Sensory Shelf Life of Dulce de Leche

L. Garitta, G. Hough, and R. Sánchez*
Instituto Superior Experimental de Tecnología Alimentaria, (6500) Nueve de Julio, Buenos Aires, Argentina

Abstract

The objectives of this research were to determine the sensory cutoff points for dulce de leche (DL) critical descriptors, both for defective off-flavors and for storage changes in desirable attributes, and to estimate the shelf life of DL as a function of storage temperature. The critical descriptors used to determine the cutoff points were plastic flavor, burnt flavor, dark color, and spreadability. Linear correlations between sensory acceptability and trained panel scores were used to determine the sensory failure cutoff point for each descriptor. To estimate shelf life, DL samples were stored at 25, 37, and 45°C. Plastic flavor was the first descriptor to reach its cutoff point at 25°C and was used for shelf-life calculations. Plastic flavor vs. storage time followed zero-order reaction rate. Shelf-life estimations at different temperatures were 109 d at 25°C, 53 d at 37°C, and 9 d at 45°C. The activation energy, necessary to calculate shelf lives at different temperatures, was 14,370 ± 2080 cal/mol.

(Key words: dulce de leche, cutoff point, sensory failure, shelf life)

Abbreviation key: DL = dulce de leche.

Introduction

Dulce de leche (DL) is a typical Argentine dairy product prepared from milk that is concentrated by evaporation and sucrose and glucose are added (Moro and Hough, 1985). It has a characteristic color and texture that consumers consider important when evaluating overall preference (Hough et al., 1986). These attributes change during storage and can reach levels that consumers could find unacceptable in relation to the fresh product. Also during storage, off-flavors are produced, such as plastic or chemical flavors, due to the packaging material, or a burnt flavor, due to caramelization or Maillard reactions, which continue after manufacture of the product.

Generally, in published sensory shelf-life studies, the critical attributes that have been considered are product defects, such as oxidized and rancid flavor in milk (Lawless and Claassen, 1993) or oxidized flavor in sunflower oil (Ramírez et al., 2001). However, the increase of defects is not always what determines the end of the shelf life of a food product. There can be changes in levels of desirable attributes, like color or texture, that lead to product rejection. The experimental design and corresponding shelf-life calculations that consider changes in desirable effects have not been published.

It is not expected that a stored product should be exactly the same as a fresh standard, rather that the sensory differences be sufficiently small so as not to alter the acceptability of the product significantly. These sensory differences would define the failure criterion. A consumer panel is the most appropriate tool for determining when a food product reaches the end of its shelf life. To repeatedly assemble consumer panels for the multiple measurements needed during shelf-life studies would be impractical and prohibitively expensive. A trained sensory panel is much simpler to assemble, but can only measure analytical attributes such as color intensity or hardness level. How big does the color change measured with a trained panel need to be for the acceptability of the product to decrease? The answer to this question can be obtained by correlating data from a consumer panel with data obtained from a trained panel and thus acquire the sensory cutoff point. Similar correlations and sensory cutoff points have been published for sensory defects in sunflower kernels (Fritsch et al., 1997), sunflower oil (Ramirez et al., 2001), and powdered milk (Hough et al., 2002). There are no published correlations for DL defects or for variations in desirable attributes in other products.

The objectives of this research were: (1) determine the sensory cutoff points for DL critical descriptors for defective off-flavors and for changes in desirable attributes during storage, and (2) to estimate the shelf life of DL as a function of storage temperature.
Table 1. Preparation of reference samples used to train the sensory panel.

<table>
<thead>
<tr>
<th>Descriptor</th>
<th>Reference solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dark color</td>
<td>6 mL of 2% SICNA (Milán, Italy) caramel coloring were mixed into 250 g of DL(^1) previously heated to 95°C in a water bath. Total solids were adjusted to 70 Brix by evaporation at 95°C.</td>
</tr>
<tr>
<td>Crystals</td>
<td>2.5 g of sterile and sieved sand (0.8 mm sieve) were mixed into 250 g of DL previously heated to 95°C in a water bath. Total solids were adjusted to 70 Brix by evaporation at 95°C.</td>
</tr>
<tr>
<td>Floury</td>
<td>35 mL of a 20% starch solution were mixed into 250 g of DL previously heated to 95°C in a water bath. Total solids were adjusted to 70 Brix by evaporation at 95°C.</td>
</tr>
<tr>
<td>Sweet</td>
<td>30 mL of a 50% sucrose solution were mixed into 250 g of DL previously heated to 95°C in a water bath. Total solids were adjusted to 70 Brix by evaporation at 95°C.</td>
</tr>
<tr>
<td>Acid</td>
<td>6.5 mL of a 10% lactic acid solution were mixed into 250 g of DL previously heated to 95°C in a water bath. Total solids were adjusted to 70 Brix by evaporation at 95°C.</td>
</tr>
<tr>
<td>Burnt</td>
<td>4 g of sugar heated till first smoke emanation were mixed into 250 g of DL previously heated to 95°C in a water bath.</td>
</tr>
<tr>
<td>Plastic</td>
<td>250 g of DL were heated in the polystyrene pot during 24 h at 80°C. Every 6 h the DL was stirred in the pot for approximately 2 min.</td>
</tr>
<tr>
<td>Metallic</td>
<td>10 mL of a 5% ferrous sulfate solution were mixed into 250 g of DL previously heated to 95°C in a water bath. Total solids were adjusted to 70 Brix by evaporation at 95°C.</td>
</tr>
<tr>
<td>Oxidized</td>
<td>100 mL of whole pasteurized milk + 22 g of butter + 0.15 mL of a 1% copper sulfate solution + a 3 × 6 cm copper plate were placed in a 45°C oven for 96 h. 10 mL of this mixture were mixed into 250 g of DL previously heated to 95°C in a water bath. Total solids were adjusted to 70 Brix by evaporation at 95°C.</td>
</tr>
<tr>
<td>Cheese</td>
<td>15 mL of a 1% Parmesan cheese essence (GIVAUDAN, Argentina) were mixed into 250 g of DL previously heated to 95°C in a water bath. Total solids were adjusted to 70 Brix by evaporation at 95°C.</td>
</tr>
</tbody>
</table>

\(^1\)DL = Commercial dulce de leche used in this project.

**MATERIALS AND METHODS**

**Samples**

Commercial home use DL provided by a leading manufacturer in Argentina was used. The product was all from the same batch, packaged in 250-g polystyrene pots. This is standard packaging for DL in Argentina and other countries. The composition listed on the label was 7% protein, 57% carbohydrates, 6% lipids, 0.5% ashes, and 29.5% water.

Dulce de Leche is stable at room temperature. The pots were stored at 5 ± 1°C until use; it was considered that changes at this temperature were negligible compared with changes at the storage temperatures studied. To determine the cutoff point, a total of 136 pots were used, and for the shelf-life study, 90 pots were used.

**Ethical Considerations**

The Ethical Committee of our Institute concluded that all samples were adequate for testing on humans in the concentrations and quantities to be served.

**Trained Sensory Panel and Critical Descriptors**

A panel of 9 assessors was selected and trained following the guidelines of the ISO (1993) standard, including the Ishihara color test. They all had a minimum of 100 h of experience in discrimination and descriptive tests and had worked previously with DL.

For this project, assessors were initially trained to recognize the following descriptors in DL: dark color, spreadability, floury, crystals, sweet, acid, burnt, metallic, plastic, oxidized, and cheese. Most of the references for these descriptors were prepared following the IRAM 14067-9 Standard (1996). Preparation of references for dark color, spreadability, and plastic flavor are not detailed in the IRAM 14067-9 Standard (1996) and were developed separately. References are in Table 1. In the case of spreadability, where commercial brands were used, the hardest sample was presented as reference of this descriptor. Spreadability was defined as the ease in spreading the DL over a cracker-type biscuit with a knife; a high spreadability is considered easy to spread. Other descriptors are self-explanatory. To train the panel in recognizing the references, these were presented with their labels together with 4 or 5 blind-coded samples. These samples had to be correctly identified in repeated sessions to ensure that each assessor had been properly trained.

The critical descriptors used to determine the cutoff points were obtained by accelerated storage at 45°C during 30 d. Every 4 d the stored sample was compared to the fresh control sample stored at 5°C using the above descriptors. Figure 1 illustrates the ballot used by assessors during this stage of searching for critical descriptors. The descriptors that changed most during accelerated storage were plastic flavor, burnt flavor, dark color, and spreadability.
SENSORY SHELF LIFE OF DULCE DE LECHE 1603

Figure 1. Portion of the ballot used by assessors during the search of critical descriptors.

Sensory Cutoff Points

During the storage of a product like DL, sensory defects increase over time. For example, plastic flavor will be nonexistent in the fresh product and will increase to barely tolerable levels at the end of the product’s shelf life. The definition of this barely tolerable level is obtained by measuring the sensory cutoff point. The following stages are necessary to measure the cutoff point of a sensory defect: (a) preparation of a series of samples with an increasing level of the sensory defect, (b) determination of the intensity level of the samples by a trained sensory panel, (c) determination of the acceptability level of the same samples by a consumer panel, and (d) calculation of the cutoff point.

Sample preparation. To determine the sensory failure cutoff points, it was necessary to prepare samples with different levels of the descriptors to be studied.

Plastic flavor and burnt flavor stock references were prepared, as indicated in Table 1, and given scores of 100 and 80, respectively, on the 0 to 100 sensory scale. These stock references were mixed with untreated DL to obtain samples with a decreasing level of the defect. Concentrations are noted in Table 2.

For sensory defects, such as plastic flavor in DL, we suppose that all consumers will prefer the fresh product where the off-flavor is absent. For an attribute like dark color, which is typical of DL and which increases during storage, there will be consumers who prefer the lighter color of the fresh product, whereas other consumers will prefer the darker color of the stored product. There can also be consumers who would prefer a color lighter than the freshly manufactured product. Thus, for this attribute, samples with a range of colors, from lighter to darker than the fresh product, were prepared as shown in Table 2.

Table 2. Concentrations of references used to determine correlations between consumer acceptability and trained panel scores. Numbers in parentheses indicate the consensus score from the trained panel on the 0 to 100 scale.

<table>
<thead>
<tr>
<th>Concentration</th>
<th>Plastic</th>
<th>Burnt</th>
<th>Color</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>201 (10)</td>
</tr>
<tr>
<td>2</td>
<td>6 (10)</td>
<td>6 (10)</td>
<td>101 (20)</td>
</tr>
<tr>
<td>3</td>
<td>9</td>
<td>9</td>
<td>51</td>
</tr>
<tr>
<td>4</td>
<td>13</td>
<td>13</td>
<td>2.51</td>
</tr>
<tr>
<td>5</td>
<td>20 (50)</td>
<td>20 (40)</td>
<td>0.0 (50)</td>
</tr>
<tr>
<td>6</td>
<td>30</td>
<td>30</td>
<td>0.72</td>
</tr>
<tr>
<td>7</td>
<td>44</td>
<td>44</td>
<td>1.62</td>
</tr>
<tr>
<td>8</td>
<td>67</td>
<td>67</td>
<td>1.62</td>
</tr>
<tr>
<td>9</td>
<td>100 (100)</td>
<td>100 (80)</td>
<td>2.42 (100)</td>
</tr>
</tbody>
</table>

1Samples lighter than fresh dulce de leche (DL) were prepared with milk. Values indicate milliliters of milk added to 100 g of DL.

2Samples darker than fresh DL were prepared with caramel coloring (SICNA, Milan, Italy). Values indicate milliliters of 2% coloring solution added to 100 g of DL.

The spreadability descriptor was similar to dark color in the sense that some consumers would prefer a DL with less spreadability than the fresh product, others would prefer the spreadability of the fresh product, and others would prefer a DL with higher spreadability than the fresh product. During storage, spreadability of DL tends to decrease. Reference samples were initially prepared by heating the fresh product and adding milk or starch to make samples more or less spreadable, respectively. Apart from spreadability, these samples were different from normal commercial DL in color and general appearance, and it would have led to a logical error or halo effect (Meilgaard et al., 1999). Thus the trained sensory panel chose 6 samples from a range of 16 commercial brands, 3 with lower spreadability than the fresh product under study and 3 with higher spreadability than the fresh product. These 6 samples were similar in appearance to the fresh product. The brands are not identified for these references for confidentiality reasons.

Samples for sensory cutoff point determinations were prepared the day before the sensory tests and stored at 5°C. Two hours before the test, they were placed at room temperature.

Trained panel measurements. The trained sensory panel was calibrated in the critical descriptors. For plastic flavor, assessors received samples 1, 2, 5, and 9 from Table 2, identified with the letters K, A, B, and C, respectively. In the first session, assessors had to score these samples in plastic flavor intensity using a 10 cm scale ranging from 0 to 100. After discussion, a consensus score for each sample was reached, as

Journal of Dairy Science Vol. 87, No. 6, 2004
Figure 2. Scales used by sensory-trained panel to assess critical descriptors.

shown in Table 2. In successive sessions, the panel received the same 4 samples coded with 3-digit numbers and had to score them according to the consensus score, with a variation in the scale no greater than ±10 points. A similar procedure was used for calibration of the other critical descriptors. A total of twelve 45-min sessions were used to calibrate the panel.

The scales used in the training and measurement of each descriptor are shown in Figure 2. Once the panel was calibrated, samples were measured in triplicate, one descriptor at a time. For plastic flavor, burnt flavor, and spreadability, 18 g of each of the samples (see Table 2) were served in 70-mL plastic cups coded with a 3-digit number. Burnt-flavor samples were evaluated under red light to eliminate color differences between the samples. For spreadability, each sample was accompanied by a plastic knife and a cracker. For the flavor descriptors, water and peeled slices of Granny Smith apples were provided for palate cleansing between samples. For dark color evaluation, the 9 samples (Table 2) were presented in 5.0-cm diameter Petri dishes coded with 3-digit numbers, placed in a chamber equipped with artificial, daylight-type fluorescent lighting. Assessors were allowed to move the samples to adjust the viewing angle.

Consumer panel. Consumers were from the town of Nueve de Julio, Argentina. They consumed DL at least once a week, and their ages were between 18 and 50 yr. For each critical descriptor, 50 consumers were recruited. For plastic flavor, burnt flavor, and dark color, each consumer received the 9 samples corresponding to the 9 concentration levels of the descriptor (see Table 2) and evaluated these using the scales shown in Figure 3. For spreadability, consumers received 7 samples and were asked to spread DL over a cracker with a knife and evaluate it using the scale shown in Figure 3. The samples were presented monadically in random order. For plastic and burnt flavors, water and cracker-type biscuits were available. Following the test, the consumers received a bag with candy and chocolates as a reward for their participation.

Sensory cutoff point calculations. The first step in the calculations was to estimate the least significant difference in acceptability based on the consumer data and then apply the following equations:

\[ S = F - LSD = F - Z_{\alpha} \sqrt{\frac{2MSE}{n}} \text{ plastic flavor and burnt flavor} \]  

\[ I = F + LSD = F + Z_{\alpha} \sqrt{\frac{2MSE}{n}} \text{ color and consistency} \]

where:

\[ S \text{ = value below which the sensory acceptability of the stored product is significantly reduced based on the 1 to 9 acceptability scale; }\]

\[ I \text{ = value which is significantly different from ideal based on the -4 to +4 ideal-point acceptability scale; }\]

\[ F \text{ = acceptability of fresh sample; }\]

\[ Z_{\alpha} \text{ = one-tailed coordinate of the normal curve for } \alpha \text{ significance level; }\]

\[ MSE \text{ = mean square of the error derived from the analysis of variance of the consumer data.} \]
using consumer and sample as variation factors;
n = number of consumers;

Equation [1] was used for plastic and burnt flavor as we can assume that acceptability of the stored samples with these off-flavors will be below the acceptability of the fresh samples where these off-flavors are absent. Equation [2] was used for dark color and spreadability, as these attributes increase over storage time and were measured on an ideal-point scale.

For each descriptor, a linear regression of consumer acceptability (averaged over consumers) versus trained sensory panel scores (averaged over assessors) was calculated. Simple linear regressions were chosen, as they fitted the data adequately (correlation coefficients were >0.9). The cutoff point on the trained sensory-panel scale was determined by entering the S or I values (Equations [1] and [2]) in the acceptability scale and calculating the corresponding sensory coordinate from the regression equation, as shown in Figures 4 and 5. The sensory cutoff points (C values in Figures 4 and 5) are the intensity limits beyond which the product would have a reduced acceptability.

**Shelf-Life Measurements**

Pots of DL were stored at the following temperatures and times: 25°C for 200 d, 37°C for 122 d, and 45°C for 24 d. Both the 25 and 37°C samples were removed approximately every 10 d and, for the 45°C temperature, every 4 d, for sensory evaluation by the trained panel. For plastic and burnt flavors, samples were taken from the DL in contact with the polystyrene pot. Consumers usually scrape the inside of the DL pot with a spoon, so this would be the critical portion of DL as regards the plastic flavor. For dark color and spreadability, the samples were taken from the center of the pots. Coding, sample presentation, and scales used by the trained panel were the same as described above. At each session, a fresh control was presented to calibrate the panel, and a coded fresh control was also included.

Linear regressions of the critical descriptors versus storage time were used to estimate shelf life at each storage temperature. The confidence regions of these regressions were calculated using the Working-Hotelling equations (Draper and Smith, 1981). The sensory cutoff point calculated for each descriptor, as described above, was used to estimate the shelf life with corresponding confidence intervals, as illustrated in Figure 6.

**Activation Energy**

In the present study, shelf life was measured at 25, 37, and 45°C. To estimate shelf life at other storage
temperatures, the activation energy of the changes during storage is needed. A simplified approach used by many authors to calculate activation energy is to first estimate the reaction-rate constants at each temperature. Following this, and using the Arrhenius equation, activation energy is estimated from the slope of the linear regression of log (reaction rate constant) versus the inverse of absolute temperature (Labuza, 1982).

The problem with this approach is that, if there are few temperatures, as is generally the case with shelf life experiments, the confidence interval of the estimated activation energy tends to be very high.

A more rigorous approach is to estimate the activation energy from the following equation, which combines zero-order reaction rate (based on the results, it was the chosen order) with the Arrhenius model:

\[
SD = SD_0 - k_{ref} \cdot t \cdot \exp \left[ - \frac{E_a}{R} \left( \frac{1}{T} - \frac{1}{T_{ref}} \right) \right]
\]

where

- \( SD \) = sensory descriptor at time \( t \),
- \( SD_0 \) = sensory descriptor at time \( t = 0 \),
- \( k_{ref} \) = reaction rate constant at \( T_{ref} \),
- \( t \) = time,
- \( E_a \) = activation energy,
- \( R \) = gas law constant,
- \( T \) = absolute temperature, and
- \( T_{ref} \) = reference temperature (300K).

This equation is nonlinear in \( E_a \), and to calculate its parameters (\( SD_0 \), \( k_{ref} \), and \( E_a \)), nonlinear regression facilities of Genstat 6th Edition (VSN International, Oxford, UK) were used.

**RESULTS**

**Sensory Failure**

Figure 4 shows the cutoff point for plastic flavor obtained from the correlation of consumer acceptability versus plastic flavor, measured by a trained sensory panel. As expected, sensory acceptability decreased as plastic flavor increased. Burnt flavor (not shown) gave a similar regression. Table 3 summarizes the cutoff point results for plastic and burnt flavors, including the percentage of variance explained by the linear regressions of acceptability versus trained sensory-panel scores. The sensory failure cutoff points (column C in Table 3) cannot be compared directly, as the sensory scales were different. To state, for example, that consumers tolerated a higher intensity of burnt flavor than plastic flavor would imply that the burnt and plastic samples from Table 2 were somehow equivalent from a sensory point of view. Complex multimodal sensory tests would have been necessary to equate the scales. For future studies on the shelf life of DL, the cutoff points from Table 3 can be used as a guide to define sensory failure, provided the panel is trained with the same samples and scale as in the present work.

For future studies on the shelf life of DL, the cutoff points from Table 3 can be used as a guide to define sensory failure, provided the panel is trained with the same samples and scale as in the present work.

Figure 5 shows the cutoff point for the dark color descriptor obtained from the correlation of consumer acceptability versus dark color measured by a trained sensory panel. This cutoff point reflects how far the color can change during storage before the consumers say it is darker than their ideal. The DL with the ideal color had a score on the dark color trained panel scale of 49, and the product consumers found darker than their ideal had a score given by the trained panel of 67. That is, when the color measured on the trained panel scale changed by 18 points (67 to 49), consumers considered the sample as too dark in relation to their ideal. Similar considerations can be made for spreadability. Table 3 summarizes the cutoff point results for dark color and spreadability.

**Shelf Life**

At the storage temperature of 25°C, close to room temperature, the descriptor that first reached its cutoff point was plastic flavor. Thus this descriptor was chosen to determine the product’s shelf life. Dulce de leche is produced by boiling milk and sugar at approximately 100°C. Once it is ready, it is cooled to approximately 70°C and packaged in the polystyrene pots. This high packaging temperature may influence the development of plastic flavor during storage. At 37 and 45°C, the descriptor that first reached its cutoff point was dark color.

Plastic flavor versus time and log (plastic flavor) versus time were correlated to determine reaction order at 25, 37, and 45°C. At 25°C, \( R^2 \) values were 0.88 and...
Table 4. Days of sensory shelf life ±95% confidence limits for dulce de leche at different storage temperatures.

<table>
<thead>
<tr>
<th>Temperature (°C)</th>
<th>95% Lower limit</th>
<th>Estimated shelf life</th>
<th>95% Upper limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>61</td>
<td>109</td>
<td>134</td>
</tr>
<tr>
<td>37</td>
<td>39</td>
<td>53</td>
<td>66</td>
</tr>
<tr>
<td>45</td>
<td>1</td>
<td>9</td>
<td>17</td>
</tr>
</tbody>
</table>

0.85 for zero and first-order reaction rates, respectively. At 37°C, R² values were 0.89 and 0.71, and at 45°C, R² values were 0.84 and 0.69. As the fit was similar for one temperature and better for two of the temperatures, zero-order reaction rate was chosen for shelf-life calculations. Figure 6 illustrates the linear regression of plastic flavor versus storage time for DL stored at 37°C together with the shelf-life estimation using the previously calculated cutoff point. Similar regressions were obtained at 25 and 45°C. Shelf-life values and their 95% confidence intervals are noted in Table 4.

Using Equation [3], the nonlinear Arrhenius approach, the activation energy (± the 95% confidence interval) were calculated as equaling 14,370 ± 2080 cal/mol. There are no published studies on Ea values associated with plastic flavor development during storage, so it was not possible to compare this Ea value to others. Using Equation [3] shelf life at a normal room temperature of 20°C would be 146 d, approximately 5 mo. Common practice by Argentine manufacturers of DL is to allow a 6-mo shelf life, a value they should re-consider.

CONCLUSIONS

An approach to calculating cutoff points of sensory descriptors that change during storage, but that are not necessarily defects, has been developed. For DL, this approach was applied to dark color and spreadability.

For a dairy product like DL with a relatively high fat content (6%), oxidized flavor would have been expected to be critical for shelf-life estimation, but in the present study, the plastic flavor from the polystyrene package was the more critical descriptor. A number of other dairy and nondairy products are packaged in polystyrene pots, so plastic flavor should be considered in estimating their shelf life as well.

REFERENCES