Available online at www.sciencedirect.com
ScienceDirect
LWT - Food Science and Technology 41 (2008) 2070–2078
LWT
www.elsevier.com/locate/lwt

Risks and pitfalls of sensory data analysis for shelf life prediction: Data simulation applied to the case of coffee

S. Guerra [a], C. Lagazio [b,*], L. Manzocco [c], M. Barnabà [a], R. Cappuccio [a]

[a] *Illycaffè spa – Sensory Lab, via Flavia 110, 34147 Trieste, Italy*
[b] *Dipartimento di Scienze Statistiche, Università degli Studi di Udine, via Treppo 18, 33100 Udine, Italy*
[c] *Dipartimento di Scienze degli Alimenti, Università degli Studi di Udine, via Marangoni 97, 33100 Udine, Italy*

Received 5 June 2007; received in revised form 10 January 2008; accepted 21 January 2008

Abstract

Shelf life determination by means of sensory analysis is thought to be of paramount importance even in case of a microbiologically stable food. Several approaches are found in literature, both in terms of data collection and data processing. Whatever method is used, the subjectivity in the choice of some parameters for data collection and analysis can deeply influence the final result. We put in evidence some typical pitfalls that the researcher should avoid when planning the test and analysing data. A comparison between the most utilized techniques in sensory data processing for shelf life prediction is reported, taking as a *fil rouge* the case of coffee. In particular, a non-linear regression, a logistic regression and a survival models were applied to simulated data frames of coffee. We evaluated the influence of the choice of acceptability limits, as well as the effect of data variability and we found out that they strongly influence predictions, as well as the panel and the batch of product do. We suggest that in case of microbiologically stable food, like coffee, shelf life is not univocal and it is a choice of the company or the researcher, rather than the result of the interaction between product and consumer.

Keywords: Shelf life; Sensory analysis; Logistic regression; Weibull model; Coffee

1. Introduction

Sensory evaluation is a very important methodology for shelf life prediction of microbiologically stable products (Hough, Langohr, Gomez, & Curia, 2003). In literature the problem is faced with different approaches, not only in the way tests are carried out, but also in terms of data processing.

The descriptive method is based on an intensity scale relative to a sensory attribute strictly correlated to product's degradation (Al-Kadamany, Khattar, Haddad, & Toufeili, 2003; Cappuccio, Full, Lonzarich, & Savonitti, 2001; Fritsch, Hofland, & Vickers, 1997; Grosso & Resurreccion, 2002; Hough, Puglieso, Sanchez, & Mendes Da Silva, 1999; Nielsen, Stapelfeldt, & Skibsted, 1997; O'Connor-Shaw, Roberts, Ford, & Nottingham, 1994; Rustom, Lopez-Leiva, & Nair, 1996; Vallejo-Cordoba & Nakai, 1994). Collected data are generally processed using ANOVA and regression analysis. Shelf life is defined as the time by which the intensity of the attribute reaches a selected value, called acceptability limit (Meilgaard, Civille, & Carr, 1999). The most difficult problem is to find a simple mathematical relation able to describe the evolution of the chosen attribute as a function of storage time. Moreover, the underlying assumption is that sensory data analysis can predict a product's shelf life, which might not be the case.

Another widely used methodology in data processing of sensory shelf life studies is survival analysis (Al-Kadamany et al., 2003; Cardelli & Labuza, 2001; Duyvesteyn, Shimoni, & Labuza, 2001; Gacula, 1975; Gacula & Kubala, 1975; Gacula & Singh, 1984; Gimenez et al., 2007; Hough, Garitta, & Sanchez, 2004; Hough et al., 2003; Hough et al., 1999; Schmidt & Bouma, 1992; Wittinger & Smith, 1986). This technique studies the characteristics of the distribution of the time variable, i.e. the time passed before a specific event is

* Corresponding author.
E-mail address: lagazio@dss.uniud.it (C. Lagazio).

0023-6438/$34.00
doi:10.1016/j.lwt.2008.01.011

observed. In sensory studies for shelf life prediction, the event is represented by a response of unacceptability. The variable that describes the event in survival analysis must be binary, since only two states are possible (acceptable or unacceptable food product) and the change of status is irreversible. No matter how the sensory data are collected, their transformation in binary way is always possible, once acceptability limit is selected.

Though seldom used in sensory evaluation (Vaisey-Genser et al., 1994), also logistic regression (Hosmer & Lemeshow, 2000) is suitable for shelf life estimation. The model describes the probability of observing a failure as a function of storage time in a regression framework. Shelf life is usually estimated as the time by which the probability of unacceptability is equal to 50%.

Since the choice of data processing technique and of acceptability limit is subjective, it would be interesting to evaluate how different choices can influence shelf life prediction.

1.1. Aim of the paper

Coffee shelf life can be interpreted as the consequence of coffee staling, that is a change in the aroma profile due to deterioration processes, like loss of low boiling compounds and oxidation reactions. Storage temperature, moisture, penetration of oxygen in the package and loss of volatiles through diffusion are crucial for staleness (Clinton, 1980; Hinman, 1991; Leino, Kaitaranta, & Kallio, 1992; Nicoli, Innocente, Pittia, & Lerici, 1993).

Sensory methods are extremely important in determining coffee shelf life: an inaccurate estimation by a company can lead as a result to customers' dissatisfaction, changing their smile while tasting a cup into a grimace, and therefore causing complaints and quality assurance problems.

In literature, the problem of coffee shelf life has been often discussed from a chemical point of view (Czerny & Schieberle, 2001; Grosch, 1999, 2001; Holscher & Steinhart, 1992; Kallio, Leino, Koullias, Kallio, & Kaitaranta, 1990; Sanz, Pascual, Zapelena, & Cid, 2001; Shimoda & Shibamoto, 1990; Steinhart & Holscher, 1991; Vitzthum & Werkhoff, 1978, 1979). In these works, the evolution of some volatile compounds of coffee with storage time is studied using different techniques. As a result some coffee "freshness indices" are proposed, usually a ratio of two volatile compounds, whose value changes with time. No straightforward relation is discovered between a molecule, or a set of molecules, and shelf life, and therefore the study of a so-called "freshness index" is not representative of shelf life when not compared to sensory results.

From a sensory point of view, the end of coffee shelf life is a consequence of the development of a "stale" note (Cappuccio et al., 2001; Cardelli & Labuza, 2001; Clinton, 1980). We can assume that the cause of the rejection can be imputed only to one sensory attribute, that is "staleness". This assumption cannot be generalised to other food categories where a set of variables are involved, textural (e.g. in bread), aromatic (a set of off-flavours like in milk) or even colour related.

In this paper, we are going to put in evidence a number of pitfalls, which the researcher or practitioner should avoid, in order to achieve meaningful results, and which are often neglected.

We will do that by means of a careful review of the literature, pointing out the risks related to different choices in experimentation planning, data collection and data analysis. As far as data processing is regarded we will also compare different techniques by means of simulated data, with different data variability, built on an ideal profile. This profile is referred to the evolution of the perception of a stale note as a function of storage time of coffee, once its package is open, and it was obtained according to previous experimentations (Guerra, 2005). The simulated datasets were built as if they were obtained by a descriptive test. We will evaluate how the choice of acceptability limits influences shelf life value with regard to each processing technique and each scenario.

2. Materials and methods

2.1. Scenarios with different data variability

In order to evaluate how data variability influences shelf life estimates, 3000 simulated datasets were built. An ideal profile of coffee staleness evolution as a function of storage time was built, as if it were obtained by a descriptive test, based on the results of previous experiments (Guerra, 2005). In those tests, 12 trained assessors evaluated the attribute "stale" ("rancido" in Italian) on a nine points discrete scale with semantic anchors (1 = not stale, 9 = extreme stale), using a complete balanced block design with two replicates, at eight storage times (0, 20, 45, 55, 65, 70, 80, 100 days after opening). The panel was trained using five reference samples created in order to obtain a precise and repeatable staleness level (Cappuccio, Teixeira, & Teixeira, 2006). The results of this experiment showed that the evolution of the stale note as a function of storage time follows a sigmoidal trend. This profile is here assumed as the ideal one for the creation of the simulated scenarios.

On this base, 3000 simulated profiles were randomly created: 1000 with low, 1000 with medium and 1000 with high variability in the assessor's judgements (Fig. 1). In order to create the distributions, the frequencies of scores have been set. For example in the case of low variability (Fig. 1a), we supposed that the judges provide very homogeneous results, with at most 1 point difference from the expected one (on a 1–9 scale), with decreasing probability. These probabilities would be 0.5 for the expected score and 0.25 for the next and previous ones. In case of medium variability (Fig. 1b), the probabilities will be 0.0417, 0.0833, 0.2083, 0.3333, 0.2083, 0.0833, and 0.0417, respectively. From these probability distributions random numbers have been then extracted, thus being able to perform 1000 data frames per distribution. The transformation of the nine point scaled data into binary ones (required by logistic regression and survival models) was made considering as unacceptable all the samples that received a score higher than a chosen cut-off value, called acceptability

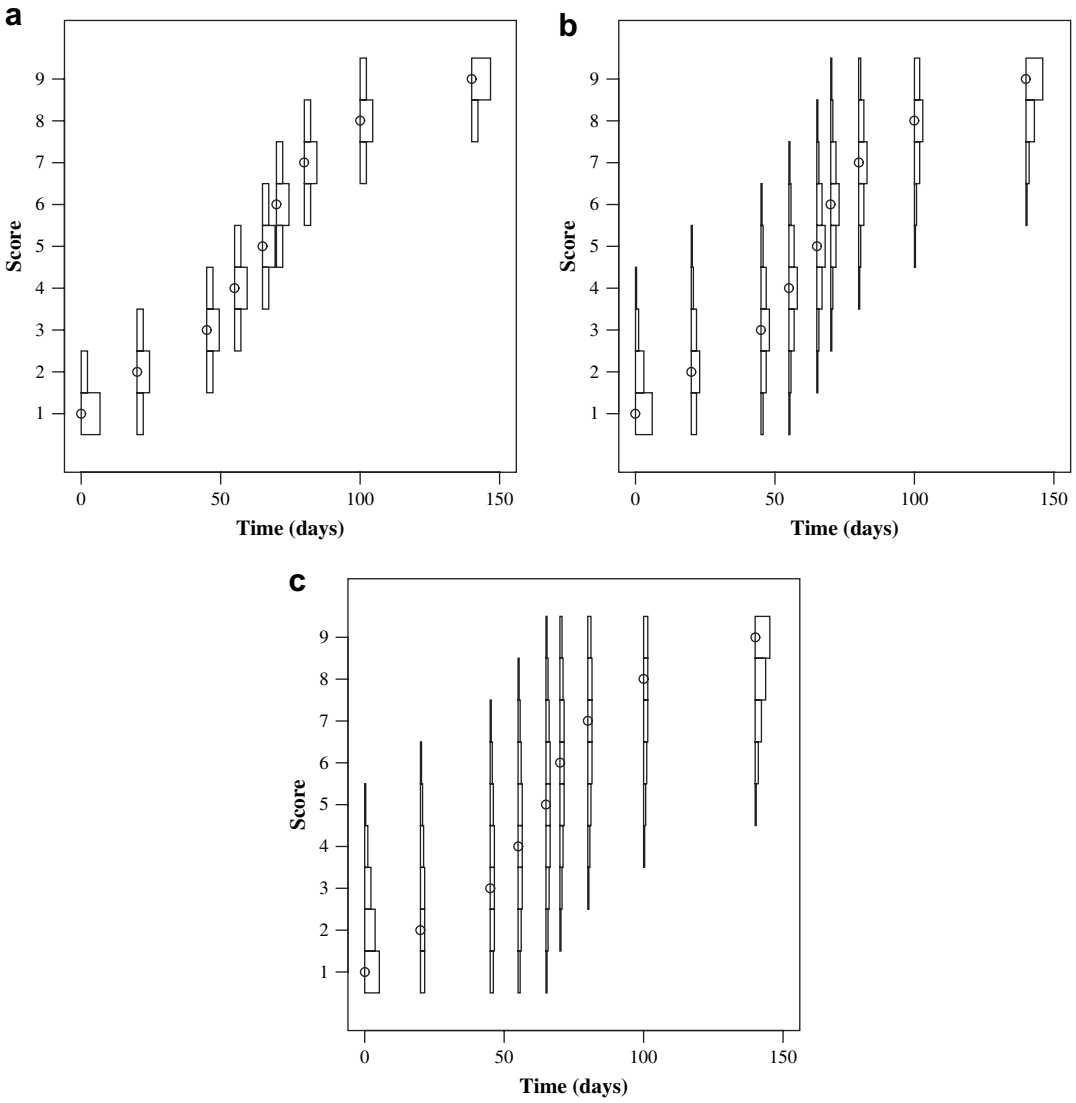

Fig. 1. Probability distributions used for the creation of the simulated datasets at (a) low variability, (b) medium variability and (c) high variability. Rectangle height is proportional to the probability assigned to each score, the circle indicates the median score.

limit. In this case two different cut-off values were chosen, according to previous results (Guerra, 2005): 2 and 3 on the nine points scale, corresponding to ''barely perceivable'' and ''perceivable'' on the staleness scale.

Finally, we point out that we have generated (and consequently also analysed) the scores as if they were completely independent, without taking into account any judge effect. When analysing real data, however, ignoring this effect can cause an underestimation of the standard errors of estimates.

2.2. Data analysis

Shelf life was estimated for each simulated data set using three different models: non-linear regression, logistic regression and survival analysis based on Weibull distribution. Other distributions, which can be found in the literature, like exponential and log-normal (Gimenez et al., 2007), have a priori been discarded, because of lack of physical meaning.

2.2.1. Non-linear regression

We supposed that staleness (s) follows a sigmoidal profile as a function of the logarithm of storage time (t) and so we chose a sigmoidal regression model:

$$s = \frac{8}{1 + \exp(-b + (\log(t) - c))} + 1 \quad (1)$$

where b and c are regression parameters to be estimated and 1 and 8 come from the limit of the function for $\log(t)$ close to 0 or ∞ in a nine point scale. The logarithm of time was used to avoid negative estimates of shelf life.

Shelf life value was obtained considering three different acceptability limits: 2, 2.5 and 3.5 on the chosen nine points scale; a score of 2 means that the staleness note is ''barely perceivable'' and a score of 3 means that the staleness is ''perceivable''.

2.2.2. Survival analysis

In survival analysis, time to failure is a random variable, and is therefore characterized by a cumulative density function (cdf, giving the probability of observing a value of time to failure lower or equal to t), or correspondingly, by the so-called survival function (probability of surviving after time t). From these, also the probability density function (pdf), the hazard function and the cumulative hazard function can be derived (Lawless, 1982).

A widely used survival model is the Weibull model, that was applied on the simulated data. The Weibull distribution (Weibull, 1951) presents a survival function characterized by two constants, the shape parameter (λ) and the scale parameter (ν) (Breyfogle, 1992).

$$S(t;\ \lambda, \nu) = \exp\left[\left(-\frac{t}{\lambda}\right)^{\nu}\right] \tag{2}$$

Parameters were estimated by maximum likelihood method. Shelf life values were obtained using three different cdf values, 0.1, 0.3 and 0.5, corresponding to a 10, 30 or 50% probability of observing a lower failure time.

2.2.3. Logistic regression

The logistic model studies the evolution of the probability of a sample being judged unacceptable (π), as a function of the logarithm of storage time (t), as reported in the following equation:

$$\ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta \log(t) \tag{3}$$

where α and β are the regression parameters.

The model was estimated using $\log(t)$ as explanatory variable. Shelf life was defined as time by which the probability of unacceptability value was equal to 0.1, 0.3 and 0.5.

2.3. Data summaries

Results of the analyses on the simulated datasets were summarised by means of box-plots (Tukey, 1970), a powerful tool to show the distribution characteristics of a quantitative variable and to compare them across different groups. Each box-plot refers to the shelf life estimates obtained with a specific combination of estimation method, cut-off and probability level. The extremes of the box correspond to the first (low) and third (high) quartile of the distribution of the results, the line in the box indicates the median. The whiskers extend from the quartiles to the lowest and highest observed values. The position of the box allows us to compare the central tendency of estimates while the width of the box and the length of the whiskers are informative about variability (the wider the box, the higher the variability).

2.4. Computation

Simulations and data analysis were carried out using R (Venables, Smith, & the R Development Core Team, 2007).

3. Results and discussion

The following results and considerations come both from a critical literature review and from the analysis of the different considered scenarios.

3.1. Risks in the choice of the test

When performing a shelf life test, a consumer or expert approach can be used. In the case of a panel of experts, as we stated before, one or more sensory attributes are evaluated, assuming that the prediction of the evolution of such attributes with time will determine the acceptability of the product. This assumption is all but straightforward, but in the case of coffee it can be valid, since the only time-related sensory variable, which can possibly lead to a rejection is staleness. On the other hand, the consumer approach brings about several problems, like the inconsistency of their judgements (Hough et al., 2003), the variability of the result when different consumer panels are used (Gimenez et al., 2007) and organisational problems for a company, since a consumer cannot evaluate more than three or four samples in one session.

3.2. Pitfalls in the choice of the model

Sensory datasets are generally incomplete, because all the observations are taken at fixed times and then times to failure cannot be observed exactly (Blischke & Murthy, 2000; Hough et al., 2003). The mechanism that prevents precise observation of times to failure is called censoring. If it is neglected a bias in shelf life estimation will be probably obtained, especially in case of staggered designed experiments (Gacula, 1975).

In order to obtain an estimate of the distribution of time to failure, non-parametric or parametric methods can be applied. Non-parametric methodologies are used in order to determine the shape of the survival (or hazard) function without assuming any particular distribution (Lawless, 1982). For example, the Kaplan–Meier technique estimates the value of the survival function at each sampling time, assuming a constant value for the function on unexplored time intervals. However, this kind of analysis is not useful when all the data are censored, and so it is unsuitable for sensory analysis.

The alternative way consists of the application of parametric techniques, which allow obtaining a specific characterisation and a parametric representation of all the functions that describe the distribution of time to failure and are based on the definition of constants that are specific of the adopted model. Many models have been used in the literature, for example exponential, log-normal or Weibull. Exponential and log-normal models are not adequate for shelf life studies because in this context they lack physical meaning: in fact the hazard function associated to the former is constant over time, while in the second model it shows a peak corresponding to the earliest values of the time variable. Both these shapes are incompatible with a food product's staling, since the hazard function is expected to increase with storage time (Gacula & Kubala, 1975; Gacula & Singh, 1984). This is why the

Table 1
Shelf life values predicted for each considered model, scenario and cut-off (acceptability limit)

Applied model	Shelf life values for low variability data (days)	Shelf life values for medium variability data (days)	Shelf life values for high variability data (days)
Logistic regression (log scale); cut-off = 2, probability level = 0.1			
Median	16	14	12
First quartile	15	12	10
Third quartile	18	16	14
Logistic regression (log scale); cut-off = 2, probability level = 0.3			
Median	23	22	21
First quartile	22	20	19
Third quartile	25	24	23
Logistic regression (log scale); cut-off = 2, probability level = 0.5			
Median	27	28	28
First quartile	26	26	25
Third quartile	29	30	30
Logistic regression (log scale); cut-off = 3, probability level = 0.1			
Median	42	27	18
First quartile	41	24	16
Third quartile	43	32	22
Logistic regression (log scale); cut-off = 3, probability level = 0.3			
Median	47	38	31
First quartile	46	35	29
Third quartile	48	41	34
Logistic regression (log scale); cut-off = 3, probability level = 0.5			
Median	49	44	40
First quartile	48	42	37
Third quartile	50	47	43
Non-linear regression (log scale); cut-off = 2			
Median	37	35	31
First quartile	36	34	29
Third quartile	38	37	34
Non-linear regression (log scale); cut-off = 2.5			
Median	42	40	37
First quartile	41	39	34
Third quartile	43	42	39
Non-linear regression (log scale); cut-off = 3.5			
Median	50	49	47
First quartile	50	48	45
Third quartile	51	50	49
Weibull model; cut-off = 2, probability level = 0.1			
Median	15	12	10
First quartile	13	10	8
Third quartile	17	14	12
Weibull model; cut-off = 2, probability level = 0.3			
Median	25	24	22
First quartile	23	21	19
Third quartile	27	26	25
Weibull model; cut-off = 2, probability level = 0.5			
Median	31	31	31
First quartile	29	28	28
Third quartile	33	33	34
Weibull model; cut-off = 3, probability level = 0.1			
Median	40	26	17
First quartile	39	23	15
Third quartile	42	29	20
Weibull model; cut-off = 3, probability level = 0.3			
Median	47	39	34
First quartile	46	37	31
Third quartile	48	42	37
Weibull model; cut-off = 3, probability level = 0.5			
Median	50	46	43
First quartile	49	44	41
Third quartile	51	48	46

Weibull distribution has been chosen for our discussion. From the estimates of the parameters based on survival data it is possible to calculate the quantiles of the time distribution. The reason why in shelf life studies a probability of acceptability equivalent to 50% (i.e. the median time) is often chosen is that if the shape parameter is large enough, the pdf tends to be symmetric, and the 50th percentile coincides with the mean value (Gacula & Kubala, 1975).

A choice of paramount importance in the estimation of the parameters is the hazard function; there has been a tendency to use the hazard value $h(t)$ (expressed in percentage) for each failure time from the expression:

$$h(t) = \frac{100}{k}$$

where k is the reverse rank assigned to each termination time (failure as well as censored) (Gacula & Kubala, 1975). Unfortunately, this method can be used only when the occurrence of more than one event (failure or withdrawal) in the same time has a negligible probability. Furthermore, only right censored data can be managed in this way. Therefore, despite it is found in the literature (Cardelli & Labuza, 2001; Duyvesteyn et al., 2001), the application of this hazard estimate is not adequate for sensory data analysis. This does not mean that Weibull is not suitable for shelf life estimation, but the problem of this method consists in the necessity of paying attention to censoring definition. A wrong definition of censoring leads necessarily to a wrong shelf life prediction. Nowadays, the use of modern statistical packages can solve this problem; nonetheless, researchers have to be aware of the consequences of all possible choices. In last years, more sophisticate and adequate techniques for parameter estimation, namely maximum likelihood (Kalbfleisch & Prentice, 1980; Lawless, 1982), were made available also to practitioners by the development of suitable and user-friendly software.

Considering the non-linear regression analysis, it is important to underline that the choice of a sigmoidal model and of the relative equation is necessarily arbitrary. But, since the underlying physical–chemical phenomenon is not completely known, whatever model would be arbitrarily chosen. In this case, the model was chosen according to suggestions from the literature (Breslin, 2001) and previous experimentations (Guerra, 2005).

It is important to point out that the model was applied on all the data. The application of the regression analysis on the median values (or on the mean ones), though widespread, is incorrect, because it causes an underestimation of the variability. In fact, if a unique value replaces 12 observations at each sampling time, the variability falls necessarily down. Thus, the most correct way for applying a non-linear regression model is to work on all the scores.

The advantage of non-linear regression is that it is not necessary to transform data in a binary way. In fact, the model is not built on probability values, but rather on the scores given by the judges. Thus, this method works like a calibration, in fact observations are taken at well-defined times in order to determine how the profile of the scores evolves with time. Then, given a score, it is possible to go back to the corresponding time. Unfortunately this method brings about a logical problem, since time is considered as explicative variable, while actually time is the variable that has to be estimated.

Another problem to be cautiously faced is the choice of the acceptability limit. This value is often calculated as the mean of a number of acceptability evaluations given on standard samples that are characterized by a well-known intensity of the sensory attribute. So it is possible (as happened in this specific case) that the chosen limit is not an integer number. In this way scores are considered as a continuous variable, even if they are not. This conceptual problem can be overcome working on binary scores and on probabilities.

Eventually, a comment should be done on the nature of the time variable. We applied the non-linear regression and the logistic models using a logarithmic form for the time variable. We recommend to work in the second way, since the use of non-logarithmic times can lead to negative shelf life values, especially in case of high variability data.

3.3. Variability due to the choice of the cut-off limit or the noise in the data

In Table 1 and in Figs. 2–4 results are given for the three considered models, scenarios and cut-offs (acceptability limits). Table 1 lets us notice some clear trends. Non-linear regression seems to be unaffected by the change in the parameters, logistic regression and Weibull models give very similar predictions, both in terms of mean value and of variability of estimates. A comparison between these models and non-linear regression is quite difficult because non-linear regression is not built on probability values. The comparison between logistic regression and survival analysis puts in evidence that the choice of the model does not affect the final result, when the other parameters are set: the differences are small when compared with sampling variability, always within the dispersion.

Also the variability in the data (that is the degree of difference in the assessors' judgement) does not affect considerably the result, and depends on the value of the probability of unacceptability. A high probability level (0.5) leads to differences within 20% for both models, whereas a low probability level (0.1) leads to shelf life differences of 50% when comparing a panel with low and high variability. Anyway, the advantage of working with a trained panel leads to the possibility of obtaining data affected by lower variability, and therefore more precise shelf life predictions. Data variability affects shelf life estimates not only in terms of precision (width of the boxplot), but also in terms of accuracy (position of the centre of the box): indeed, as variability increases, shelf life estimates tend to decrease. In case of consumer studies, the panel should be carefully recruited with regard to the type of product, and large enough to allow a correct data analysis even in case of elimination of part of the panel because of lack of consistency.

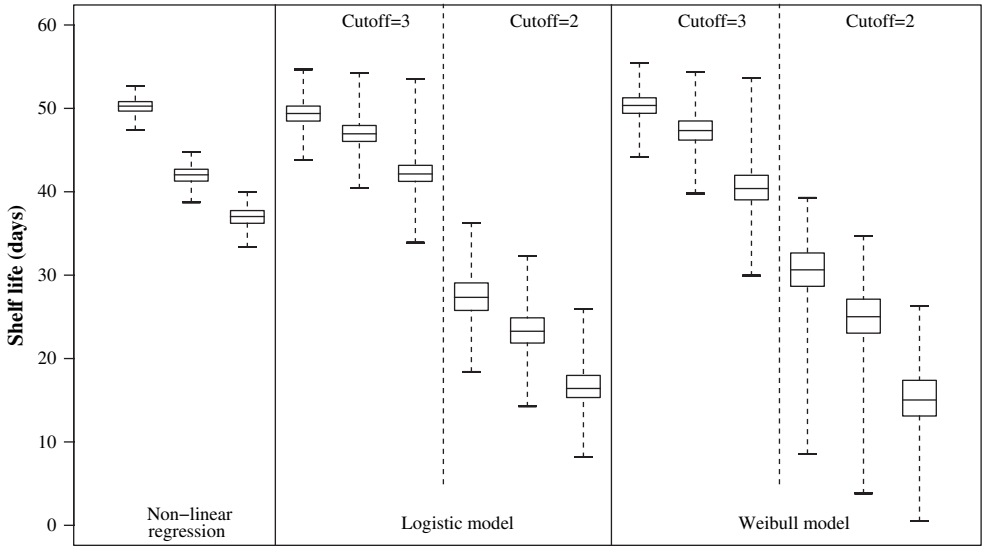

Fig. 2. Box-plot relative to data at low variability. Non-linear regression: the three boxes show predictions obtained considering, respectively, a score of 3.5, 2.5 and 2 as acceptability limit. Logistic model: the first three boxes are obtained with a cut-off = 3 and a probability level of, respectively, 0.5, 0.3 and 0.1. The second three boxes represent prediction obtained with cut-off = 2 and a probability level of, respectively, 0.5, 0.3 and 0.1. Weibull model: the first three boxes are obtained with a cut-off = 3 and a probability level of, respectively, 0.5, 0.3 and 0.1. The second three boxes represent prediction obtained with cut-off = 2 and a probability level of, respectively, 0.5, 0.3 and 0.1.

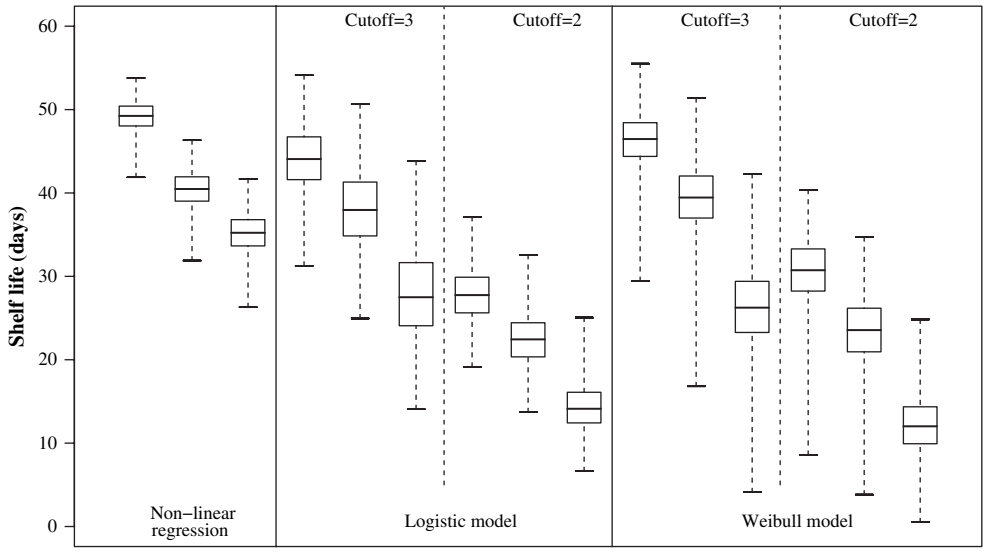

Fig. 3. Box-plot relative to data at medium variability. Non-linear regression: the three boxes show predictions obtained considering, respectively, a score of 3.5, 2.5 and 2 as acceptability limit. Logistic model: the first three boxes are obtained with a cut-off = 3 and a probability level of, respectively, 0.5, 0.3 and 0.1. The second three boxes represent prediction obtained with cut-off = 2 and a probability level of, respectively, 0.5, 0.3 and 0.1. Weibull model: the first three boxes are obtained with a cut-off = 3 and a probability level of, respectively, 0.5, 0.3 and 0.1. The second three boxes represent prediction obtained with cut-off = 2 and a probability level of, respectively, 0.5, 0.3 and 0.1.

As regards the value of the probability of unacceptability, literature tells us that the median time is most often chosen. That means that if 100 items are stored for a period of time equal to shelf life, we expect that 50 of them have already failed at that period and 50 are still surviving. The choice of this level is arbitrary, and shelf life estimate is strongly influenced by it. So it is very important to evaluate and to justify every choice. Table 1 suggests that the choice of the probability level can easily affect the final value by some 20% (e.g. 23 and 27 days in case of logistic regression with low variability data, cut-off value of 2 and comparing probability levels of 0.3 and 0.5). The differences become considerable for low cut-off limits, reaching 100% for the Weibull model (e.g. 15 and 31 days in case of Weibull model with low variability data, cut-off value of 2 and comparing probability levels of 0.1 and 0.5).

Finally, the choice of the acceptability limit affects the result dramatically. A difference of only 1 point on a nine point scale (2 or 3 in our case), can lead to final results which differ by 160% in case of low probability level, irrespective of the method. That is a difference of more than 20 days (16 and

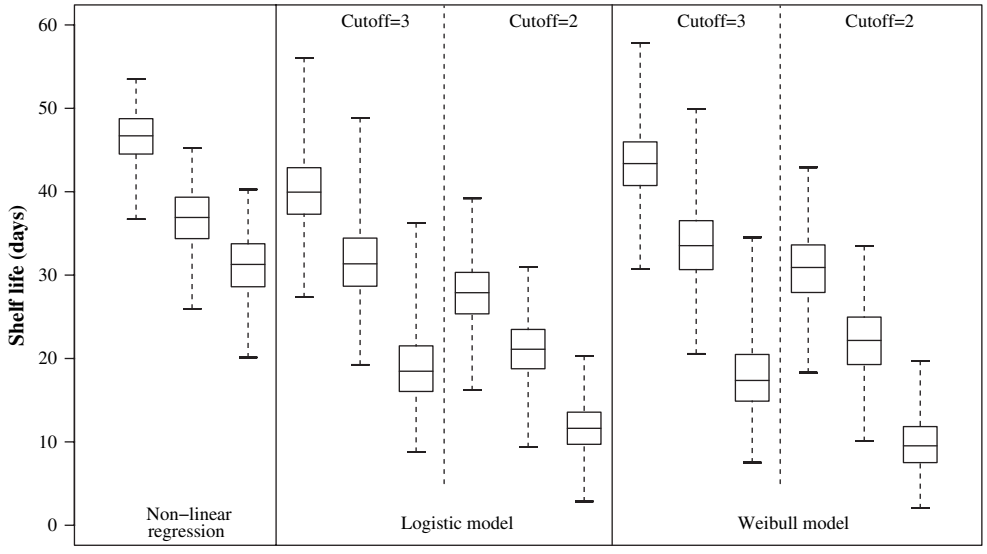

Fig. 4. Box-plot relative to data at high variability. Non-linear regression: the three boxes show predictions obtained considering, respectively, a score of 3.5, 2.5 and 2 as acceptability limit. Logistic model: the first three boxes are obtained with a cut-off = 3 and a probability level of, respectively, 0.5, 0.3 and 0.1. The second three boxes represent prediction obtained with cut-off = 2 and a probability level of, respectively, 0.5, 0.3 and 0.1. Weibull model: the first three boxes are obtained with a cut-off = 3 and a probability level of, respectively, 0.5, 0.3 and 0.1. The second three boxes represent prediction obtained with cut-off = 2 and a probability level of, respectively, 0.5, 0.3 and 0.1.

42 days in case of logistic regression, 15 and 40 days in case of survival analysis) for the estimate of the same product.

We can conclude that in case of the use of a trained panel, the variability in the assessors' judgements can be handled, while the choice of the cut-off value by the researcher or the company can change the result by 160%. Also the choice of the probability level is crucial, and this choice is needed both in case of a trained panel or a consumer test.

4. Conclusions

This study demonstrates that there is not a univocal method suitable for the determination of sensory shelf life of microbiologically stable products. There is a plethora of choices, which lead to the most different results: the choice of the panel of consumers (that should adequately represent the target market); the training of the panel of experts (causing more or less variability in the data); the choice of the acceptability limit and of the probability level, which are arbitrarily selected by the researcher or by the company; the model used and the way sensory data are processed. Some of the models are patently wrong, but they have been used; most of them are no better no worse than others, but their use cannot be generalised. Therefore, it is important to pay attention to the planning of the test sessions, to the recruitment and/or training of the panel and, in the phase of data analysis, to the choice of processing methods, not to run into incorrect predictions. Given the number of arbitrary choices, we can conclude that the shelf life concept for microbiologically stable food products is more company or researcher driven than product or consumer dependent.

Acknowledgements

The Authors wish to thank the reviewers for their careful review and insightful comments, which helped us improve the paper.

References

Al-Kadamany, E., Khattar, M., Haddad, T., & Toufeili, I. (2003). Estimation of shelf-life of concentrated yogurt by monitoring selected microbiological and physicochemical changes during storage. *Lebensmittel Wissenschaft und Technologie, 36*(4), 407–414.

Blischke, W. R., & Murthy, D. N. P. (2000). *Reliability: Modelling, prediction, and optimisation*. New York: John Wiley & Sons.

Breslin, P. A. S. (2001). Human gustation and flavour. *Flavour and Fragrance Journal, 16*, 439–456.

Breyfogle, F. W. (1992). *Statistical methods for testing, development, and manufacturing*. New York: John Wiley & Sons.

Cappuccio, R., Full, G., Lonzarich, V., & Savonitti, O. (2001). Staling of roasted and ground coffee at different temperatures: combining sensory and GC analysis. In: *Proceedings of the19th international scientific colloquium on coffee*. Trieste: ASIC.

Cappuccio, R., Teixeira, A. A., & Teixeira, R. (2006). The effect of black bean, black–green bean and immature bean defects in espresso coffee: one single bean can spoil one cup. In: *Proceedings of the 21th international scientific colloquium on coffee*. Montpellier: ASIC.

Cardelli, C., & Labuza, T. P. (2001). Application of Weibull hazard analysis to the determination of the shelf-life of roasted and ground coffee. *Lebensmittel Wissenschaft und Technologie, 34*(5), 273–278.

Clinton, W. P. (1980). Consumer and expert evaluations of stored coffee products. In: *Proceedings of the 9th international scientific colloquium on coffee*. London: ASIC.

Czerny, M., & Schieberle, P. (2001). Changes in roasted coffee aroma during storage – influence of the packaging. In: *Proceedings of the 19th international scientific colloquium on coffee*. Trieste: ASIC.

Duyvesteyn, W. S., Shimoni, E., & Labuza, T. P. (2001). Determination of the end of shelf-life for milk using Weibull hazard method. *Lebensmittel Wissenschaft und Technologie, 34*(3), 143–148.

Fritsch, C. V., Hofland, C. N., & Vickers, Z. M. (1997). Shelf-life of sunflower kernels. *Journal of Food Science, 62*(2), 425–428.

Gacula, M. C. (1975). The design of experiments for shelf life study. *Journal of Food Science, 40*, 399–404.

Gacula, M. C., & Kubala, J. J. (1975). Statistical models for shelf life failures. *Journal of Food Science, 40*, 404–409.

Gacula, M. C., & Singh, J. J. (1984). Shelf life testing experiments. In M. C. Gacula, & J. J. Singh (Eds.), *Statistical methods in food and consumer research* (pp. 274–312). New York: Academic Press.

Gimenez, A., Varela, P., Salvador, A., Ares, G., Fiszman, S., & Garitta, L. (2007). Shelf life estimation of brown pan bread: a consumer approach. *Food Quality and preference, 18*(2), 196–204.

Grosch, W. (1999). Key odorants of roasted coffee: evaluation, release, formation. In: *Proceedings of the 18th international scientific colloquium on coffee*. Helsinki: ASIC.

Grosch, W. (2001). Chemistry III – volatile compounds. In R. J. Clarke, & O. G. Vitzthum (Eds.), *Coffee: Recent developments* (pp. 68–90). Oxford: Blackwell Science.

Grosso, N. R., & Resurreccion, A. V. A. (2002). Predicting consumer acceptance ratings of cracker-coated and roasted peanuts from descriptive analysis and hexanal measurements. *Journal of Food Science, 67*(4), 1530–1537.

Guerra, S. (2005). Comparazione di metodi sensoriali per la stima della shelf-life secondaria di caffè tostato. MSc thesis in Food Science and Technology, University of Udine.

Hinman, D. C. (1991). Rates of oxidation of roast and ground coffee and the effect on shelf life. In: *Proceedings of the 14th international scientific colloquium on coffee*. San Francisco: ASIC.

Holscher, W., & Steinhart, H. (1992). Investigation of roasted coffee freshness with an improved headspace technique. *Zeischift für Lebensmittel Untersuchung und Forschung, 195*, 33–38.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression*. New York: John Wiley & Sons.

Hough, G., Garitta, L., & Sanchez, R. (2004). Determination of consumer acceptance limits to sensory defects using survival analysis. *Food Quality and Preference, 15*(7–8), 729–734.

Hough, G., Langohr, K., Gomez, G., & Curia, A. (2003). Survival analysis applied to sensory shelf life of foods. *Journal of Food Science, 68*(1), 359–362.

Hough, G., Puglieso, M. L., Sanchez, R., & Mendes Da Silva, O. (1999). Sensory and microbiological shelf-life of a commercial Ricotta cheese. *Journal of Dairy Science, 82*(3), 454–459.

Kalbfleisch, J. D., & Prentice, R. L. (1980). *The statistical analysis of failure time data*. New York: John Wiley & Sons.

Kallio, H., Leino, M., Koullias, K., Kallio, S., & Kaitaranta, J. (1990). Headspace of roasted ground coffee as an indicator of storage time. *Food Chemistry, 36*(2), 135–148.

Lawless, J. F. (1982). *Statistical models and methods for lifetime data*. New York: John Wiley & Sons.

Leino, M., Kaitaranta, J., & Kallio, H. (1992). Comparison of changes in headspace volatiles of some coffee blends during storage. *Food Chemistry, 43*(1), 35–40.

Meilgaard, M., Civille, G. V., & Carr, B. T. (1999). *Sensory evaluation techniques*, (3rd ed.). Boca Raton: CRC Press.

Nicoli, M. C., Innocente, N., Pittia, P., & Lerici, C. R. (1993). Staling of roasted coffee: volatile release and oxidation reactions during storage.

In: *Proceedings of the 15th international scientific colloquium on coffee*. Montpellier: ASIC.

Nielsen, B. R., Stapelfeldt, H., & Skibsted, L. H. (1997). Early prediction of the shelf-life of medium-heat whole milk powders using stepwise multiple regression and principal component analysis. *International Dairy Journal, 7*(5), 341–348.

O'Connor-Shaw, R. E., Roberts, R., Ford, A. L., & Nottingham, S. M. (1994). Shelf-life of minimally processed honeydew, kiwifruit, papaya, pineapple and cantaloupe. *Journal of Food Science, 59*(6), 1202–1206.

Rustom, I. Y. S., Lopez-Leiva, M. M., & Nair, B. M. (1996). UHT-sterilized peanut beverages: kinetics of physicochemical changes during storage and shelf-life prediction modeling. *Journal of Food Science, 61*(1), 198–203.

Sanz, C., Pascual, L., Zapelena, M. J., & Cid, M. C. (2001). A new "aroma index" to determine the aroma quality of a blend of roasted coffee beans. In: *Proceedings of the 19th international scientific colloquium on coffee*. Trieste: ASIC.

Schmidt, K., & Bouma, J. (1992). Estimating shelf-life of cottage cheese using hazard analysis. *Journal of Dairy Science, 75*(11), 2922–2927.

Shimoda, M., & Shibamoto, T. (1990). Isolation and identification of headspace volatiles from brewed coffee with an on-column GC/MS method. *Journal of Agriculture and Food Chemistry, 38*, 802–804.

Steinhart, H., & Holscher, W. (1991). Storage-related changes of low-boiling volatiles in whole coffee beans. In: *Proceedings of the 14th international scientific colloquium on coffee*. San Francisco: ASIC.

Tukey, J. W. (1970). *Exploratory data analysis*. [Limited preliminary edition]. Reading, MA: Addison Wesley.

Vaisey-Genser, M., Malcomsom, L. J., Ryland, D., Przybylski, R., Eskin, A. M., & Armstrong, L. (1994). Consumer acceptance of canola oils during temperature-accelerated storage. *Food Quality and Preference, 5*(4), 237–243.

Vallejo-Cordoba, B., & Nakai, S. (1994). Keeping-quality assessment of pasteurized milk by multivariate analysis of dynamic headspace gas chromatographic data. 1. Shelf-life prediction by principal component regression. *Journal of Agricultural and Food Chemistry, 42*(4), 989–993.

Venables, W. N., & Smith, D. M.the R Development Core Team (2007). An introduction to R. <http://cran.r-project.org/doc/manuals/R-intro.pdf> Accessed 18.11.07.

Vitzthum, O. G., & Werkhoff, P. (1978). Aroma analysis of coffee, tea and cocoa. In G. Charalambous (Ed.), *Analysis of food and beverages: Headspace techniques* (pp. 115–133). New York: Academic Press.

Vitzthum, O. G., & Werkhoff, P. (1979). Messbare Aromaveraenderungen bei Bohnenkaffee in sauerstoffdurchlaessiger Verpackung. *Chemie Mikrobiologie Technologie der Lebensmittel, 6*, 25–30.

Weibull, W. (1951). A statistical distribution function of wide applicability. *Journal of Applied Mechanics, 18*(3), 293–297.

Wittinger, S. A., & Smith, D. E. (1986). Effect of sweeteners and stabilizers on selected sensory attributes and shelf life of ice cream. *Journal of Food Science, 51*(6), 1463–1466.