Validating and comparing predictive models

József Baranyi[a,*], Carmen Pin[b], Thomas Ross[c]

[a]*Institute of Food Research Reading Laboratory, Earley Gate, Whiteknights Rd, Reading, RG6 6BZ, UK*
[b]*Departamento de Nutricion y Bromatologia III, Facultad de Veterinaria, Universidad Complutense de Madrid, 28040 Madrid, Spain*
[c]*School of Agricultural Science, University of Tasmania, GPO Box 252-54, Hobart 7001, Tasmania, Australia*

Abstract

The bias and accuracy factors introduced by Ross [Ross, T., 1996. Indices for performance evaluation of predictive models in food microbiology. J. Appl Bacteriol. 81, 501–508] for the evaluation of the performance of models in 'predictive food microbiology' are refined by basing the calculation of those measures on the mean square differences between predictions and observations. The use of the indices is extended by presenting formulae and methods which enable evaluation of the difference between alternative models for growth of an organism of interest over a domain of environmental factors. This is done by calculating the integral mean of the square differences between the models under investigation over the domain of the environmental variables common to those models, or a sub-region of it. The use of the techniques is exemplified by evaluating the difference between four published models for the growth rate of psychrotrophic pseudomonads. © 1999 Elsevier Science B.V. All rights reserved.

1. Introduction

Predictive microbiology is based upon the premise that the responses of populations of micro-organisms to environmental factors are reproducible, and that, by characterising environments in terms of identifiable, dominant factors which control growth responses, it is possible, from past observations, to predict the responses of those micro-organisms in other, similar, environments (Ross and McMeekin, 1994). Proponents claim that predictive microbiology offers many benefits to the practice of food microbiology, such as providing tools for HACCP implementation and for exposure assessment in microbial risk assessment (Cassin et al., 1998), and decision support in many aspects of the management of microbial food safety and quality. (Ross and McMeekin, 1994; Whiting and Buchanan, 1996). Reflecting the development of the field from a largely descriptive exercise into a more investigative scientific pursuit, McMeekin et al. (1997) proposed the term 'quantitative microbial ecology of foods' as a more apt, albeit less wieldy, description of this field of research.

Predictive microbiology models are, typically, generated in liquid laboratory culture media. Ross (1996) proposed measures of the performance of mathematical models used in predictive microbiol-

*Corresponding author. Tel.: +44-118-935-7135; fax: +44-118-935-7222.

E-mail address: jozsef.baranyi@bbsrc.ac.uk (J. Baranyi)

ogy. The objective of those performance measures was to enable the assessment of the reliability of such models when compared to observations *not* used to generate the model, particularly in foods, and hence to evaluate their utility to assist in food safety and quality decisions. A further objective was to provide a simple and quantitative measure of model reliability. Those indices were termed the bias factor and accuracy factor, but were suggested 'as a first step towards the development of an objective and useful definition of the term 'validated model' ''. Ross (1996) also foreshadowed that modifications and refinements to those indices might be proposed.

In this paper, the accuracy and bias indices are modified and generalised to enable comparison of growth models with each other as well as with observations. The model to be analysed will be that of Pin and Baranyi (1998), for organisms causing the aerobic spoilage of refrigerated meat. It is a generic spoilage model describing the maximum specific growth rates of the dominant organisms (*Pseudomonas* spp.) during spoilage, as a function of temperature and pH. In addition, the difference between that model and four other published models for the growth rate of *Pseudomonas* spp. (Davey, 1989; Fu et al., 1991; Gill and Jones, 1992; Neumeyer et al., 1997) will be evaluated.

When comparing these models for the growth rate of the same group of organisms, supported by observations made in different laboratories and under various experimental conditions, one develops a picture of the variability of the effects of those factors (e.g. composition of media). Those factors are sometimes ignored in model development, either because they are considered insignificant, because they are inherently random or because our present knowledge is inadequate to quantify them with sufficient accuracy. A classification of the sources of inaccuracy in predictive models can be found in Baranyi and Roberts (1995).

2. Materials and methods

2.1. Quantification of discrepancy between model and observation

It is common to plot predicted growth parameters against their observed values to demonstrate the performance of a model (Fig. 1). If the points
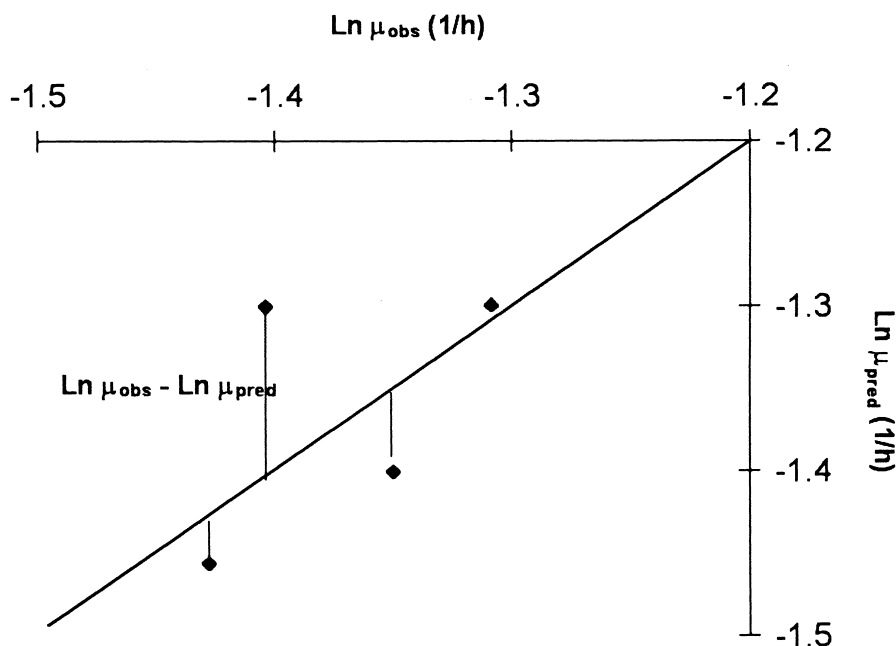

Fig. 1. Typical representation of predicted against observed values. The difference of a plotted point from the line of equivalence is an accuracy measure of that particular prediction.

obtained this way are on the line of equivalence then the predictions are exact. The difference between a point and the line of equivalence is a measure of the inaccuracy of the respective prediction.

Because the concept of relative error (the ratio between the error and the value of the prediction) is more frequently used in practice, the 'predictions versus observations' are commonly plotted on the log scale. Thus, we define our performance indicators by means of the (natural) logarithm transformation. The growth parameter chosen for comparison is the maximum specific growth rate, μ, as a function of the environmental factors which are the independent variables of the model.

Suppose that the specific growth rate is modelled by the function, $f(\mathbf{x})$, where

$$\mathbf{x} = [x_1, x_2, \ldots, x_n]$$

is the vector of the environmental factors whose effects are modelled, and f is defined in a region, R, of those environmental factors. An example for $n = 2$ factors is: $x_1 = $ temperature, $x_2 = $ pH.

We seek to quantify how well f approximates a set of μ values which are derived from a set of m observations: $\mu^{(1)}, \mu^{(2)}, \ldots, \mu^{(m)}$, at the

$$\mathbf{x}^{(i)} = [x_1^{(i)}, x_2^{(i)}, \ldots, x_n^{(i)}] \quad (i = 1, 2, \ldots, m)$$

environmental factor combinations (discrete case), or by another function, $\mu = g(\mathbf{x})$, interpreted in the same environmental region (continuous case).

Consider the environmental vector $\mathbf{x}$ as a random variable distributed in the region R. Let the random variable X be defined as:

$$X = \text{Ln } f(\mathbf{x}) - \text{Ln } \mu$$

The definition of the accuracy factor A_f of Ross (1996) is equivalent to the formula $A_f = \exp(E(|X|))$, where E(.) denotes the expected value (i.e. the mean) of the argument in parentheses. We suggest a modification of that definition to:

$$A_f = \exp\left(\sqrt{E(X^2)}\right)$$

The formula we suggest for the Bias Factor (B_f) is equivalent to that of Ross (1996), i.e:

$$B_f = \exp(E(X))$$

The advantage of the modified definition is that it

is consistent with the least squares algorithm of fitting models to observed values of Ln μ, as shown below.

2.1.1. Discrete case

If m observations, $\mu^{(1)}, \mu^{(2)}, \ldots, \mu^{(m)}$ are made randomly in the environmental region R, then the accuracy factor for the model f can be estimated by:

$$A_f = \exp\left(\sqrt{\frac{\sum_{k=1}^{m}\left(\text{Ln } f(\mathbf{x}^{(k)}) - \text{Ln } \mu^{(k)}\right)^2}{m}}\right) \tag{1a}$$

Note that if the Ln $f(\mathbf{x})$ function was fitted to the Ln $\mu^{(i)}$ observations by the least squares method, then Ln A_f is equal to, apart from the degree of freedom in the denominator, the standard error of that fit.

Our estimation of the bias factor is compatible with that of Ross (1996), and for the model f is given by:

$$B_f = \exp\left(\frac{\sum_{k=1}^{m}\left(\text{Ln } f(\text{x}^{(k)}) - \text{Ln } \mu^{(k)}\right)}{m}\right) \tag{1b}$$

It is worth noting that if the doubling time $=$ Ln $2/\mu$ is modelled in a similar way, the best fit is given by the Ln $2/f(\mathbf{x})$ model and the accuracy and bias factors do not change.

2.1.2. Continuous case

Suppose also that another model $\mu = g(\mathbf{x})$ is given to predict the maximum specific growth rate of an organism. In this case, the accuracy factor of f compared to g can be estimated by the integral mean of the square differences between f and g:

$$A_{f,g} =$$

$$\exp\left(\sqrt{\frac{\int_R (\text{Ln } f(x_1, \ldots, x_n) - \text{Ln } g(x_1, \ldots, x_n))^2 \, dx_1 \ldots dx_n}{V(R)}}\right) \tag{2a}$$

where $V(R)$ is the volume of the R region: $V(R) = \int_R 1 \, dx_1 \ldots dx_n$. For example, if R is a temperature interval, $R = [T_1, T_2]$, then $V(R) = T_2 - T_1$.

In the continuous case, the definition of the bias factor for f compared to g leads to the formula:

$$B_{f,g} = \exp\left(\frac{\int_R (\text{Ln } f(x_1, \ldots, x_n) - \text{Ln } g(x_1, \ldots, x_n)) \, dx_1 \ldots dx_n}{V(R)}\right)$$

(2b)

For an accuracy factor A_f, let the value

$$\%D_f = (A_f - 1) \cdot 100\%$$

be called the '*per cent* discrepancy' between the model and observations (discrete case); or, between two models (continuous case).

If a bias factor B_f is given, then let

$$\%B_f = \text{sgn}(\text{Ln } B_f) \cdot (\exp|\text{Ln } B_f| - 1) \cdot 100\%$$

be called the '*per cent* bias', where sgn( ) is the sign function interpreted as:

$$\text{sgn}(b) = \begin{cases} +1 & \text{if } b > 0 \\ 0 & \text{if } b = 0 \\ -1 & \text{if } b < 0 \end{cases}$$

The role of the sgn(Ln B) factor is to indicate whether the overall bias is negative or positive. If $\%B_f > 0$ then, on average, f predicts faster growth than the observations (discrete case) or the g model (continuous case).

For practical situations, the above integrals can be most readily calculated by numerical algorithms. A program from the Numerical Recipes (Press et al., 1990) was used to estimate the $A_{f,g}$ and $B_{f,g}$ values presented in this paper.

The per cent discrepancy cannot be negative, and it is zero if and only if all the predictions are equal to the observations (perfect fit), when the per cent bias is zero, too. The per cent bias, however, can be zero even if the accuracy is very poor.

2.2. Preparation of naturally contaminated meat

Three samples of minced beef were purchased from three local butchers' shops. Each sample was divided into units of 5 g which were packaged in sterile bags and sealed. Twenty-five to thirty 5g sub-samples were prepared for each growth curve determination.

For each sample of minced beef the pH, a_w and initial bacterial load were determined from one of the 5 g samples. The remaining 5g sample units were incubated at 2, 5, 8 and 11°C. Samples units were withdrawn at appropriate intervals, dependent upon the incubation temperature, and bacteria enumerated by plating on Tryptone Soya Agar (TSA, Oxoid CM 131) for total viable counts, and Cetrimide Fusidin Cephaloridine agar (CFC, Oxoid CM559, SR 103) for *Pseudomonas* spp. Colonies were visible on these media after 24 h at 25°C, and increased only in size, not number, with longer incubation. Thus, plates were incubated for at least 24 h, but up to 72 h, at 25°C.

2.3. Fitting growth curves

The time dependencies of the bacterial loads on the meats, i.e. the bacterial growth curves, were fitted by the model of Baranyi and Roberts (1994).

2.4. Measuring the discrepancy from M(G)

The model of Pin and Baranyi (1998), presented below, was denoted by M(G) in that paper and we continue to use that notation in what follows. That model was found general enough to satisfactorily describe the growth of pseudomonads both in isolation and mixed with other spoilage organisms in tryptone soya broth. The newly proposed discrepancy and bias indices were calculated for M(G) compared with models for the growth rates of *Pseudomonas* spp. presented by:

- Davey (1989), based on growth on ox muscle;
- Fu et al. (1991), where the growth substrate was dehydrated non-fat dried milk;.
- Gill and Jones (1992), derived from growth on meat, in milk and in laboratory media; and
- the model of Neumeyer et al. (1997), based on growth in laboratory media.

From those published models, the maximum specific growth rates (μ) were calculated, when necessary. The formulae derived are:

Pin and Baranyi (1998):

$$\mu = \exp(-12.65 + 0.004234 \times Temp^2 - 0.3024$$
$$\times pH^2 + 0.01535 \times Temp \times pH - 0.004356$$
$$\times Temp + 3.467 \times pH)$$

(3)

Davey (1989):

$$\mu = \exp\left[-919.54 + 1.6033 \times \frac{10^5}{Temp + 273} \right.$$

$$\left. -2.3784 \times \frac{10^7}{(Temp + 273)^2} + 1317.7 \times a_w \right.$$

$$\left. -669 \times a_w^2 \right] \tag{4}$$

Fu et al. (1991):

$$\mu = [0.0306 \times (Temp + 7.85)]^2 \tag{5}$$

Gill and Jones (1992):

$$\mu = \ln 2 \times (0.033 \times Temp + 0.27)^2 \tag{6}$$

Neumeyer et al. (1997):

$$\mu = \ln 2 \times \left[0.1539 \times (Temp + 7.7) \times \sqrt{a_w - 0.947} \right]^2 \tag{7}$$

The common environmental domain, R, used for this comparative study, was the temperature interval from 2 to 11°C. The pH and a_w were fixed at 5.8 and 0.995, respectively, i.e. the measured values of pH and water activity of the minced meat.

The predictions given by M(G) were also compared with the maximum specific growth rates of pseudomonads occurring in natural contaminated meat. In this case, the per cent discrepancy and bias (%D, %B) were calculated as defined for the discrete case using the Eqs. 1a and 1b.

3. Results

Pin and Baranyi (1998) fitted the natural logarithm of their maximum specific growth rate data against temperature and pH, and obtained 0.126 as the standard error of the fit. They fitted 6 parameters on 36 observations, therefore the accuracy factor with respect to the data they based their model on was

$$A_f = \exp\left(\sqrt{\frac{0.126^2 \cdot 30}{36}} \right) = 1.12$$

Accordingly, the per cent discrepancy between M(G) and the data used to create M(G) was 12%. The R^2 percentage of that fit was good (94%), so we deduce that the maximum specific growth rate can be

measured, by the viable counts technique, at about 10% accuracy. This is in accord with the estimated standard errors of the maximum specific growth rates obtained when the individual growth curves were fitted by the model of Baranyi and Roberts (1994). Those standard errors were always 5–15% of fitted value of the parameter, i.e. se(μ)/$\mu \approx 0.1$.

The per cent bias between M(G) and the data used to create M(G) was zero. This is a consequence of the well-known properties of the least squares method for fitting models to data. Namely, because our definition for the bias indicator uses logarithm transformations and the regression of Pin and Baranyi (1998) was carried out minimising the sum of squared differences between the logarithms of model predictions and observations, so the bias to the data used for building the model must, by definition, be zero.

Table 1 shows the per cent discrepancy between M(G) and the other models for *Pseudomonas* spp. The model of Neumeyer et al. (1997) and Gill and Jones (1992) were quite close to M(G) with *per cent* discrepancies of 8.8% and 12.9%, respectively. Note that these discrepancies are inside or around the 12% error margin of the model of Pin and Baranyi (1998) to the data used to create it. Thus, these three models are very consistent.

The model M(G) deviated most (43.6%) from that of Davey (1989). M(G) showed negative bias from all the other models but in the case of Neumeyer et al. (1997) and Gill and Jones (1992), the growth rates predicted by the other models were only slightly faster than those predicted by M(G), as might be expected from the %B values. The predictions of Davey (1989) and Fu et al. (1991) were much faster.

The per cent discrepancy between the model M(G) and the growth rates in natural meat (Table 2) was 46.3%, while the per cent bias was 35.5%

Table 1
Per cent discrepancy and bias between the model M(G) of Pin and Baranyi (1998) and other published models for the maximum specific growth rates of pseudomonads

Models described by	% $D_{f,g}$	% $B_{f,g}$
Davey (1989)	44.5	−43.6
Fu et al. (1991)	32.6	−31.8
Gill and Jones (1992)	12.9	−11.4
Neumeyer et al. (1997)	8.8	−6.3

Table 2
Initial logcounts in three samples of naturally contaminated beef mince plated on TSA and on CFC

	Sample 1 (log cfu/ml)	Sample 2 (log cfu/ml)	Sample 3 (log cfu/ml)
TSA	7.59	7.25	7.04
CFC	5.55	6.43	5.34

Table 3
Per cent discrepancy and bias between the model M(G) of Pin and Baranyi (1998) and the maximum specific growth rates observed in three samples of naturally contaminated meat

	% $D_{f,g}$	% $B_{f,g}$
Meat sample 1	67.2	61.8
Meat sample 2	41.3	37.3
Meat sample 3	25.3	11.9
Average	46.3	35.5

(Table 3). Therefore the laboratory-media-produced M(G) overestimates the growth rates observed in meat (see Fig. 2). This overestimation is even more pronounced for the other models.

The growth rates of pseudomonads seem to be higher than those of the total viable population. Note, however, that the initial counts of the three samples of naturally contaminated meat on TSA were higher than 10^7 cfu/ml while the initial counts on CFC were about 5.10^5 cfu/ml in two samples and 5.10^6 cfu/ml in the third case (Table 2).

4. Discussion

The discrepancies between M(G) and the models of Neumeyer et al. (1997) and of Gill and Jones (1992) were small. We have not been able to identify the reasons why the other two models give 30–40% higher growth rates.

Predictive models based on microbial growth data measured in laboratory media commonly overestimate the growth rates of micro-organisms observed in food, and we assume the laboratory media is, in most cases, optimal for growth. Nonetheless, the % Discrepancy and % Bias indices for the M(G) model compared to naturally contaminated meat merit comment. Commercially available minced meat at retail often has a high initial concentration of spoil-
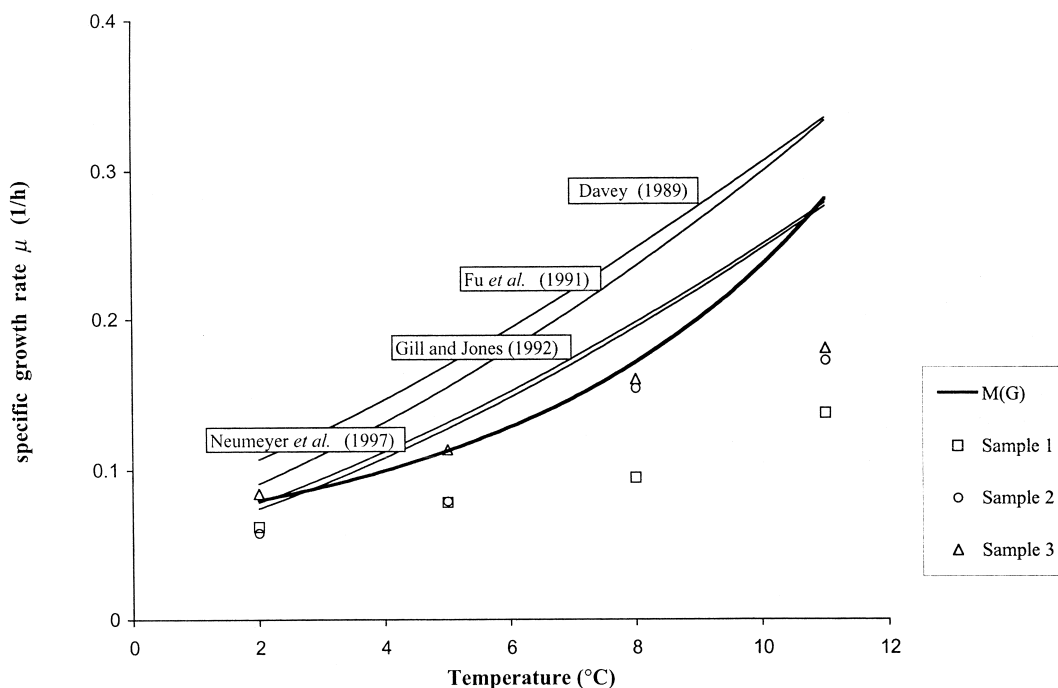

Fig. 2. Specific growth rate model M(G) of Pin and Baranyi (1998) and other published models for pseudomonads (continuous lines) against observations in three samples of naturally contaminated meat.

age organisms. This is shown in Table 2. Under these conditions, the meat may contain growth-retarding levels of metabolites. The data in Table 2 also show that pseudomonads were only a small element of the initial microbiota and it should be noted that the growth rate of pseudomonads in the three samples used in this study (see Table 4) was, generally, inversely related to the initial level of contamination of those samples. Taken together these observations suggest that the growth rate of pseudomonads in the samples was limited by the activity or metabolites of the high numbers of other organisms, and may explain the values of the % Discrepancy and % Bias indices for M(G) in this case.

Both the accuracy and bias factors are, in some sense, average values, but they are defined using different concepts of 'average'. The accuracy factor is based on 'mean square differences', while the bias factor is based on the 'arithmetical mean of the differences'. One advantage of the new definition of the accuracy factor is that, applied to the data on which the model is based, it directly relates to the goodness of fit, inasmuch as the least squares method was used to fit the logarithm of the specific rates. This did not hold for the accuracy factor definition of Ross (1996). However, with the new definition, accuracy and bias factors cannot be the same values, unlike in the case of Ross (1996), where they are identical, if the sign of the difference (prediction-observation) is constant in the studied region.

It is well known that for a random variable, X, that if

$$0 \leq \text{Var}(X) = E([X - (E(X))]^2) = E(X^2) - [E(X)]^2$$

therefore

$$[E(X)]^2 \leq E(X^2).$$

Consequently, if one model consistently overestimates the observations, or provides higher predictions than another model, then the absolute values of the discrepancy and bias indicators are close to each other, the latter being somewhat smaller. This was observed in all our comparisons.

The integral formula for comparing two predictive models provides a good theoretical basis but, in practice, is rarely worth calculating explicitly. If the user has access to spreadsheet software, the following method provides a reasonable approximation.

Choose a matrix and imagine the values of the environmental factors as forming a grid of equal intervals. Consider the g-model predictions at the grid-values of the environmental factors as observations. Calculate the discrepancy factor according to the first, discrete definition (Eq. 1a). Repeat the procedure with double the number of values of each variable in the grid. If the new accuracy factor is close enough to the first one, then accept the latter one as a sufficiently good approximation of the integral.

In the definitions of discrepancy and bias indices, we used the inverse function-pairs, $\text{Ln}(x)$ and $\exp(x)$. This is a question of personal preference, the result is the same as using the pairs $\log_{10}(x)$ and 10^x as demonstrated by Ross (1996).

In conclusion, the definitions of the new bias and accuracy factors are consistent with established measures of goodness-of-fit. The modified definitions

Table 4
Maximum specific growth rates (μ) in three samples of naturally contaminated meat plated on TSA and CFC

	Temperature							
	2°C		5°C		8°C		11°C	
	$\mu(1/h)$	SE	$\mu(1/h)$	SE	$\mu(1/h)$	SE	$\mu(1/h)$	SE
Sample 1:								
TSA	0.03	0.003	0.03	0.004	0.04	0.005	0.05	0.005
CFC	0.06	0.004	0.08	0.003	0.09	0.004	0.14	0.004
Sample 2:								
TSA	0.04	0.003	0.04	0.003	0.06	0.006	0.11	0.012
CFC	0.06	0.003	0.08	0.008	0.15	0.021	0.17	0.013
Sample 3:								
TSA	0.03	0.003	0.04	0.004	0.06	0.007	0.10	0.006
CFC	0.08	0.009	0.11	0.014	0.16	0.013	0.18	0.007

have similar meaning and utility to the original bias and accuracy factors. These indices were developed to enable direct comparison of models to one another, rather than to a specific, independent, data set which may, or may not, be representative of 'true' behaviour. Such direct comparison of models can reveal whether models are significantly different to one another, or describe the same behaviour equally well within the limits of the error in the data. As with the original definitions of the bias and accuracy factors, systematic deviations between predicted and observed responses in one part of the response surface may be 'balanced' by deviations in another part of the response surface. Such behaviour might be suggested by a larger accuracy factor, but it is perhaps still important to plot the predicted and observed values to guard against such systematic deviations. The modified bias and accuracy factors are suggested as another step towards the establishment of objective and useful definitions of the term 'validated model'.

References

Baranyi, J., Roberts, T.A., 1994. A dynamic approach to predicting bacterial growth in food. Int. J. Food Microbiol. 23, 277–294.

Baranyi, J., Roberts, T.A., 1995. Mathematics of predictive food microbiology. Int. J. Food Microbiol. 26, 199–218.

Cassin, M.H., Lammerding, A.M., Todd, E.C.D., Ross, W., McColl, R.S., 1998. Quantitative risk assessment for Escherichia coli O157-H7 in ground beef hamburgers. Int. J. Food Microbiol. 41, 21–44.

Davey, K.R., 1989. A predictive model for combined temperature and water activity on microbial growth during the growth phase. J. Appl. Bacteriol. 67, 483–488.

Fu, B., Taoukis, P.S., Labuza, T.P., 1991. Predictive microbiology for monitoring spoilage of dairy products with time-temperature integrators. J. Food Sci. 56, 1209–1215.

Gill, C.O., Jones, T., 1992. Assessment of the hygienic efficiencies of two commercial processes for cooling pig carcass. Food Microbiol. 9, 335–343.

McMeekin, T.A., Brown, J., Krist, K., Miles, D.M., Neumeyer, K., Nichols, D.S., Olley, J., Presser, K., Ratkowksy, D.A., Ross, T., Salter, M., Soontranon, S., 1997. Quantitative microbiology; a basis for food safety. Emerg. Infect. Dis. 3, 541–549.

Neumeyer, K., Ross, T., McMeekin, T.A., 1997. Development of a predictive model to describe the effects of temperature and water activity on the growth of spoilage pseudomonads. Int. J. Food Microbiol. 38, 45–54.

Pin, C., Baranyi, J., 1998. Predictive models as means to quantify the interactions of spoilage organisms. Int. J. Food Microbiol. 41, 59–72.

Press, W.H., Flannery, B.P., Teukolsky, S.A., Vetterling, W.T., 1990. Numerical Recipes, Cambridge University Press, Cambridge.

Ross, T., 1996. Indices for performance evaluation of predictive models in food microbiology. J. Appl. Bacteriol. 81, 501–508.

Ross, T., McMeekin, T.A., 1994. Predictive microbiology – a review. Int. J. Food Microbiol. 23, 241–264.

Whiting, R.C., Buchanan, R.L., 1996. Predictive modeling. In: Doyle, N., Beuchat, N., Montville, N. (Eds.), Food Microbiology: Fundamentals and Frontiers, ASM Press, Washington, DC, pp. 728–739.