SADD: Stochastic Analysis of Destructively Measured Data - Possibilities to Include Biological Variations in Statistical Analysis

L.M.M. Tijskens[1], G. Jongbloed[2] and M. Kessler[3]
[1] Wageningen UR, Group Horticultural Supply Chains, Wageningen, The Netherlands
[2] Delft University of Technology, Institute of Applied Mathematics, Delft, The Netherlands
[3] Technical University of Cartagena, Group Applied Mathematics & Statistics, Cartagena, Spain

Abstract

Three techniques are presented to include the structural variation always present in measured data in statistical analysis. The methods are investigated and compared using cross sectional data, generated based on an exponential model as if gathered by destructive measuring methods. All three methods are based on optimising objective functions based on the data and the biological shift model. These objective functions are calculated for each separate measuring point in time either according the specific density function belonging to the model applied, or after conversion into biological shift factors (also according to the model applied) according to a Gaussian distribution. The procedures used need to be improved, embedded in the existing statistical framework and all available statistical expertise and skills need to be combined into robust procedures capable of analysing everyday data.

INTRODUCTION

Usually in postharvest research data are gathered using destructive measuring techniques (leading to cross sectional data). Only in special cases non-destructive techniques can be used (leading to longitudinal data) that allow proper analysis including the always present biological variation using mixed-effects non-linear regression techniques. The success and the power of this last methodology can be taken from recent reports (Hertog et al., 2002, 2004; Tijskens et al., 2003, 2005, 2007, 2008a,b; Schouten et al., 2004, 2007; De Ketelaere et al., 2006). Until now, however, no such methodology exists for data gathered using destructive measuring techniques, which are used in horticultural research in about 80 to 90% of the cases.

From the information and understanding of the dynamics of biological variation in time for different batches of product obtained from the analysis of longitudinal data, some rules and plausible assumptions can be deduced that should also apply to cross sectional data. In this paper three techniques are presented to include information on the dynamics of biological variation in the analysis of cross sectional data: optimising the fidelity to the Gaussian distribution of the biological shift factor using the Shapiro-Wilk test statistic, optimising the distribution of the measured data directly on the density function for that model using the Kolmogorov-Smirnov test, and optimising the log likelihood, also based on the density function.

Setting Up an Example

Exponential behaviour is one of the most frequently encountered models in day to day physiological research. Chemically, it is based on a simple first order reaction. A generic formulation of this type of behaviour is shown in Equation 1.

$$y = y_{min} + (y_0 - y_{min}) \cdot e^{-k \cdot t} \quad (1)$$

Converting the pre-exponential factor to the biological shift factor notation, expressing the state of development as a difference in time (Tijskens et al., 2005), one gets:

$$\text{with } dt = -\frac{\ln\left(\frac{y_{min} - y_0}{y_{ref} - y_{min}}\right)}{k} \qquad y = y_{min} + (y_{ref} - y_{min}) \cdot e^{-k\cdot(t+dt)} \tag{2}$$

In these equations, y is the measured variable (e.g., firmness), k the rate constant, t the time and dt the biological shift factor (normally distributed), while subscript *0* refers to initial conditions, *min* to the asymptotic value and *ref* to an arbitrarily chosen reference value.

The assumption that the biological shift factor dt should be normally distributed is supported by analysing longitudinal data applying mixed effects non-linear regression analysis on individual fruit (Hertog et al., 2002, 2004; Schouten et al., 2004, 2007; De Ketelaere et al., 2006; Tijskens et al., 2007, 2008a,b). Applying this assumption, the density function for this model can be derived (Schouten et al., 2004; Hertog et al., 2004) as shown in Equation 3.

$$p(y) = \frac{1}{2}\frac{\sqrt{2}\ \mathbf{e}^{\left(-\frac{1}{2}\frac{\left(-\frac{k\ t + \ln\left(\frac{y - ymin}{yref - ymin}\right)}{k} - \mu\right)^2}{\sigma^2}\right)}}{\sqrt{\pi}\ \sigma\ (y - ymin)\ k} \tag{3}$$

The density function is closely related to the frequency as encountered in histograms. It describes in the formal statistical framework the stochastic distribution of data (y) at any moment (t) that change in time according to an exponential behaviour.

All mathematical deductions and conversions were conducted using Maple 10 (MapleSoft, Waterloo Maple Inc, Waterloo, Canada). All simulations, analyses and graphs were using R (R Development Core Team, 2005). Based on these equations, data were simulated using the exponential dynamics (Eqs. 1 and 2) including normally distributed random variation by means of the biological shift factor (mean μ=0, standard deviation σ). To mimic also unstructured variation a small technical random error was added to the variable y directly (mean 0, standard deviation ε). In Table 1 the values of the input variables used to generate the data are shown. In Figure 1 left, the behaviour of the simulated data is shown. In Figure 1 right, the distribution is shown, changing shape with increasing time.

Destructive Measurements in Stochastic Terms

Following a product property, e.g., firmness of fruit, using a destructive technique (e.g., compression until rupture), for every repetition at every measuring point in time a new sample is taken out of a mother population with variation in that property. That means that the measured firmness changes between the samples as a consequence of drawing at random another sample from the mother population. But the distribution of the repetitions should all follow the same pattern (Eq. 3). So, when analysing the measured data points, information is available in the variation that is not fully used in standard regression analysis. Not only should the measured data points (either mean value or individual values) follow the same kinetic behaviour (regression analysis), but also the distribution of the measured points, either expressed as measured (y), or converted into the biological shift factor (dt), should obey the same underlying distribution pattern (defined by the applied model), with the same mean value and standard deviation.

Three methods will be explored to include the variation in the mother population

in the regression analysis to obtain information on the variation in that mother population. The results will be compared to the standard regression analysis without taking the variation into account.

Normality of the Biological Shift Factor Distribution

The first method is based on the indication that the biological shift factor (Tijskens et al., 2005) should be distributed according to a normal or Gaussian distribution. Within the repetitions of the optimisation procedure (optim procedure in R), the biological shift factor for each data point is calculated from the actual kinetic parameters values using Equation 2. The standard Shapiro-Wilk test on normality (standard available in every statistical package) is applied to this calculated shift factor. The higher the p-value of this test, or the lower the D value, the more the distribution tested can be considered normal. Optimising the model parameters with respect to the p- or D-value of the normality test of Shapiro-Wilk on the calculated biological shift factor (calculated according to Eq. 2), would then deliver a reliable indication of the model parameters (y_{min}, k) and the standard deviation in the mother population (σ). The p- and D- values are normalised between 0 and 1, which easily allows combined loss functions for the optimisation procedure.

Optimising the Distribution of the Property Directly

The second method is based on the theoretical density function for the model applied without converting the measured y-values into individual biological shift factors (Eq. 3). For every time point in the data, the distribution of the y-data is compared to the theoretical distribution function using the Kolmogorov-Smirnov procedure. Maximising the p-value (or, equivalently, minimising the D-value) with respect to the parameters in the model will then deliver not only estimated values for the model parameters but also a characterisation of the biological variation present. The p- and D- values are normalised between 0 and 1, which easily allows combined loss functions for the optimisation procedure.

Optimising the Log Likelihood

A well known general method in statistics, related to the second method just described, is based on optimising the log likelihood. This quantity is the sum of the logarithm of the density function (Eq. 3) over all times and repetitions of measurement. Maximising this log likelihood directly as the objective function delivers an estimate of the model parameters, taking into account that the data have to be distributed according to the density function, which changes with time (Eq. 3). A problem that arises using this system is that the range of log likelihood is not normalised between 0 and 1, and strongly depends on the number of data points used. Consequently, combined loss functions are more difficult to apply, because the relative importance will depend on the actual data.

Combined Loss Functions and Numerical Problems

Relying solely on optimising stochastic criteria (p-, D- and log likelihood values), the analysis completely neglects the kinetic behaviour. To ascertain that the kinetic behaviour (simple non linear regression) is not too much mutilated by the stochastic estimation systems, loss functions can be combined, e.g., 20% of pure kinetics combined with 80% of stochastic as applied in Table 2.

A problem that arises using the third, log likelihood method, since the range of the parameter is not standardised, and strongly depends on the number of data points used. Consequently, no combined loss functions can be applied (yet).

All approaches applied rely somehow directly (dt) or indirectly (density function, log likelihood) on data transformation. With an exponential behaviour as in this example, an asymptote is present, very strongly determining the calculated dt in its neighbourhood. As soon as data points are present at the wrong side that asymptote, due to variation in the data not covered by the assumptions (ε), numerical problems arise (e.g., logarithm of a negative number). At this moment no clear strategy is available for dealing with this

problem. For the time being these points are just disregarded. This will however, affect the results of the analysis. More study is required to solve this problem properly.

RESULTS AND DISCUSSION

The results of the analyses using the three estimation methods, as well as the values of the input parameters, used to generate the data, and the simple non linear regression are shown in Table 2. The estimated behaviour for the analyses are shown in Figure 1 (Left), while the density function for the model used with the input parameter values are show in Figure 1 (Right). The estimations shown in Table 2 are clearly different. Estimation directly on p(y).p and p(y).D in combination with pure kinetics seem to provide the most reliable results, considering the values for the asymptote (y_{min}) and the rate constant (k) compared to the input values. The estimation based on p(dt).p, p(dt).D and LogLik seem to overemphasise the value of the asymptote. Probably these latter methods are more sensitive to the numerical problems mentioned above. In Figure 2 the histogram of y values at 4 time points is shown for input, regression and the 5 stochastic analysing methods. In Figure 2 one can clearly see that at low values of t, the agreement between the methods and the data is generally better. That is also an indication of the major influence of the asymptotic value y_{min} on behaviour and estimation.

When repeating the analyses using newly generated data based on the same input values, the results are of course every time slightly different. The general pattern however is the same. How to present the overall 'goodness of fit', that is the kinetic explained part of variation combined with the explained part of the stochastic estimation is not yet known. Statistical expertise, knowledge and skills are needed to develop these rudimentary systems into robust and applicable procedures.

All methods presented heavily rely on the model formulation or dynamics of change in the property under study. The behaviour and shape of the data distributions is specific for each particular mechanism. Irrespective of the problems and difficulties in the statistical area (how to do the analysis technically), the main problem analysing cross sectional data is to find the proper model mechanism for the property under study. Due to the always present and mostly huge variation, that mechanism is most of the time not known and certainly not readily found. With these analysing techniques available and working, we can start a search for proper models for destructively measured properties in the postharvest sector.

ACKNOWLEDGEMENT

The financial support of EU COST 924 for a Short Term Scientific mission is gratefully acknowledged.

Literature Cited

De Ketelaere, B., Stulens, J., Lammertyn, J., Cuong, N.V. and de Baerdemaeker, J. 2006. A methodological approach for the identification and quantification of sources of biological variance in postharvest research. Postharvest Biol. Technol. 39:1-9.

Hertog, M.L.A.T.M. 2002. The impact of biological variation on postharvest population dynamics. Postharvest Biol. Technol. 26:253-263.

Hertog, M.L.A.T.M., Lammertyn, J., Desmet, M., Scheerlinck, N. and Nicolaï, B.M. 2004. The impact of biological variation on postharvest behaviour of tomato fruit. Postharvest Biol. Technol. 34:271-284.

R Development Core Team. 2005. R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, Freely available at: http://www.R-project.org.

Schouten, R.E., Jongbloed, G., Tijskens, L.M.M. and van Kooten, O. 2004. Batch variability and cultivar keeping quality of cucumber. Postharvest Biol. Technol. 32:299-310.

Schouten, R.E., Huijben, T.P.M., Tijskens, L.M.M. and van Kooten, O. 2007. Modelling quality attributes of truss tomatoes: linking colour and firmness maturity. Postharvest

Biol. Technol. 45:298-306.
Tijskens, L.M.M., Konopacki, P. and Simcic, M. 2003. Biological variance, burden or benefit? Postharvest Biol. Technol. 27:15-25.
Tijskens, L.M.M., Heuvelink, E., Schouten, R.E., Lana, M.M. and van Kooten, O. 2005. The biological shift factor. Biological age as a tool for modelling in pre- and postharvest horticulture. Acta Hort. 687:39-46.
Tijskens, L.M.M., Eccher Zerbini, P., Schouten, R.E., Vanoli, M., Jacob, S., Grassi, M., Cubeddu, R., Spinelli, L. and Torricelli, A. 2007. Assessing harvest maturity in nectarines. Postharvest Biol. Technol. 45:204-213.
Tijskens, L.M.M., Konopacki, P.J., Schouten, R.E., Hribar, J. and Simčič, M. 2008a. Biological variance in the colour of Granny Smith apples. Modelling the effect of senescence and chilling injury. Postharvest Biol. Technol. 50:153-163.
Tijskens, L.M.M., Dos-Santos, N., Jowkar, M.M., Obando, J., Moreno, E., Schouten, R.E., Monforte, A.J. and Fernández Trujillo, J.P. 2008b. Postharvest firmness behaviour of near-isogenic lines of melon. Postharvest Biol. Technol. (in press DOI: 10.1016/j.postharvbio.2008.06.001).

Tables

Table 1. Input values to generate simulated data.

Parameter	Value	Meaning
y_{min}	10	asymptote: y value at + infinite time
y_0	50	initial y value
y_{ref}	50	referene y value
k	0.1	rate constant of the process
σ	5	standard deviation biological shift factor (dt) of the mother population
ε	2	standard deviation of real measuring (technical) error
n_{tim}	11	number of times in a time series
n_{rep}	60	number of repetitions at one point in time

Table 2. Analysis results of the simple regression and the five stochastic methods.

Code	Stoch. Crit.	Kin. Crit.	y_{min}	k	μ	σ	R^2_{adj}
input	-	-	10.00	0.100	0.00	5.00	-
regres.	0.80	0.2	12.22	0.117	1.54	6.00	0.56
p(dt).D	0.80	0.2	1.56	0.134	-2.33	4.71	0.15
p(dt).p	0.80	0.2	3.78	0.185	-4.43	4.22	0.37
p(y).D	0.80	0.2	10.05	0.112	-0.83	5.44	0.81
p(y).p	0.80	0.2	9.71	0.110	-0.67	5.53	0.51
LogLik	0.02	0.2	3.84	0.154	-3.33	4.39	0.28

Figures

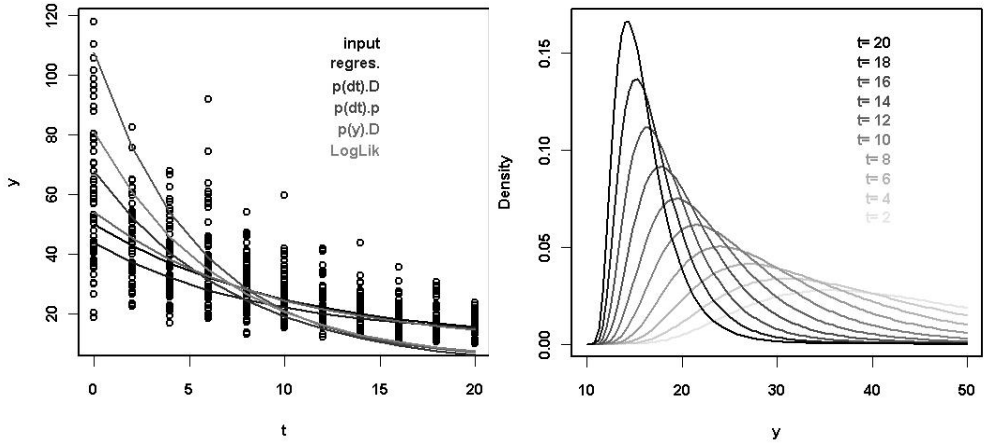

Fig. 1. Left: Behaviour of simulated data, together with the results of the 6 analysing techniques applied. Right: Dynamic behaviour of the distribution in time of the same data as a function of time.

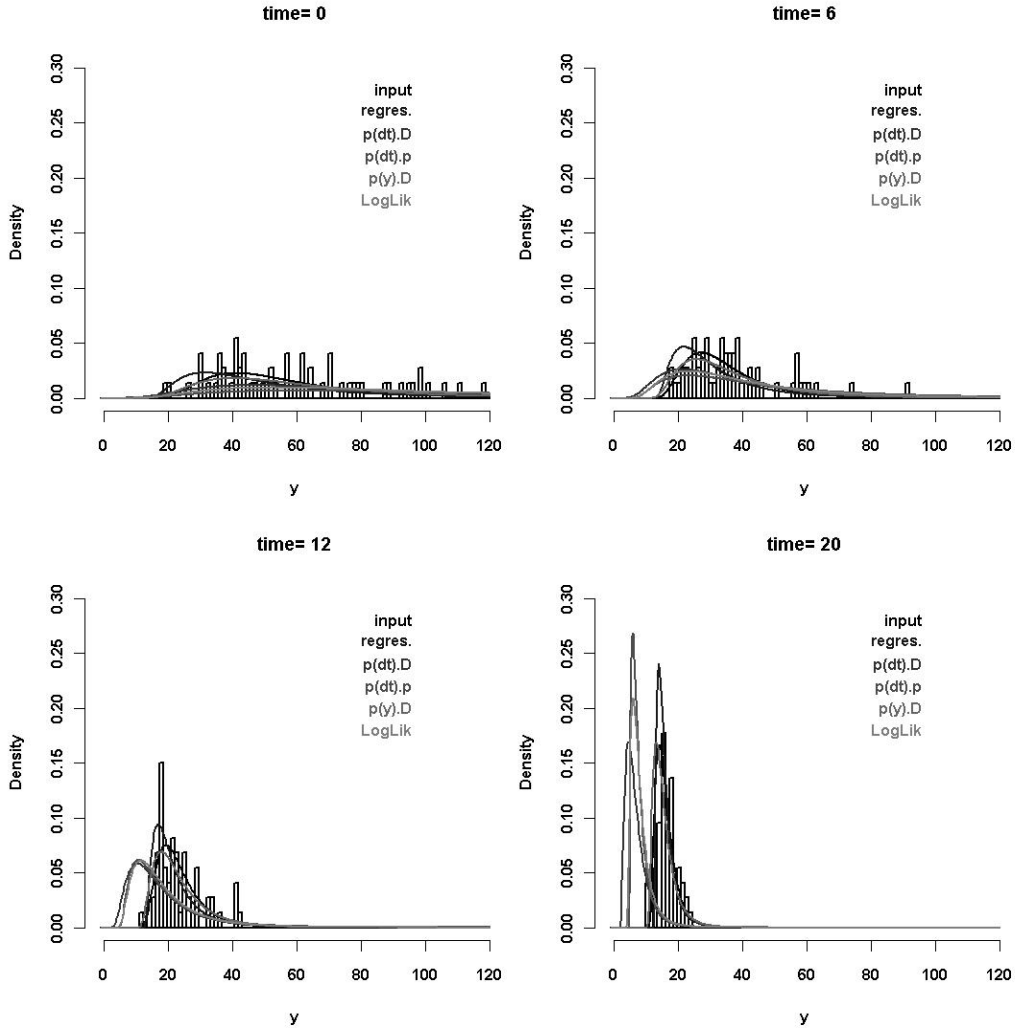

Fig. 2. Measured and simulated distributions at 4 different time points, based on the estimated parameter values in Table 2 and the density function (Eq. 3).