ISSN 1946-7664. MCFNS 2018
Vol. 10, Issue 1, pp. 1–5
*Mathematical and Computational Forestry & Natural-Resource Sciences*

AVAILABLE ONLINE AT HTTP://MCFNS.COM

Submitted: Jan. 31, 2018
Accepted: Mar. 12, 2018
Published: Mar. 30, 2018
Last Correction: Apr. 4, 2018

# REVERSE CAUSALITY IN SIZE-DEPENDENT GROWTH

Oscar García
*Dasometrics, Concón, Chile*

Abstract. *Size-dependent growth* is likely to be *growth-dependent size* instead. Larger organisms do not necessarily grow faster, but faster-growing ones always tend to be larger. This fact has been generally ignored. Correct causality structures are essential for plausible predictions outside the range of the data. Some techniques potentially useful for studying these issues are briefly described. In forestry, the relevance of multiple size measures like volume, height, diameter and basal area complicates the picture. Additionally, purely mathematical sources of growth-size correlations arise. Physiological considerations suggest avoiding stem thickness measures as explanatory variables in growth equations.

**Keywords:** Confounding, bias, consistency, path analysis, structural equation modelling, mixed effects, endogenous variables, instrumental variables, allometry.

## 1 Introduction

Growth models typically utilize regressions of growth rate on size, possibly with additional independent variables (Weiskittel et al. 2011). A causal relationship is usually implied, with growth rate depending on current size. Size-dependent growth is also the subject of theoretical generalizations that have attracted much attention (e.g. Damuth 2001, Sheil et al. 2017, and references therein). However, size is an accumulation of past growth, and therefore faster growth causes larger size, not necessarily the other way around. Although once stated this observation seems obvious, the only reference related to biological growth that I have found is in Perry (1985): "past competitive interactions are integrated in current tree size." The statistical difficulties have been recognized in econometrics where, for a linear discrete-time growth model, least-squares estimation is known to be biased and inconsistent (Bun and Sarafidis 2015, this is discussed in detail below in Section 3).

Does it matter? The direction of causality may not be important for prediction in populations similar to the one originating the data. For instance, for yield forecasting of unmanaged or lightly managed forest stands, as in many growth model applications (Weiskittel et al. 2011). Or with intensive management where sufficient data is available, so that growing conditions are largely interpolated rather than extrapolated (Goulding 1994). On the other hand, causality rather than just correlation is crucial for understanding process dynamics, and for prediction under circumstances not represented in the data.

The section that follows explains further the statistical confounding arising from causal ambiguity. Sections 3, 4 and 5 sketch some techniques that might help in understanding those issues, and perhaps eventually in dealing with them more effectively. All that assumes that "size" is expressed as a single number, typically biomass or volume. This is the case most commonly found in the literature, and demonstrates most clearly the main problems. Although a one-dimensional size can be appropriate for animals or annual plants, tree size is more accurately described by at least two dimensions, height and radius (or diameter, circumference or cross-sectional area). Longitudinal and radial growth derive from different meristems, and respond differently to various growing conditions. This multivariate description complicates things, giving rise to additional, purely mathematical, sources of correlation that are discussed in Section 6. The paper closes with a Conclusions section.

## 2 The problem

A minimal example may be useful for exposing the essentials (Lee and García 2016). Assume that tree annual volume increment $\Delta v$ in a forest stand is constant over time, and that it is not affected by tree size. Nevertheless, $\Delta v$ varies among trees due to genetics, microsite, competition, or other factors. Take 3 trees with increments of 2, 5, and 10 $dm^3$ / year. The tree volumes $v$ at

age 10 are 10 times the increment, i.e., 20, 50, and 100 $dm^3$. Plotting increment over size (Figure 1) indicates a perfect regression model $\Delta v = 0.1\, v$ (or more generally, $\Delta v = v/t$). This model produces exact predictions for trees of any size, but it is biologically "wrong", in the sense that in this instance $\Delta v$ does not causally depend on $v$. The example could be embellished by introducing growth variability and measurement error, and by using larger samples of trees. Then the regression would not be error-free, but the predictions would still be better than those from any model not including $v$ as a predictor. The wrong model is best.

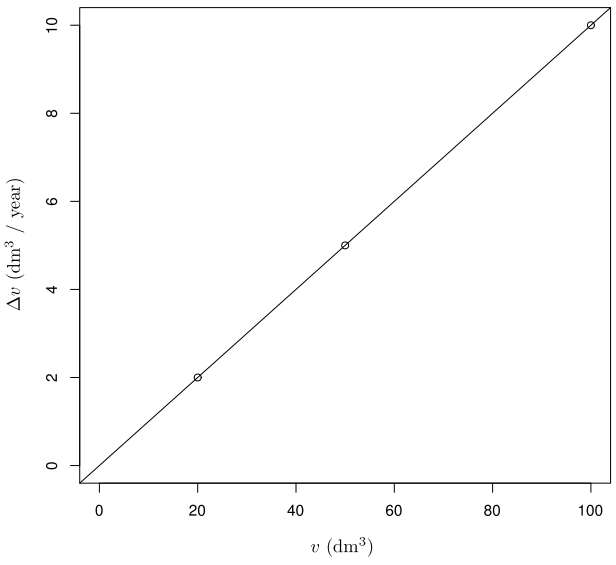


Figure 1: Artificial example of a perfect non-causal relationship, estimating tree volume growth increment from current tree volume (see text).

Lee and García (2016) analyzed real growth data from spruce-hardwoods mixtures in British Columbia. The best tree volume growth rate estimates not using current volume had $R^2$ values of 0.62 with spatially explicit competition indices, and 0.75 with distance independent competition indices[1]. In contrast, a simple linear regression $\Delta v = \beta_0 + \beta_1 v$ had $R^2 = 0.86$.

In general, consider a regression for growth rate

$$_{\Delta}y = f(y, \boldsymbol{x}) \,, \tag{1}$$

where $y = y(t)$ is current size, and $\boldsymbol{x}$ is a vector of additional independent variables. The dependent variable $\Delta y$ could be an annual increment $y(t+1) - y(t)$, a periodic increment $y(t+k) - y(t)$ or $[y(t+k) - y(t)]/k$, or even an instantaneous increment in continuous time, $\mathrm{d}y/\mathrm{d}t$. In forestry, these regressions have been called *self-referencing* models (Northway 1985, Strub and Cieszewski 2012). Because $y$ is an accumulation of past values of $\Delta y$ for each individual, in a heterogeneous population the two variables are correlated. That can produce good fit statistics, even if $\Delta y$ is not causally dependent on $y$. The fit is good not necessarily because larger individuals grow faster, but because faster-growing individuals tend to be larger. Predictions are based essentially on an extrapolation of past growth rates (Lee and García 2016). The extrapolation can fail if there is a change in the growing conditions that prevailed in the sample (e.g., Russell et al. 2015).

Of course, students of statistics are taught that *correlation is not causation*, and that models should not be used outside the range of the data. But the first point is often forgotten in the interpretation of research results. And models are invariably pushed beyond the comfort zone. After all, if we had enough data for all the conditions of interest there would be little need for models.

## 3 The dynamic panel data model in econometrics

Panel data consists of observations at consecutive times $t = 1, 2, \ldots, T$ on each of $N$ items or individuals $i = 1, 2, \ldots, N$. The (linear) dynamic panel data model can be written as

$$y_{it} = \alpha y_{i,t-1} + \boldsymbol{\beta}' \boldsymbol{x}_{it'} + u_{it} \,. \tag{2}$$

Here the vector $\boldsymbol{x}_{it'}$ may include various regressors observed at times such as $t' = t$, $t' = t-1$, $t' = t-2$, etc. The error term $u_{it}$ varies across individuals as

$$u_{it} = \lambda_i + \epsilon_{it} \,, \tag{3}$$

where $\lambda_i$ is *unobserved individual heterogeneity*, and the $\epsilon_{it}$ are errors with mean 0 and equal variance, independent across individuals and times.

The $y_{it}$ depend on the value of $\lambda_i$, so that in particular the regressor $y_{i,t-1}$ and the error $u_{it}$ in eq. (2) are correlated. In econometric terminology $y_{i,t-1}$ is *endogenous* (correlated with the error term), as opossed to *exogenous* (independent of the error term as assumed in standard linear regression). Consequently, it is found that the ordinary least-squares (OLS) estimate of $\alpha$ is biased. Worse, OLS is inconsistent, that is, estimates do not converge to the true values as $N \to \infty$ for fixed $T$.

Dynamic panel data models have been used to study size-dependent growth of firms and other organizations. To put them in the notation of Section 2, re-label eq. (2)

[1] Unlike most competitions indices, those used in the study did not contain embedded tree diameter measurements. Presumably the perfect plasticity assumption behind the aspatial indices is better than the assumption of no plasticity in the spatial ones (Strigul et al. 2008).

as

$$y_{i,t+1} = \alpha y_{it} + \boldsymbol{\beta}' \boldsymbol{x}_{it} + u_{it} \ , \qquad (4)$$

making use of the fact that $u_{it}$ has the same distribution for all $t$. Then,

$$\Delta y_{it} = y_{i,t+1} - y_{it} = (\alpha - 1) y_{it} + \boldsymbol{\beta}' \boldsymbol{x}_{it} + \lambda_i + \epsilon_{it} \ . \qquad (5)$$

Bun and Sarafidis (2015) review estimation methods for the dynamic panel data model, and point out that there is an implicit stationarity assumption. That might not be suitable for biological growth modelling, where interest focuses on transients far from a steady state. Econometric methods may or may not be useful for growth parameter estimation, but the main point is the recognition that OLS fails when size appears on the right-hand side of the growth rate regression.

## 4 A mixed effects view

A key property of the dynamic panel data model is the presence of individual heterogeneity, $\lambda_i$. More generally, the regression model of eq. (1) may be written as

$$\Delta y_{it} = f(y_{it}, \boldsymbol{x}_{it}, \boldsymbol{\gamma}, \boldsymbol{\lambda}_i, \epsilon_{it}) \ , \qquad (6)$$

for observations on individuals $i = 1, \ldots, N$ at times $t = t_{i1}, \ldots, t_{i,T_i}$. In growth data the number and timing of measurements on each individual and the time intervals can vary. The *local* parameters $\boldsymbol{\lambda}_i$ are specific to individual $i$, while the *global* parameters $\boldsymbol{\gamma}$ are common to all. In econometrics, locals are called incidental parameters, globals are called structural, and $\epsilon_{it}$ is an *idisioncratic error*.

For instance, the minimal example of Section 2 might be written as

$$\Delta v_{it} = \gamma v_{it} + \lambda_i + \epsilon_{it} \ , \qquad (7)$$

where we assumed $\gamma = 0$, $\lambda_1 = 2$, $\lambda_2 = 5$, and $\lambda_3 = 10$, and $\epsilon_{it}$ was ignored.

The model of eq. (6) could be estimated, for instance, by maximum likelihood. The locals can be considered either as fixed unknown individual-specific parameters, or as random variables representing sampling from some hypothetical meta-population of individuals (García 2017b, Sec. 3.3). The random locals alternative is far more popular nowadays, corresponding to a mixed effects model. The parameters could be estimated with mixed-effects software. Note however that, as indicated in Section 2, standard fit statistics will be worse than those for the "wrong" model that ignores individual heterogeneity.

## 5 Path analysis and structural equation models

Path analysis is a technique for studying causal relationships, developed by Sewell Wright in the 1920's (Wright 1921, Bollen 2005a). Later, inference methods were refined in Structural Equation Modelling (SEM; Bollen 2005b, Fox 2006, Umbach et al. 2017). Applications have been mostly in the social sciences, but more recently biological uses have increased (Iriondo et al. 2003, Lamb et al. 2011).

The approach allows for testing postulated causal models, assessing if they are consistent with the data. The model is commonly visualized in a path diagram. Figure 2 shows a tentative path diagram for the example of Section 2. Variables are classified as observed, unobserved (also called *latent*), or disturbances (errors). Each observed variable is enclosed in a box. Latent variables appear in ellipses or ovals. Errors are not enclosed in either, or are sometimes placed in ovals or circles since they are also unobserved. Arrows between variables indicate direct causal influences (*paths*). In path analysis, a *path coefficient* associated to each path is calculated.

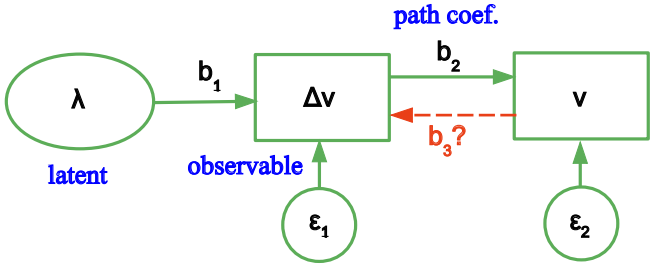


Figure 2: A path diagram for the model behind the example of Section 2. The dashed arrow represents a possible "real" size effect.

For the example, $v$ and $\Delta v$ are the observed variables size and growth rate, respectively. The unobserved variable $\lambda$ represents the growth rate intrinsic to each tree. Together with random variation (assumed to be 0 in the numerical example), it determines the actual growth increment. Growth increments cause the size to increase. A size observation error has been included in the diagram. SEM could be used to evaluate the consistency of this model with observed data, and to compare it to an alternative that adds a direct causal effect of size on growth, represented by the dashed arrow ($\gamma \neq 0$ in eq. (7)).

## 6 Which size?

In trees, growth in stem volume or biomass is the result of two fundamentally different processes, growth in height, and growth in radius, diameter or cross-sectional area. Height (and branch-length) growth results from the activity of apical meristems, while radial growth is generated in the cambium. It is well-known that height growth is less sensitive to competition than radial growth, so that height-diameter relationships vary

depending on past stand densities. Therefore, allometric relationships that estimate volume or biomass from diameter can only be accurate for growing conditions similar to those in the data from which they are derived[2].

It follows that uni-dimensional growth-size relationships are unsatisfactory for trees and forests. It is necessary to take into account simultaneously the dimensional components: at least height, diameter, and in forest stands, number of trees.

Confusion also arises from the use of different variables in growth-size relationships: biomass or volume, or the more easily obtained diameter or basal area. In fact, the behavior of their increments is quite different, as shown by Assmann (1970, p. 151). His result for stand volume *vs* basal area can be easily proved in continuous time, with instantaneous increments denoted as $\dot{V} = \mathrm{d}V/\mathrm{d}t$. Assume that $V = \alpha + \beta BH$, where $V$ is stand volume, $B$ is basal area per hectare, and $H$ is mean or top height (for simplicity we avoid the volume as product of basal area and form-height used by Assmann). Then, differentiating,

$$\dot{V} = \beta(\dot{B}H + B\dot{H}) \,. \tag{8}$$

This is essentially equivalent to Assmann's equation. Solving for $\dot{B}$ gives

$$\dot{B} = \frac{\dot{V}}{\beta H} - \frac{\dot{H}}{H} B \,, \tag{9}$$

which shows that even if $\dot{V}$ is independent of size, $\dot{B}$ does depend on $B$ (and $H$). Similarly, Lee and García (2016) show that if a tree volume is approximated in terms of tree dbh $d$ and height $h$ by $\alpha + \beta d^2 h$,

$$\dot{d} = \frac{\dot{v}}{2\beta h d} - \frac{\dot{h} d}{2h} \,. \tag{10}$$

For models to have a chance of performing acceptably outside the range of the data, they need to reflect the causal logic of the biological processes. Good fit statistics for purely empirical relationships are not sufficient. It makes sense to have growth in biomass, or in its proxy stem volume, as a dependent variable reflecting carbon capture and accumulation. Height influences the costs of evapotranspiration, and also dominance relationships in the case of individual trees or cohorts, so that it is a reasonable predictor. Stand density is also an important factor. Stem diameter, however, reflects the accumulation of mostly dead xylem on the stems, and there is little physiological justification for including it on the equation right-hand sides; the same is true for volume or basal area (García 2017a).

[2] Allometry is used here in the original sense of Huxley (1932), a (usually power) function of a single independent variable. The term is often misused as referring to any arbitrary volume or biomass function.

## 7 Conclusions

*Size-dependent growth* may actually be *growth-dependent size*. The direction of causality is not important for management that does not deviate markedly from the conditions represented in the data. Empirical extrapolation of past growth can be highly effective. Increasingly though, models are applied to new situations, including natural or management disturbances and environmental change. Understanding the system and proper causal modelling are then essential.

It can be difficult to disentangle the true causal effects, although mixed-effects modelling, path analysis and structural equation models seem promising tools. My presentation of those topics has been brief and tentative. I admit not fully understanding all aspects of the problem and of the techniques, and that some of the details might not be entirely correct. Obviously, more research is needed. But I believe that the important point is to get researchers to recognize that there is a problem, something that has not happened.

With trees and forests, simple one-dimensional growth-size models are unsatisfactory, because size components like height and diameter are important and react differently to growing conditions. Purely mathematical sources of correlation from these variables complicate the picture. Multidimensional systems of growth equations are needed. Still, adequate causal structures are essential if forecasts for previously unobserved situations are desired. In particular, it is suggested that diameter and basal area should be banished from the right-hand side of growth equations (García 2017a). Experiments might be useful where tree size and growing conditions are uncoupled, e.g., through randomized (not selective) thinning.

## Acknowledgements

I am grateful to anonymous reviewers for suggestions that contributed to improve the text.

## References

Assmann, E., 1970. The Principles of Forest Yield Study. Pergamon Press, Oxford, England. 506 p.

Bollen, K. A., 2005a. Path analysis. In Encyclopedia of Biostatistics, volume 6, Armitage, P., and T. Colton, eds., second edition, pp. 3973–3977. Wiley.

Bollen, K. A., 2005b. Structural equation models. In Encyclopedia of Biostatistics, volume 6, Armitage, P., and T. Colton, eds., second edition, pp. 5269–5278. Wiley.

Bun, M. J. G., and V. Sarafidis, 2015. Dynamic panel data models. In The Oxford Handbook of Panel Data,

Baltagi, B. H., ed., chapter 3, pp. 76–110. Oxford University Press.

Damuth, J., 2001. Scaling of growth: Plants and animals are not so different. Proceedings of the National Academy of Sciences 98(5):2113–2114.

Fox, J., 2006. Teacher's corner: Structural equation modeling with the sem package in R. Structural Equation Modeling 13(3):465–486.

García, O., 2017a. Cohort aggregation modelling for complex forest stands: Spruce-aspen mixtures in British Columbia. Ecological Modelling 343:109–122.

García, O., 2017b. Estimating reducible stochastic differential equations by conversion to a least-squares problem. ArXiv e-prints (arXiv:1710.06021 [stat.ME]). URL: https://arxiv.org/abs/1710.06021.

Goulding, C. J., 1994. Development of growth models for *Pinus radiata* in New Zealand — experience with management and process models. Forest Ecology and Management 69(1–3):331–343.

Huxley, J. S., 1932. Problems of Relative Growth. Methuen & Co., London. (Second Edition, Dover 1972).

Iriondo, J. M., M. J. Albert, and A. Escudero, 2003. Structural equation modelling: an alternative for assessing causal relationships in threatened plant populations. Biological Conservation 113(3):367–377.

Lamb, E., S. Shirtliffe, and W. May, 2011. Structural equation modeling in the plant sciences: An example using yield components in oat. Canadian Journal of Plant Science 91(4):603–619.

Lee, M. J., and O. García, 2016. Plasticity and extrapolation in modeling mixed-species stands. Forest Science 62(1):1–8.

Northway, S. M., 1985. Notes: Fitting site index equations and other self-referencing functions. Forest Science 31:233–235.

Perry, D. A., 1985. The competition process in forest stands. In Attributes of Trees as Crop Plants, Cannell, M. G. R., and J. E. Jackson, eds., chapter 28, pp. 481–506. Institute of Terrestrial Ecology, Abbots Ripton, Hunts, England.

Russell, M. B., A. W. D'Amato, M. A. Albers, C. W. Woodall, K. J. Puettmann, M. R. Saunders, and C. L. VanderSchaaf, 2015. Performance of the Forest Vegetation Simulator in managed white spruce plantations influenced by Eastern spruce budworm in Northern Minnesota. Forest Science 61(4):723–730.

Sheil, D., C. S. Eastaugh, M. Vlam, P. A. Zuidema, P. Groenendijk, P. van der Sleen, A. Jay, and J. Vanclay, 2017. Does biomass growth increase in the largest trees? Flaws, fallacies and alternative analyses. Functional Ecology 31(3):568–581.

Strigul, N., D. Pristinski, D. Purves, J. Dushoff, and S. Pacala, 2008. Scaling from trees to forests: Tractable macroscopic equations for forest dynamics. Ecological Monographs 78(4):523–545.

Strub, M., and C. Cieszewski, 2012. The comparative $R^2$ and its application to self-referencing models. Mathematical and Computational Forestry & Natural-Resource Sciences (MCFNS) 4(2):73–76.

Umbach, N., K. Naumann, H. Brandt, and A. Kelava, 2017. Fitting nonlinear structural equation models in R with package nlsem. Journal of Statistical Software 77(7):1–20.

Weiskittel, A. R., D. W. Hann, J. John A. Kershaw, and J. K. Vanclay, 2011. Forest Growth and Yield Modeling. Wiley-Blackwell. 430 p.

Wright, S., 1921. Correlation and causation. Journal of Agricultural Research 20(7):557–585.