A new method of model factor clustering based on second-order sensitivity index

Wuhan Institute of Physics and Mathematics
Chinese Academy of Sciences

Wu Qiongli

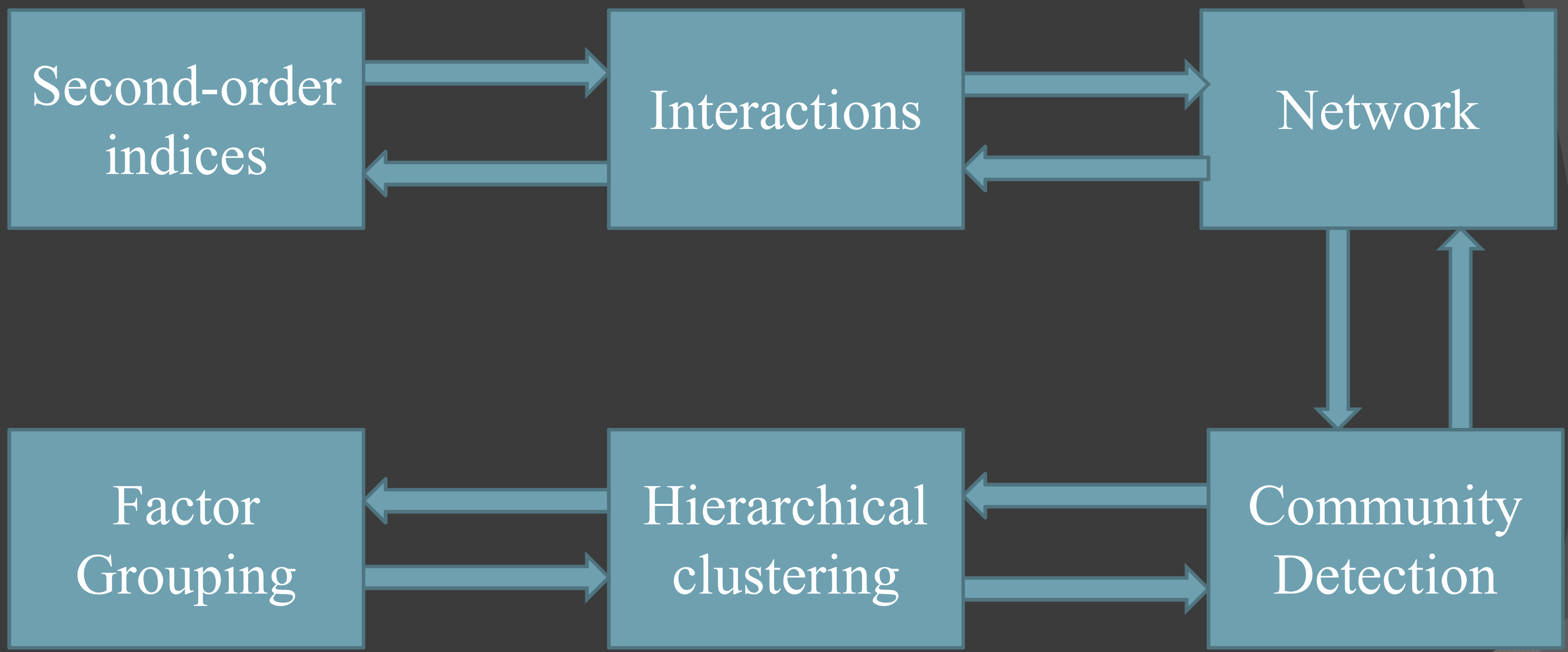
- Introduction
- Methods
- Verification
- Test cases
- Application to ecological model

Introduction

- Sensitivity Analysis (SA) methods are invaluable tools for model validation, model optimization and model diagnosis…
- Second-order Sobol's indices: interaction between parameters

How to quantitatively understand and use the interactions between parameters indicated by second-order Sobol's indices?

Our idea


Second-order indices
Interactions
Network
Factor Grouping
Hierarchical clustering
Community Detection

Methods: Second-order Sobol's indices

$$\mathrm{Y} = f(X_1, \dots, X_k)$$

$$\mathrm{V(Y)} = \sum_{j=1}^{k} V_j + \sum_{1 \le j < l \le k} V_{jl} + \cdots + V_{1,2,\dots,k}$$

The first-order sensitivity index S_i for factor X_i:

$$S_i = \frac{V_i(E_{-i}(Y|X_i))}{V(Y)}$$

The second-order sensitivity index S_{ij} for factor X_i and X_j:

$$S_{ij} = \frac{V(E(Y|X_i, X_j))}{V(Y)} - S_i - S_j$$

Methods: Group indices

$S^{g}_{\Omega_m}$ main effect of module m and helps to rank the group's importance, is defined as:

$$S^{g}_{\Omega_m} = \frac{V^{c}_{i_1,i_2,\dots,i_s}}{V(Y)} = \frac{V(E(Y|X_{i_1}, X_{i_2}, \dots, X_{i_s}))}{V(Y)}$$

$ST^{g}_{\Omega_m}$ provides the total effect of group m, is defined as:

$$ST^{g}_{\Omega_m} = 1 - \frac{V^{c}_{-i_1,i_2,\dots,i_s}}{V(Y)} = 1 - \frac{V(E(Y|X_{l_1}, X_{l_2}, \dots, X_{l_{k-s}}))}{V(Y)}$$

$S^{g}_{\Omega_{mn}}$ is the second-order interaction between a pair of groups, is defined as:

$$S^{g}_{\Omega_{mn}} = \frac{V(E(Y|\Omega_m, \Omega_n))}{V(Y)} - S^{g}_{\Omega_m} - S^{g}_{\Omega_n}$$

Methods: Clustering detection

Hierarchical clustering algorithms

- Does not require a preliminary knowledge on the number and size of the clusters.

➢ Agglomerative algorithms: clusters are iteratively merged if their interaction is sufficiently high

➢ Divisive algorithms: clusters are iteratively divided if their interaction is sufficiently low

[Hastie et al., 2001] Hastie, T., Tibshirani, R., and Friedman, J. (2001). The elements of statistical learning-data mining, inference, and prediction.

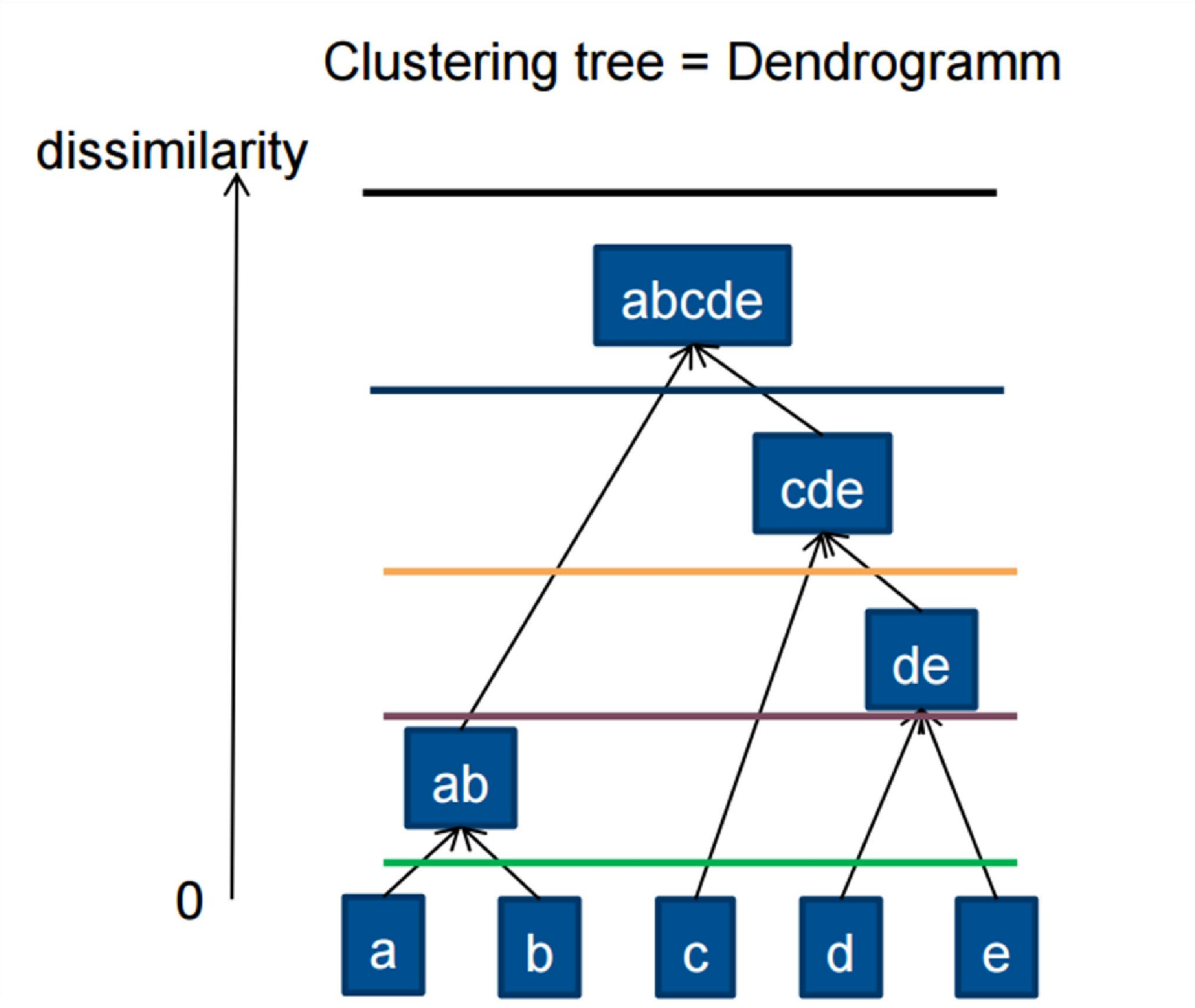
Methods: Clustering detection

Average dissimilarity between the groups is defined:

$$d_{GA}(G,H) = \frac{1}{N_G N_H} \sum_{i \in G} \sum_{i' \in H} d_{ii'}$$

- G and H represent two groups
- $d_{GA}(G,H)$is the average dissimilarity
- $d_{ii'}$ is the set of pairwise observation dissimilarities, while one member of the pair i is in group G and the other i' is in group H.
- N_G and N_H are the respective number of observations in each group

Methods: Clustering detection



Verification: Based on G-function

Objectives:

- The second-order sensitivity indices represent the interaction between parameters
- The magnitude of the indices reflect the strength of the interaction.

$$g(x_1, x_2, \dots, x_k) = \prod_{i=1}^{k} g_i(x_i)$$

$$g_i(x_i) = \frac{|4x_i - 2| + a_i}{1 + a_i}$$

$$x_i \sim U(0,1)$$

$$a = [0.2,0.5,0.8,0.9,1]$$

Verification: Based on G-function

$$\mathrm{F(X)} = g(x_1, x_2, \dots, x_5) + 0.5g(x_6, x_7, \dots, x_{10}) + 0.2g(x_{11}, x_{12}, \dots, x_{15}) \tag{1}$$

- The interaction between the parameters of the function F(X) are limited within each g function
- Interaction is determined by the structural properties of each g function.

	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14	x15
x1	0	2.9929	1.1975	2.3829	1.0739	0	0	0	0	0	0	0	0	0	0
x2	2.9929	0	0.96	1.2934	1.2812	0	0	0	0	0	0	0	0	0	0
x3	1.1975	0.96	0	0.2406	0.8549	0	0	0	0	0	0	0	0	0	0
x4	2.3829	1.2934	0.2406	0	0.5986	0	0	0	0	0	0	0	0	0	0
x5	1.0739	1.2812	0.8549	0.5986	0	0	0	0	0	0	0	0	0	0	0
x6	0	0	0	0	0	0	0.6199	0.5744	0.6882	0.464	0	0	0	0	0
x7	0	0	0	0	0	0.6199	0	0.2234	0.0216	0.4241	0	0	0	0	0
x8	0	0	0	0	0	0.5744	0.2234	0	0.0216	0.2265	0	0	0	0	0
x9	0	0	0	0	0	0.6882	0.0216	0.0216	0	0.0216	0	0	0	0	0
x10	0	0	0	0	0	0.464	0.4241	0.2265	0.0216	0	0	0	0	0	0
x11	0	0	0	0	0	0	0	0	0	0	0	0.0479	0.2077	0.0216	0.0986
x12	0	0	0	0	0	0	0	0	0	0	0.0479	0	0.0216	0.0646	0.2838
x13	0	0	0	0	0	0	0	0	0	0	0.2077	0.0216	0	0.0867	0.0255
x14	0	0	0	0	0	0	0	0	0	0	0.0216	0.0646	0.0867	0	0.1298
x15	0	0	0	0	0	0	0	0	0	0	0.0986	0.2838	0.0255	0.1298	0

Table 1:Second-order sensitivity indices for function (1)(percentage)

$$S_{ij} = \begin{cases} 0 & |S_{ij}| \le 10^{-6} \\ \min(S_{ij} > 0) & S_{ij} < 0 \& |S_{ij}| > 10^{-6} \end{cases}$$

Test cases

The steps of our test cases are as following:

- Calculate the second-order sensitivity indices by Sobol's
- Interaction matrix of second-order indices
- Obtain 0-1 matrix by taking a certain threshold value to the interaction matrix
- Clustering detection, get the grouping result

Test case 1

$$F = g(x_1, x_2, \ldots, x_5) + g(x_6, x_7, \ldots, x_{10}) + g_1(x_1) \cdot g_8(x_8) + g_3(x_3) \cdot g_7(x_7) + g_4(x_4) \cdot g_{10}(x_{10}) \quad (2)$$

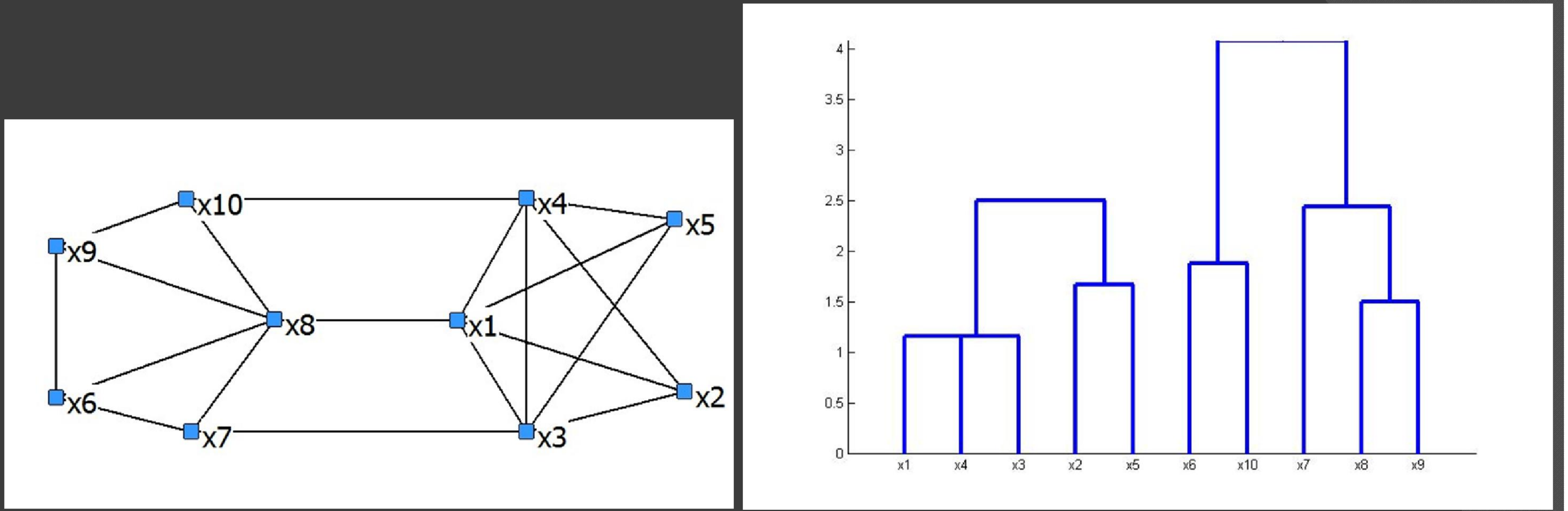
- Linear combination of two g-functions
- Interaction items between two factors

	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
x1	0	0.9027	1.0048	0.3695	0.4228	0	0	1.5308	0	0
x2	0.9027	0	0.3857	0.1778	0.0779	0	0	0	0	0
x3	1.0048	0.3857	0	0.1072	0.1836	0	0.2988	0	0	0
x4	0.3695	0.1778	0.1072	0	0.2033	0	0	0	0	0.204
x5	0.4228	0.0779	0.1836	0.2033	0	0	0	0	0	0
x6	0	0	0	0	0	0	0.4668	0.1894	0.502	0.0375
x7	0	0	0.2988	0	0	0.4668	0	0.2489	0.0375	0.0375
x8	1.5308	0	0	0	0	0.1894	0.2489	0	0.1383	0.357
x9	0	0	0	0	0	0.502	0.0375	0.1383	0	0.1387
x10	0	0	0	0.204	0	0.0375	0.0375	0.357	0.1387	0

Table 2: second-order sensitivity indices for function (2)

	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
x1	0	1	1	1	1	0	0	1	0	0
x2	1	0	1	1	0	0	0	0	0	0
x3	1	1	0	1	1	0	1	0	0	0
x4	1	1	1	0	1	0	0	0	0	1
x5	1	0	1	1	0	0	0	0	0	0
x6	0	0	0	0	0	0	1	1	1	0
x7	0	0	1	0	0	1	0	1	0	0
x8	1	0	0	0	0	1	1	0	1	1
x9	0	0	0	0	0	1	0	1	0	1
x10	0	0	0	1	0	0	0	1	1	0

Table 3: 0-1 matrix of Table 2



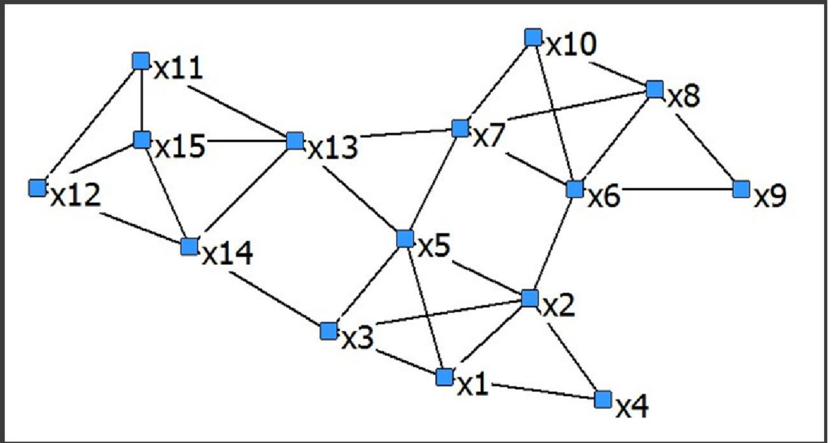
(a) (b)

Figure 1: Network structure and grouping result of the model function (2)

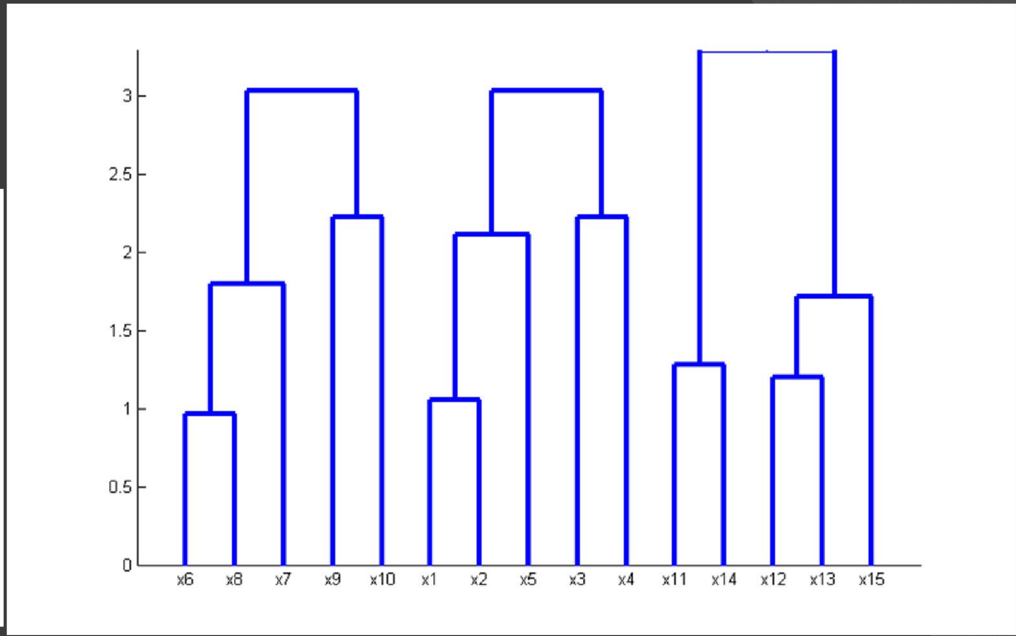
Test case 2

$$F = g(x_1, x_2, \dots, x_5) + g(x_6, x_7, \dots, x_{10}) + g(x_{11}, x_{12}, \dots, x_{15}) + g_2(x_2) \cdot g_6(x_6) + g_5(x_5) \cdot g_7(x_7) \cdot g_{13}(x_{13}) + g_3(x_3) \cdot g_{12}(x_{12}) \cdot g_{14}(x_{14}) \quad (3)$$

- Linear combination of three g-functions
- Interaction items between two/three factors



(a)



(b)

Figure 2: Network structure and grouping result of the model function (3)

Application to ecological model: Lotka-Volterra

The three species Lotka-Volterra competitory ecological model equations are as following:

$$x' = a_1 x - b_1 x^2 - c_1 xy$$
$$y' = -a_2 y + c_2 xy - b_2 y^2 - c_3 yz$$
$$z' = -a_3 z + c_4 yz - b_3 z^2$$

Variables x, y, z denote the size of the plant, herbivore and carnivores populations.

coefficients	distribution	meaning
a1	U[0.5,1.5]	intrinsic rate of plant nature increase
a2	U[0.18,0.42]	mortality of herbivore
a3	U[0.28,0.52]	mortality of carnivores
b1	U[0.015,0.045]	internal interaction in plant
b2	U[0.012,0.028]	internal interaction in herbivore
b3	U[0.007,0.013]	internal interaction in carnivores
c1	U[0.04,0.16]	proportionality constant linking plant mortality to the number of plant and herbivore
c2	U[0.025,0.075]	proportionality constant linking increase in predator to the number of plant and herbivore
c3	U[0.01,0.03]	proportionality constant linking herbivore mortality to the number of herbivore and carnivores
c4	U[0.036,0.084]	proportionality constant linking increase in predator to the number of herbivore and carnivores

Table 4: The distribution of the factors

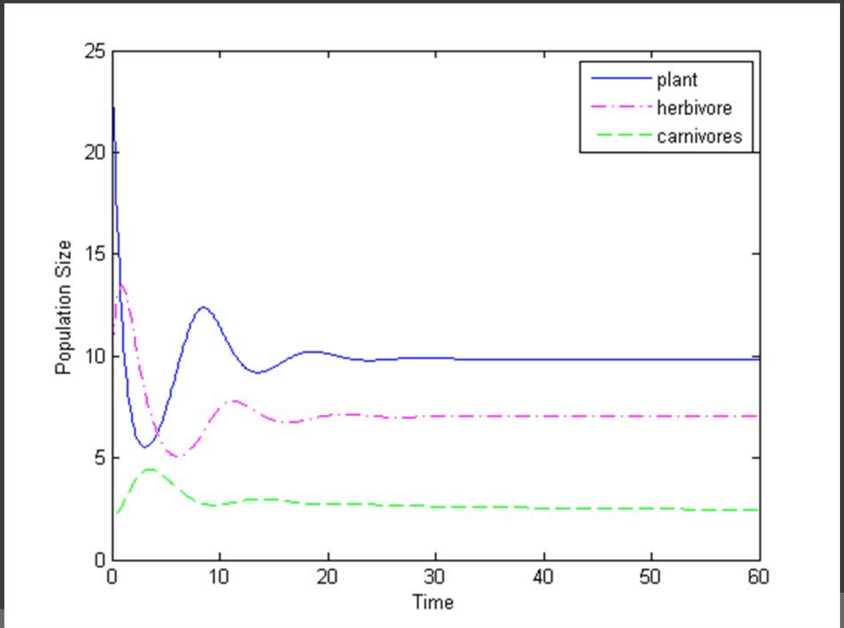
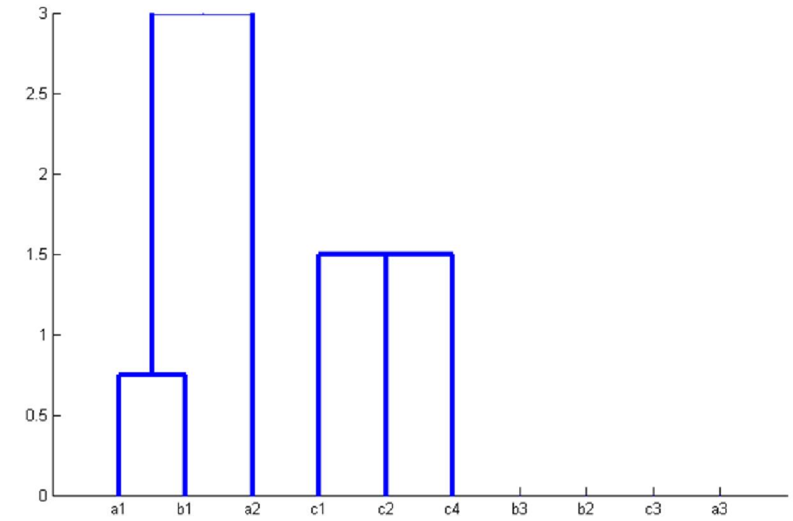
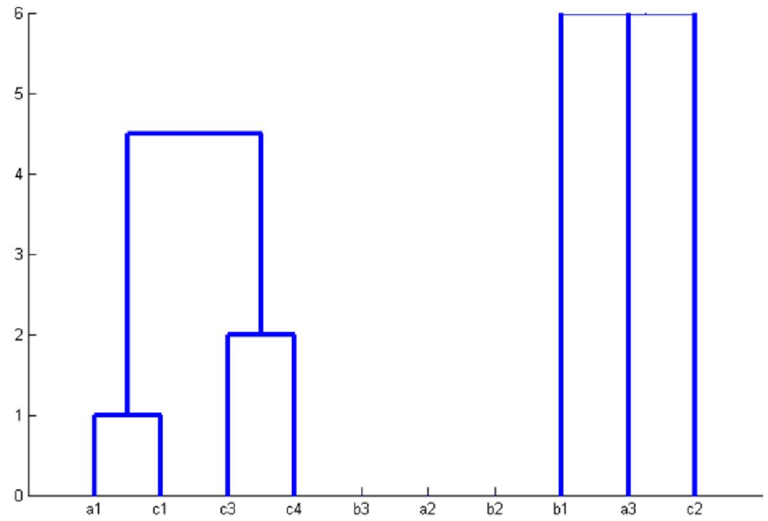


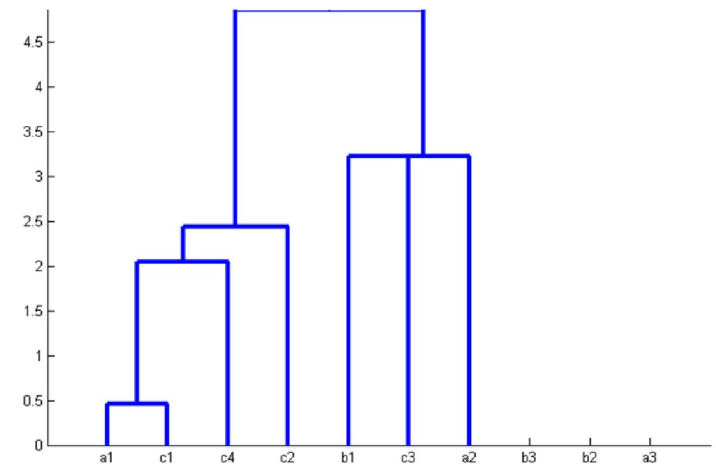
Figure 3: Stimulated evolution of Lotka-Volterra model



(a)



(b)



(c)

Figure 4:

(a): Take x: the amount of plant as the model output

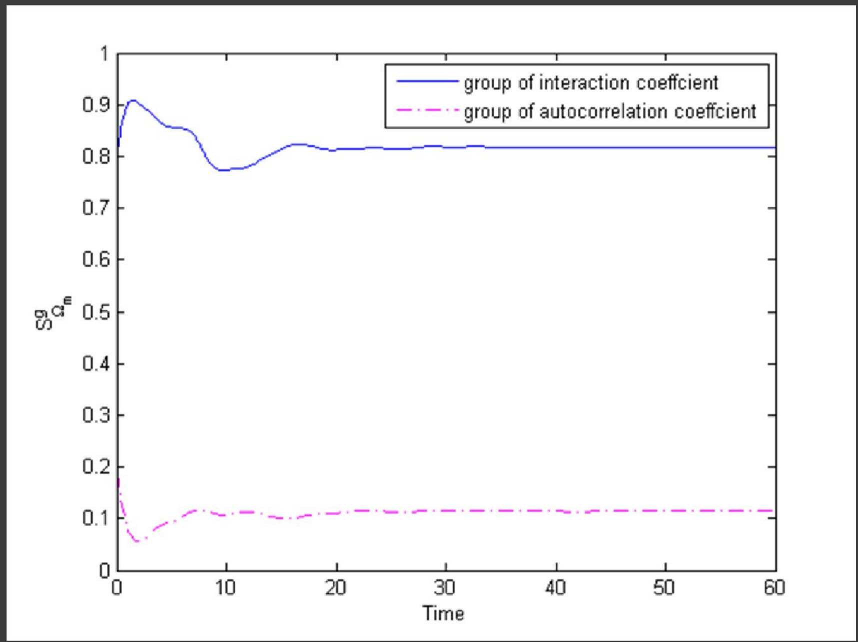
(b): Take y: the amount of herbivore as the model output

(c): Take z: the amount of carnivores as the model output

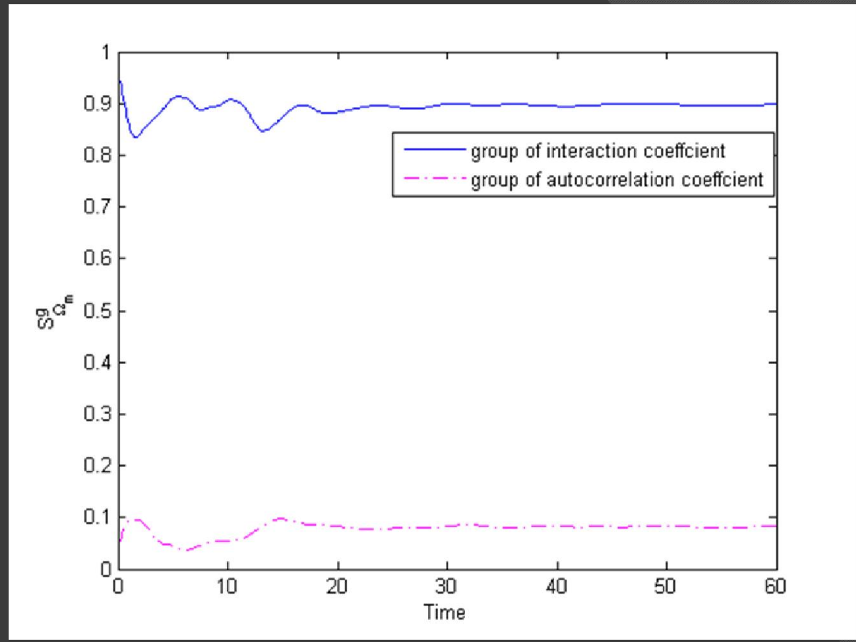
Application to ecological model: Lotka-Volterra-group analysis

- Group index $S^{g}_{\Omega_m}$ which represent the main effect of the group to the corresponding output.
- The comparison of the three outputs for indices $S^{g}_{\Omega_m}$ and $ST^{g}_{\Omega_m}$.
- Comparison of the interaction inter(a. $S^{g}_{\Omega_{mn}}$) and inner(b. $S^{g}_{\Omega_{m-in}}$) groups for three outputs.

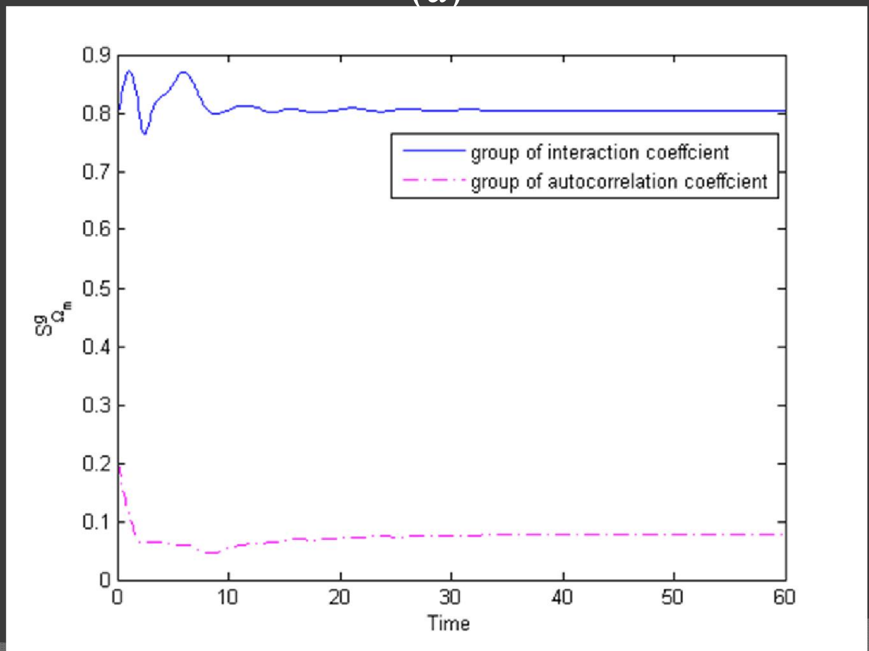
[Wu and Cournede, 2014] Wu, Q. and Cournede, P.-H. (2014). A comprehensive methodology of global sensitivity analysis for complex mechanistic models with an application to plant growth. Ecological Complexity, 20:219-232.



(a)

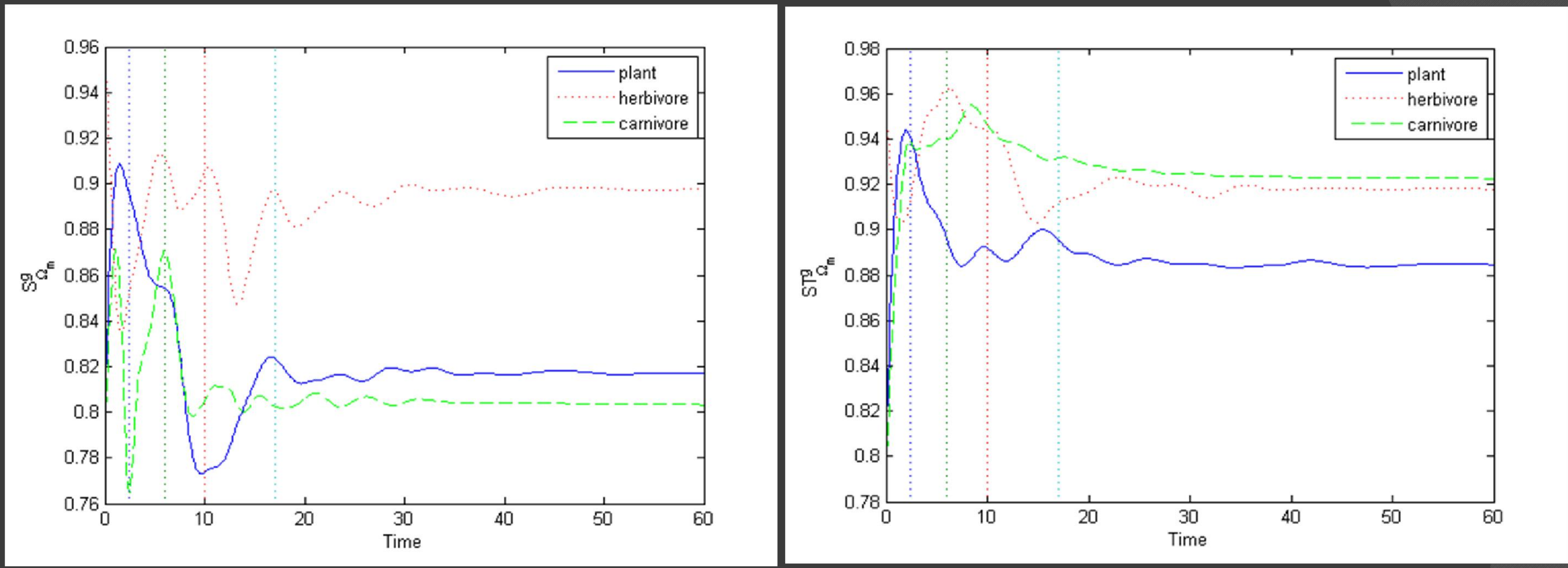


(b)



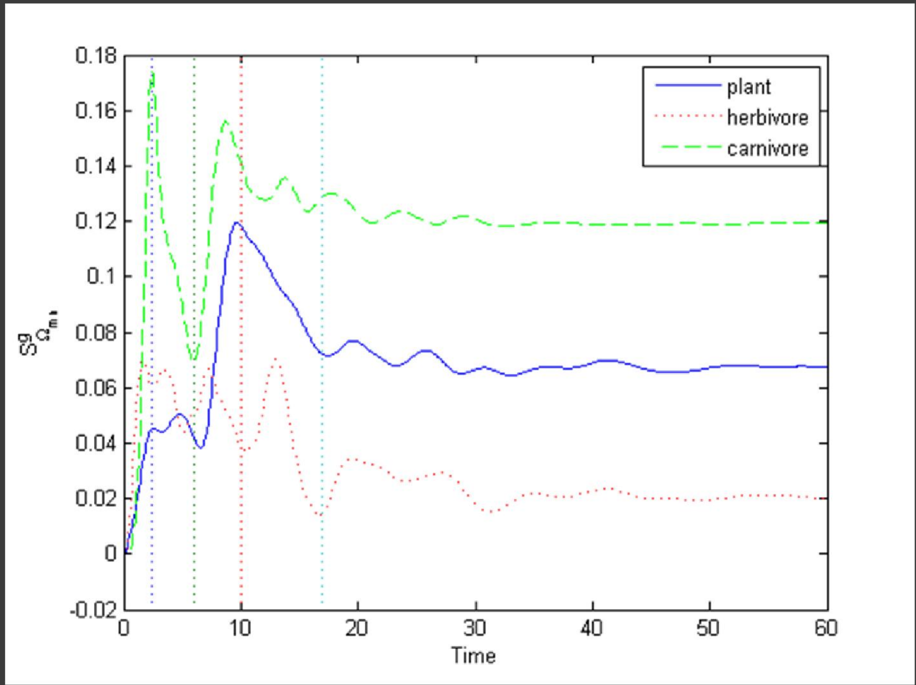
(c)

Figure 5: The main effect $S^g_{\Omega_m}$ of each group for three outputs of model.
(a) output x: plant.
(b) output y: herbivore.
(c) output z:carnivore.

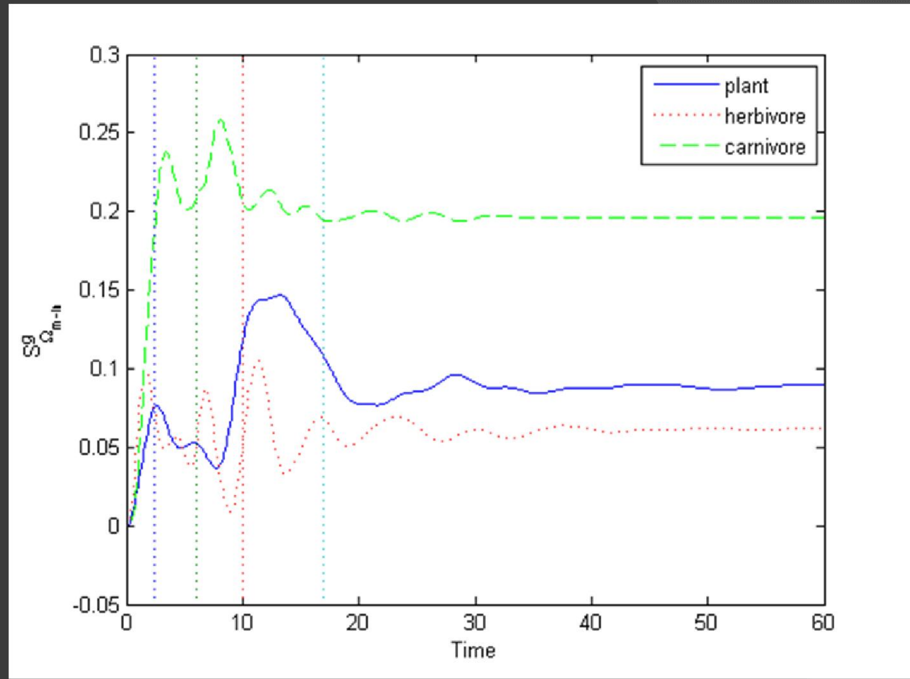


(a) (b)

Figure 6: Comparison of the group main effect, total effect of the three outputs, take the group of interaction coefficient in the case.



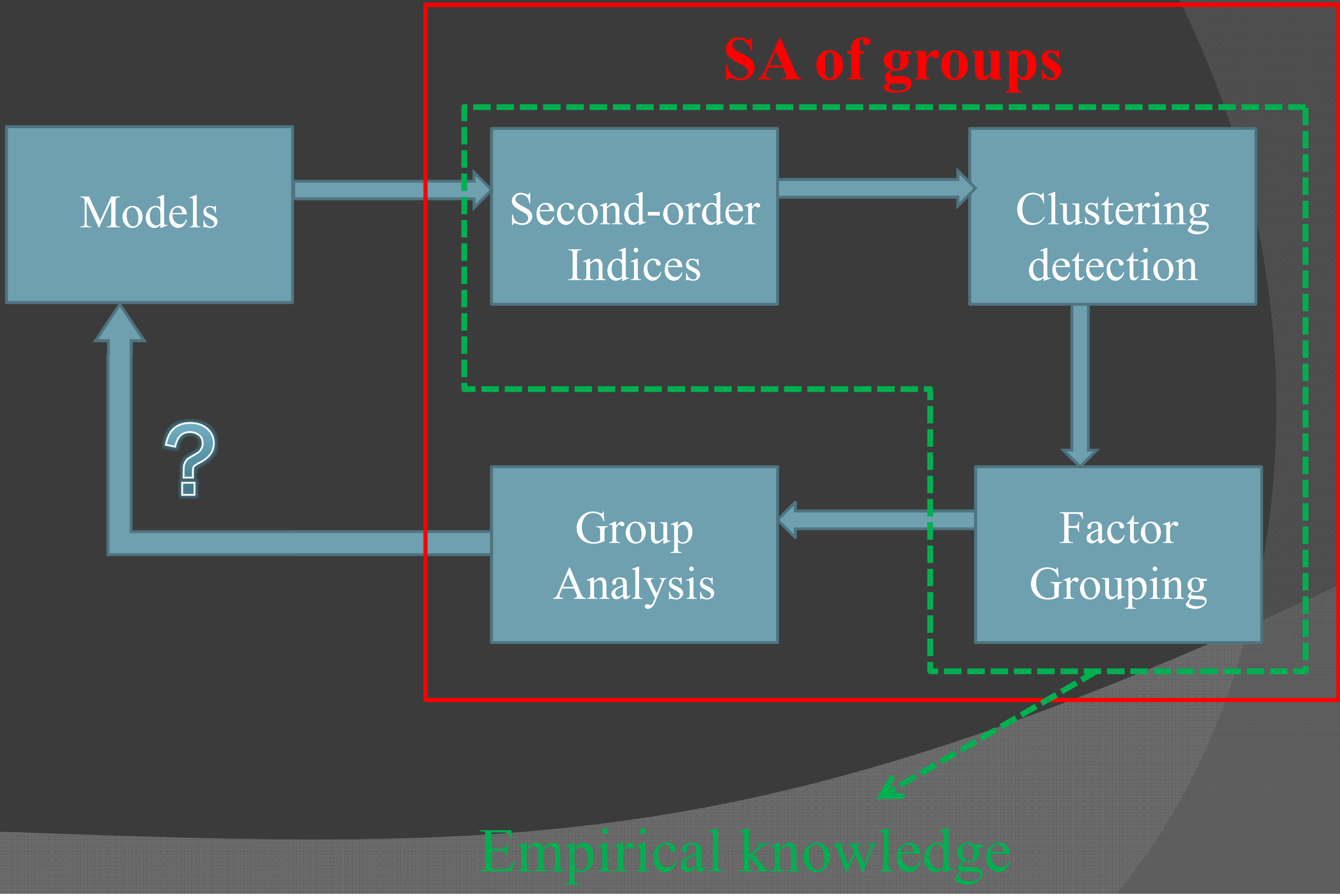
(a)



(b)

Figure 7: Comparison of the interaction inter and inner groups for three outputs, take the group of interaction coefficient in the case.

Conclusion



SA of groups
Models
Second-order Indices
Clustering detection
Factor Grouping
Group Analysis
?
Empirical knowledge


清華
汇智译丛
全局敏感度分析
GLOBAL SENSITIVITY ANALYSIS
A.Saltelli 等⊙著
吴琼莉 丁文明 易 鸣⊙译
清華大学出版社
全局敏感度分析
A.Saltelli 等⊙著
吴琼莉 丁文明 易 鸣⊙译
清華大学出版社
全局敏感度分析
GLOBAL SENSITIVITY ANALYSIS
上架建议：经济学
定价：00.00元
清华大学出版社数字出版网站
WQBook
www.wqbook.com


荆楚网
www.cnhubei.com

Thank you !

Qiongli Wu
Associate Research Fellow
Lab of Data Analysis and Statistical Computation
Wuhan Institute of Physics and Mathematics (WIPM)
Chinese Academy of Sciences (CAS)
Wuhan, 430071, China

wuqiongli@wipm.ac.cn