Techniques to assess biological variation in destructive data[a]

L.M.M. Tijskens[1,b], R.E. Schouten[1], G. Jongbloed[2] and P.J. Konopacki[3]

[1]Horticulture and Product Physiology, Wageningen University, the Netherlands; [2]Delft University of Technology, Institute of Applied Mathematics, Delft, the Netherlands; [3]Research Institute of Horticulture, Skierniewice, Poland.

Abstract

Variation is present in all measured data, due to variation between individuals (biological variation) and variation induced by the measuring system (technical variation). Biological variation present in experimental data is not the result of a random process but strictly subject to deterministic rules as found on non-destructive data. The majority of data obtained in research are obtained by destructive techniques. The rules on behaviour and magnitude of variation should however, also apply to these cross sectional data. New techniques have been developed for analysing cross sectional data including the assessment of variation: 1) Probelation. In a set of cross-sectional data, the individual with the highest value at some point in time will resemble the individual with the highest value at previous or future times, and the second highest the second highest at previous times, and so on. One can assign an identification number based on the sorted order of the measured values per measuring point in time. This number can be used as a pseudo fruit number in indexed or mixed effects regression analysis, similar to the data analysis of longitudinal data; 2) Density assessment. For not too complex kinetic processes the density function can be deduced. Measuring a large number of individuals (on a single point in time) provides the possibility to assess directly the variation in the data; 3) Quantile regression. This technique also relies on ranking the data per measuring time. The probelation number is now converted into a probability, and the mean and standard deviation is estimated directly along with the kinetic parameter, using simple non-linear regression. Based on simulated data sets, all three techniques are demonstrated, and the results compared with the input values. Explained parts (R^2_{adj}) obtained are generally well over 90%, provided that the technical variation is not excessively large.

Keywords: biological variation, technical variation, cross-sectional data, non-destructive data

INTRODUCTION

Everywhere in nature, variation is present. All living creatures differ from each other in one way or another. We are proud to be different from our neighbour. We are even different from the selves we were say 20 years ago. For horticultural and agricultural produce that is not different. So, in all measured data, variation is present. This reflects the variation between individuals (biological variation) or the variation induced by the measuring system (technical variation). The presence of variation makes data analysis difficult, sometimes preventing a useful analysis 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.

For data obtained by non-destructive measuring techniques (longitudinal data),

[a]To the memory of Marjan Simčič (1960-2016), believer in and promotor of studying biological variation from the start.

[b]E-mail: Pol.Tijskens@wur.nl

Acta Hortic. 1194. ISHS 2018. DOI 10.17660/ActaHortic.2018.1194.194
Proc. VIII International Postharvest Symposium:
Enhancing Supply Chain and Consumer Benefits – Ethical and Technological Issues
Eds.: F. Artés-Hernández et al.

statistical procedures have been developed and applied (non-linear mixed effect regression, non-linear indexed regression, quantile regression). Studies on longitudinal data (Hertog, 2002; Hertog et al., 2004, 2007; De Ketelaere et al., 2006; Schouten et al., 2004, 2007; Tijskens and Wilkinson, 1996; Tijskens et al., 1999, 2003, 2007, 2008, 2009b, 2015a, b, 2016; Unuk et al., 2012) have proven that biological variation is not at all the result of a random process, but of distinct interactions of underlying kinetic processes. In other words, biological variation present in experimental data is strictly subject to deterministic rules.

The majority of data obtained in research are, however, obtained by destructive techniques using new samples from a large population at every measuring point in time. The rules on behaviour and on magnitude of variation should however, also apply to these cross sectional data. New techniques have been developed for analysing cross sectional data including the assessment of variation. The aim of this paper is to present, explain and highlight the working of these novel techniques to analyse cross-sectional data (i.e., obtained by destructive measuring techniques), while taking proper care of the existing variation. With exponential (decay and production) as examples, the techniques will be elucidated based on simulated data.

MATERIALS AND METHODS

The models

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). The model formulations, expressed in biological shift factor notation (Δt) are shown in Equation 1 (exponential decay), exponential production (Equation 2). Detail on the deduction and formulation of the three models can be found in Tijskens et al. (2015b, 2016).

$$S(t) = (S_{ref} - S_{min}) \cdot e^{-k\cdot(t+\Delta t)} + S_{min} + \varepsilon \quad (1)$$

$$P(t) = P_0 + (S_{ref} - S_{min}) \cdot \left(e^{-k\cdot(t+\Delta t)} - e^{-k\cdot\Delta t}\right) + \varepsilon \quad (2)$$

with t representing time, S the substrate, P the product and k the reaction rate constant. S_{ref} is an arbitrarily chosen value (preferably around the midpoint of the overall range of change in substrate), used as a reference value for the biological shift factor. Subscripts min refers to the lower asymptote. Δt is the biological shift factor, a stochastic variable expressing the stage of development of individual fruit ($\approx \mathbf{N}(\mu_{\Delta t},\sigma_{\Delta t})$) and ε a stochastic variable ($\approx \mathbf{N}(0,\sigma_{\varepsilon})$) expressing the technical variation or measuring error.

Part of the origin of biological variation as well as the meaning, working and application of the biological shift factor (Δt) is explained in Figure 1.

The methods

Three methods have been developed to analyse destructive data while taking care of the existing variation.

1. Probelation.

PROBing variance for PROBabilty. In a set of cross-sectional data, the individual with the highest value at some point in time will resemble the individual with the highest value at previous or future times, and the second highest the second highest at previous times, and so on. One can assign an identification number based on the sorted order of the measured values per measuring point in time. This (probelation) number can be used as a pseudo fruit number in indexed or mixed effects regression analysis, similar to the data analysis of 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, in other words simulating non-destructive data. After probelation data can be

estimated using indexed non-linear regression ('nlsi').

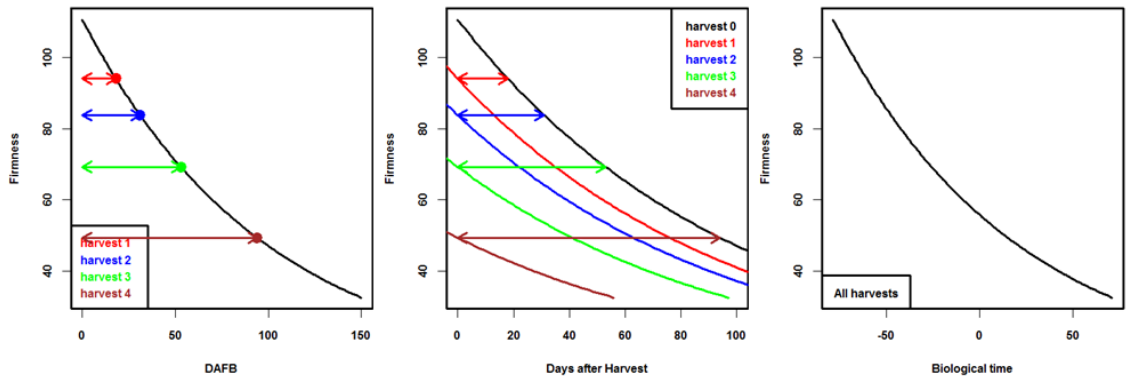

Figure 1. Explaining the biological shift factor. Left: Generic development of individuals expressed in time after full bloom (DAFB). Centre: expressed in time after harvest (as usual). Generic development appears shifted. Right: Estimate that shift (= biological shift factor) and express time as biological time (t+Δt) to find again the original generic development curve.

2. Quantile regression.

This technique also relies on ranking the data per measuring time. The probelation number is now (linearly) converted into a probability (Equation 3), which can be used to estimate the mean and standard deviation of the underlying property based on the inverse of the cumulative distribution (here a normal one on Δt, procedure 'qnorm' in R), directly along with the kinetic parameter using simple non-linear regression, indicated in the figures as 'nlsQ' (Jordan and Loeffen, 2013; Tijskens et al., 2015b).

$$\mathrm{Pr} = \frac{\mathrm{PN} - ½}{\mathrm{n_{obs}}} \tag{3}$$

3. Density assessment.

For not too complex kinetic processes, the density function can be deduced that describes at any time of development the distribution in measured values changing shape during development (Hertog et al., 2004; Schouten et al., 2004, 2010). For exponential decay, it is shown in Equation 4. When a large number of individuals is measured in time or even at a single point in time, that density function can be used to assess directly the magnitude and behaviour of variation in the data. Data analysis using this method (procedure 'fitdistr' in the Mass package of R), is indicated as 'dens' in the figures below.

$$\mathrm{p(S)} = \frac{\mathrm{e}^{\left(-\frac{1}{2\cdot(\mathrm{k}\cdot\sigma)^2}\cdot\left(\ln\left(\frac{\mathrm{S}-\mathrm{S_{min}}}{\mathrm{S_{ref}}-\mathrm{S_{min}}}\right)+\mathrm{k}\cdot(\mathrm{t}+\mu_{\Delta\mathrm{t}})\right)^2\right)}}{\sqrt{2\cdot\pi}\cdot\sigma_{\Delta\mathrm{t}}\cdot(\mathrm{S}-\mathrm{S_{min}})\cdot\mathrm{k}} \tag{4}$$

As minimising criterion, the log Likelihood is used:

$$\mathrm{logLik} = \sum_{\mathrm{i}=1}^{\mathrm{N_{obs}}} \ln(\mathrm{p_{y_i}}) \tag{5}$$

Data generation

Experimental data, either longitudinal or cross-sectional, are not very useful to establish the power and possibilities of analysing techniques: neither the underlying kinetic mechanism nor the parameters are known. When using simulated data, estimated parameter values can be compared to the values of the input parameters. Stochastic simulation is a very important tool in assessing the quality of statistical procedures. Data were generated using Equations 1 and 2 as model with the parameters as shown in Table 1. These values are completely arbitrary, but reflect behaviour as can be encountered in real time experiments.

Table 1. Values for the input parameters for the models used in simulated data generation (Equations 1 and 2).

Parameter	Exp. decay	Exp. prod
P_0	na	10
S_{min}	10	0
S_{ref}	50	50
k	0.05	0.05
$\mu_{\Delta t}$	-10	-20
$\sigma_{\Delta t}$	from .3 to 8.0	
σ_{ε}	from .1 to 8.0	
n_{rep}	100	100
n_{times}	21	21

na: not applicable.

RESULTS

Simulated data

An example of the raw simulated data for both exponential models is shown in Figure 2 (top row). This representation provides a fair overview of the behaviour of the models and the induced variation, both biological and technical. Clearly, it can be observed that the variation in data values, i.e., the width of the cloud of points around the mean behaviour, changes with time. For the exponential decay model the largest variation is at the start of the process, for the exponential production model at the end of the process. That the standard deviation of the distribution changes with time, can be observed directly in the simulated data.

In Figure 2 (bottom row), the same examples are shown but now standardised (versus biological time (t+Δt) for exponential decay (left) and relative to the asymptotic value for exponential production (right)) to elucidate the power of including biological variation in the analysis. To assess the effect of the magnitude of variation, biological (Δt) as well as technical (ε), data were simulated with a range of values for their standard deviation ($\sigma_{\Delta t}$, σ_{ε} in Table 1) and analysed using simple non-linear regression and the three systems mentioned above. The results for exponential decay are graphically presented in Figures 3 and 4.

Provided the biological variation ($\sigma_{\Delta t}$) and the technical variation (σ_{ε}) are relatively small, up to about 2.5% of the total range of change (here 100), the analysing systems that include variation, estimate the kinetic parameters rather well. Even the simple regression system ('nls') estimates the kinetic constants rather well, considering no variation at all is taken into account. At higher variation (higher $\sigma_{\Delta t}$ and σ_{ε}), the estimates tend to show larger differences with the input value. Most susceptible to the magnitude of variation is the asymptote variable (S_{min}), especially as a function of the technical variation. The density analysis seems to be more erratic than the regression system both with regard the technical and the biological variation. Since the only assumption underlying this system is the correctness of the model structure, reflected in the density function, the possibilities to

assign variance at different parameters is much larger. In other words, the framework for this system is less rigid than for the regression systems, resulting in less performance compared to the probelation and quantile regression methods.

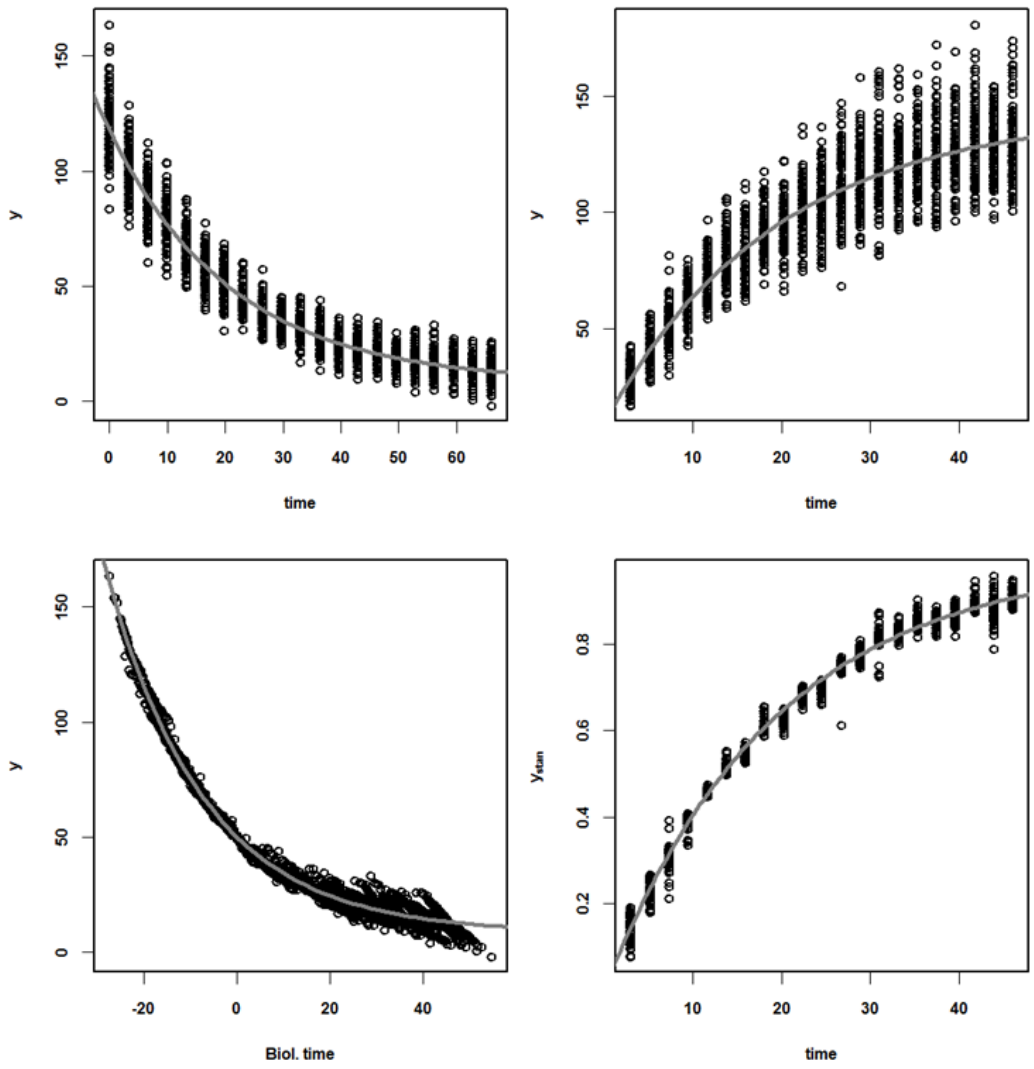

Figure 2. Simulated data. Top row: behaviour of the raw data; bottom row: standardised data as a function of estimated biological time (t+Δt). Left column: exponential decay; right column: exponential production. the input parameters as shown in Table 1, with the technical error (σε) 5 and the biological variation (σδt) 2.5. lines represent the behaviour estimated with indexed non-linear regression.

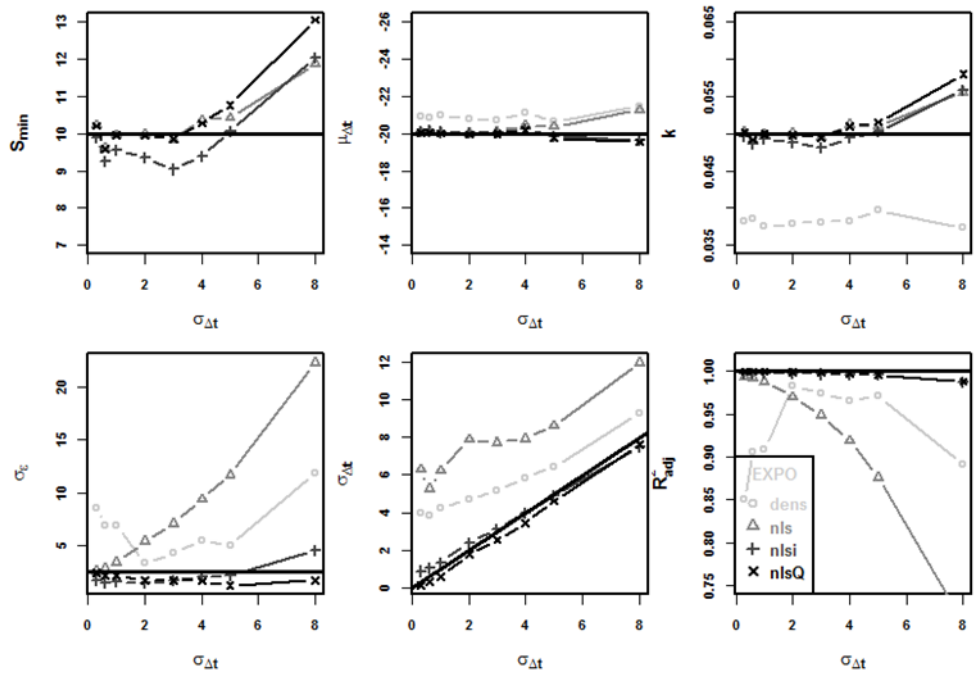

Figure 3. Results for the different methods of analysis for the model first order decay as a function of increasing biological variation. The input value of the variables is represented by the full black line. For clarity, the y-axis in the upper row is limited to ±30% of the input value.

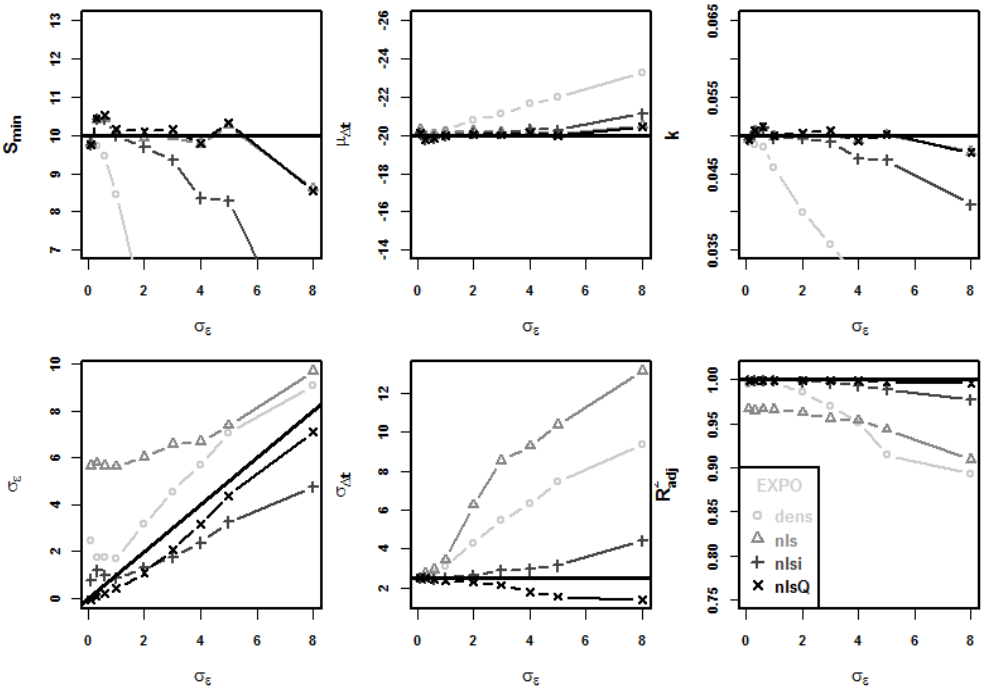

Figure 4. Results for the different methods of analysis for the model first order decay as a function of increasing technical variation. The input value of the variables is represented by the full black line. For clarity, the y-axis in the upper row is limited to ±30% of the input value.

For the estimation of the variances ($\sigma_{\Delta t}$ and σ_{ε}) and the reliability (R^2_{adj}), the picture is completely different (bottom rows in Figures 3 and 4). As a function of technical variation (Figure 4), the estimated technical variation of the analysis without taking care of the biological variation starts rather high, and the explained parts (R^2_{adj}) is initially quite high, but drops quickly as a function of the technical variation. When the technical variation is low all analysis systems that take care of the biological variation estimate $\sigma_{\Delta t}$ and σ_{ε} rather well. At higher variation the estimated values start to deviate from the thick black horizontal lines which indicate the input values Density analysis under-estimates, while the regression systems tend to over-estimate. The obtained explained parts, when taking care of the biological variation is always extremely high, well over 98%.

When the simulations and analyses are conducted as a function of biological variation (Figure 3), a similar pattern is obtained for the estimated technical variation and explained parts. For the exponential decay model, the 'nls' and the density system provides a considerable over estimation of biological variation. Results for exponential production and for logistic behaviour reflect similar behaviour.

CONCLUSIONS

Probelation combined with analysis systems that take variation into account proves to be a very powerful tool to enable the analysis of destructively obtained (cross-sectional) data. In other words, the rules and relations, derived in many studies on non-destructive data are applicable on destructive data as well. Splitting up the total variation in the data in a biological part, determined completely by the occurring processes, and a technical part is the key issue to make the analyses more reliable.

The regression systems, indexed and quantile regression, delivered the most reliable estimates. Technical variation, i.e., measuring errors and human errors, were shown to be the major source of unreliable estimates but the biological variation is estimated rather well. The density analysis puts the major emphasis on the distribution of the measured variables, rather than on the kinetics of the process (based on sum of squares). Explained parts (R^2_{adj}) obtained are generally well over 90%, provided that the technical variation is not excessively large.

ACKNOWLEDGEMENTS

Parts of the presented work were developed for application in the framework of the project MONALISA, funded by the Autonomous Province of Bolzano, and in the framework of the large ISAFRUIT EU project (Contract no. FP6-FOOD 016279-2).

Lierature cited

De Ketelaere, B., Howarth, M.S., Crezee, L., Lammertyn, J., Viaene, K., Bulens, I., and De Baerdemaker, J. (2006). Postharvest firmness changes as measured by acoustic and low-mass impact devices: a comparison of techniques. Postharvest Biol. Technol. *41* (*3*), 275–284 https://doi.org/10.1016/j.postharvbio.2006.04.008.

Hertog, M.L.A.T.M. (2002). The impact of biological variation on postharvest population dynamics. Postharvest Biol. Technol. *26* (*3*), 253–263 https://doi.org/10.1016/S0925-5214(02)00044-3.

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* (*3*), 271–284 https://doi.org/10.1016/j.postharvbio.2004.05.014.

Hertog, M.L.A.T.M., Lammertyn, J., De Ketelaere, B., Scheerlinck, N., and Nicolaï, B.M. (2007). Managing quality variance in the postharvest food chain. Trends Food Sci. Technol. *18* (*6*), 320–332 https://doi.org/10.1016/j.tifs.2007.02.007.

Jordan, R.B., and Loeffen, M.P.F. (2013). A new method for modelling biological variation using quantile functions. Postharvest Biol. Technol. *86*, 387–401 https://doi.org/10.1016/j.postharvbio.2013.07.008.

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* (*3*), 299–310 https://doi.org/10.1016/j.postharvbio.2003.12.005.

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* (*3*), 298–306 https://doi.org/10.

1016/j.postharvbio.2007.03.011.

Schouten, R.E., Natalini, A., Tijskens, L.M.M., Woltering, E.J., and van Kooten, O. (2010). Modelling the firmness behaviour of cut tomatoes. Postharvest Biol. Technol. *57* (*1*), 44–51 https://doi.org/10.1016/j.postharvbio.2010.02.001.

Tijskens, L.M.M., and Wilkinson, E.C. (1996). Behaviour of biological variability in batches during post-harvest storage. Paper presented at: AAB Modelling Conference: Modelling in Applied Biology - Spatial Aspects (Uxbridge, UK: Brunel University).

Tijskens, L.M.M., Hertog, M.L.A.T.M., van Kooten, O., and Simčič, M. (1999). Advantages of non-destructive measurements for understanding biological variance and for modelling of the quality of perishable products. Paper presented at: 34. Vortragstagung der DGQ: Zurstörungsfreie Qualitätsanalyse (Freising Weihenstephan).

Tijskens, L.M.M., Konopacki, P., and Simčič, M. (2003). Biological variance, burden or benefit? Postharvest Biol. Technol. *27* (*1*), 15–25 https://doi.org/10.1016/S0925-5214(02)00191-6.

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* (*2*), 204–213 https://doi.org/10.1016/j.postharvbio.2007.01.014.

Tijskens, L.M.M., Konopacki, P.J., Schouten, R.E., Hribar, J., and Simčič, M. (2008). Biological variance in the colour of Granny Smith apples. Modelling the effect of senescence and chilling injury. Postharvest Biol. Technol. *50* (*2-3*), 153–163 https://doi.org/10.1016/j.postharvbio.2008.05.008.

Tijskens, L.M.M., Dos-Santos, N., Jowkar, M.M., Obando-Ulloa, J.M., Moreno, E., Schouten, R.E., Monforte, A.J., and Fernández Trujillo, J.P. (2009a). Postharvest firmness behaviour of near-isogenic lines of melon. Postharvest Biol. Technol. *51* (*3*), 320–326 https://doi.org/10.1016/j.postharvbio.2008.06.001.

Tijskens, L.M.M., Unuk, T., Tojnko, S., Hribar, J., and Simčič, M. (2009b). Biological variation in the colour development of golden delicious apples in the orchard. J. Sci. Food Agric. *89* (*12*), 2045–2051 https://doi.org/10.1002/jsfa.3689.

Tijskens, L.M.M., Schouten, R.E., Walsh, K.B., Zadravec, P., Unuk, T., Jacob, S., and Okello, R.C.O. (2015a). Harvesting quality, where to start? Acta Hortic. *1099*, 269–276 https://doi.org/10.17660/ActaHortic.2015.1099.30.

Tijskens, L.M.M., Schouten, R.E., Konopacki, P., and Jongbloed, G. (2015b). Basic principles of analysing biological and technical variation in non-destructive data. Comput. Electron. Agric. *111*, 121–126 https://doi.org/10.1016/j.compag.2014.12.022.

Tijskens, L.M.M., Unuk, T., Okello, R.C.O., Wubs, A.M., Šuštar, V., Šumak, D., and Schouten, R.E. (2016). From fruitlet to harvest: modelling and predicting size and its distributions for tomato, apple and pepper fruit. Sci. Hortic. (Amsterdam) *204*, 54–64 https://doi.org/10.1016/j.scienta.2016.03.036.

Unuk, T., Tijskens, L.M., Germšek, B., Zadravec, P., Vogrin, A., Hribar, J., Simčič, M., and Tojnko, S. (2012). Effect of location in the canopy on the colour development of three apple cultivars during growth. J. Sci. Food Agric. *92* (*12*), 2450–2458 https://doi.org/10.1002/jsfa.5651. PubMed