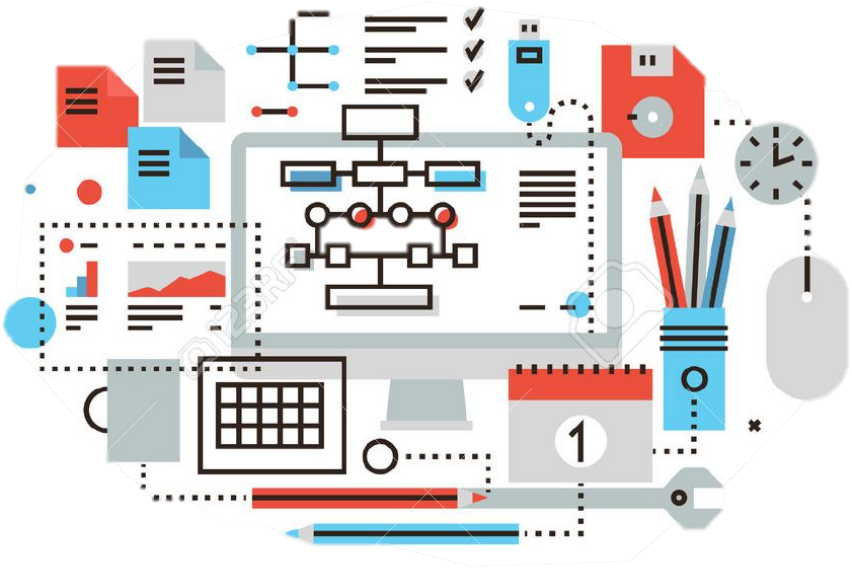
Is it possible to extract causal relationships from data analysis?

Yaman Barlas

Nefel Tellioğlu



Karmaşık Sistemler ve

Veri Bilimi Çalıştayı

Bilgi Üniversitesi

26 Mayıs 2018

Outline

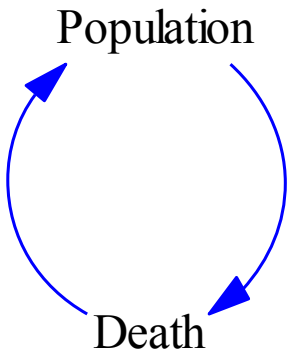
1. Problem Introduction
2. Research Questions
3. Methodology
4. Discovering Signs of Causal Relations
5. Discovering State Variables in Differential Equations
6. Discovering Causal Equations
7. Ongoing Research
8. Conclusion

1. Problem Introduction

System Dynamics

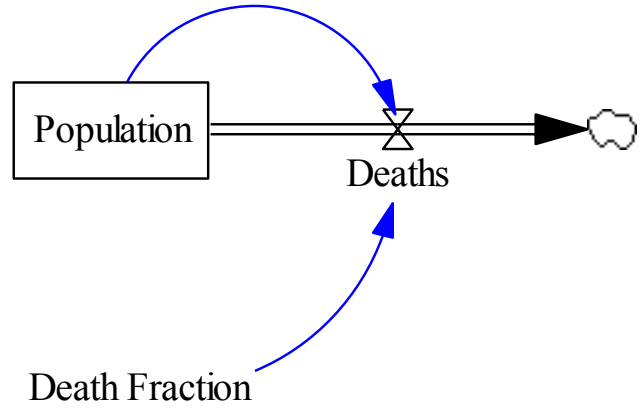
-A methodology to frame, analyze and seek solutions for complex dynamic policy problems.
-A technique to construct 'theory-like' descriptive/causal dynamic models

2 main tools:

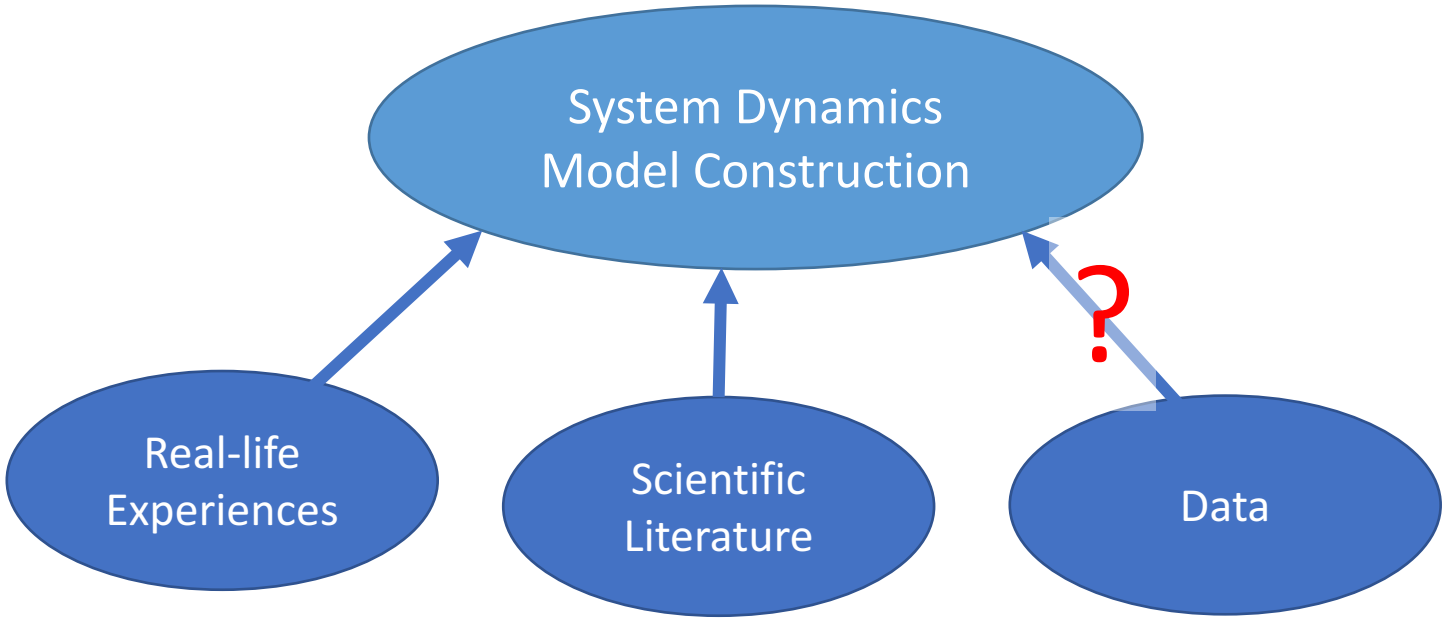


Causal-Loop Diagram

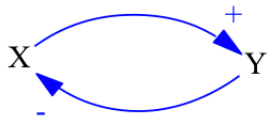
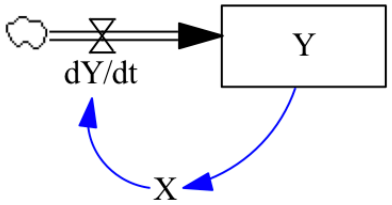
Stock-Flow Diagram



1. Problem Introduction



2. Research Questions

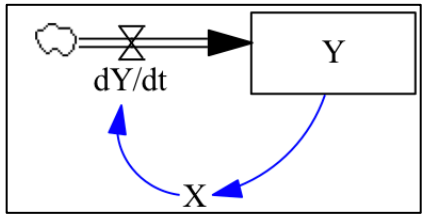
Discovering the Directions of Relations	Discovering State Variables	Discovering Mathematical Expression of Relations
		$dY/dt = f_1(X)$ $X = f_2(Y)$

-Can we discover the signs (polarity) of causal relations?
-Can we discover which variables are the state (stock) variables?
-Can we discover the causal equations?

What are the limits of data analysis?

3. Methodology

SD Model Construction



Data Generation

t	X	Y	dY/dt
1	..	..	..
2	..	..	..
3	..	..	..

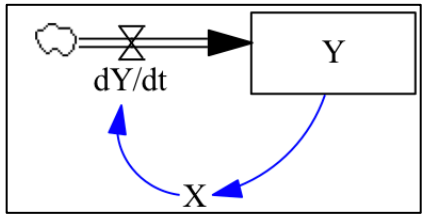
Non-experimental
Real-life like data

Drawing Conclusions about Structure

(Re)discovery of
some features of
the model structure?

3. Methodology

SD Model Construction



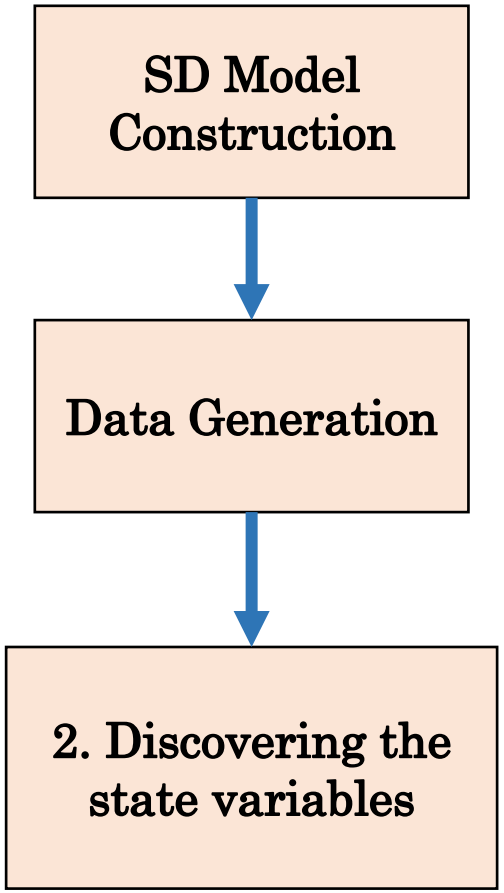
Data Generation

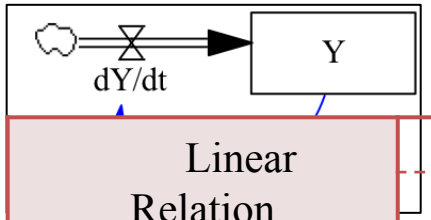
t	X	Y	dY/dt
1	..	..	..
2	..	..	..
3	..	..	..

1. Discovering the signs of relations

Correlation Analysis: Spearman Correlation

3. Methodology





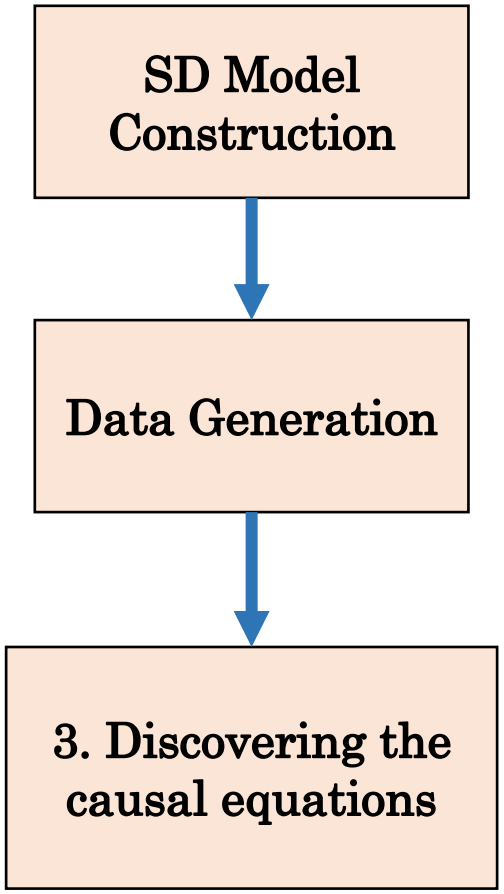
Curve Fitting

Linear Relation	gnls(y~ a+ b*x , start= c(a=0, b=1))
	lm(y~ x)
Exponential Relation	gnls(log(y)~a+b*x, start=c(a=0, b=1))
S-shaped Relation	nls(y~ SSlogis(x, Asym, xmid, scal))
	gnls(y ~ Asym/(1+exp((xmid-x)/scal)), start = coef(above function))
	nls(y~ SSasymp(x, Asym, R0, lrc))
	gnls(y ~max(y)/(1+exp((xmid-x)/scal)), start = c(xmid=1, scal=1))
	gnls(y ~max(y)-(max(y)-min(y))/(1+exp((xmid-x)/scal)), start = c(xmid=1, scal=1))
	gnls(y ~min(y) +(max(y)-min(y))/(1+exp((xmid-x)/scal)), start = c(xmid=1, scal=1))

Root mean square error (RMSE)
Mean absolute percentage error (MAPE)

Performance Measures

3. Methodology



Curve Fitting is applied

3. Methodology

- Correlation Analysis: Spearman correlation

- non-parametric; no assumption for underlying distributions

- the monotonic relationship between variables

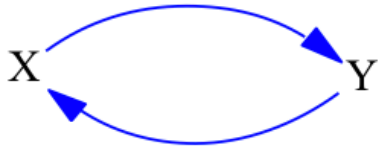
- Spearman coefficient (ρ), which is between [-1, 1]:

$$\rho = 1 - \frac{6\Sigma d_i^2}{n(n^2-1)}$$

$$\rho_{12.3} = \frac{\rho_{12} - \rho_{13}\rho_{23}}{\sqrt{(1-\rho_{13}^2)(1-\rho_{23}^2)}}$$

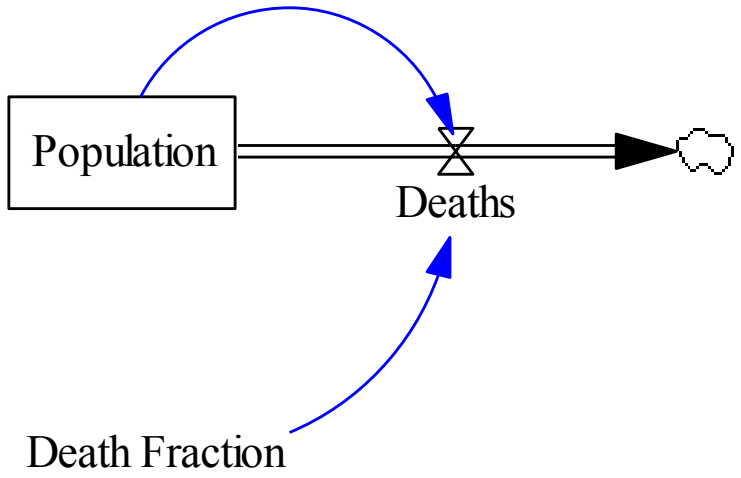
4. Discovering the Signs of Relations

- Correlation Analysis



4. Discovering the Signs of Relations

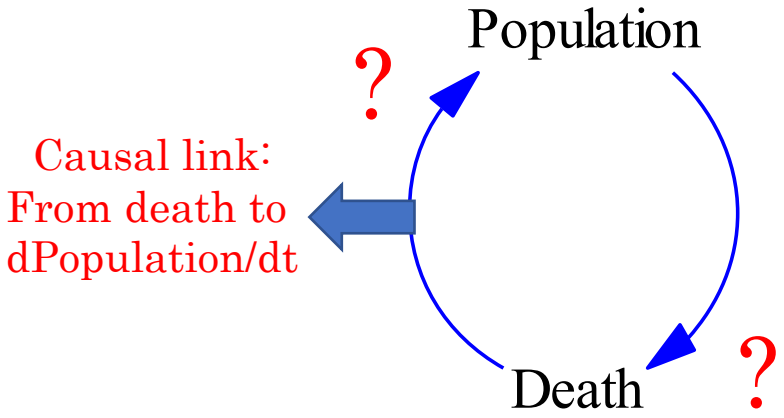
- Correlation Analysis: An Example

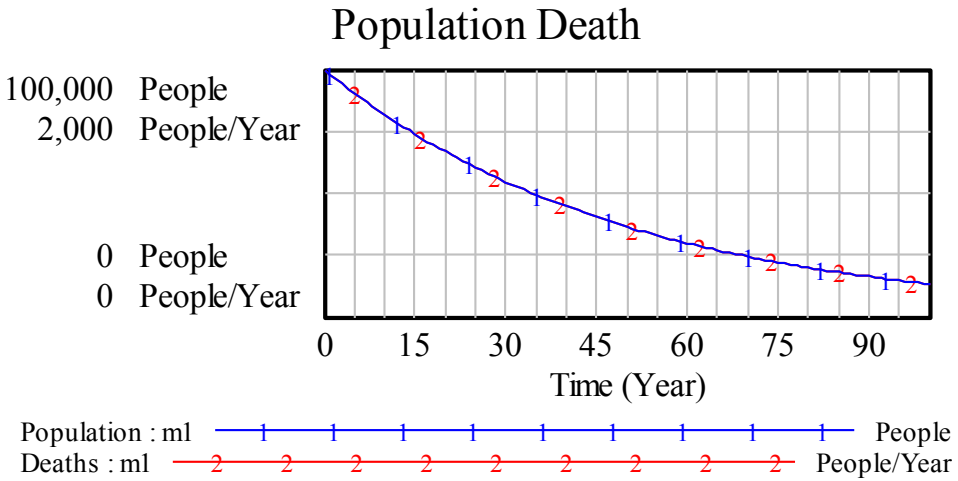


dPopulation/dt = - Death
Death = 0.2 * Population

4. Discovering the Signs of Relations

- Correlation Analysis: An Example





	Death	Population
Death	1	1
Population	1	1

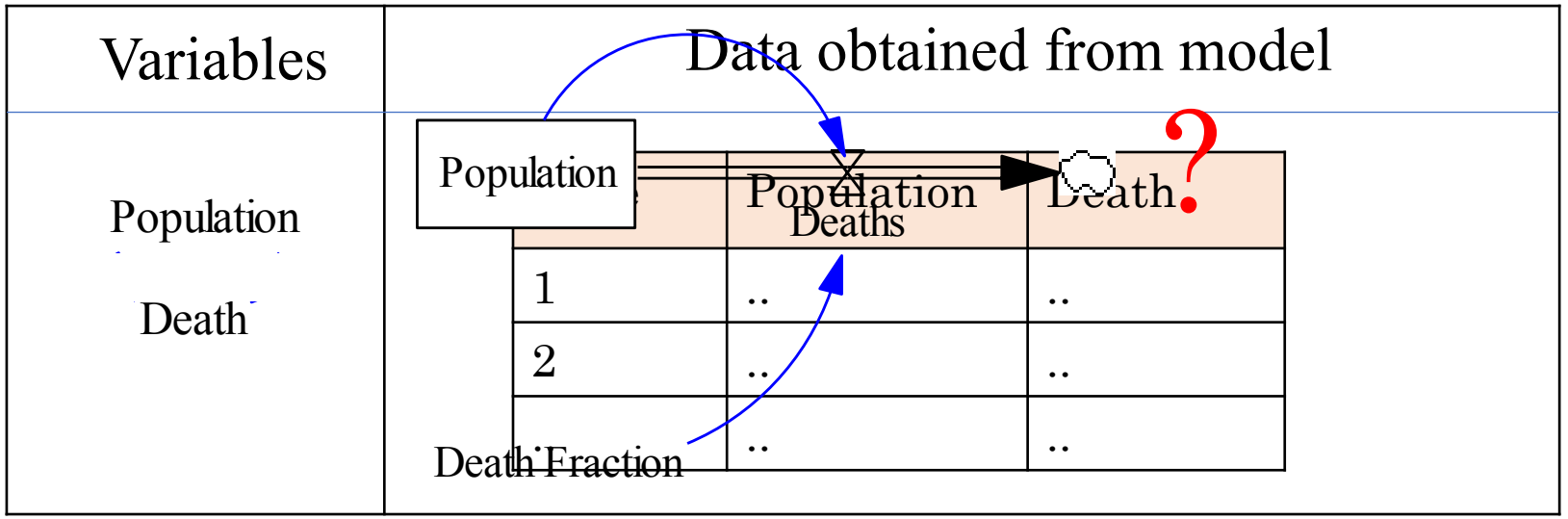
No information for the state and its rate of change

5. Discovering State (Stock) Variables

- Curve Fitting is applied

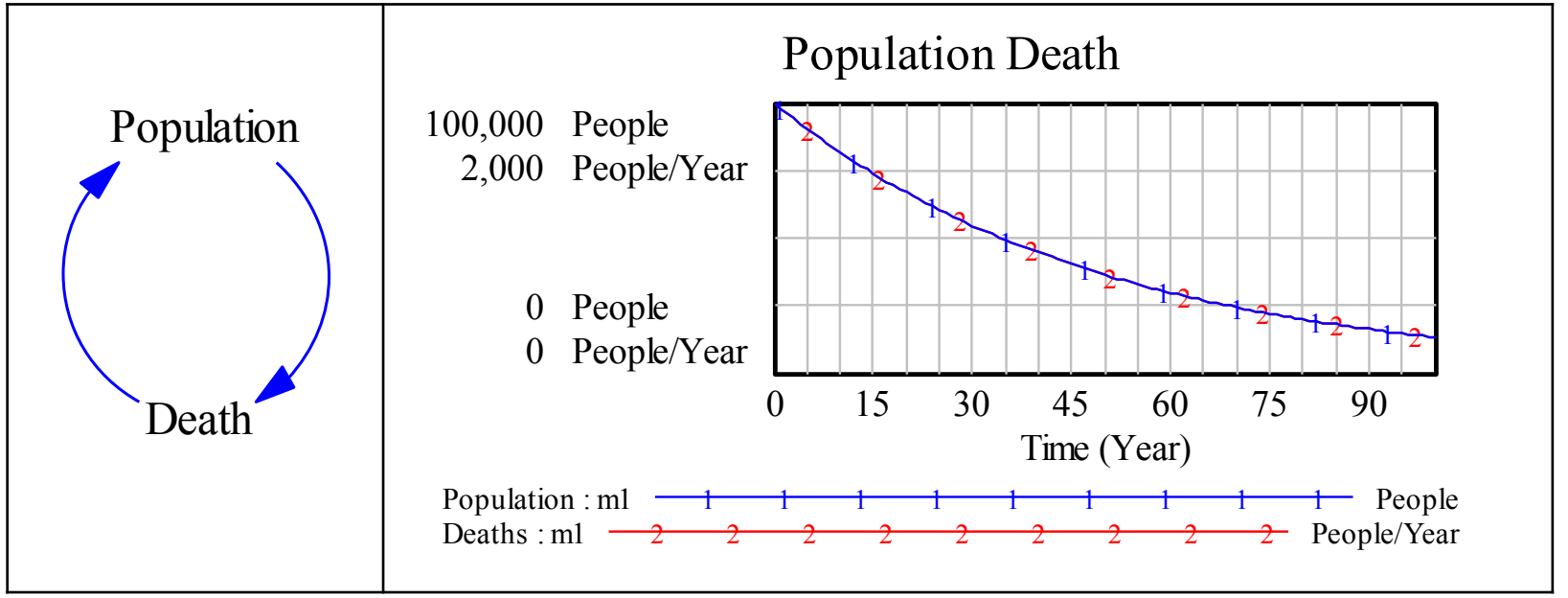
5. Discovering State Variables

- Example *dPopulation/dt =f(Death) ?*



5. Discovering State Variables

- Example *dPopulation/dt =f(Death) ?*



5. Discovering State Variables

- Example *dPopulation/dt =f(Death) ?*

Curve Fitting from Death to Population	Curve Fitting from Population to Death
(1) Population= f(Death)	(3) Death= f(Population)
(2) dPopulation/dt =f(Death)	(4) dDeath/dt= f(Population)

5. Discovering State Variables

Curve Fitting Results from Death(x) to Population(y)

	Rmse	Mape	Method	Function Relation	
1	**0.0339**	**0.0001**	**LinFunc**	**dy~ a+ b*x**	**XtoDY**
2	**0.0967**	**0.0001**	**LinFunc**	**y~ a+b*x**	**XtoY**
3	2318.4037	3.2818	ExpFunc	log-dy~ a+b*x	XtoDY
4	8422.5490	13.0940	ExpFunc	logy~a+b*x	XtoY
5	15571.5583	4.6847	SShaped	y~ SSlogis(x, Asym, xmid, scal)	XtoY

Population= 50*Death
dPopulation/dt = -1*Death

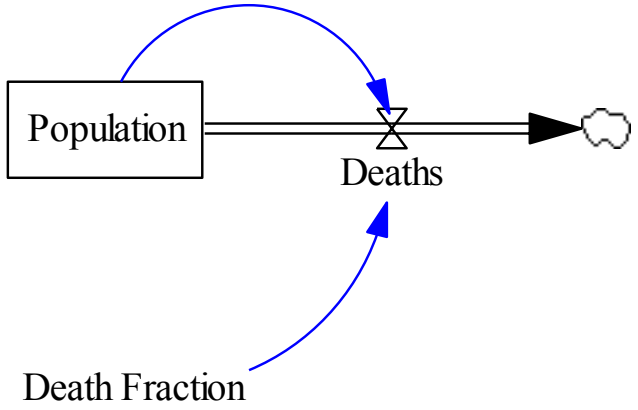
Curve Fitting Results from Population(x) to Death(y)

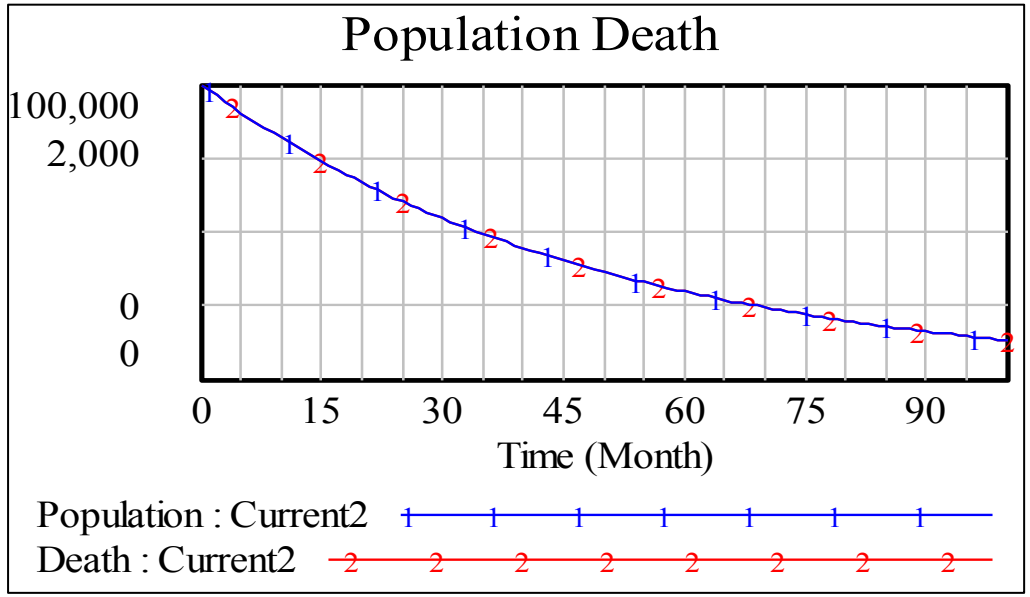
	Rmse	Mape	Method	Function Relation	
1	**0.00188**	**0.00010**	**LinFunc**	**dy~ a+ b*x**	**XtoDY**
2	**0.00193**	**0.00014**	**LinFunc**	**y~ a+b*x**	**XtoY**
3	46.36758	3.28176	ExpFunc	log-dy~ a+b*x	XtoDY
4	168.45130	13.09397	ExpFunc	logy~a+b*x	XtoY
5	311.43145	4.68475	SShaped	y~ SSlogis(x, Asym, xmid, scal)	XtoY

Death= 0.02*Population
dDeath/dt= -0.0004*Population

dPopulation/dt = -1*Death

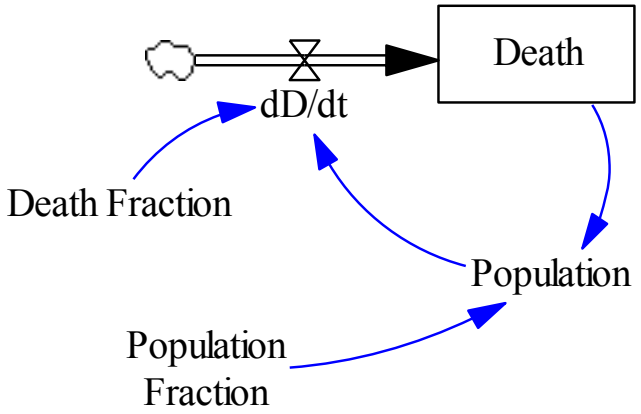
Death= 0.02*Population





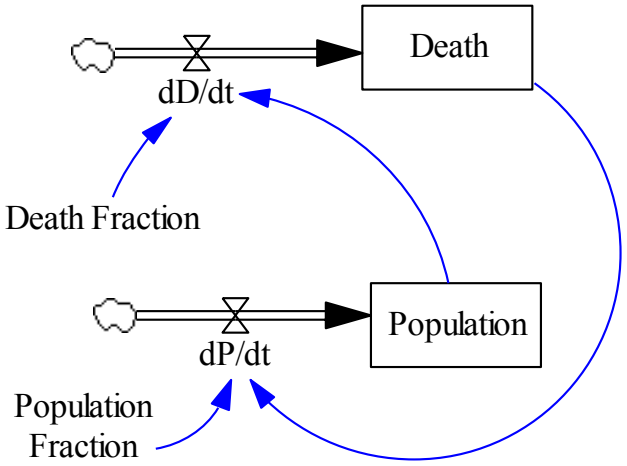
dDeath/dt=-0.0004*Population

Population= 50*Death



dPopulation/dt = -1*Death

dDeath/dt= -0.0004*Population



Impossible to derive unique and correct state variable identification from data analysis

6. Discovering Causal Equations

-Single Cause Variable

-Multiple Cause Variables

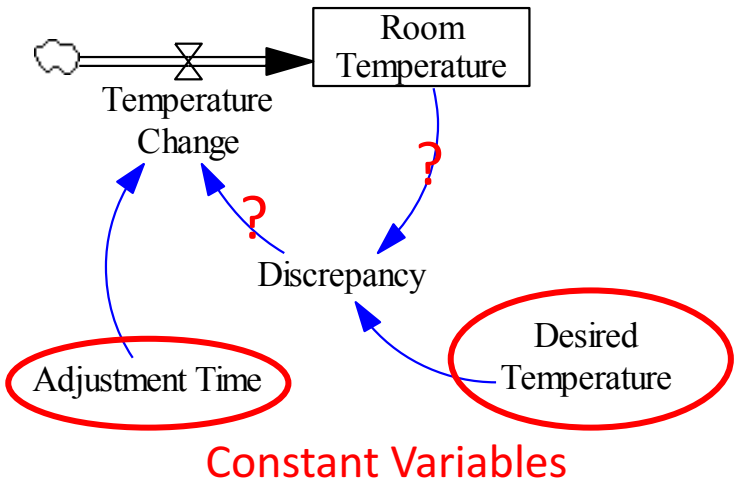
6. Discovering Causal Equations

-Single Cause Variable

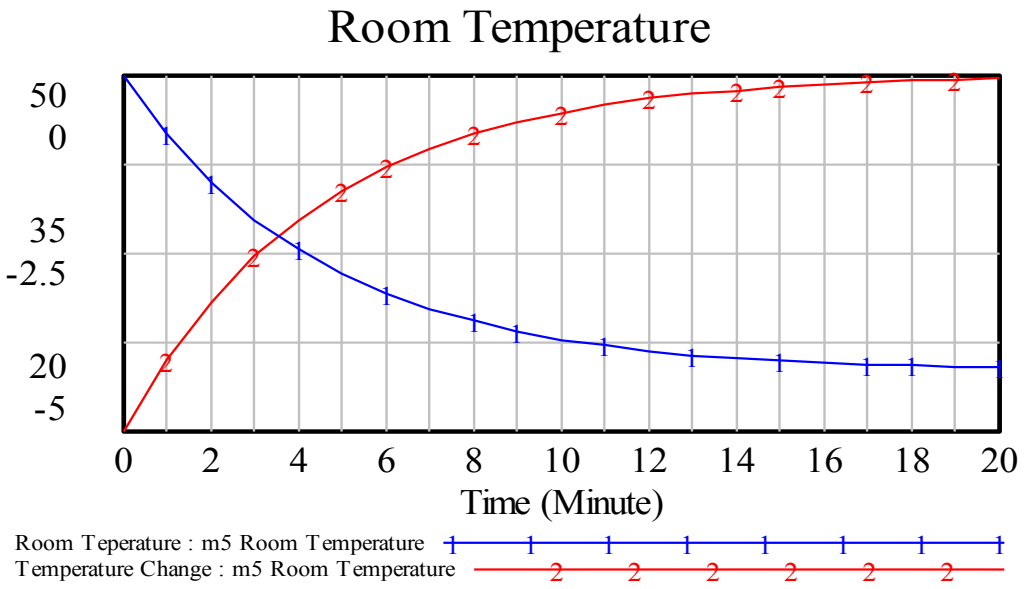
Curve Fitting with possible linear and nonlinear relationships

6. Discovering Causal Equations

-Single Cause Variable: Example



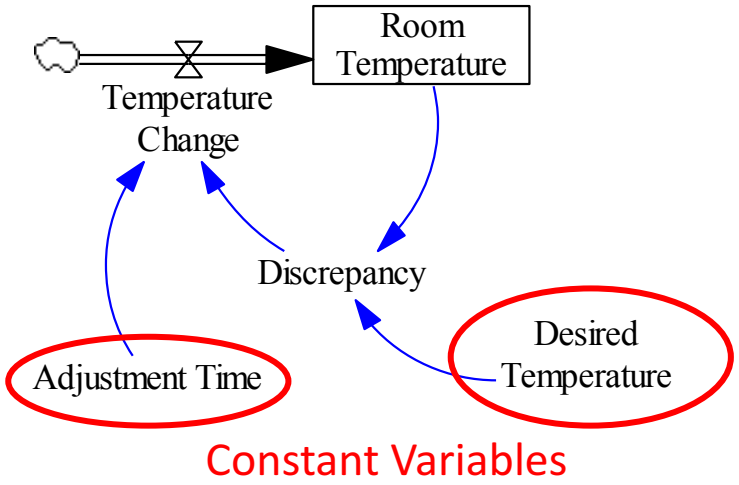
Constant Variables



Temperature Change= f(Discrepancy)
Discrepancy= g(RoomTemperature)

6. Discovering Causal Equations

-Single Cause Variable: Example



Temperature Change= 0.2*Discrepancy

Discrepancy= 25 – Room Temperature

Curve fitting can correctly estimate the underlying equation with a single input variable

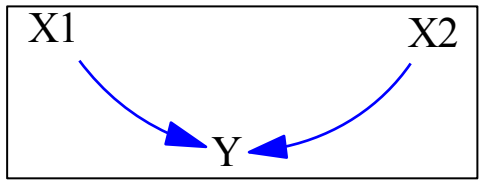
7. Ongoing Research

- Multiple Cause Variables

- Multiple Cause Variables

7. Ongoing Research

-Multiple Cause Variables



$$Y = Y^* + f_1(X_1) + f_2(X_2)$$

For additive formulation, if X functions are linear and there is no multicollinearity between X_1 and X_2, then linear regression can be used.

$$Y = Y^* * f_1(X_1) * f_2(X_2)$$

For multiplicative formulation, regression cannot seperate the f() functions even if they are linear.

8. Conclusion

To discover the signs of causal relations:

- Correlation analysis can be applied. However, we must know which variables are the state & rate variables.

8. Conclusion

To discover the state variables:

- Data analysis cannot return correct/reliable results. Real-life experience, reasoning, and scientific literature must be used.

8. Conclusion

To discover the causal equations:

- In the case of one cause variable: Curve fitting approach gives consistent results with the underlying structure.
- In the case of multiple cause variables:
  - For additive formulation, if the functions f(x) are linear and there is no multicollinearity, linear regression can be used.
  - For multiplicative formulation, regression cannot be used to estimate the effects of X variables.