Computers and Electronics in Agriculture 111 (2015) 121–126

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

# Basic principles of analysing biological and technical variation in non-destructive data

L.M.M. Tijskens [a,*], R.E. Schouten [a], P. Konopacki [b], G. Jongbloed [c]

[a] Horticulture and Product Physiology, Wageningen University, The Netherlands
[b] Research Institute of Horticulture, Skierniewice, Poland
[c] Delft University of Technology, Institute of Applied Mathematics, Delft, The Netherlands

ARTICLE INFO

*Article history:*
Received 19 January 2014
Received in revised form 20 October 2014
Accepted 26 December 2014

*Keywords:*
Biological variation
Technical variation
Statistical analysis
Mixed effects regression
Indexed regression

ABSTRACT

More and more, the omnipresent variation between individual items in a batch is taken into account by using special analysing techniques like mixed effects and indexed regression. In this paper the assumptions upon which these techniques are based, are explained, based on a simulated data set. The most important issue is the separation of biological variation and the technical variation (or measuring error). The techniques are further elucidated with some examples from practise (skin colour apples in storage, softening of melons in storage and water loss in plums, melons and mandarins), to show the increased reliability of the analyses. Since variation is always present in any measured data set, the techniques are applicable in all fields working with living material.



## 1. Introduction

Life is based on variation. Genetic variation, variation in conditions (climate, soil) as well as in properties (firmness, colour, weight, length). That is what makes individuals unique. Both in the animal kingdom, as in the plant world. So, the uniqueness of individuals is based on variation. That variation is therefore always present in properties whether measured or not. To gain a clear picture of the changes in properties, this so-called biological variation between individuals has to be taken into account. On top of the biological variation, a technical variation exists in measured data. Technical variation is caused by flaws in the measuring system, either by an imperfect measuring system or by human error. This technical variation is therefore assumed independent of the properties of the measured individuals.

More and more, experimental data are gathered using non-destructive measuring techniques to assess properties such as colour, firmness and weight loss. By using these techniques, it becomes possible to follow units of a batch over time, and follow the kinetics of change in these properties on an individual level. The resulting data are called longitudinal. Much more information can be extracted from this type of data than is possible on destructive data (so-called cross-sectional data), where one inevitably has to deal with mean values in some form. To gain information on the variance in properties between individuals, the analysis of longitudinal data, however, requires a special approach and the use of special analysing techniques.

The principles and procedures for mixed effects models (linear as well as non-linear) are described by Pinheiro and Bates (2000). The problems and possibilities of including the variation in the product and data are reported by Tijskens and Wilkinson (1996), (Tijskens et al., 2003), Hertog (2002), (Hertog et al., 2004), De Ketelaere et al. (2006) and Lammertyn et al. (2003). Examples of practical application of the technology can be found in Schouten et al. (1997, 2002) and Tijskens et al. (2005). Hertog et al. (2007) provided an extended overview of the possibilities for including biological variation into the conception of models focussing on the propagation of the observed variation in measured properties. Hertog et al. (2011) elaborated on combining genetic information with kinetic modelling, including inter-individual variation.

The increasing number of recent publications (see References) applying the same or similar technology indicates that the method is very useful in gaining more insight in the processes occurring in the produce.

In this paper, the difference between biological and technical variation is explicitly described mathematically in terms of the statistical models involved. The main aim of this paper is to provide some guidelines to connect expert knowledge on the field of application with available statistical procedures to analyse

* Corresponding author.
*E-mail address:* Pol.Tijskens@wur.nl (L.M.M. Tijskens).

http://dx.doi.org/10.1016/j.compag.2014.12.022
0168-1699/

experimental data. The benefits of longitudinal data and their analysis using mixed effects and indexed nonlinear regression for extracting information on maturity and biological variance within a batch, is highlighted based on a number of examples covering water loss in plums, mandarins and melons, firmness in near isogenic lines of melons and skin colour of apples in storage and during growth.

## 2. Basic principles

All experimental data contain inevitably variation. In fact, the presence of biological variation is suspected to be the major source of difficulties in understanding the processes occurring in natural product. Without proper understanding of these processes, converting scientific knowledge into practical guidelines will remain cumbersome and very limited to the actual conditions of the experiment.

Where that variation resides in the data is not always clear. In regression analyses, applying basic statistical knowledge and procedures, it is explicitly assumed that the "noise" ($\varepsilon$) in the measured property is additive and distributed according to a centred Gaussian (normal) distribution. In Eq. (1) an example model equation of this type is shown where the dynamics of the quantity of interest $y(t)$ is exponential towards an asymptotic value $y_{\min}$. Exponential behaviour is a direct consequence of first order reactions, and is frequently encountered in agricultural and horticultural research. The model equation reads:

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

where $y$ is the measured variable under study, $t$ the time, $k$ the rate constant of the process and $\varepsilon$ the additive noise. Subscript 0 refers to the initial state at time zero, and min to the amount of $y$ that is still present at infinite time (asymptote).

The basis of the technique presented in this paper is that the residual variation ($\varepsilon$) is split up in two parts. The first one is the variation caused by the measuring technique and/or equipment (the technical error or random noise $\varepsilon_{\text{tech}}$). Technical error is therefore to be expected at the level of the measured variable and has (expectedly for this exponential model) a normal distribution.

The second one is the biological variation in product properties. Due to different exposure to light (e.g. location in the canopy), and the inevitable small variations in weather conditions (microclimate), soil structure and fertilisation, plants and plant parts do not grow and ripen in the same time. Harvest procedures are another source of variation, and add to previously described sources of variation. In other words: the overwhelming majority of observed variation in measured variables may be considered as the result of differences in biological age of the individual entities. That really represents the biological variation present in any batch of produce, whether still on the plant or post harvest. That means that the initial condition ($y_0$) is bound to be different for each individual fruit or entity in a batch. So, in short, the biological variation has to be put where it belongs: at one of the model parameters to be estimated, e.g., at the initial value $y_0$ or at the difference of that initial value with the limiting value ($y_0 - y_{\min}$).

This total variation ($\varepsilon$ in Eq. (1)) has (expectedly for this exponential model) a lognormal distribution. The lognormal distribution can be converted into a normal distribution by taking the logarithm. Equivalently, a normal distribution can be converted into a lognormal distribution by exponentiation.

In a mathematical sense, the biological variation in the equation (e.g. Eq. (1)), should be put exactly there where it belongs: at the time variable ($t$). When the pre-exponential factor in Eq. (1) ($y_0 - y_{\min}$) is converted into an exponential form (Eq. (2)), the resulting expression is as shown in Eq. (3). This conversion also transforms part of the total variation (expectedly lognormal distributed) into a normal distribution.

$$y_0 - y_{\min} = (y_{\text{ref}} - y_{\min}) \cdot e^{-k \cdot \Delta t} \tag{2}$$

$$y(t) = (y_{\text{ref}} - y_{\min}) \cdot e^{-k \cdot (t + \Delta t)} + y_{\min} + \varepsilon_{\text{tech}} \tag{3}$$

where $y_{\text{ref}}$ is a chosen value (within the range of change in the measured data) as a reference point to express the biological shift factor $\Delta t$, a stochastic variable different for each individual in the set. $\varepsilon_{\text{tech}}$ is the technical variation or measuring error (random noise), a stochastic variable that is assumed to be normally distributed with mean zero. This deduction arrives at the same relation as for the statistical conversion the lognormal distribution to a normal distribution. Rearranging Eq. (2), the relationship between $\Delta t$, $y_0$ and $y_{\min}$ is described by:

$$\Delta t = -\frac{\log\left(\frac{y_0 - y_{\min}}{y_{\text{ref}} - y_{\min}}\right)}{k} \tag{4}$$

The conversion of a model formulation into the $\Delta t$ notation (Eq. (2)) is for other types of models not always straightforward. A nice way to arrive at that is to change the integration condition of the differential equation $y(0) = y_0$ (at time equal to zero the initial condition equals $y_0$) into $y(-\Delta t) = y_{\text{ref}}$ (at time equal to $-\Delta t$ the value of $y$ equals $y_{\text{ref}}$).

Both the biological shift factor $\Delta t$ and the technical error $\varepsilon_{\text{tech}}$ in Eq. (3) are (assumed to be) normally distributed random variables with a mean and a standard deviation ($\Delta t \approx \mathcal{N}(\mu, \sigma_{\Delta t})$ and $\varepsilon_{\text{tech}} \approx \mathcal{N}(0, \sigma_{\varepsilon\text{tech}})$). For the technical error, that is the standard assumption in regression analyses, for the biological shift fact it was indicated in all analyses of real time data. It is clear that normality of $\Delta t$ corresponds to lognormality of $(y_0 - y_{\min})$.

Furthermore, $\Delta t$ and $\varepsilon_{\text{tech}}$ are independent variables. When applying different temperatures during storage, the rate constant $k$ depends on the applied temperature (according to Arrhenius' law). As a consequence, the biological shift factors estimated at different temperature conditions will also show different values, even for the same fruit. To compare biological shift factors at different temperatures, one can convert them to a dimensionless expression by multiplying with the actual rate constant at that particular temperature: $\Delta t^* = \Delta t \cdot k$.

Modern measuring techniques are, in general, most of the time accurate and reliable. That means that the technical error is relatively small. If not, the technique of measurement will never provide informative data and will in the long run be considered as unreliable and disappear completely.

So, Eq. (3) expresses the exponential behaviour including (or pointing to) the variation in the data. Of course for more complex formulations of the exponential behaviour, e.g., exponential increase towards an asymptote, or for completely different type of behaviour, the deduction has to be conducted anew. The line of reasoning, however, remains exactly the same: put the biological variation at the time variable as a biological shift factor.

## 3. Data analysis

Some data were generated using Eq. (3) with the input values as shown in Table 1 (21 measuring times for 15 repetitions) and will be used to explain and elucidate the technology. A script to run in the statistical package R is available in the Supplementary Material. This script can be used to analyse simulated new data with different input values as well as the data used here, with the three mentioned regression systems.