Talanta 116 (2013) 45–49

Contents lists available at ScienceDirect

Talanta

journal homepage: www.elsevier.com/locate/talanta

Short communication

The connection between inverse and classical calibration

Emili Besalú *

Department of Chemistry and Institute of Computational Chemistry and Catalysis, Universitat de Girona, Av. Montilivi s/n, 17071 Girona, Spain

ARTICLE INFO

Article history:
Received 18 May 2012
Received in revised form
17 April 2013
Accepted 24 April 2013
Available online 2 May 2013

Keywords:
Classical calibration
Inverse calibration
Linear equation
Regression towards the mean

ABSTRACT

Within the context of the simple classical linear calibration procedure (regression of y on x), here it is shown how a distinction between the distributions of the observed dependent variable (y_{obs}) and the calculated (fitted) one (y_{calc}) leads to the following counterintuitive approach: in order to get the independent x values with lesser systematic deviations, do not identify as direct inputs in the classical calibration equation the new y observed ones (experimentally acquired), but instead the transformed ones by means of a regression towards the mean effect correction. It is shown how the conjunction of both steps, i.e., first the transformation of observed values and then the ulterior use in the classical calibration equation, corresponds to an operation totally equivalent to the direct implementation of the inverse calibration equation (regression of x on y in a single step). The reasoning given here explains in a simple manner why the inverse calibration numerically performs usually better for predictions of interpolated x values. Results are accompanied with the analysis of both theoretical and experimental data.

1. Introduction

In classical calibration, the dependent variable y_{obs} (for instance, an experimentally acquired absorbance or signal) is regressed on the independent one x_{obs} (usually a concentration). It is assumed that the dependent variable has a statistical uncertainty and that the errors are normally distributed around the true values with constant variance, whereas x_{obs} is error-free or, at least, it has a substantial lesser error than the dependent variable. The regression of n (x_{obs},y_{obs}) data points provides with the classical linear model useful to obtain the expected value for y_{obs} variable from the knowledge of x_{obs}

$$E(y_{obs}|x_{obs}) = y_{calc} = a + b\,x_{obs} \tag{1}$$

being a and b the estimators of the true unknown parameters α and β which express the underlying relationship among the x and y variables (e.g., Beer–Lambert law). The model is also characterized by the determination coefficient r^2 indicating the fraction of data variance explained by Eq. (1). It has not to be understood that previous Eq. (1) stands for the relationship $y_{obs}=a+bx_{obs}$. The observed y_{obs} variable and the calculated one, y_{calc}, are not interchangeable entities. This article concerns about this distinction.

On the other side, the linear inverse calibration consists into regress the x_{obs} variable on y_{obs}, usually disregarding which one bears errors or which one is a random independent error-free variable, thus violating basic assumptions underlying least squares regression [1]. The inverse model obtained with the same previous set of n available points is of the form

$$E(x_{obs}|y_{obs}) = x_{calc} = a' + b'y_{obs} \tag{2}$$

Models (1) and (2) are not homologous nor interchangeable, but are related in such a way that the knowledge of a model will automatically lead to the other equation (see next paragraph). The two equations share the same determination coefficient and, as much as the determination coefficient tends to 1, both equations will tend to be the same, minimizing the difference between the classical and inverse procedures [2].

The numerical relationship between models (1) and (2) obtained from the same set of n points is very simple [3]: the product of the slopes b and b' is r^2, and both equations are satisfied by the data points center of mass ($\bar{x}$,$\bar{y}$). Note that, for the n points data set, $\overline{y_{calc}} = \overline{y_{obs}} \equiv \bar{y}$, as it is also well known. Consequently, once the regression parameters a, b and r are known from the classical fitting (1), the ones appearing in (2) can be obtained either by the inverse regression procedure or directly by the following relationships:

$$\begin{cases} a' = \frac{(1-r^2)\bar{y}-a}{b} \\ b' = \frac{r^2}{b} \end{cases} \tag{3}$$

Conversely, a similar transformation exists in order to infer the equation parameters of (1) from the previous knowledge of those of (2). We will call here *conjugated* a couple of regression equations (the classical and the inverse) arising from the same set of fitted points and linked by the aforementioned rule (3).

* Tel.: +34 972 41 8875; fax: +34 972 41 8356.
E-mail address: emili.besalu@udg.edu

0039-9140/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.talanta.2013.04.054

Within the classical approach to calibration, Eq. (1) serves to estimate the x of an unknown sample from its measured y_{obs} value, the one provided by the experiment. Along this process, the acquired variable y_{obs} is implicitly identified as being the calculated variable y_{calc}. That's a mistake because doing that it is *incorrectly* assumed that, once one instance of y_{obs} is known, the corresponding expected value for y_{calc} is y_{obs} itself. Below it will be shown how the expression $E(y_{calc}|y_{obs})=y_{obs}$ does *not* hold in general. In spite of this, the two-step process usually followed to obtain x from the knowledge of y_{obs} is described by the codification

$$\begin{cases} 1. \quad y_{obs} \rightarrow y_{calc} \\ 2.\ (y_{calc}-a)/b \rightarrow x_{obs} \end{cases} \tag{4}$$

where in step (4.1) the erroneous identification $E(y_{calc}|y_{obs})=y_{obs}$ is being assumed, and then x is derived isolating from (1) by means of step (4.2).

The process of inverse calibration, i.e., to rely on Eq. (2), in some circumstances performs better for predictions of x of new known experimental y values [4–9]. This constitutes a statistical trend but not a general situation because the performance depends on the kind of data we are manipulating, on the distribution of x variable, on the number of replicates, on the distance of the predicted points from the mean x and y values and, among others, on the followed criteria to define superiority of a method. Here, by better we mean that the sum of quadratic errors (mean squared errors, MSE) between obtained x values and the actual ones is lesser. In many practical situations, inverse approach performs better for interpolations of x values within the explored interval [4]. This feature was noticed early and discussed in the context of Montecarlo simulations [5,6] and the interest continued during the time until now (see for instance references in [8]). Investigations have been carried on dealing with many aspects of this situation, as for instance the asymptotic performance of the inverse calibration for $n\rightarrow\infty$ [1,9] or showing that the inverse procedure if favored respect to the classical one even if the number of data point samples is small [8]. Shukla [7] shows that, despite one method is not in all the cases superior to the other, the inverse approach gives lesser MSE when a single value of y is available and the inferred value of x lies near the mean of the population sample (e.g. for interpolations). Modern approaches take into account a balance of practical factors and favor the inverse approach respect to the classical one [8].

This paper exposes an elemental relationship which links inverse and classical procedures and that explains in a simple manner why many times inverse calibration performs better than the classical one when inferring interpolated values of x. The reasoning deals with the distinction which has to be made between calculated y values (y_{calc}) and observed ones (y_{obs}). The key idea is that the knowledge of a particular value of the variable y_{obs} has not to be directly identified with the same numerical value of the variable y_{calc}. This seems counterintuitive because, according to (4), the usual classical approach consists into get y_{obs} values from the population sample and identify them with y_{calc} ones by plugging the former numerical value directly into Eq. (1) to get x.

2. Results

The exposition below is based on a general theorem described within the multiple linear regression framework [10,11] and is related to the regression towards the mean effect [12]. The theorem applies for any arbitrary and finite data set disregarding particular statistical distributions of x and y variables and their correlation values. The exposition will be initially illustrated using an arbitrary artificial data set. In order to improve the visualization of some features, the artificial set shows a small correlation value despite this is not the case in general calibration or analytical purposes. Afterwards, data coming from real experiments will be also analyzed.

2.1. Artificial data set

An artificial large set of $n=3000$ (x_{obs},y_{obs}) sample points has been generated exhibiting a low correlation value ($r^2=0.80$). Without loose of generality, the low correlation value and other data parameters have been chosen in order to made graphically evident the explored features. The variable x is normally distributed ($\mu={}^9/_2$, $\sigma={}^7/_6$) and the y data points obey to the inner (true) linear relationship $y=1+2x$ (i.e., $\alpha=1$ and $\beta=2$), but modified by a Gaussian noise error function having mean zero and a fixed arbitrary variance (1.3535) which leads to the aforementioned coefficient of determination. Once the classical linear regression equation is calculated, the fitting line $y_{calc}=0.995+2.00x_{obs}$ is obtained, which stands for Eq. (1). Fig. 1 shows the representation of the calculated values (y_{calc}) against the observed ones (y_{obs}). The diagonal solid line corresponds to the equation bisector $y=x$. It has *not* to be assumed that the cloud of points depicted in Fig. 1 is symmetrically distributed along the bisector equation. Sometimes that's the erroneous underlying idea which leads to apply the procedure (4). The assumption that the numerical values of the variables y_{obs} and y_{calc} are interchangeable is incorrect.

Fig. 1 reveals the evidence of a regression towards the mean effect [12] due to an artifact of the linear model construction (in fact, due to the asymmetry in x and y variables treatment). As it can be seen, the point cloud is not symmetrically distributed along the bisector line but rotated around the point cloud center of mass. As a consequence, for a given experimental value, for instance the observation $y_{obs}=15$ depicted in Figure 1, the expected corresponding y_{calc} one *is not* the same number in general (unless for the particular cases for which $r^2=1$ or y_{obs} coincides with the sample mean value $\bar{y}$). In Fig. 1 it is also shown the distribution of y_{calc} values attached to the particular observed result $y_{obs}=15$. It is graphically revealed how from the knowledge that the value $y_{obs}=15$ has been experimentally acquired, the corresponding mean (expected) value of variable y_{calc} is *distinct* than 15. This artifact is overlooked many times. The expected y_{calc} value is depicted at the center of the distribution curve and lies in the diagonal dashed line. Fig. 1 reveals that, within the context of

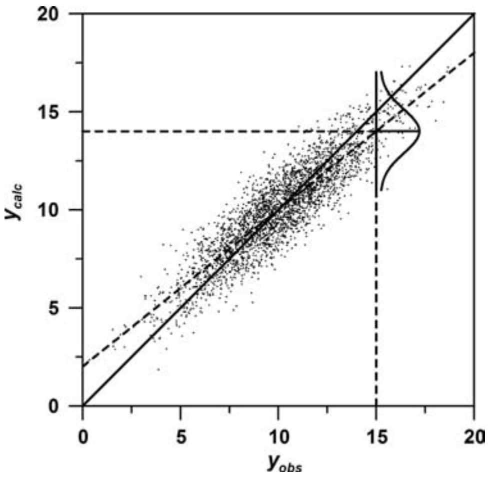

Fig. 1. Representation of calculated y values (y_{calc}) by means of the classical regression Eq. (1) against the original observed (experimental) ones (y_{obs}). Solid diagonal line is the quadrant bisector whereas the dashed diagonal one corresponds to the linear regression of calculated values on the observed ones. The other dashed lines and the Gaussian show how the expected y_{calc} value attached to an observed one ($y_{obs}=15$) are not coincident.

linear regression, given an observed value y_{obs}, in general it does not coincide with the corresponding expected value of y_{calc} variable in Eq. (1)

$$E(y_{calc}|y_{obs}) \neq y_{obs} \text{ (unless } y_{obs} = \overline{y} \text{ or } r^2 = 1) \quad (5)$$

If y_{obs} is greater than the population mean value $\overline{y}$, the corresponding expected y_{calc} value is lesser than y_{obs}. For instance, for the particular example which is explored here (see Fig. 1)

$$E(y_{calc}|y_{obs} = 15 > \overline{y}) = 14 < y_{obs},$$

according to inequality (5). Conversely, if y_{obs} is lesser than the sample y mean, the corresponding expected y_{calc} value is greater than y_{obs}. As said, the numerical identification $y_{obs}=y_{calc}$ can be stated only for the particular cases for which $r^2=1$ (perfect correlation) or if y_{obs} coincides with $\overline{y}$.

2.2. Experimental data set

A collection of experimental calibration sets was analyzed in order to show up to which extend the inverse regression performs better than the classical one. In order to reveal a statistical trend, a quite big amount of data is needed. As a data source it was considered the database contents of a sequential inductively coupled plasma atomic emission spectrometer (Sequential ICP-AES, Liberty RL, Varian). A total of 500 data files were manually retrieved and prepared containing 1738 stored linear calibration regressions dealing with a variety of elements found in food, water, soils, plants, and construction materials samples, mainly. None of the stored calibration sets where prepared by the author. All data were generated prior to the design of this study and by independent researchers.

Afterwards, the files containing semiquantitative data or less than four regression points were discarded. The linear calibration equations attached to a regression coefficient lesser than 0.99 were also discarded. Due to these constrains, a total of 274 sets of regression points were compiled (see Table 1). Along the calibration equations, the elements concentration (independent variable) ranged from 0 up to 30 ppm and the number of points ranged from 4 up to 14 (see Table 2). Despite of previous data filtering, in all the regressions the range of concentrations did not follow a uniform or Gaussian distribution. Due to the usual preparation procedure followed by the machine users (usually they did successive dilutions coming from an original sample), a non-uniform geometric progression pattern of concentrations was found. Heteroscedasticity was also present in the data. Hence, the experimental data of this section, after filtering, do not constitute an idealized repository, but a practical set found in real analysis. Afterwards, a test was performed to measure up to which extend a regression method (classical or inverse) is superior to the other. The test was applied to every one of the 274 calibration sets. It consisted into perform a leave-one-out prediction for the interpolated $n-2$ points in each regression (initial and final concentration points were discarded for prediction in each set). In other words, given a calibration set of n points, for each interpolated regression point it was simulated to hide it, do the regressions with the remaining $n-1$ points (in direct and in inverse modes) and predict the concentration value (also in direct and in inverse modes) assuming that the hidden experimental x value is the true one. Then, the sum of mean squared concentration errors for each regression and for each mode was retrieved. In 107 cases the direct method gave less error than in the inverse one, and for the other 167 remaining cases the inverse method resulted to give lesser mean errors. Considering as null hypothesis a theoretical balance of 137 cases, the χ^2 test reveals that the inverse method is superior with a significance p value of 0.0003. Despite such a result, the mean error differences are small, of the order of 0.002 ppm favoring the inverse case.

Previous results are sensitive to the particular features present in the collected experimental data (presence of heteroscedasticity, number of calibration points, non-uniform x values distributions, value of the determination coefficient,…). In order to check the influence of the data structure, a simple computational experiment has been reproduced using a population space of 3000 pairs of data points with the same characteristics of the one considered in Section 2.1 (homoscedastic Gaussian distribution) but this time exhibiting a value of $r^2=0.98$ ($r=0.99$). The tests consisted in generating 274 sets of calibration equations involving 10 random points each having a minimal value of $r=0.99$ and then make predictions by the inverse and direct methods of the x variable for 8 interpolated points also taken randomly from the pool. After 10^5 iterations the mean number of cases for which the inverse method gave a lesser error than the classical one were 172.1 against 101.9. The attached significance p value is 0.00002. For the classical method the mean error in prediction was 0.174 (a maximum value of 0.77 was found), whereas for the inverse procedure it was 0.171 (maximum of 0.67). The differences in mean and maximal errors are found at the thousandths and at the tenths, respectively. Among others, this bootstrapping result is sensitive to the required

Table 1
Number of accepted calibration equations per element.

Atomic Number	Element	Equations
2	Li	2
5	B	1
11	Na	7
12	Mg	16
13	Al	20
14	Si	4
15	P	7
19	K	8
20	Ca	4
22	Ti	3
23	V	4
24	Cr	8
25	Mn	9
26	Fe	6
28	Ni	21
29	Cu	22
30	Zn	30
33	As	23
38	Sr	1
39	Y	2
45	Rh	1
46	Pd	14
48	Cd	15
56	Ba	5
78	Pt	19
79	Au	7
82	Pb	15
	Total	274

Table 2
Distribution of the number of experimental data points (n) in the valid calibration equations.

Number of points, n	Frequency
4	79
5	105
6	54
7	29
8	1
10	4
12	1
14	1
Total	274

r value of the calibration equations. If the same experiment is repeated requiring an r value of 0.90 for both the sample pool and the calibration equations, then the mean number of cases favoring the inverse method against classical one are 228.8 against 45.2. The attached p value is almost zero indicating a very strong deviation favoring the inverse procedure. In this case, the mean prediction errors were 0.575 (classical) and 0.525 (inverse), and the respective maximal errors found were 3.3 and 2.7 units. Now, the differences in mean and maximal errors are of the order of hundredths and a half unit, respectively.

3. Discussion

In Fig. 1, the diagonal dashed straight line (the set of expected y_{calc} values) coincides with the linear regression line of y_{calc} on y_{obs} obtained from the n (y_{obs},y_{calc}) pairs arising from (1). Within the general field of multiple linear regressions, a theorem has been demonstrated [10,11] which states that this regression equation corresponds to the one passing across the sample mean point and has an slope of r^2. Consequently, the dashed line in Fig. 1 is described by the following equation:

$$y_{calc}-\bar{y}=r^2(y_{obs}-\bar{y}) \tag{6}$$

Relationship (6) holds for any arbitrary finite set of y_{calc} values obtained by a simple or multiple linear regression equation disregarding the particular statistical distribution the original variables follow. From this knowledge and within the context of classical regression, it arises a proposal in order to infer with lesser expected error the corresponding variable value x_{obs} of a particular observed y one: first, it is necessary to transform the experimentally acquired y_{obs} value into the appropriate expected calculated one y_{calc} and then use the latter in Eq. (1). The complete two-steps algorithmic procedure substitutes Eq. (4) which now become

$$\begin{cases} 1.\ E(y_{calc}|y_{obs})=r^2y_{obs}+(1-r^2)\bar{y}\rightarrow y_{calc} \\ 2.\qquad\qquad (y_{calc}-a)/b\rightarrow x_{obs} \end{cases} \tag{7}$$

The first transformation (7.1) arises from (6) and it furnishes the input value to enter in Eq. (1) which is being isolated in (7.2) in the same way as it is done in (4.2). Procedure (7) seems to be counterintuitive because the numerical value arising straight from the experiment (y_{obs}) is not directly identified as to be y_{calc}, the one to be plugged into the original classical linear equation to infer x_{obs}. Instead of that the previous transformation or pre-treatment in (7.1) has to be considered.

Equations in (7) can be recast into a single one-step procedure, a compact linear equation. After arranging terms, this new equation is

$$x_{obs}=\frac{(1-r^2)\bar{y}-a}{b}+\frac{r^2}{b}y_{obs} \tag{8}$$

Surprisingly, it is now immediate to check that (8) is the conjugated equation of (1), as it is build with the same relationships codified in (3). This means that Eq. (8), and hence the procedure (7), is tantamount to Eq. (2), the inverse calibration expression. In this way it has been demonstrated that the inverse calibration equation is totally equivalent to rely on the classical one (1) isolating x_{obs} from the knowledge of y_{calc} but always performing previously the data pre-treatment (7.1) getting from the knowledge of y_{obs} the expected value $E(y_{calc}|y_{obs})$. The inverse calibration method corresponds to consider the classical approach but previously 'rotating' and correcting the acquired y_{obs} values. That is the existing link between classical and inverse procedures. The rotation described here is implicit in the inverse method and produces systematic shifts on y variable which, in turn, leads to systematic shifts on the inferred x values respect to the classic treatment. That is the main qualitative difference between classic and inverse approaches.

Within the context of the classical regression, the knowledge of x allows to calculate y_{calc} by means of Eq. (1) and that is equivalent to obtain the corresponding estimation of y_{obs}. This is so because, following this operational sequence, it is found an equivalence between expected values:

$$E(y_{calc}|x_{obs})=E(y_{obs}|x_{obs})\ \text{(for classical regression)}$$

allowing for the numerical identification y_{calc}=y_{obs} if these values are inferred from the previous knowledge of x_{obs}. Paradoxically, the reverse process does not work symmetrically as it could be expected. An observed y value cannot be plugged directly into Eq. (1) to obtain the expected x. If this is done, the estimations for x will carry statistical systematic errors due to the intrinsic regression towards the mean effect. A more convenient procedure is to rely on the inverse relationship (2). It has been demonstrated here that this procedure is totally equivalent to previously transform y_{obs} value into the corresponding expected y_{calc} one, i.e. to retrieve $E(y_{calc}|y_{obs})$ by means of equation (7.1), and then substitute it into Eq. (1).

Regarding the results of Section 2.2, the consideration of a set of experimental data allowed to check how the inverse regression performed better in interpolations than the classical one. This behavior can be partially explained by the aforementioned regression towards the mean effect correction, which allows to minimize systematic errors. The conclusion is that the trend is to favor the inverse method, mainly at the qualitative level when regarding the number of times one method is better than the other (above giving p values of 0.0003 or lesser). Following the same trend, the mean and maximal errors in x prediction are smaller for the inverse case and sensitive to the correlation coefficient among variables, as the bootstrapping shows: The difference of mean and maximal errors were 0.003 and 0.1 in the first bootstrapping experiment and about 0.05 and 0.5 in the second. The results corroborate a conclusion of Tellinghuisen [8]: the balance of factors tends to favor the inverse approach, but with small improvements for the prediction of x respect to the classical paradigm. The arguments given here also justify an affirmation of Centner and collaborators [1]: the improvement found in inverse calibration will increase with decreasing precision of the measurements, i.e., with a decreasing value of r^2, as it is found when comparing the two bootstrapping results. This is so because the effects of regression towards the mean are magnified as r^2 decreases, and the two diagonal lines in Fig. 1 become more dissimilar. Under such circumstances, the transformational step (7.1) becomes more necessary and influent.

The inverse calibration process constitutes a fitting procedure being optimal in the sense that it provides with the minimum sum of quadratic errors when adjusting the set of n known x values. This feature heuristically correlates with the concept of errors minimization for new interpolated values for x. In any case, such assertion has to be understood as to describe only a trend and not a systematic behavior, as it has been checked in the above results. Additionally, the bootstrapping experiments revealed the tendency to get greater errors (in both methods but favoring the inverse one) when predicting interpolated values if the data exhibit a lesser correlation. This feature indicates that in calibration processes where high values of r are desired, the tendency for a better performance of the inverse method in interpolations is strong but at the same time that the quantitative differences between inverse and classical approaches in predicting interpolated values are less relevant. The above results point to the fact that notable and very clear differences are to be found in cases for which the correlation coefficient is small, but those are of minor interest in the analytical field.

A linear equation fit is reversible in the sense that, mathematically, one can express one variable in terms of the other. Despite of that, fitting procedures bear an inherent asymmetry by construction and in variables treatment, and the direction in which the variables are plugged in and obtained from becomes relevant. Classical and inverse equations are intended to be applied in a precise particular direction, the former putting x and getting y and the later putting y and getting x, but not in the reverse one unless an additional transformational step is being also considered.

4. Conclusions

It has been justified why inverse calibration has the tendency to predict interpolated concentrations better than classical calibration. It has been shown how the two procedures are related, being an important concept that the variables y_{calc} and y_{obs} are not always equivalent, nor interchangeable. It has been shown how the regression towards the mean effect is responsible of systematic deviations, and also how its compensation allows to relate the classical and inverse regression equations: the former corresponds to the later if previously the observed data is transformed. The study of a real experimental data set lead to quantify the degree of superiority of inverse regression revealing how this approach has a tendency to give lesser mean quadratic errors in interpolations. A simple bootstrapping numerical experiment showed how the effects favoring the inverse method are increased if the correlation among variables decrease.

Acknowledgments

The author thanks the Institute of Computational Chemistry and Catalysis and the University of Girona for financial support by means of the ASING2012 action. Two anonymous referees are acknowledged. Their comments and suggestions clearly improved the contents of this work. Profs. M. Iglesias and C. Fontàs are greatly acknowledged because they provided the idea and the resources to access to the experimental data.

References

[1] V. Centner, D.L. Massart, S. de Jong, Fresenius J. Anal. Chem. 361 (1998) 2–9.
[2] F.H. Walters, G.T. Rizzuto, Anal. Lett. 21 (11) (1988) 2069–2076.
[3] M.R. Spiegel, Statistics, McGraw-Hill, New York, 1988.
[4] J.J.Z. Liao, Stat. Probab. Lett. 56 (2002) 271–281.
[5] R.G. Krutchkoff, Technometrics 9 (1967) 425–439.
[6] R.G. Krutchkoff, Technometrics 11 (1969) 605–608.
[7] G.K. Shukla, Technometrics 14 (1972) 547–553.
[8] J. Tellinghuisen, Fresenius J. Anal. Chem. 368 (2000) 585–588.
[9] J. Berkson, Technometrics 11 (4) (1969) 649–660.
[10] E. Besalú, J.V. de Julian-Ortiz, M. Iglesias, L. Pogliani, J. Math. Chem. 39 (2006) 475–484.
[11] E. Besalú, J.V. de Julian-Ortiz, L. Pogliani, J. Chem. Inf. Model. 47 (2007) 751–760.
[12] F. Galton, J. Anthrop. Inst. 15 (1886) 246–263.