 mm.
- For more measurements, the errors need to be added to obtain the total error. Errors are added by adding the variances (the squares of the error). So, the total variance or square sum, $S^2$, is:

$$S^2 = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (x^* - x_i)^2 \qquad (2)$$

where $x_i$ are the measured diameters and $n$ is the number of measurements.

- The best value for $x^*$ is the one for which the total variance is smallest, so $S^2$ needs to be minimized. (This is where the term 'Least squares' comes from: The 'squares' are the variances, the squares of the errors. 'Least' is used because the total of the squares needs to be minimized.). This minimization is done by differentiation with respect to the model value $x^*$:

[1] Please refer to this technical note as: "Statistical Quantification of LEIS Data", Rik ter Veen, Tascon Technical Note, 2016

$$\frac{\partial S^2}{\partial x^*} = \frac{\partial \sum_{i=1}^{n}(x_i - x^*)^2}{\partial x^*} = 0 \qquad (3)$$

$$\frac{\partial S^2}{\partial x^*} = -2\sum_{i=1}^{n}(x_i - x^*) = 0$$

$$\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x^* = nx^*$$

$$x^* = \frac{\sum_{i=1}^{n} x_i}{n} \qquad (4)$$

This is the average value, which explains why it (usually) is the best guess: because the total model error will be smallest.

Figure 2 shows the exact same data set as in figure 1, but this time error bars have been added to the measurements. It turns out that the measurements were performed by two different students. The most assertive of the two got to use the lab's Vernier caliper, the appropriate tool to measure a diameter. He provided the odd numbered measurements. The other had to use the ruler that he had on his desk. He brought the even numbered measurements. The dashed and dotted lines show the two averages of the two sets of three measurements, at 84 and 90 mm, respectively.

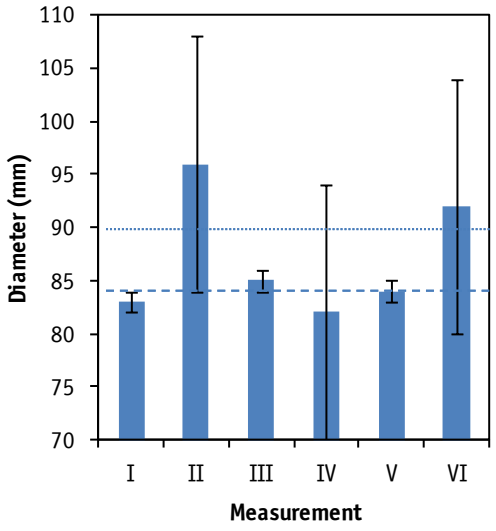


*Figure 2: Results of the measurement of the diameter of a ball*

This new knowledge about the accuracy of the measurements will change our best guess of the ball's diameter. The ball is still the same but, from now on, we will think that its diameter is closer to 84 mm where before (see figure 1) we thought it was 87 mm.

The reason is that we don't think it is "strange" that the values measured with a ruler are further from the real value than the values measured with the caliper. How "strange" the difference between a measured value and a true value is, depends on what difference we would expect. For an inaccurate measurement we expect a larger difference than for an accurate measurement. Mathematically, this means that the model error needs to be normalized with respect to the expected measurement error, $\sigma$:

$$\frac{\varepsilon}{\sigma} = \frac{x - x^*}{\sigma} \qquad (5)$$

For a set of measurement values with different measurement errors, it can be shown that

$$x^* = \frac{\sum_{i=1}^{n} \frac{x_i}{\sigma_i^2}}{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}} \qquad (6)$$

This is a weighted average, where the weighting factors are the inverses of the expected variances for the individual measurements. An accurate measurement, with a small $\sigma$, will have a larger weighting factor than an inaccurate measurement.

## Linear regression

The previous example was simple. It involved only one variable, the diameter of one ball.

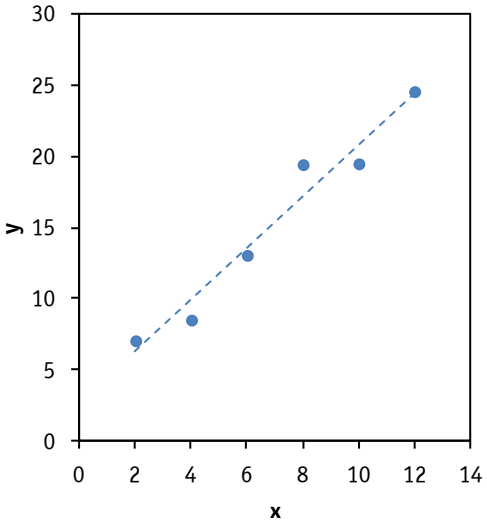


*Figure 3: A set of (x,y) data points with a regression line*

Figure 3 shows a two dimensional data set with a regression line, calculated in the way that most scientists will use: The data points were entered in a program, e.g. Excel, and it calculated a line. This line is calculated using linear regression.
Before, it was assumed that the ball had a given, but unknown, diameter. Similarly, linear regression assumes that there is a given, but unknown, linear relation between $x$ and $y$:

$$y^* = a + bx \qquad (7)$$

The '*' superscript for the $y$ variable denotes that this is the 'model $y$ ', in contrast to the 'measured $y$': simply '$y$'. The model error for data point $i$, $\varepsilon_i$, will be given by equation 8.

$$\varepsilon_i = y_i^* - y_i = a + bx_i - y_i \qquad (8)$$

Again, we have more data points, and, therefore, more errors to consider. We need to square them first and then add them up to obtain the square sum, $S^2$:

$$S^2 = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (a + bx_i - y_i)^2 \qquad (9)$$

This time, we have to optimize two parameters: $a$ and $b$. Again, we take the derivatives and set them to 0.

$$\frac{\partial S^2}{\partial a} = 0, \quad \frac{\partial S^2}{\partial b} = 0 \qquad (10)$$

With a little work this leads to:

$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$
$$a = \frac{\sum y - b\sum x}{n} \qquad (11)$$

But, again, we need to keep in mind that this least squares method assumes that all measurements are equally accurate. Figure 4 shows, not entirely surprisingly, the same data as figure 3, but now with added error bars. Given the difference in measurement accuracy, it is clear that the trend line calculated by the software is not the best guess for a linear relation anymore.

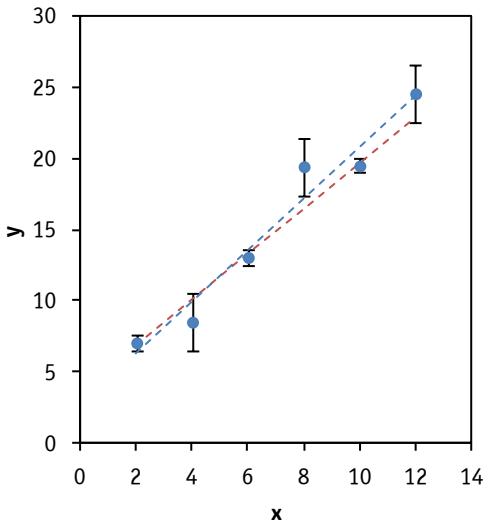


*Figure 4: A set of (x,y) data points with error bars and regression lines, obtained by standard least squares linear regression (--) and by error weighted linear regression (--)*

Again, the solution is to compare the model error to the measurement error. So, rather than to make $\varepsilon$ the "driving force" for the fit, we can make $\varepsilon/\sigma_y$ the driving force, where $\sigma_y$ is the measurement error. We define $\chi^2$ as:

$$\chi^2 = \sum_{i=1}^{n} \frac{\varepsilon_i^2}{\sigma_{y_i}^2} = \sum_{i=1}^{n} \frac{(a + bx_i - y_i)^2}{\sigma_{y_i}^2} \qquad (12)$$

It is equivalent to $S^2$, but it has been normalized on the expected measurement errors. Minimizing $\chi^2$ leads to:

$$b = \frac{\sum \frac{1}{\sigma_y^2} \sum \frac{xy}{\sigma_y^2} - \sum \frac{x}{\sigma_y^2} \sum \frac{y}{\sigma_y^2}}{\sum \frac{1}{\sigma_y^2} \sum \frac{x^2}{\sigma_y^2} - \left(\sum \frac{x}{\sigma_y^2}\right)^2}$$
$$a = \frac{\sum \frac{y}{\sigma_y^2} - b\sum \frac{x}{\sigma_y^2}}{\sum \frac{1}{\sigma_y^2}} \qquad (13)$$

This results in the red line in figure 4. The similarities and the differences between equation 11 and 13 are clear.

**Fitting LEIS data**

There are two instances in the quantification of LEIS data where fitting is used: in the determination of the peak areas and in the conversion from peak areas to concentrations. In both cases, the measurement error is not constant. However, the measurement error is known reasonably well. The LEIS signal is based on count rates and is, therefore, governed by Poisson statistics. This means that the expected variance is equal to the number of counts.
The number of counts, $N_c$, is given by:

$$N_c = yq$$

with $q$ the measurement charge in nC and $y$ the LEIS signal in counts/nC. Hence,

$$y = \frac{N_c}{q}$$

$$\sigma_y = \frac{\sigma_{N_c}}{q}$$

$N_c$ is governed by Poisson statistics. Therefore,

$$\sigma_{N_c} = \sqrt{N_c} = \sqrt{yq}$$

$$\sigma_y = \frac{\sqrt{yq}}{q} = \sqrt{\frac{y}{q}}$$

$$\sigma_y^2 = \frac{y}{q} \qquad \textbf{(14)}$$

**Determining the peak areas**

A model typically is the sum of some functions (e.g. a Gaussian and an error function) with parameters.

$$y^*(x) = aG_{\mu,\sigma}(x) + bE_{\mu,\sigma}(x) \qquad \textbf{(15)}$$

Where $y^*(x)$ represents the model signal in cts/nC, $G_{\mu,\sigma}(x)$ denotes a Gaussian with parameters $\mu$ and $\sigma$ and $E_{\mu,\sigma}(x)$ an error function with different parameters $\mu$ and $\sigma$. The Gaussian and error function contribute to the signal with factors $a$ and $b$.
So, $\chi^2$, taking the measurement error into account, will be given by equation 16.

$$\chi^2 = \sum \frac{\left(aG_{\mu,\sigma}(x) + bE_{\mu,\sigma}(x) - y\right)^2}{\sigma_y^2}$$

$$\chi^2 = \sum \frac{q\left(aG_{\mu,\sigma}(x) + bE_{\mu,\sigma}(x) - y\right)^2}{y} \qquad \textbf{(16)}$$

Since $\chi^2$ will be minimized the exact value of $q$ is irrelevant, as long as it is constant.

Equation 16 is not linear and there is no analytical solution. However, $\chi^2$ can be minimized numerically, e.g. using the Solver procedure in Excel. This means that starting values are needed. It normally isn't difficult to estimate starting values by looking at the spectra.
There is, however, a small practical problem. For some energies the measured intensity may be equal to 0. To prevent a 'division by zero' error a small, constant offset can be added to the denominator. Its value will typically be the equivalent of 1 count in the spectrum. When the data are expressed in counts/nC then this 1 count equivalent equals $1/q$.

$$\chi^2 = q\sum \frac{\left(aG_{\mu,\sigma}(x) + bE_{\mu,\sigma}(x) - y\right)^2}{y + 1/q} \qquad \textbf{(17)}$$

**Conversion of peak areas to concentrations**

In LEIS, peak areas can be converted to fractions by fitting the data to equation 18.

$$1 = \sum_{j=1}^{N} x_j = \sum_{j=1}^{N} a_j Y_j \qquad \textbf{(18)}$$

The sum of the fractions for the individual elements, $x_j$, must be equal to 1. Given that the fraction is proportional to the intensity $Y_j$, by a factor of $a_j$, it follows that $a_j$ can be fitted, as long as a few conditions are met:

- The number of spectra is much larger than the number of elements.
- The concentrations of elements are independent.
- There is enough variation in the signals.

(The capital $N$ is chosen to signify the number of elements, with $j$ as a running variable. The lower case $n$ is used for the number of samples with $i$ as a running variable.)
Again, it is important to take the measurement error into account.
Figure 5 shows a set of constructed concentration profiles for 5 elements, A-E. (This analysis is valid for any series of samples. Profiles were chosen since they can be visualized easily.)

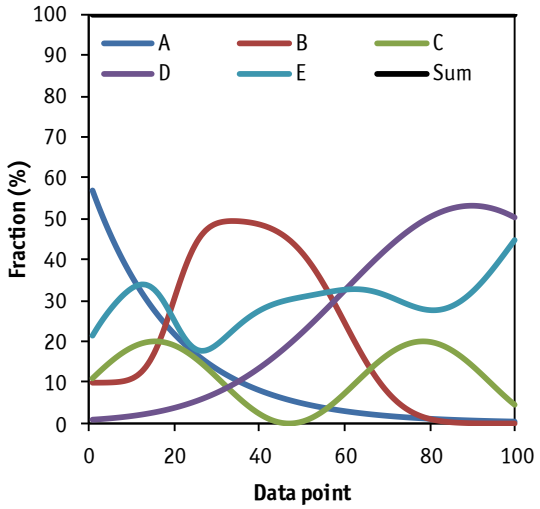


*Figure 5: Constructed concentration profiles for 5 elements*

Each profile was multiplied by a sensitivity factor to yield ideal, noise free, LEIS intensity profiles. Then Poisson noise was added to the intensity profiles. The resulting, noisy profiles are shown in figure 6.

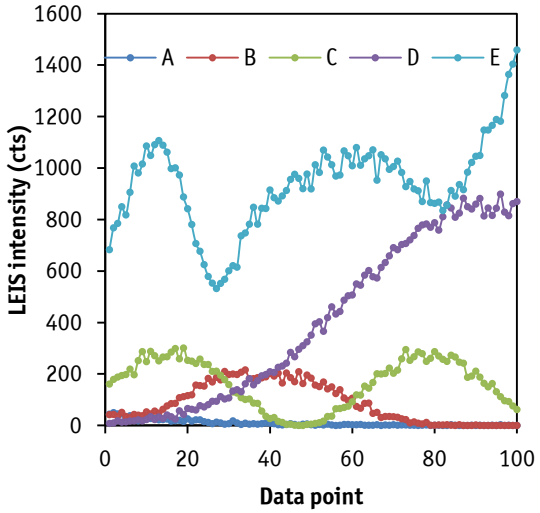


*Figure 6: LEIS intensity profiles with Poisson noise from the concentration profiles of figure 5*

Figure 6 depicts the data that the LEIS user would typically try to evaluate: Noisy intensity profiles that need to be converted to the concentration profiles of figure 5.
The standard way is to use linear regression analysis of the data. Such a procedure is available in Excel: the 'x data' are the LEIS intensities, the 'y data' are set equal to 1, the offset is set to 0 and the parameters are fitted. This yields the parameters $a_j$ in equation 18.
Figure 7 shows the results of the standard regression analysis from the signals in figure 6.

They are compared to the original concentration profiles from figure 5.

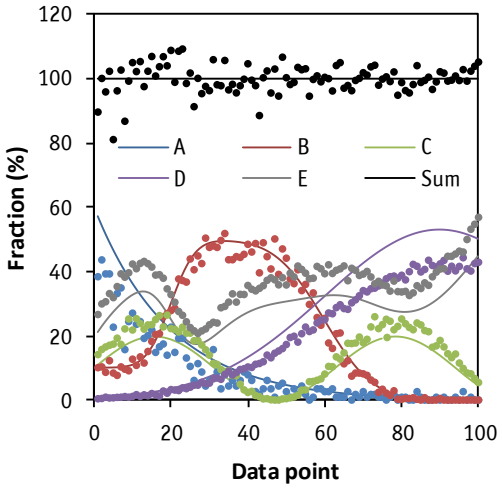


*Figure 7: Concentration profiles, derived from the LEIS intensity profiles from figure 6 by standard regression analysis compared to the concentration profiles of figure 5*

It is clear that the standard regression analysis does a fair job: It gets the trends correct and the sum of the concentrations is close to 100%. However, when it comes to an absolute quantification, clearly the concentration for A is underestimated whereas for E it is overestimated. The sensitivity factors are not entirely correct.
This deviation is caused by the fact that the measurement accuracy was not taken into account. It is possible to adapt equation 18 to fit the parameters with the correct measurement accuracy.

$$1 = \sum_{j=1}^{N} x_j = \sum_{j=1}^{N} a_j Y_j \qquad \textbf{(18)}$$

$$\varepsilon_{model} = 1 - \sum_{j=1}^{N} a_j Y_j$$

$$\varepsilon_{model}^2 = \left(1 - \sum_{j=1}^{N} a_j Y_j\right)^2$$

This describes the error in the model for $1 - \sum_{j=1}^{N} a_j Y_j$. The equivalent measurement error in the measurement of $1 - \sum_{j=1}^{N} a_j Y_j$ is given by:

$$\sigma^2 = \sum_{j=1}^{N} a_j^2 \sigma_{Y_j}^2$$

And for data with Poisson noise, $\sigma_{Y_j}^2 = Y_j/q_j$ , so, for constant $q$, we obtain

$$\sigma^2 = \frac{1}{q}\sum_{j=1}^{N} a_j^2 Y_j$$

We need to minimize $\chi^2$:

$$\chi^2 = \sum_{i=1}^{n} \frac{\varepsilon_{model,i}^2}{\sigma_i^2}$$

$$\chi^2 = q\sum_{i=1}^{n} \frac{\left(1-\sum_{j=1}^{N} a_j Y_{ij}\right)^2}{\sum_{j=1}^{N} a_j^2 Y_{ij}} \qquad \textbf{(19)}$$

Equation 19 can be minimized numerically. This procedure requires starting values for the fit parameters, $a_j$. Obviously, the results of the standard regression analysis are good starting values.

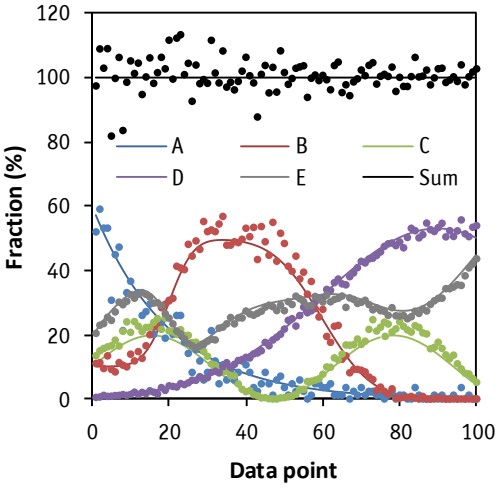


*Figure 8: Concentration profiles, derived from the LEIS intensity profiles from figure 6 by least squares minimization, with the measurement error included compared to the concentration profiles of figure 5*

Figure 8 shows the result of the data evaluation of the concentration profiles of figure 5, when the measurement accuracy is taken into account.
It is clear that this fits the original concentration profiles better than the linear regression analysis from figure 7. The fraction for element A is not underestimated and the fraction for element E is not overestimated.
Table 1 shows a comparison of the parameters $a$ that were used to generate the data and the results of the two different data analysis methods. The error in the two methods is shown in figure 9.

*Table 1: The determined parameters compared to the parameters that were used to generate the data*

| | Parameter $a$ | | |
|---|---|---|---|
| Element | Input | Standard | With error |
| A | 0.8 | 1.15 | 0.85 |
| B | 4 | 4.17 | 3.81 |
| C | 14 | 11.39 | 11.97 |
| D | 16 | 20.32 | 16.15 |
| E | 32 | 25.69 | 33.48 |

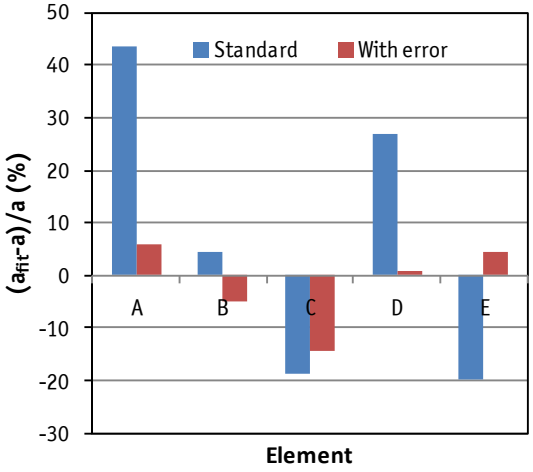


*Figure 9: Comparison of the relative error for the determination of the fit parameters ($\frac{a_{fit}-a_{input}}{a_{input}}$) for the two methods*

It may be of interest to note that, though the individual concentrations are clearly estimated more accurately in figure 8 than in figure 7, the sum of the concentrations in figure 8 shows more scatter than the sum of the concentrations in figure 7. This may seem counterintuitive, but upon reflection it is easy to explain: In figure 7, the individual concentrations were determined by minimizing the error in the sum of the concentrations. Then, it is obvious that in that case the scatter in the sum should be smaller than for any other method, even if that other method is more accurate at determining the individual concentrations.

**Practical issues**

When quantifying LEIS data, it is perfectly possible to compare data from different types of measurements. If a sample surface has been analyzed with 3 keV $^{4}He^{+}$, 7 keV $^{4}He^{+}$ and 5 keV $^{20}Ne^{+}$, then the data from all measurements can be used. After all, the surfaces are still the same and the sum of the surface fractions still equals 1. However, there are some things to consider.

The different measurements often use different measurement charges. This means that $q$ is no longer constant and needs to be added in the calculation.

Usually $He^{+}$ data have larger errors than $Ne^{+}$ data. Apart from the fact that the counts in a $He^{+}$ spectrum are usually lower, $He^{+}$ spectra also have higher backgrounds. So, usually it is better to rely on $Ne^{+}$ data than on $He^{+}$ data.

**Co-linearity and lack of variation**

When data are co-linear or when there is a lack of variation, these methods cannot be used as is. Suppose one wants to analyze a depth profile of NaCl layer on gold. The Na and Cl concentration profiles will be co-linear. This means that the Na concentration cannot be determined independently from the Cl concentration. If one attempts to do this, the mathematics will search for trends in the data that aren't real, but simply statistical fluctuations, to separate the two signals.

The way around this is to use the available physical/chemical knowledge. It is a fact that the fraction of Na must be equal to the fraction of Cl. In the analysis, Na and Cl are lumped together by adding the two signals and giving them one multiplier.

Something similar is going on when one of the signals doesn't vary. This can, e.g., be the case in a series of samples that have been treated with oxygen atoms. In that case, the oxygen fraction doesn't vary, because the surface is saturated with it. Therefore, we shouldn't use the oxygen signal in the fit. (The mathematics would look for patterns that aren't real.)

Instead of establishing the absolute surface fractions of all elements, the sum of the other elements will be set to 1. This will yield the contributions of the other elements as "fractions of their oxides".

**Summarizing**

Statistical methods are very powerful to evaluate LEIS data. The best results are obtained when measurement errors are taken into account and when one is aware of the pitfalls, such as co-linearity and lack of variation.