ESTIMATING CAUSAL RELATIONS OF DYNAMIC MODELS FROM REAL-LIFE DATA

Nefel Tellioğlu
nefel.tellioglu@gmail.com
Boğaziçi University
Industrial Engineering Department
Bebek, Istanbul, Turkey

Yaman Barlas
ybarlas@boun.edu.tr
Boğaziçi University
Industrial Engineering Department
Bebek, Istanbul, Turkey

With major advances in data analysis tools in recent years, constructing models grounded in data has become an increasingly promising topic. There are four categories in model construction where data may be used: discovering causal relations, the polarity of relations, stock variables and mathematical expressions (see Figure 1). In this work, we focus on discovering (1) polarity of relations, (2) stock variables, and (3) mathematical expressions, by applying correlation analysis, curve fitting, and structural equation modeling (SEM) on "simulation-generated" data.

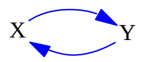
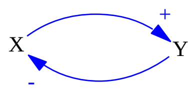
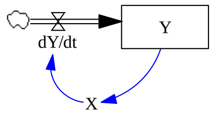
Discovering Causal Relations	Discovering the Directions of Relations	Discovering Stock Variables	Discovering Mathematical Expression of Relations
			$dY/dt = f_1(X)$ $X= f_2(Y)$

Figure 1: Types of Causality Inferences in Model Construction

In this study we assume that we have real-life (non-experimental) dynamic data about multiple system variables. 'Synthetic' data generated by Vensim model simulations (with DT=1). Secondly, it is assumed that the boundary of the system is well-chosen for the problem. Finally, it is assumed that the collected data range is enough to represent the overall causal relation. To perform the analysis, R programming language is used.

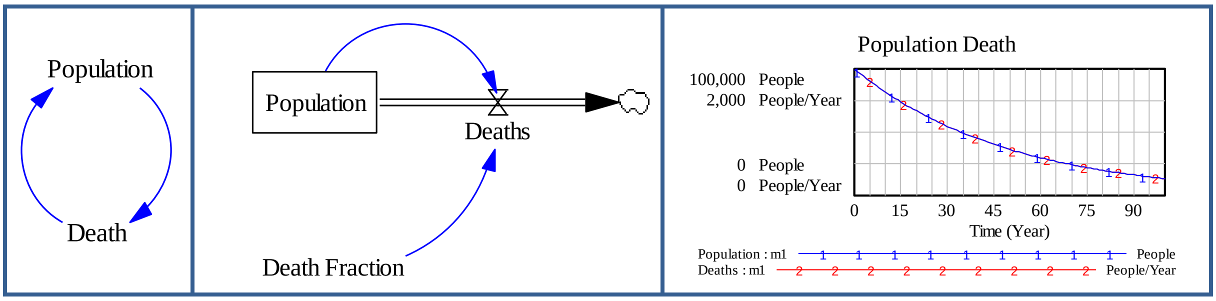


Figure 2: CLD, SFD and Behavior of Population-Death Model

Spearman correlation analysis is used to discover the signs of the causal relations. We consider a population model with only death outflow (see Figure 2). A perfect positive correlation, 1, between "death and population" is obtained. This is misleading, as death does not have a positive effect on the population. The reason for this spurious situation is the lack of stock-flow information in causal loop diagram (CLD). When the stock-flow information is included in the method, a perfect *negative* correlation, -1, is correctly obtained between death and the *rate of change* of population. However, correlation analysis with SFD can still produce misleading results when there are more than one causal links affecting the same variable with a perfect multicollinearity. To resolve the perfect collinearity, *noise* can be introduced (just like in real data) and then, the partial correlation analysis can be applied.

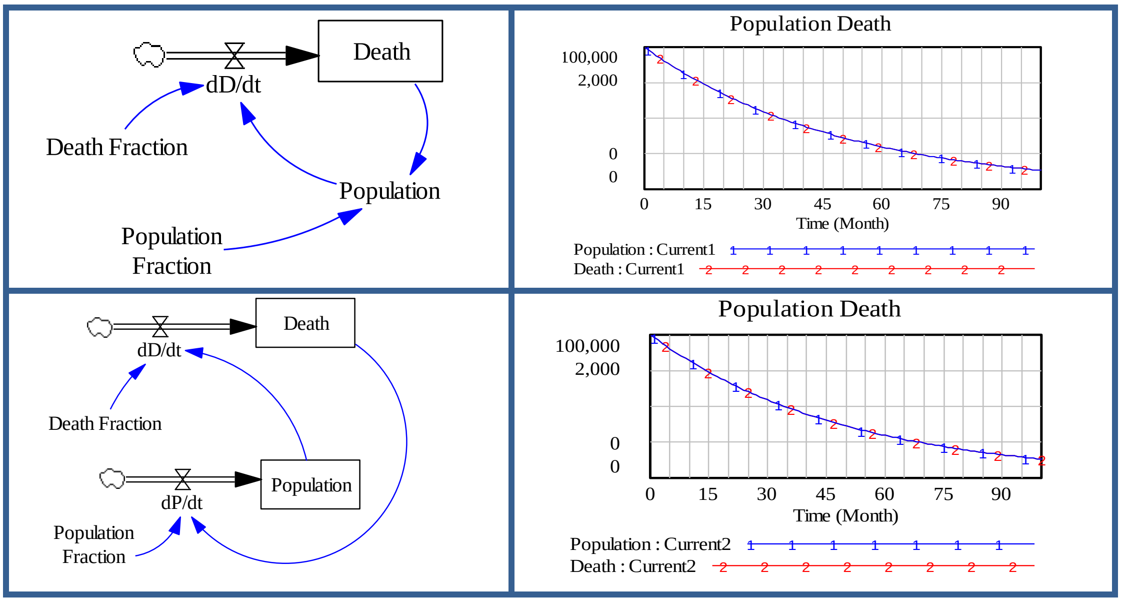


Figure 3: Alternative SFDs and their behaviors, derived from population-Death Data

Discovering SFD using a known CLD is investigated by curve fitting. Briefly, possible monotonic relations are fitted to both "*x→y*" and "*x→dy/dt*" for each "*x→ y*" in CLD. Then, the fits are compared according to the performance measures root mean square error (RMSE) and mean absolute percentage error (MAPE). The best fit is checked whether it is directly fitted to the effect variable, "*y*", or to the rate of change of the effect variable, "*dy/dt*". This procedure is applied to the population-death model of Figure 2. By keeping at least one variable as stock, we obtain additional two SFDs by curve fitting (see Figure 3). Obviously, the two SFDs in Figure 3 have nothing to do with the true structure. Therefore, even for very simple models, automatic determination of stock variables from data is not possible.

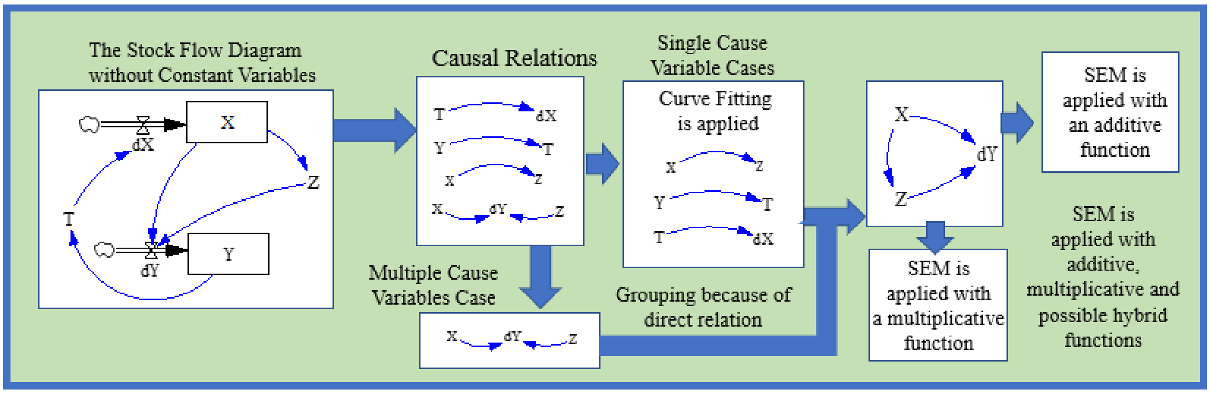


Figure 4: A Procedure of Discovering Mathematical Expressions

Finally, *Mathematical expressions* are estimated based on an assumed stock-flow diagram. The whole procedure is represented in Figure 4. Curve fitting for a single causal link is promising, although modelers must make sense of the constant values in the equations. For multiple causal links, SEM can only be applied when the effect functions are linear. But when there are perfect linear deterministic relationships between variables, SEM cannot converge because of multiple solutions and/or zero-variance problem, which do not exist when the data involve randomness/noise. There are many other difficult research problems that must be tackled in order to advance automated model construction from data.

Research supported by Boğaziçi University Research Grant.

REFERENCES

Abdelbari, H., Shafi, K., 2017, A computational Intelligence-based Method to 'Learn' Causal Loop Diagram-like Structures from Observed Data, System Dynamics Review, 2017 vol: 33 (1) pp: 3-33.

Boker, S.M., 2001, Differential Structural Modeling of intraindividual variability, In L. Collins & A. Sayer (Eds.), New Methods for the Analysis of Change (pp. 3-28), Washington, DC, Psychological Association.

Butner, J., Polemnia G. A., and Genna M. M., 2005, Multilevel Modeling of Two Cyclical Processes: Extending Differential Structural Equation Modeling to Nonlinear Coupled Systems, Psychological Methods, 10 June, 159-177.

Drobek M, Gilani W, Soban D. 2014. A data-driven and tool supported CLD creation approach. In Proceedings of the International System Dynamics Conference, Delft, Netherlands.

Hosseinichimeh, N., Rahmandad, H., Jalali, M. S., Wittenborn, A. K., 2016, Estimating the parameters of system dynamics models using indirect inference, System Dynamics Review 32 (2): 156–180.

Jalali, S. M., Rahmandad, H., Ghoddusi, H., 2015, Using the Method of Simulated Moments for System Identification, In H. Rahmandad, R. Oliva, & N. D. Osgood (Eds.), Analytical Methods for Dynamic Modelers, Cambridge, MA: The MIT Press, p. 39-70.

Lacerda, G., Spirtes P., Ramsey J., Hoyer P.O., 2008, Discovering cyclic causal models by independent components analysis, In Proc. 24th Conf. Uncertainty in Artificial Intelligence (UAI2008), Helsinki, Finland.

MacCallum, R., Austin, J., 2000, Applications of Structural Equation Modeling in Psychological Research, Annual Review of Psychology.51: 201–226.

Medina-Borja A., Pasupathy K.S., 2007, Uncovering complex relationships in system dynamics modeling: exploring the use of CART, CHAID and SEM. In Proceedings of the International System Dynamics Conference.

Pruyt, E., Cunningham, S. C., Kwakkel, J. H., de Bruijn, J. A., 2014, From data-poor to data-rich: system dynamics in the era of big data, In: Proceedings of the 32nd international conference of the System Dynamics Society, Delft, The Netherlands.

Struben, J., Sterman, J., Keith, D., 2015, Parameter Estimation Through Maximum Likelihood and Bootstrapping Methods, In H. Rahmandad, R. Oliva, & N. D. Osgood (Eds.), Analytical Methods for Dynamic Modelers, Cambridge, MA: The MIT Press, p. 3-38.

Takahashi, Y., 2006, Stock Flow Diagram Making with Incomplete Information about Time Properties of Variables, In Proceedings of the 24th International Conference of the System Dynamics Society, page 122, Nijmegen, Netherlands.