Journal of Applied Bacteriology 1996, **81**, 501–508

Indices for performance evaluation of predictive models in food microbiology

T. Ross

Department of Agricultural Science, University of Tasmania, Hobart, Tasmania, Australia

5536/11/95: received 2 November 1995, revised 27 March 1996 and accepted 29 March 1996

T. ROSS. 1996. Two complementary measures are proposed as simple indices of the performance of models in predictive food microbiology. The indices assess the level of confidence one can have in the predictions of the model and whether the model displays any bias which could lead to 'fail-dangerous' predictions. The use of the indices is demonstrated using data collated from independent and published literature. This analysis supports previous reports that evaluation of predictive models by comparison to published microbial growth rate data may be inappropriate because of limitations in that data. The indices may fail to reveal some forms of systematic deviation between observed and predicted behaviour. It is concluded, however, that the indices provide an objective and readily interpreted summary of model performance and may serve as a first step towards the development of an objective and useful definition of the term 'validated model' in predictive food microbiology.

INTRODUCTION

Predictive food microbiology is based upon the premise that the responses of populations of micro-organisms to environmental factors are reproducible and that it is possible, from past observations, to predict the responses of micro-organisms in a particular environment. In general, a reductionist approach is adopted and microbial responses are measured under defined and controlled laboratory conditions, often in liquid media. The results are summarized in the form of mathematical equations which, by interpolation, can be used to predict responses to sets of conditions which were not specifically tested.

Methods for comparing the goodness-of-fit of competing models to the data used to generate them, or to determine whether a fitted model is statistically acceptable relative to the measuring error inherent in the data, have been used in the predictive microbiology literature (Adair *et al.* 1989; Zwietering *et al.* 1990, 1994) and general methods are described in statistical texts (e.g. Draper and Smith 1981). Before they can be used in practice, however, predictive models must be shown to predict accurately the behaviour of micro-organisms in foods during processing, storage and distribution. Demonstration of this ability, a process generally termed 'validation', remains an ill-defined aspect of predictive microbiology but must involve comparison of predicted responses to observations in product, independent of those used to generate the model. Typically, growth rates or generation times predicted by the model are compared to those observed for the same organism in food (Gibson *et al.* 1988; Buchanan *et al.* 1993; McClure *et al.* 1993; Wijtzes *et al.* 1993; Sutherland *et al.* 1994).

Wijtzes *et al.* (1993) plotted literature values for the generation time of *Listeria monocytogenes* against the corresponding predictions of a model derived from studies in laboratory broth. From this plot, predictions which would be unsafe in practice could be visualized readily, and the overall reliability of the model assessed. Duh and Schaffner (1993) developed predictive equations for *Listeria* growth rate based on measurements in brain heart infusion broth. Complementary literature values for the growth of the organism in food were then added to the data set and regression analysis of the supplemented data set performed. The close similarity in MSE (mean-square error) and r^2 values of the equations fitted to either data set was taken as an indication of the reliability of the models when applied to foods. Another measure of the accuracy of predictive equations was introduced by McClure *et al.* (1993) who compared their models on the basis of the sum of the squares of the differences of the natural logarithm of observed and predicted values:

$$\Sigma (\ln (GT_{\text{published}}) - \ln (GT_{\text{predicted}}))^2. \quad (1)$$

Correspondence to: Dr Thomas Ross, Department of Agricultural Science, University of Tasmania, GPO Box 252-54, Hobart, Tasmania 7001, Australia (e-mail: Tom.Ross@agsci.utas.edu.au).

A smaller value indicates a model which, on average, better predicts the observed response.

It is important that users of models have both an understanding of the range of applicability of their predictive models, and also the limits of their performance. When considering performance, of most immediate interest is whether the model is 'fail-dangerous', i.e. whether it produces estimates that underestimate the risk of spoilage or extent of pathogen growth, but models should also predict as closely as possible the observed behaviour to avoid wastage of product. To date, no standard method or set of criteria has been published by which a model can be said to have been validated. Ideally, such measures would be readily interpretable and dimensionless so that the performance of models could be readily summarized and compared. In this paper simple indices of performance for kinetic models are developed and their utility to measure the reliability of models is assessed.

MATERIALS AND METHODS

Development of indices

Consistent with the desire to develop readily interpretable indices, measures based on an average deviation between predicted and observed generation times were sought. Generation times may vary greatly in magnitude depending upon the organism and its environment. Thus, the *absolute* deviation, i.e. the difference between the predicted and observed response, for large values will have greater influence in any averaging process, requiring that some measure of *relative* average deviation be developed. The simplest relative measure is a *ratio* of the predicted and observed generation times and, thus, was adopted to 'standardize' the deviation. The implications of this assumption are considered later.

The ratio alone, however, may be misleading because, for example, a 'factor of 10' overprediction (predicted/observed = 10) will have more weight than a 'factor of 10' underprediction of generation time (predicted/observed = 0·1) in the calculation of a mean. It was considered desirable to give equal weight to over- and underprediction. Thus, the logarithm of the ratio was chosen so that over- and underprediction were given equal weight in determining the average deviation. The antilogarithm of this value (average relative deviation) may be interpreted as the average ratio of the predicted and observed generation times, i.e. the geometric mean of the ratios. For convenience, this value will be termed the 'bias factor', and is defined:

$$\text{bias factor} = 10^{(\Sigma \log (GT_{predicted}/GT_{observed})/n)} \quad (2)$$

where $GT_{predicted}$ is the predicted generation time, $GT_{observed}$ is the observed generation time, and n is the number of observations used in the calculation.

Perfect agreement between predictions and observations will lead to a bias factor of 1. This is, perhaps, contrary to the more intuitive and established understanding of zero bias indicating perfect agreement. In developing this index, however, the objective was to derive a simple and readily interpretable quantitative measure of bias, and it was considered that bias measured on a logarithmic scale would be less immediately interpretable. Instead, by taking the antilogarithm, the value obtained is a multiplicative factor by which the model, on average, over- or under-predicts, hence the terminology bias *factor*. Thus, a bias factor of 1·1 indicates not only that the model is 'fail-dangerous' because it predicts longer generation times than are observed, but also that the predictions exceed the observations, on average, by 10%. Conversely, a bias factor less than one indicates that a model is, in general, 'fail-safe', but a bias factor of 0·5 indicates a poor model that is overly conservative because it predicts generation times, on average, half of that actually observed.

Under- and over-prediction will tend to 'cancel out' in this measure because the logarithm of the ratios will have opposite signs. Consequently, eqn 2 provides no indication of the average accuracy of estimates. Thus, the average of the *absolute* values of the logarithm of the ratio (similar to eqn 1 in which the *square* of the ratio makes all values positive) was calculated. The antilogarithm of this value will always be greater than or equal to one. This value will be termed the 'accuracy factor', and is defined:

$$\text{accuracy factor} = 10^{(\Sigma |\log (GT_{predicted}/GT_{observed})|/n)} \quad (3)$$

where the terms are as previously defined.

The larger the value, the less accurate is the average estimate. As with the bias factor, the accuracy factor is a simple multiplicative factor indicating the spread of results about the prediction. Thus, an accuracy factor of two indicates that the prediction is, on average, a factor of two different from the observed value, i.e. either half as large or twice as large, while a value of one indicates that there is perfect agreement between all predicted and observed values.

Note that eqns 2 and 3 can equally well be used for any time-based response, e.g. lag time, time to an n-fold increase, maximum specific growth rate. If rate values are compared, however, a bias factor less than 1 indicates a 'fail-dangerous' model.

Data sources and bases of comparison

Model. A model derived to predict the growth rate of *Staphylococcus aureus* 3b as a function of temperature and water activity, due to NaCl as the humectant, was chosen to exemplify the use of the indices. A data set of 211 growth curves for *Staph. aureus* 3b, measured by optical density methods, was generated in the temperature range 8·9–36·2°C and water

activity range 0·860–0·997. Generation times (GT) were derived from the parameters of a Gompertz-like equation (McMeekin *et al.* 1993; App. 2 A.9) fitted to the growth curve data. Growth rate (1/GT) data were fitted to a combined temperature–water activity model of the type introduced by McMeekin *et al.* (1987). Full details of the generation of the data and derivation of the model are given in Ross (1993).

Data sources. Data for the growth of the modelled micro-organism in foods, independent of that used to generate the model, were obtained from three sources:

(i) by determination of the rates of growth of *Staph. aureus* 3b inoculated on to foods of varying water activity and stored at various temperatures (Ross 1993);
(ii) by reference to published data for the growth of the *Staph. aureus* and including data collated from the literature by Sutherland *et al.* (1994); and
(iii) unpublished data of Dr Isabel Walls.

The model does not include a pH term. Therefore, only data in the pH range 6·0–7·0 were used in the comparison on the basis of the reported optimal pH range for *Staph. aureus* (Banwart 1989, Table 4.6; Jay 1992, p. 459). Other results (Aalberts, unpublished) confirm that variations in pH in this range have a negligible effect on the growth rate of *Staph. aureus* 3b.

For publications in which a range of generation times were recorded for one or more strains, the shortest recorded generation time was chosen as the basis for comparison to assess the ability of the models to make 'fail-safe' predictions.

RESULTS

Table 1 demonstrates, in detail, the calculation of the bias and accuracy factors for a small data set.

Table 2 presents the predicted and observed generation times of *Staph. aureus* 3b inoculated into a range of food types at temperatures and water activities covering a wide range of the response surface encompassed by the predictive equation, and the bias and accuracy factors appropriate to that data for the model.

Table 3 shows observed generation times of *Staph. aureus* reported in the literature for a variety of foods, temperatures and water activities covering a wide range of the response surface encompassed by the predictive equation. The predicted generation times are also shown, as are the bias and accuracy factors appropriate to that data for the model. The data in Table 3 are also presented, in the manner of McClure *et al.* (1993), in Fig. 1.

The bias factor and accuracy factor determined for the data set used to generate the predictive model were found to be 1·00 and 1·20, respectively.

DISCUSSION

The bias and accuracy factors may be interpreted as quantitative summaries of the type of plot, shown in Fig. 1, used by several groups (McClure *et al.* 1993; Wijtzes *et al.* 1993; Sutherland *et al.* 1994; Bhaduri *et al.* 1994) to evaluate the performance of predictive food microbiology models. The bias factor answers the question whether, on average, the observed values lie above or below the line of equivalence and, if so, by how much. Thus it assesses whether the model is 'fail-safe'. The bias factor can also be calculated from that plot by fitting to the data the equation:

$$y = x + c$$

where y is the logarithm of the predicted response time, x is the logarithm of the observed response time, and c is the logarithm of the bias factor.

The accuracy factor averages the minimum 'distance' between each point and the line of equivalence as a measure of how close, on average, predictions are to observations. The accuracy factor is, thus, a measure of average deviation and may be used as a simple measure of the level of confidence one may have in the model's predictions.

Equation 3, though developed entirely pragmatically, is analogous to existing statistical measures of 'goodness-of-fit' and is a member of a general class of measures of 'performance' which can be summarized as:

$$\text{measure of performance} = \sqrt[m]{\frac{\sum_{i=1}^{n} |(\text{predicted} - \text{observed})|^m}{n}} \qquad m = 1, 2, 3 \ldots \quad (4)$$

where n is the number of comparisons, e.g. $m = 2$ corresponds to the root mean square error (RMSE). The larger the value of m the greater the influence of predictions which deviate widely from the observed result, resulting in larger values of the 'error' estimate. RMSE is a widely used measure of 'goodness-of-fit', and can be used to derive a measure analogous to the accuracy factor (e.g. accuracy factors based on eqn 4, using log predicted and log observed values and with $m = 2$, are 1·24, 1·39 and 1·75 for the data of Tables 1, 2 and 3, respectively). That eqn 3 is similar to existing statistical measures is a simple consequence of a common objective. The novelty of the proposed approach lies mainly in the use of eqn 2 in combination with eqn 3, and also in their intended application.

Provided that the bias factor is close to 1, the accuracy factor is almost equivalent to the 50% confidence interval calculated from the standard deviation (S.D.) of the ratios. (The 50% confidence intervals, i.e. 0·67449 × S.D., of the ratios for data in Tables 1, 2 and 3 are 1·17, 1·25 and 1·46,

Table 1 Demonstration of the calculation of bias and accuracy factors (data of Ross 1993)

Food type	Variables: Temperature (°C)	Variables: Water activity	Observed GT (h)	Predicted GT (h)	predicted/ observed	log (pred/obs)	absolute value
Smoked salmon	12·5	0·965	11·5	17·4	1·52	0·18	0·18
Smoked salmon	17·5	0·965	4·05	3·89	0·96	−0·02	0·02
Smoked salmon	22·5	0·975	1·65	1·52	0·92	−0·04	0·04
Smoked salmon	25·0	0·955	1·90	1·34	0·71	−0·15	0·15
Smoked salmon	27·5	0·975	0·73	0·84	1·15	0·06	0·06
Smoked salmon	32·5	0·965	0·57	0·58	1·02	0·01	0·01
Smoked salmon	35·0	0·955	0·50	0·53	1·06	0·03	0·03
	Mean					0·01	0·07
	⇒ **bias factor** (= antilog_{10} 0·01)						**1·02**
	⇒ **accuracy factor** (= antilog_{10} 0·07)						**1·17**

GT, Generation time.

respectively.) Based on the accuracy factor for the data in Table 2 the predictions are, on average, within 26% of the observation. Thus, for a predicted GT of 100 min, the bounds expected to encompass approximately half of the observations would be 79–126 min (i.e. 100 ÷ 1·26, 100 × 1·26). Note that the bounds are asymmetrical about the prediction. A similar calculation was shown by McClure *et al.* (1993) based on the RMSE of the model to the data used to generate it. The standard deviation, however, calculates deviation from the average ratio of values. Thus, for a model which consistently over- or underpredicted the observed response, an accuracy estimate based on the standard deviation could be misleading. The accuracy factor, however, will reflect the extent of the bias of the model.

The values of the indices reflect the level of 'fidelity' of the different data types. The accuracy factor is best, i.e. closest to 1, for the data set used to generate the model for *Staph. aureus* 3b. Those data are derived from the simplest and most homogenous system, and represents the highest degree of experimental control. The bias factor is 1 also for that data set, which is expected because of the minimization-of-errors strategies used to fit the equation to the data.

Also as expected, Tables 2 and 3 suggest that model accuracy decreases as the degree of experimental control is reduced. The data in Table 2 represent a lower level of experimental control than those upon which the model was based because non-sterile, inhomogenous foods were used as the growth medium. It is reasonable to assume, however, that it is the growth of *Staph. aureus* 3b that is being measured, and that the temperature and water activity are as reported. The data in Table 3, however, represent a yet lower level of data fidelity. For example, some of the values in Table 3 were estimated from published growth curves and must be considered as approximate values. Furthermore, methods of growth rate determination were not uniform and relevant information (i.e. pH, temperature, a_w) upon which to base predictions was not documented in all publications. Nonetheless, that the bias factor remains close to unity for all these data sets is a reassuring feature of the predictive equation generated, and supports the validity of the predictive modelling approach in food microbiology.

Measures of performance previously employed in the predictive microbiology literature have been used, typically, to evaluate the goodness of fit of a model or models, to the data used to generate them. Equations 2 and 3 are primarily intended to be used to evaluate the performance of a model by comparison to *novel* data. Traditional statistical methods are based upon the deviation between the observed and mean response and are inappropriate for evaluating model performance by comparison to novel data because the mean response is not known. The bias and accuracy factors test, in effect, the hypothesis that the model under evaluation predicts the true mean, or that it represents it better than some other model. Inherent in the development of the bias and accuracy factors is the assumption that the ratio of predicted to observed generation time, or equivalently the difference between the logarithms of the predicted and observed times, is independent of the magnitude of the generation time. Consequently, the distribution of results about the mean must be considered so that the reliability of the indices themselves can be assessed.

McMeekin *et al.* (1993, p. 130) suggested that the variance in bacterial growth response times is either Inverse Gaussian or Gamma distributed. These distributions require that the

Table 2 Evaluation of a model for *Staphylococcus aureus* 3b growth by comparison with novel data

Food type	Variables: Temperature (°C)	Variables: Water activity	Predicted GT (h)	Observed GT (h)
Milk (Whole)	12·5	0·995	13·5	16·5
	12·5	0·995	13·5	7·27
(UHT)	12·5	0·995	13·5	7·51
(Whole)	17·5	0·995	3·02	1·57
	17·5	0·995	3·02	8·31
(UHT)	17·5	0·995	3·02	2·69
(Whole)	22·5	0·995	1·30	1·51
	22·5	0·995	1·30	1·25
(UHT)	22·5	0·995	1·30	1·29
(Whole)	27·5	0·995	0·72	0·89
	27·5	0·995	0·72	0·80
(UHT)	27·5	0·995	0·72	0·93
(Whole)	32·5	0·995	0·45	0·44
	32·5	0·995	0·45	0·54
(UHT)	32·5	0·995	0·45	0·51
(Whole)	37·5	0·995	0·31	0·37
	37·5	0·995	0·31	0·40
(UHT)	37·5	0·995	0·31	0·51
Prawns	32·5	0·995	0·45	0·40
(Cooked)	30·0	0·995	0·56	0·57
	30·0	0·995	0·56	0·47
(Uncooked)	25·0	0·995	0·94	0·79
(Cooked)	20·0	0·995	1·89	1·80
(Uncooked)	20·0	0·995	1·89	1·64
(Uncooked)	17·5	0·995	3·02	2·48
(Uncooked)	12·5	0·995	13·53	6·12
Smoked salmon	12·5	0·965	17·4	11·5
	17·5	0·965	3·89	4·05
	22·5	0·975	1·52	1·65
	25·0	0·955	1·34	1·90
	27·5	0·975	0·84	0·73
	32·5	0·965	0·58	0·57
	35·0	0·955	0·53	0·50
Smoked salmon	12·6	0·920	29·1	43·3
	17·5	0·920	6·80	7·93
	22·5	0·920	2·91	3·78
	27·5	0·920	1·61	1·58
	32·5	0·920	1·02	0·99
	bias factor			**1·00**
	accuracy factor			**1·26**

square root of the reciprocal or the logarithmic transformation of the response time is used to homogenize the variance in the data for the purposes of model fitting or statistical evaluation. However, it has not yet been shown clearly, which distribution, if either, is most appropriate (Ratkowsky *et al.* 1991; Alber and Schaffner 1993; Zwietering *et al.* 1994; Schaffner 1994) and in practice the difference may be slight (Ratkowksy *et al.* 1996). Thus, the logarithmic transformation employed in the calculation of the performance factors also has the effect of, at least partly, homogenizing the variance in the data. A further property of Gamma distributions is that the standard deviation is proportional to the mean response. Thus, the ratio will be independent of the conditions under which it was measured and the accuracy factor will reliably

Table 3 Evaluation of a model for *Staphylococcus aureus* 3b generation time (GT) by comparison with independent food-based data

Food type	Temperature	a_w	Predicted GT	Observed GT	Reference
Sterile baby food	12·0	0·993	17·5	9·50	Walls (unpublished)
	12·0	0·993	17·5	6·35	
	12·0	0·965	22·0	8·18	
	12·0	0·965	22·0	8·81	
	20·0	0·993	1·93	1·65	
	20·0	0·993	1·93	1·79	
	20·0	0·924	3·99	3·60	
	20·0	0·924	3·99	3·67	
	35·0	0·993	0·38	0·50	
	35·0	0·993	0·38	0·54	
	35·0	0·924	0·79	1·48	
	35·0	0·924	0·79	1·11	
Shrimp slurry	37·0	0·991	0·33	0·50	†Troller and Stinson 1975
	37·0	0·950	0·49	1·00	
Potato dough	37·0	0·970	0·40	2·44	
	37·0	0·931	0·61	2·30	
Sterilized milk	25·0	0·997	0·93	0·63	†Yotis and Teodoro 1957
	30·0	0·997	0·55	0·51	
	33·0	0·997	0·43	0·44	
	37·0	0·997	0·32	0·39	
UHT milk plus glucose	23·2	0·980	1·33	1·12	Broughall *et al.* 1983
	20·0	0·980	2·13	1·24	
	16·4	0·980	4·35	3·04	
	12·3	0·980	16·7	6·41	
	23·2	0·960	1·59	1·02	
	20·0	0·960	2·55	1·60	
	16·4	0·960	5·22	3·31	
	12·3	0·960	20·0	8·61	
	26·2	0·900	2·77	2·74	
	20·0	0·900	6·39	12·2	
	20·0	0·880	12·8	15·3	
	16·4	0·880	26·1	34·4	
	20·0	0·930	3·65	4·95	
	16·4	0·930	7·46	9·45	
	12·3	0·930	28·6	34·4	
UHT milk + NaCl	19·0	0·997	2·22	2·44	Rajkowski *et al.* 1994
	19·0	0·994	2·27	1·82	
	19·0	0·970	2·76	2·08	
	28·0	0·997	0·67	0·39	
	28·0	0·994	0·69	0·55	
	28·0	0·970	0·83	0·65	
	37·0	0·997	0·32	0·31	
	37·0	0·994	0·33	0·51	
	37·0	0·970	0·40	0·38	
Egg noodles	15·0	0·997	5·49	3·39	†Gockler *et al.* 1988
	20·0	0·997	1·87	2·19	
	25·0	0·997	0·93	1·17	
	30·0	0·997	0·55	1·13	
	37·0	0·997	0·32	0·58	
			⇒ bias factor	**1·01**	
			⇒ accuracy factor	**1·53**	

† Reported in Sutherland *et al.* (1994).

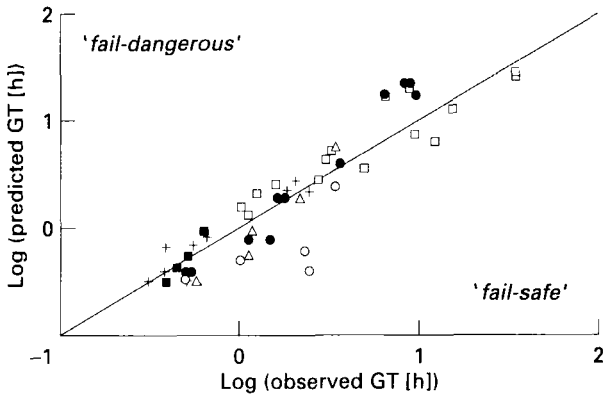

Fig. 1 Predictions of the model of Ross (1993) for the generation time of *Staphylococcus aureus* 3b compared to independently derived data for the growth of *Staph. aureus* in foods. The diagonal line is the line of identity. Points above this line represent predictions which are longer than the observed generation time and are thus 'fail-dangerous'. Conversely, points below the line of identity are 'fail-safe' predictions. The data are detailed in Table 3. Data of: Troller and Stinson (1975) (○); Yotis and Teodoro (1957) (■); Walls (unpublished) (●); Broughall *et al.* (1993) (□); Rajkowski *et al.* (1994) (+); Göckler *et al.* (1988) (△)

represent an average proportional deviation between model predictions and observed responses throughout the response surface.

Taken together, eqns 2 and 3 can be used to compare objectively the performance of different models to different data sets. It must be remembered, however, that eqns 2 and 3 do not generate absolute measures of performance: unless perfectly representative 'test' data sets can be collated, the values of the indices will be specific to the data sets used to evaluate them. This raises the issue of whether comparison to literature data is an appropriate means of model validation. Sutherland *et al.* (1994) discussed the limitations and difficulty of using growth rate data obtained from the literature, and Fig. 1 suggests that some data sets are 'anomalous'. The predictions of the model used in this paper agree closely with those of the model assessed by Sutherland *et al.* (1994) even though the models were derived completely independently and using different strains of *Staph. aureus* and different experimental approaches. Neither model, however, performed well when assessed against the literature data set collated by those authors. To evaluate model performance it may be more appropriate to use data derived under well controlled conditions, so that the model's performance is not unfairly prejudiced by comparison to unrepresentative data or data collected under inadequately controlled or defined conditions. Cole *et al.* (1994) observed that validation with literature data can be useful but often reveals marked deficiencies in that literature itself. In this context, the indices may also have utility as a measure of 'cleanness' of data or to compare methods of data collection.

The accuracy values calculated in this exercise are consistent in magnitude with the experience of other workers (Walker and Jones 1994; Sutherland *et al.* 1994) but the tabulated data suggest that some predictions (e.g. lower temperatures in Table 3) appear to be dangerously erroneous. The poorer predictions may simply be a consequence of the inherently greater variability in responses as conditions become less favourable for growth, i.e. those points may represent the extremes of the distribution of possible response times. Alternatively, a weakness in the measures proposed is that the bias factor may fail to reveal some forms of systematic bias, e.g. systematic overprediction in one region of the response surface may be balanced by systematic underprediction in another region. Other means of assessment, such as examination of the signs of the residuals, are needed to verify the bias factor.

Plots of observed *vs* predicted responses, such as Fig. 1, may reveal systematic deviations. While Fig. 1 reveals poor agreement of the model to specific data sets, overall there is no evidence of systematic over- or underprediction as a function of the response time. More searching analysis could be undertaken by plotting the ratio as a function of temperature, or as a function of water activity, to ascertain whether there is a systematic error in different regions of the response surface. Useful methods of examination and analysis of residuals, in the context of predictive microbiology, were presented by Bratchell *et al.* (1990). If there are no trends in the pattern of the residuals, however, the bias factor will be a reliable measure.

The bias and accuracy factors provide an objective summary of the performance of predictive models in food microbiology. They are insufficient on their own because the bias factor, as an average, may obscure systematic deviations between predicted and observed responses in one part of the response surface if they are 'balanced' by deviations in another part of the response surface. Such behaviour might be signalled by a larger accuracy factor, but it is still important to plot the predicted and observed values to guard against such systematic deviations. Similarly, higher values of m in eqn 4 may prove to generate more useful indices of accuracy. Nonetheless, though imperfect, the bias and accuracy factors are suggested as a first step towards the development of an objective and useful definition of the term 'validated model'.

ACKNOWLEDGEMENTS

The author gratefully acknowledges the support provided to him by the Rural Industries Research and Development Corporation and the Meat Research Corporation, and thanks Drs J. Baranyi, D. Ratkowsky and P. Dalgaard for helpful discussions, and Dr Isabel Walls of the National Food Pro-

cessors Association for permission to use the data presented in Table 3.

REFERENCES

Adair, C., Kilsby, D.C. and Whittall, P.T. (1989) Comparison of the Schoolfield (non-linear Arrhenius) model and the Square Root model for predicting bacterial growth in foods. *Food Microbiology* **6**, 7–18.

Alber, S.A. and Schaffner, D.W. (1992) Evaluation of data transformations used with the square root and Schoolfield models for predicting bacterial growth rate. *Applied and Environmental Microbiology* **58**, 3337–3342.

Banwart, G.J. (1989) *Basic Food Microbiology*. New York: Van Nostrand Reinhold.

Bhaduri, S., Turner-Jones, C.O., Buchanan, R.L. and Philips, J.G. (1994) Response surface model of the effect of pH, sodium chloride and sodium nitrite on growth of *Yersinia enterocolitica* at low temperatures. *International Journal of Food Microbiology* **23**, 333–343.

Bratchell, N., McClure, N.J., Kelly, T.M. and Roberts, T.A. (1990) Predicting microbial growth: graphical methods for comparing models. *International Journal of Food Microbiology* **11**, 279–288.

Broughall, J.M., Anslow, P. and Kilsby, D.C. (1983) Hazard analysis applied to microbial growth in foods: development of mathematical models describing the effect of water activity. *Journal of Applied Bacteriology* **55**, 101–110.

Buchanan, R.L., Smith, J.L., McColgan, C., Marmer, B.S., Golden, M. and Dell, B. (1993) Response surface models for the effects of temperature, pH, sodium chloride, and sodium nitrite on the aerobic and anaerobic growth of *Staphylococcus aureus* 196E. *Journal of Food Safety* **13**, 159–175.

Cole, M.B., McClure, P.J., Stephens, P.J. and Davies, K.W. (1994) Model validation (and confidence in models) – an industry perspective. Abstract 102, *Abstracts of Papers to be Presented at the 81st Annual Meeting of IAMFES*, San Antonio, TX, USA, 31 July–3 August, 1994.

Draper, N.R. and Smith, H. (1981) *Applied Regression Analysis*, 2nd edn. New York: John Wiley.

Duh, Y.-H. and Schaffner, D.W. (1993) Modeling the effect of temperature on the growth rate and lag time of *Listeria innocua* and *Listeria monocytogenes*. *Journal of Food Protection* **56**, 205–210.

Gibson, A.M., Bratchell, N. and Roberts, T.A. (1988) Predicting microbial growth: growth responses of salmonellae in a laboratory medium as affected by pH, sodium chloride and storage temperature. *International Journal of Food Microbiology* **6**, 155–178.

Göckler, L., Notermans, S. and Kramer, J. (1988) Production of enterotoxins and thermonuclease by *Staphylococcus aureus* in cooked noodles. *International Journal of Food Microbiology* **6**, 127–139.

Jay, J.M. (1992) *Modern Food Microbiology*, 4th edn. New York: Van Nostrand Reinhold.

McClure, P.J., Baranyi, J., Boogard, E., Kelly, T.M. and Roberts, T.A. (1993) A predictive model for the combined effect of pH, sodium chloride and storage temperature on the growth of *Brocothrix thermosphacta*. *International Journal of Food Microbiology* **19**, 161–178.

McMeekin, T.A., Chandler, R.E., Doe, P.E., Garland, C.D., Olley, J., Putro, S. and Ratkowsky, D.A. (1987) Model for the combined effect of temperature and water activity on the growth rate of *Staphylococcus xylosus*. *Journal of Applied Bacteriology* **62**, 543–550.

McMeekin, T.A., Olley, J., Ross, T. and Ratkowsky, D.A. (1993) *Predictive Microbiology: Theory and Application*. Taunton, UK: Research Studies Press.

Rajkowski, K.T., Calderone, S.M and Jones, E. (1994) Effect of polyphosphate and sodium chloride on the growth of *Listeria monocytogenes* and *Staphylococcus aureus* in ultra-high temperature milk. *Journal of Dairy Science* **77**, 1503–1508.

Ratkowsky, D.A., Ross, T., McMeekin, T.A. and Olley, J. (1991) Comparison of Arrhenius-type and Belehrádek-type models for prediction of bacterial growth in foods. *Journal of Applied Bacteriology* **71**, 452–459.

Ratkowsky, D.A., Ross, T., Macario, N., Dommett, W.T. and Kamperman, L. (1996) Choosing probability distributions for modelling generation time variability. *Journal of Applied Bacteriology* **80**, 130–137.

Ross, T. (1993) A philosophy for the development of kinetic models in predictive microbiology. PhD Thesis, University of Tasmania, Hobart, Australia.

Schaffner, D.W. (1994) Application of statistical bootstrapping technique to calculate growth rate variance for modelling psychrotrophic pathogen growth. *International Journal of Food Microbiology* **24**, 309–314.

Sutherland, J.P., Bayliss, A.P. and Roberts, T.A. (1994) Predictive modelling of the growth of *Staphylococcus aureus*: the effects of temperature, pH and sodium chloride. *International Journal of Food Microbiology* **21**, 217–236.

Troller, J.A. and Stinson, J.V. (1975) Influence of water activity on growth and enterotoxin production by *Staphylococcus aureus* in foods. *Journal of Food Science* **40**, 802–804.

Walker, S. and Jones, J. (1994) Microbiology modelling and safety assessment. *Food Technology International, Europe*, 25–29.

Wijtzes, T., McClure, P.J., Zwietering, M.H. and Roberts, T.A. (1993) Modeling bacterial growth of *Listeria monocytogenes* as a function of water activity, pH and temperature. *International Journal of Food Microbiology* **18**, 139–149.

Yotis, W. and Teodoro, R. (1957) The growth and resistance of *Staphylococcus aureus* in milk and broth substrates. *Journal of Food Science* **30**, 351–358.

Zwietering, M.H., Jongenburger, I., Rombouts, F.M. and van't Riet, K. (1990) Modeling of the bacterial growth curve. *Applied and Environmental Microbiology* **56**, 1875–1881.

Zwietering, M.H., Cuppers, H.G.S.M., de Witt, J.C. and van't Riet, K. (1994) Evaluation of data transformations and validation of a model for the effect of temperature on bacterial growth. *Applied and Environmental Microbiology* **60**, 195–203.