Postharvest Biology and Technology 132 (2017) 31–42

Contents lists available at ScienceDirect

Postharvest Biology and Technology

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

Assessing biological and technical variation in destructively measured data

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

[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
[d] Postharvest Technology Innovation Centre (King Mongkut's University of Technology Thonburi), Office of the Higher Education Commission, Thailand

ARTICLE INFO

Keywords:
Biological variation
Technical variation
Cross-sectional data
Non-destructive data
Statistical analysis

ABSTRACT

The majority of experimental data are obtained by destructive measuring techniques. Inevitably, in all these data variation is present, sometimes small and negligible, sometimes large, preventing proper analysis and extraction of meaningful information by traditional statistical techniques altogether. In this paper, three systems are presented to analyse destructive (cross-sectional) data, including biological as well as technical variation. The first system involves ranking the data per measuring point in time which provides a pseudo fruit number that can be used in non-linear indexed regression analysis similar as for non-destructive (longitudinal) data. The rationale behind this is that the individual with the highest value at some point in time will resemble the most another individual with the highest value at previous or future times, and the second highest the second highest at previous times, and so on. The second system also relies on this ranking number, but is now converted into a probability, which is used in non-linear regression analysis with quantile functions. The third system is based on optimising the log likelihood of the density function derived from the applied model (i.e., the expected distribution) over the measured data. Simulated data are used to elucidate the power of the three systems. A dataset on mango colour is used to validate the systems on a real-world data set. Although all three systems perform satisfactorily with percentages variability accounted for (R^2_{adj}) well over 90%, a clear preference cannot be given since the choice of the proper analysis system depends on the experimental conditions (number of data, individuals and sampling points in time). Non-linear indexed and non-linear regression with quantile functions delivered the most reliable estimates. The three systems open up the possibility to analyse and reanalyse destructively measured data providing a sufficient large number of individuals and a clear indication of the kinetic model is available.

1. Introduction

Variation in experimental data is always present, either biological or technical variation, or both. The presence of variation invariably creates difficulties in data interpretation. Sometimes the difficulties are latent and not well recognised, sometimes they are huge, and prevent extracting useful information altogether. Technical variation is the result of systematic errors, random errors and blunders, while biological variation originates from the properties of the measured produce that are different due to differences in stage of development. Classical statistical analysis applies robust statistical procedures developed over more than hundred years, but does not aim at a thorough mechanistic interpretation of the variation present. Standard statistical procedures that deal with variation assume that the observed variation is normally distributed, and if not, choose from a range of transformations to obtain normality. There is, however, substantial information to be gained by a proper analysis of the variation. In the last two decades, several reports have been published to address the biological variation in longitudinal data, i.e., data obtained by non-destructive measuring techniques, repeatedly measuring the same individual items over time, applying models based on confirmed or plausible reaction mechanisms and the rules of chemical kinetics (Hertog, 2002; Schouten et al., 2004, 2007; Tijskens and Wilkinson, 1996; Tijskens et al., 2007, 2008, 2009b, 2015b, 2016; Unuk et al., 2012). The importance of dealing with biological variation has been indicated (Hertog et al., 2004; Tijskens et al., 2003) and reviewed (Hertog et al., 2007; Jordan and Loeffen, 2013). Part of the statistical background has been reported (De Ketelaere et al., 2006; Tijskens et al., 1999, 2015b), all dealing with longitudinal data. All these studies on longitudinal data have proven that biological variation is not the result of random processes, but of distinct interactions between the underlying kinetic processes and therefore subject to deterministic rules. Unravelling these deterministic rules is the true task of

* Corresponding author at: Group Horticulture and Product Physiology, Wageningen University, Droevendaalse steeg 1, 6708 PB Wageningen, The Netherlands.
E-mail address: Pol.Tijskens@wur.nl (L.M.M. Tijskens).

http://dx.doi.org/10.1016/j.postharvbio.2017.05.013
Received 24 February 2017; Received in revised form 21 May 2017; Accepted 23 May 2017

modellers and data analysts.

The majority of experimental data, however, do not consist of longitudinal, but of cross-sectional data, i.e. obtained by destructive measuring methods using new samples from a large population at every measuring point in time. The deterministic rules of the behaviour of biological variation derived from analysing longitudinal data, should however, also apply to cross-sectional data: the way samples are taken does neither affect the processes involved, nor the resulting effects. In this paper, three methods for analysing cross-sectional data, taking biological and technical variation into account, are presented and compared. The first two systems rely on ranking measured data at each measuring time. The first system uses this ranking number as a pseudo fruit number, mimicking longitudinal data in indexed non-linear regression. The second system uses non-linear regression with quantile functions (Jordan and Loeffen, 2013) based on a probability derived from the ranking number. In the remainder of this paper this system will be called QF regression. The third method directly fits the distribution of measured data on the expected distribution for these measured data based on the assumed, plausible or proven model structure, optimising the log-likelihood (Schouten et al., 2010). This method follows the dynamics of the distributions rather than of individuals. Finally, the three systems are applied on simulated and experimental data. The results are compared to a standard non-linear regression analysis, which does not consider biological variation at all. The implications of the new cross sectional analysis techniques are discussed.

2. Material and methods

The methods will be demonstrated using simulated data applying models frequently encountered in horticulture: exponential behaviour (decay to a lower asymptote and production towards an upper asymptote) and logistic behaviour. The simulations mimic destructive measuring techniques: at every sampling point in time, new samples are taken at random based on the assumption that the biological shift factor (Δt, which expresses the state of development in the time dimension (Tijskens et al., 2005)) is distributed according to a normal distribution. This assumption has been found to be valid in all cited references using longitudinal data.

2.1. Model development

2.1.1. Exponential behaviour

Exponential behaviour, both decay as well as production, is frequently encountered in experimental data, e.g., firmness (Schouten et al., 2007, 2010; Tijskens et al., 2009a). Exponential behaviour is the result of a first order reaction (Eq. (1)).

$$S \xrightarrow{k} P \tag{1}$$

where S is the substrate, P the product and k the reaction rate constant. At constant external conditions (mainly temperature) and assuming an asymptotic value in substrate (S_{min}), indicating that part of S that is not accessible for breakdown, the analytical solution is shown in Eq. (2).

$$\begin{aligned} S(t) &= (S_0 - S_{min})\cdot e^{-k\cdot t} + S_{min} \\ P(t) &= P_0 + (S_0 - S_{min})\cdot(1 - e^{-k\cdot t}) \end{aligned} \tag{2}$$

where t represents time. The subscript 0 refers to the initial state and min to the lower asymptote. P_0 expresses the amount of product present at the start of the experiment. The asymptotic value for P at +infinite time is $S_0 - S_{min} + P_0$. That signifies that the asymptotic value of the product depends on the initial amount of substrate present (S_0). So, the more initial substrate, the higher the asymptote will be. This behaviour is frequently encountered in horticultural data sets (Tijskens et al., 2015a, 2016).

The model, based on exponential production (P(t) in Eq. (2)) was first proposed by von Bertalanffy to describe length increase as part of his General Systems Theory (von Bertalanffy, 1938) and further developed and promoted by Kooijman (1986, 1988), not only for growth in size and weight in plant material but especially for animal growth. Schouten et al. (2002) applied a similar model for the elongation of chrysanthemum internode length.

Assuming variation exists in the initial amount of substrate (S_0) these equations can be converted into the biological shift factor notation using Δt, the biological shift factor (Tijskens et al., 2005). Adding a random error (ε), indicating the technical or measuring error, one arrives at Eq. (3).

$$\begin{aligned} S(t) &= (S_{ref} - S_{min})\cdot e^{-k\cdot(t+\Delta t)} + S_{min} + \varepsilon \\ P(t) &= P_0 + (S_{ref} - S_{min})\cdot(e^{-k\cdot(t+\Delta t)} - e^{-k\cdot\Delta t}) + \varepsilon \end{aligned} \tag{3}$$

with Δt the biological shift factor, a stochastic variable expressing the stage of development of individual fruit ($\approx\mathcal{N}(\mu_{\Delta t}, \sigma_{\Delta t})$) and ε a stochastic variable ($\approx\mathcal{N}(0, \sigma_\varepsilon)$) expressing the technical variation or measuring error. S_{ref} is an arbitrarily chosen value (preferably around the midpoint of the overall range of change in substrate S), used as a reference value for the biological shift factor.

2.1.2. Logistic behaviour

Sigmoidal behaviour is often described by a logistic function. This equation can be derived (Schouten et al., 2007) from a massive simplification of an autocatalytic reaction mechanism (Eq. (4)).

$$S + E \xrightarrow{k} 2\cdot E + P \tag{4}$$

where S and P are the substrate and the product respectively, k the reaction rate constant while E represents the catalyst (e.g., an enzyme) in the autocatalytic reaction mechanism. From this mechanism, the differential equations can be deduced by applying the rules of chemical kinetics and solved at constant external conditions, yielding the logistic equation. Expressing the equation in the biological shift factor notation, assuming an asymptotic value in substrate (S_{min}) and adding a random error (ε) yields Eq. (5).

$$S(t) = \frac{S_{max} - S_{min}}{1 + e^{-k\cdot(S_{max} - S_{min})\cdot(t + \Delta t)}} + S_{min} + \varepsilon \tag{5}$$

where the subscript max refers to the upper asymptote and min to the lower asymptote. Δt is the biological shift factor, a stochastic variable expressing the stage of development of individual fruit. The reference point for the biological shift factor is taken at the midpoint of the sigmoidal curve.

2.2. Probelation: ranking data

Longitudinal data analysis exploits the major resemblance between repeated measurements on individual units over time. For cross-sectional data, this method cannot be applied since the sample is destroyed at every measuring point in time, and new samples have to be used. So, the same individual cannot be used anymore for future measurements. However, continuing on this line of reasoning on time related similarities, one could postulate that in a set of cross-sectional data the individual with the highest value at some point in time will resemble the most another individual with the highest value at previous or future times, and the second highest the second highest at previous times, and so on until the lowest value. One could assign an identification number based on the sorted order of the measured values per measuring point in time (a process called probelation: PROBing variance for PROBabilty).

Purely based on statistical reasoning, Gilchrist (1997, 2000, 2008) applied the same technology, tracing back the idea of ranking observations to Galton (1883). These authors call ordered numbers 'rankits', and connect that order number to a probability, used in QF regression (see below). So, not only from a physiological point of view, but also from a statistical point of view this type of ranking of measured