Journal of Data Science **4**(2006), 233-246

Statistical Functional Modeling of Quality Changes of Garlic under Different Storage Regimes

E. T. Castano[1], E. S. Mercado[1], F. G. Leon[1], C. H. Gorrostieta[1], J. J. Chamorro[1], E. B. Vazquez[1] and V. T. Aguirre[2]
[1] *Universidad Autonoma de Queretaro and*
[2] *Instituto Tecnologico Autonomo de Mexico*

Abstract: In this paper we analyze the weight loss behaviour of Mexican garlic under different storage conditions. Garlic is an important Mexican export product. Quality losses during storage are important to understand due to cost and sale opportunity implications. Weight losses profiles for each experimental conditions, represented as functions, are modeled by means of functional linear models and hypotheses tests are performed to compare treatments. Monte Carlo sampling version of permutation tests are used to obtain p-values. Using the functional approach clearly defined storage regimes that significantly decrease the speed of deterioration of the product relative to traditional Mexican agricultural practices.

Key words: Garlic, functional data, linear model, permutation test.

1. Introduction

Garlic (Allium sativum L.) is one of the most important products from Mexican agriculture. Mexico has the second place in the Americas as garlic producer cultivating 9,850 hectares with 75,810 ton per year (Heredia, 1995; SAGAR, 1997); garlic in Mexico is important from an economic perspective because 26% from its total production goes to USA markets (Heredia, 2000). Its production cycle goes from February to August; during the period of low availability, harvested garlic is strored without any control. It is important to understand the effects of different storage conditions on its quality features.

The global objective of this research was to longitudinally study quality changes in garlic due to six different storage regimes (treatments), $0°C$, $20°C$, $30°C$, $5°C$, $0°C$ / 70% of relative humidity, and no control, this last one being the traditional storage condition of garlic in Mexican agriculture. In this paper we concentrate in weight loss changes; for each storage regime a batch of garlic (with 360 bulbs) was screened during 190 days. From each batch, three 5-bulbs sets were used as three replicates, therefore 18, the total number of experimental units, were assigned completely at random to the different storage regimes. Weight loss of every experimental unit was repeatedly measured every 10 days (from 0 to 190 days).

Then the corresponding experimental design is a completely randomized one factor design with three replicates. The response is the weight loss profile along 190 days.

In the context of the described experiment, there exist multiple alternatives for the statistical analysis. In this paper we show an application of what is called a functional linear model to estimate contrasts of interest among storage conditions. We will make statistical inference by means of permutation tests (Good, 2000). In the next section basic ideas of functional data analysis and linear modelling are presented. In section 3 we comment on the alternatives to carry out hypothesis testing in the context of the functional linear model. Section 4 is devoted to the application and discussion of the results about the comparison of storage regimes.

2. Functional Modeling

2.1 Functional representation of discrete data

Functional Data Analysis (*FDA*) is a useful approach to study variation of responses such as the one described above. Let N be the number of experimental units, n the number of times that the weight loss was recorded and y_{ij} the weight loss values for $i = 1, \dots, N$ and $j = 1, \dots, n$. These values could be thought of as a discrete manifestation of a weight loss function $y_i(t)$, such that

$$y_{ij} = y_i(t_j) + \varepsilon_{ij},\ \varepsilon_{ij} \backsim \left(0, \sigma^2\right) \text{ independent,}$$

t_j, $j = 1, \dots, n$, the so-called knots.

Then the first step in *FDA* is to represent these values by a function $y_i(t)$, $t \in [0, 190(= T)]$. Assuming observational errors, the estimation process of $y_i(t)$ involves smoothing techniques. A popular criterion (Green and Silverman, 1994) to estimate $y_i(t)$ is minimizing

$$\sum_{j=1}^{n} \{y_{ij} - y_i(t_j)\}^2 + \lambda_i \int_0^T \{y_i''(t)\}^2 \, dt \tag{2.1}$$

where $\lambda_i \geq 0$ is a smoothing parameter which represent a trade-off between the goodness of fit $y_i(t)$ to the data, the first term in (2.1), and its smoothness measure by the integral of $y_i''(t)$, the second derivative of $y_i(t)$ used as a measure of its smoothness.

Given (2.1) an important point is where to search for y_i; a searching strategy having a sound theoretical background (see for instance Wahba, 1990) is to think of $y_i\left(t\right)$ as

$$y_i\left(t\right)=\sum_{m=1}^{M} c_{im}B_m\left(t\right)$$

where $\left\{B_m\left(t\right)\right\}_1^M$ is a set of basis functions to represent y_i, and $\left\{c_{im}\right\}$ the corresponding set of coefficients of y_i. Under this representation, (2.1) can be written as

$$\sum_{j=1}^{n}\left\{y_{ij}-\sum_{m=1}^{M} c_{im}B_m\left(t_j\right)\right\}^2+\lambda_i\int_0^T\left\{\sum_{m=1}^{M} c_{im}B_m''\left(t\right)\right\}^2 dt$$

$$\left\|\mathbf{y}_i-B\mathbf{c}_i\right\|^2+\lambda_i\mathbf{c}_i^TR\mathbf{c}_i$$

where $\mathbf{y}_i=\left(y_{i1},\ldots,y_{in}\right)^T$, $\mathbf{c}_i=\left(c_{i1},\ldots,c_{i_M}\right)^T$, B a matrix with columns $B_m\left(t_j\right), m=1,\ldots,M$, and $R=\left\{R_{mu}\right\}$,

$$R_{mu}=\int_0^T B_m''\left(t\right)B_u''\left(t\right)dt.$$

The solution to this minimization problem corresponds to

$$\hat{\mathbf{c}}_i=\left(B^TB+\lambda_iR\right)^{-1}B^T\mathbf{y}_i$$

Then

$$\hat{y}_i\left(t_j\right)=\left(B_1\left(t_j\right),\ldots,B_M\left(t_j\right)\right)\hat{\mathbf{c}}_i.$$

The estimated function resulting from the above describe process belongs to the so-called natural cubic smoothing splines (Green and Silverman, 1994). A set of basis functions that efficiently expands a natural cubic spline from the computational perspective are the so-called B-splines. The specification of a B-spline basis requires augmentation of knots $t_1<\ldots<t_n$. Let t_0 and t_{n+1} denote boundary knots. The augmented knot sequence is defined as follows;

$$\begin{aligned}\tau_1 &\leq \tau_2\leq\ldots\leq\tau_M\leq t_0;\\ \tau_{j+M} &= t_j, \qquad j=1,\ldots,n\\ t_{n+1} &\leq \tau_{n+M+1}\leq\tau_{n+M+2}\leq\ldots\leq\tau_{n+2M}\end{aligned}$$

The values of the additional knots are arbitrary, and it is customary to fix them all the same and equal to t_0 and t_{n+1}, respectively. Denote by $B_{u,m}(x)$ the B-spline function of order m for the sequence of knots τ, $m \leq M$, recursively defined as:

$$B_{u,1} = \begin{cases} 1 & \text{if } \tau_u \leq t \leq \tau_{u+1} \\ 0 & \text{otherwise} \end{cases}$$

for $u = 1, \ldots, n + 2M - 1$;

$$B_{u,m}(t) = \frac{t - \tau_u}{\tau_{u+m-1} - \tau_u} B_{u,m-1}(t) + \frac{\tau_{u+m} - t}{\tau_{u+m} - \tau_{u+1}} B_{u+1,m-1}(t),$$

for $m = 1, \ldots, n + 2M - m$. Thus with $M = 4$, $B_{u,4}(t)$, $u = 1, \ldots, n + 4$ are the $n + 4$ cubic B-spline basis functions for the knot sequence $t_j's$. So defined each B-spline function has a compact support; this implies computational efficiency; for further details see for instance Hastie, Tibshirani and Friedman (2001).

The smoothing parameter λ_i is a key component in the estimation process because if $\lambda_i \to 0$, $\hat{y}_i(t)$ will be just as rough as $y_{ij}, j = 1, \ldots, n$, while if λ_i increases without limit, $\hat{y}_i(t)$ will be forced to a linear function $(\hat{y}_i''(t) = 0)$. One way to choose λ_i is by minimizing the cross - validation score

$$CV(\lambda_i) = n^{-1} \sum_{j=1}^{n} \left\{ y_{ij} - \hat{y}_i^{(-l)}(t_j; \lambda_i) \right\}^2,$$

where $\hat{y}_i^{(-l)}(t_j; \lambda_i)$ represents the estimated function once y_{il} has been omitted from the estimation process.

2.2 A functional linear model

Being $y_i(t)$ the functionally represented weight loss function for experimental unit i along storage time, the objective is to study changes in $\{y_i(t)\}_1^N$ due to the above mentioned six storage conditions. Ramsay and Silverman (1997) (see also Ramsay and Silverman, 2002), introduced the following linear model

$$y_i(t) = \mathbf{x}_i^T \beta(t) + \epsilon_i(t), \tag{2.2}$$

where $\beta(t)$ represents a vector of parameter functions of interest, $\epsilon_i(t)$ an experimental error function. In our case the purpose is to compare all storage treatments versus the no control condition; therefore model (2.2) is used in this application taking the following form

$$y_{kl}(t) = \alpha(t) + \beta_k(t) + \epsilon_{kl}(t), k = 1, \ldots, 6; l = 1, \ldots, 3; \tag{2.3}$$

where $\beta_6\,(t) = 0$ and $\beta_j\,(t)$ is interpreted as the difference effect of treatment j and treatment 6 (no control), $j = 1, \ldots, 5$; $\alpha\,(t)$ represents the expected functional response from treatment 6, and then it may be interpreted as a baseline against which the rest of the treatments are going to be compared.

An estimation criterion of $\beta\,(t)$ is to minimize in model (2.2)

$$\sum_{i=1}^{N} \int_{0}^{T} \left\| y_i(t) - \mathbf{x}_i^T \beta\,(t) \right\|^2 dt.$$

If $Y\,(t)$ represents a $N-$ vector with elements $y_i\,(t)\,, i = 1, \ldots, N$, the N profiles from N experimental units, and X an $N \times q$ design matrix of full rank $(q = 6$ in this application) with rows $\mathbf{x}_i'$, $i = 1, \ldots, N$, for actual computation of $\hat{\beta}\,(t)$ an unconstrained minimum of a sum of squares

$$\left\| Y(t) - X\beta\,(t) \right\|^2 \tag{2.4}$$

should be obtained. Taking advantage of the basis representation of each $y_i\,(t)$ described in the previous section,

$$Y(t) = \mathbf{Y}B\,(t)$$

where B represents a M-vector of basis functions and $\mathbf{Y}$ gives the coefficients of the observed vector Y of functions. Expand the estimated parameter vector $\hat{\beta}\,(t)$ in terms of the same basis, that is,

$$\hat{\beta}\,(t) = \mathbf{B}B\,(t)$$

for a $q \times M$ matrix $\mathbf{B}$. Then $\mathbf{B}$ can be obtained from

$$X^T X\mathbf{B} = X^T\mathbf{Y}.$$

Evaluation of the resulting fit can be done by using the following functions:

$$\begin{aligned} SSE(t) &= \sum_{i=1}^{N} \left[y_i(t) - \mathbf{x}_i^T \hat{\beta}\,(t) \right]^2, \\ MSE(t) &= SSE(t)/(N - \#\text{independent parameters}) \end{aligned}$$

the functional versions of the sum of squared errors and the mean square error.

3. Hypotheses Testing in Functional Models

There are many alternatives to test hypothesis about $\beta(t)$ in model (2.2) . One of these alternatives is to use analysis of variance of data from a repeated measures design, measuring the experimental unit over time where time is one of the factors in the treatment structure of the experiment. By measuring the experimental unit at several different times, the experimental unit is essentially split into parts (time intervals) and response is measured on each part, appearing a split plot structure along time. Time is not randomly assigned, of course, and then the usual analysis of variance may not be valid, because the errors corresponding to the respective experimental units may have a covariance matrix that does not conform to those for which the usual split plot analysis is valid. When usual assumptions do not hold there exist different approaches to adjust the analysis of variance; imposing a working assumption on the covariance matrix of errors (usually compound symmetry on the experimental unit errors or the Huynh - Feldt condition on errors within each experimental unit and for each experimental unit). See for further details Milliken and Johnson (1992). Nevertheless the emphasis in this approach is on means comparisons, therefore leaving aside the analysis of each complete experimental unit profile as a whole.

Another alternative is to carry out a two - step modelling process (also called a hierarchical modelling); in the first step it is proposed for the response a parsimonious parametric linear or nonlinear model on time and then, in the second step, modelling the estimated parameters as a function of the treatment structure of the experiment. There is a vast literature on the use of this modelling strategy, being representative Davidian and Giltinan (1995), Diggle, Liang and Zeger (1994), Verbeke and Molenberghs (2000). Most of the corresponding inferential procedures are based on an important number of assumptions and results are approximate.

Another form to contrast hypotheses of interest in the context of model (2.2) is to consider multivariate analysis based - methods, considering a grid of values along the domain of the functional observations; as it is explained by Faraway (1997), likelihood ratio statistics will become dominated by terms representing unimportant sources of variation as soon as the size grid becomes large, as it is the natural case with functional data. Faraway (1997) also proposed bootstrap testing methods from the computation of residual curves under the null hypothesis of interest.

An additional approach is testing hypotheses by a permutation based testing approach. Ramsay and Silverman (1997) proposed the usage of this approach in the context of functional linear modelling. Nichols and Holmes (2001) have reported applications of this kind of tests for functional neuroimaging.

Permutation tests were proposed in the early twentieth century, but now with powerful computers, are feasible (Good, 2000). In an experimental context in which treatments are compared, if treatment randomization was done, under a null hypothesis of no treatment differences, the observed data can be observed under any treatment labeling. Given a meaningful statistic $\mathcal{T}$ to test a null hypothesis, effectively permuting the treatment labels, and computing for each permutation $\mathcal{T}$, gives us a way to calculate a p-value under a specific cdf F governing the actually observed data under H_0, F being completely specified. If F is not specified under H_0, the empirical cdf $\hat{F}$ is minimal sufficient for F. For instance in the case of the comparison of two means, $H_0 : \mu_1 = \mu_2$ vs $H_0 : \mu_1 \geq \mu_2$, where μ_1 and μ_2 represent the means for the respective populations, H_0 does not specify F; however if we believe that cdf´s F_1 and F_2 have the special forms

$$F_1(\cdot) = G(\cdot - \mu_1), \; F_2(\cdot) = G(\cdot - \mu_2),$$

for some unknown G, then H_0 implies a common cdf for the populations; under H_0 $\hat{G}$ as sufficient statistics, is comprised of the ordered set of the pooled sample, that is, the sufficient statistics S is the set of order statistics for the pooled sample . Therefore the p-value is calculated as

$$p = \Pr(\mathcal{T} \geq t \,|S = s, H_0). \tag{3.1}$$

Actual computation of (3.1) , when $S = s$, implies that the pooled sample must form a permutation of s. When H_0 is true all such permutations are equally likely and then

$$p = \frac{\#\text{ of permutations such that } \mathcal{T} \geq t}{\text{total }\#\text{ of permutations}}.$$

In a functional context, a nonparametric permutation test represents an important approach due to the difficulty to specify under H_0 a non - stationary stochastic process generating the observed functional data.

In the garlic experiment experimental units were assigned at random to the different regimes, and then a permutation test is applicable permuting labels of the functional observations.

The null hypotheses of interest are

$$H_{0j} : \beta_j(t) = 0, \; j = 1, \ldots, 5. \tag{3.2}$$

A permutation test to be done requires the following tasks (Good, 2000):

1. Analyze the problem.

2. Choose a test statistic.

3. Compute the test statistic from the original labeling of the observations.
4. Rearrange (permute) the labels and recompute the test statistic for the rearranged labels. Repeat until you obtain the distribution of the test statistic for all possible permutations.
5. Accept or reject the hypothesis using this permutation distribution as a guide.

In step 2, following Ramsay and Silverman (1997), we proposed a permutation test using as the test statistic, avoiding the problem of multiplicity of testing (Nichols and Holmes, 2001),

$$S_j = \sup_t \left| \hat{\beta}_j(t) \,/\, \left[a_j \sqrt{MSE(t)} \right] \right|, j = 1, \ldots, 5, \tag{3.3}$$

where

$$a_j^2 = \mathbf{u}_j^T \left(X^T X \right)^{-1} \mathbf{u}_j,$$

and $\mathbf{u}_j$ a $q-$ vector with 1 in position j and zero elsewhere. We additionally used the statistic

$$I_j = \int_0^T \left| \hat{\beta}_j(t) \right| / \left[a_j \sqrt{MSE(t)} \right] dt. \tag{3.4}$$

In step 4, among other approaches, we follow a Monte Carlo strategy to fix the number of permutations (500) in order to build the permutation distributions of above mentioned statistics.

4. Data Analysis

All computations were done using *S-PLUS 6 for Windows* and the subroutines developed by Ramsay and Silverman (1997); beginning with these subroutines, ad hoc modifications were done to set the specific functional model, building statistics and to get permutation tests results.

Using M equal to 20 cubic B-spline functions were use to represent functionally each weight loss profile, using cross validation to choose λ_i, $i = 1, \ldots, 18$.

In Figure 1 are shown the $N = 18$ weight loss profiles with $n = 20$ data each, of the 3 replicates from each storage regime. Descriptively, from this figure there are clear differences among treatment conditions. There are 3 groups: $0°C$ and $0°C$ / $70\%RH$ with linear profiles and weight loss less than 10% until 190 days; $5°C$ with the worst weight loss profile reaching around 30% of weight loss; and $30°C$, $20°C$ and no - control with moderate weight losses at the of the period but with different trends to reach their 190-day state, linear trend for $30°C$, quadratic trends for $20°C$ and no control.

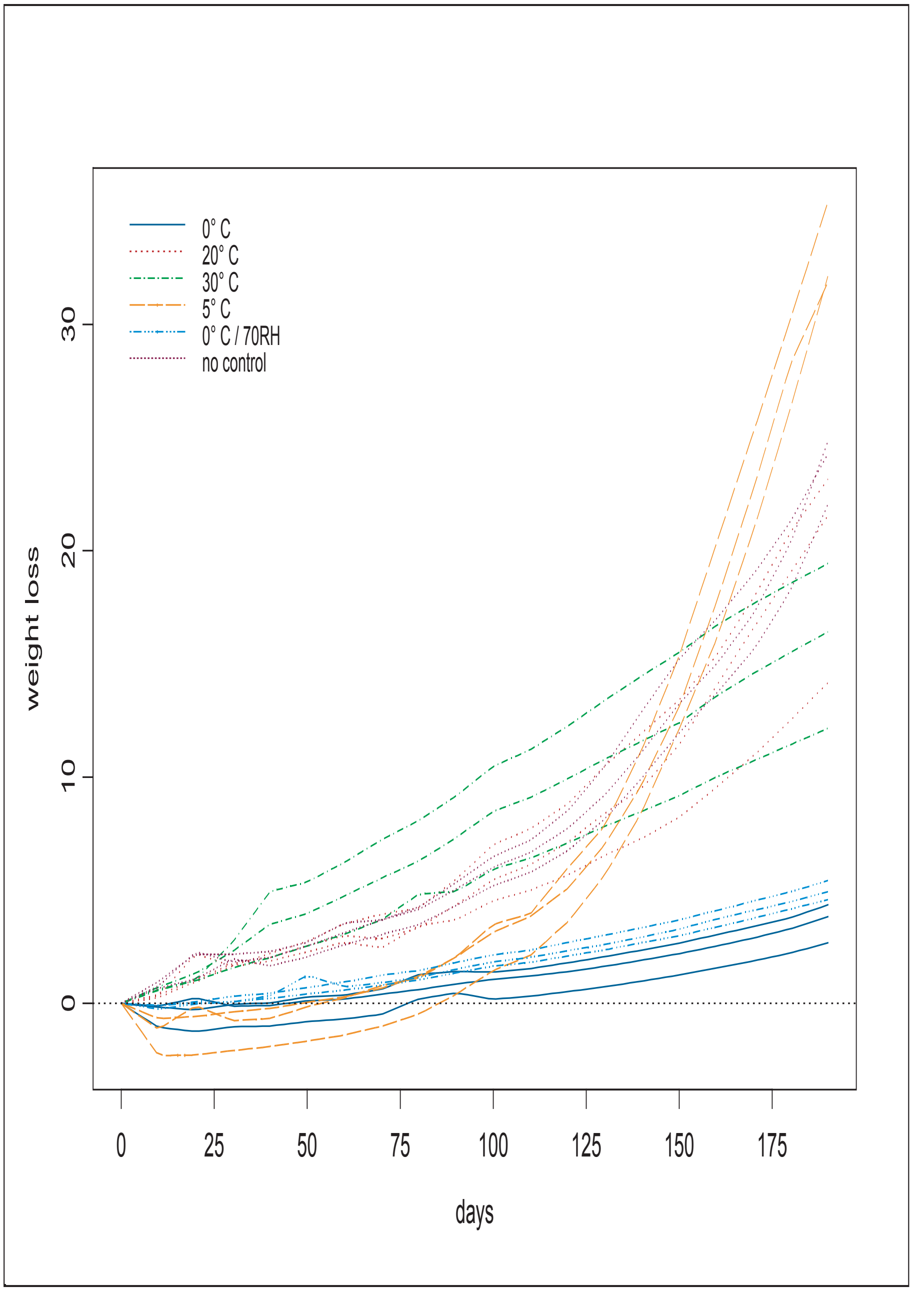

Figure 1: Weight loss profiles under different storage regimes

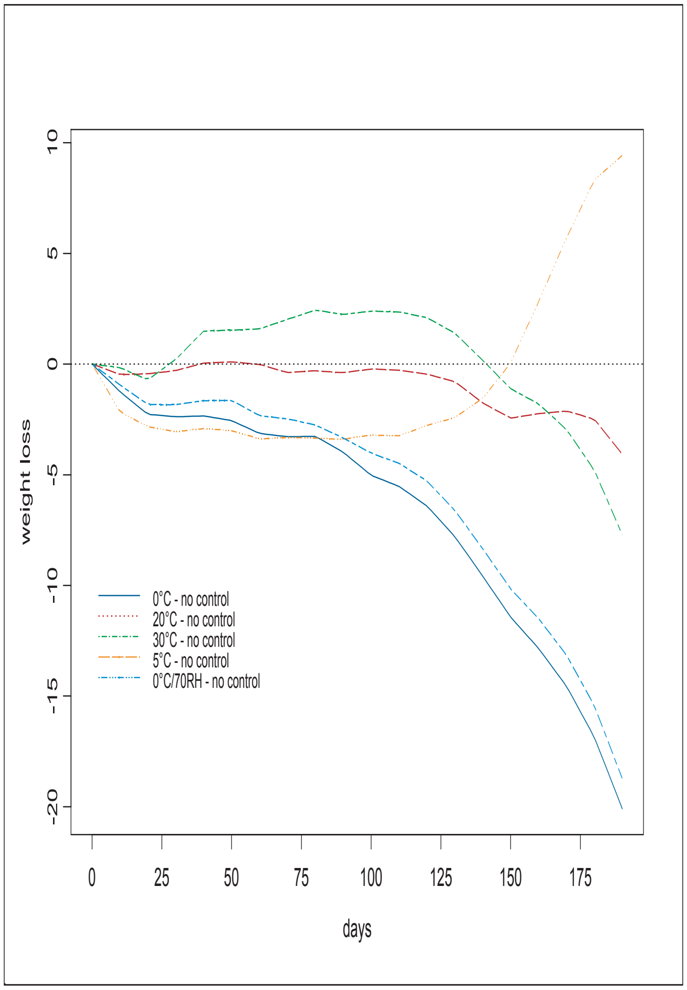

Figure 2: Estimated effects $\hat{\beta}_j(t)$

Table 1: p-values from permutation tests on S_j and $I_j, j = 1, \ldots, 5$

storage condition	S_j	I_j
0°C – no control	0	0
20°C – no control	.284	.506
30°C – no control	.026	.046
5°C – no control	.002	.002
0°C/70%RH – no control	0	0

Estimated $\beta_j(t)$ from model (2.3), are shown in Figure 2. In Table 1 are shown the estimated p-values corresponding to the permutation distributions of statistics (3.3) and (3.4) to test hypotheses (3.2). Comparing to the no control condition, both treatments under $0°C$ induce a highly significant reduction in weight loss during the storage period, specially after 75 days. In comparison to the no control condition, $5°C$ induces an important increment in weight loss after 150 days. Storage condition under $30°C$ is different from no control condition inducing a greater loss weight until 150 days, and then inducing a gradual decrement in the weight loss. Finally with $20°C$ the loss weight pattern is statistically equivalent to the no control storage condition.

In Figure 3 are shown the average derivatives of the weight loss functions showed in Figure 1. We can observe that the derivatives of weight loss functions at $5°C$ and no control show the most important changes in speed especially after 100 days. The other treatments show relatively constant speeds along the observation period.

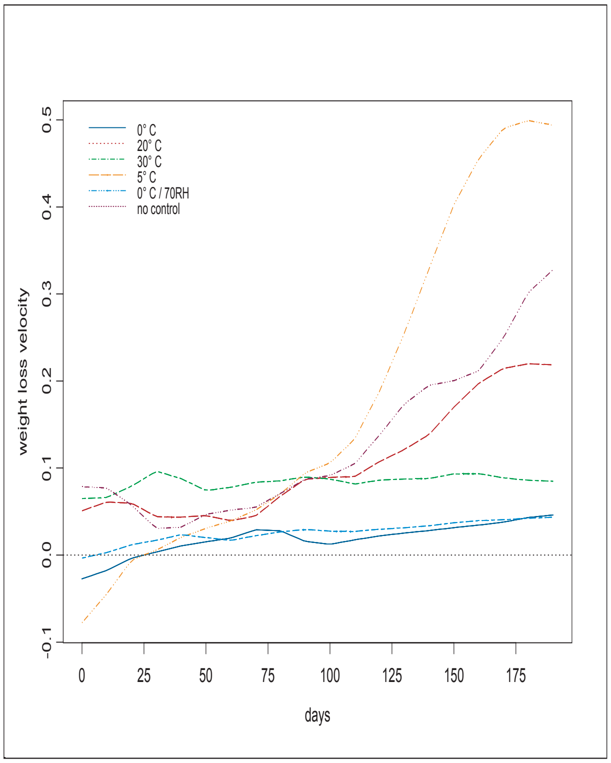

Figure 3: Derivatives of weight loss functions

Using the corresponding functional linear model (2.3) but with the velocities of weight loss as functional responses, in Figure 4 are shown the corresponding functional estimated effects of $\frac{d}{dt}\beta_j(t)$, $j = 1, \ldots, 5$. We can observe that all treatments, except $20°C$, show important differences from the no control condition especially after 100 days of storage; $5°C$ showing the biggest speed after 100 days; $30°C$ and both treatments with $0°C$ are showing a lower speed in weight loss after 100 days.

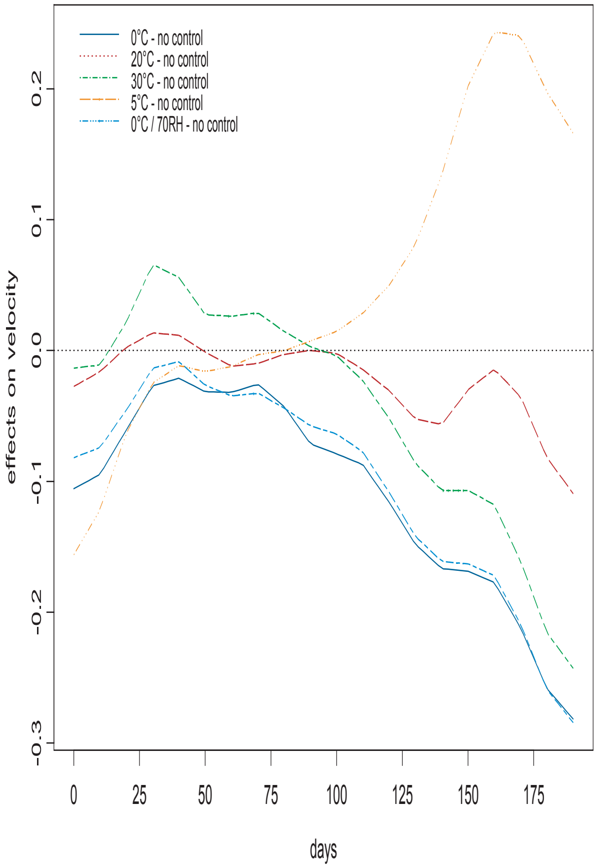

Figure 4: Functional effects on velocity of weight loss

Ackowledgements

ETC, CHG and JJC would like to thank the Mexican Council for Science and Technology (CONACyT grant 36616-E) for financial support. VAT received partial support by Asociación Mexicana de la Cultura, A.C.

References

Davidian, M. and Giltinan D. M. (1995). *Nonlinear Models for Repeated Measuement Data*. Chapman and Hall.

Diggle, P. J., Liang K. Y. and Zeger S. L. (1994). *Analysis of Longitudinal Data*. Oxford University Press.

Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models: A roughness penalty approach.* Chapman and Hall.

Good, P. (2000). *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*, second edition. Springer.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistiscal Learning: Data Mining, Inference and Prediction*. Springer.

Heredia-García, E. (2000). Cosecha. In *El ajo en México: origen, mejoramiento genético y tecnología de producción*, 102-***(give end page)*** (edited by Heredia-García, E. and F. Delgadillo Sanchez). Celaya, Gto., México. SAGAR, INIFAP, Campo Experimental Bajío. (Libro Técnico Núm. 3).

Heredia, Z. A. (1995). *Guía para cultivar ajo en el Bajío.* Folleto para productores No. 1. México. CEBAJ - INIFAP - SARH. Celaya, Gto.

Milliken, G. A and Johnson, D. E. (1992) *Analysis of Messy Data. Vol I: Designed Experiments*. Chapman and Hall.

Nichols, T. E. and Holmes, A. P. (2001). Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Human Brain Mapping,* f**15**, 1-25.

Ramsay, J. O. and Silverman, B. W. (1997). *Functional Data Analysis*. Springer.

Ramsay, J. O. and Silverman, B. W. (2002). *Applied Functional Data Analysis: Methods and Case Studies*. Springer.

SAGAR Secretaría de Agricultura y Desarrollo Rural. (1997) *Sistema Producto Ajo.* Dirección General de Política Agrícola. México.

Verbeke, G. and Molenberghs, G. (2000) *Linear Mixed Models for Longitudinal Data.* Springer.

Wahba, G. (1990) *Spline Models for Observational Data.* Philadelphia: SIAM.

Wu, C. F. J., and Hamada, M. (2000). *Experiments: Planning, Analysis, and Parameter Design Optimization*. John Wiley.

Received August 21, 2004; accepted January 20, 2005.

E. Castano - Tostado
Universidad Autonoma de Queretaro
Facultad de Quimica
Queretaro Qro 76010, MEXICO
ecastano@uaq.mx

E. Mercado - Silva
Universidad Autonoma de Queretaro
Facultad de Quimica
Queretaro Qro 76010, MEXICO
mercado@uaq.mx

F. Leon - Gonzalez
Universidad Autonoma de Queretaro
Facultad de Quimica
Queretaro Qro 76010, MEXICO
leonfquaq@hotmail.com

C. Gorrostieta - Hurtado
Universidad Autonoma de Queretaro
Facultad de Ingenieria
Lic Matematicas Aplicadas
Queretaro Qro 76010, MEXICO
cgorrostieta@yahoo.com.mx

J. Chamorro - Jimenez
Universidad Autonoma de Queretaro
Facultad de Ingenieria
Lic Matematicas Aplicadas
Queretaro Qro 76010, MEXICO
jacque_chj@yahoo.com.mx

E. Vazquez - Barrios
Universidad Autonoma de Queretaro
Facultad de Quimica
Queretaro Qro 76010, MEXICO
tita_evb@yahoo.com

V. Aguirre - Torres
Instituto Tecnologico Autonoma de Mexico
Departamento de Estadistica
Rio Hondo 1, Col. Tizapan-San Angel
01000 Mexico, D.F., MEXICO
aguirre@itam.mx