Universe Scientific Publishing

Causal Analysis of Air Pollution Based on Mixed Relation Index

Yuming Zhang, Mingwei Li, Nan Zhang
School of Science, Northeastern University, Shenyang 110004, China.

Abstract: Granger causality theory is widely used in data-driven analysis. In multivariable systems, unique, redundant, synergistic and mixed information structures may lead to distorted or complex causality. Based on mutual information and Granger causality theory, this paper decomposed multivariate information structure, and proposed Average Granger causality index and mixed relation index to reduce the data dimension and computational complexity. The experiments examine the causality of air pollutants in Beijing and its surrounding cities, and the results show that the method is feasible.
Keywords: Granger Causality; Redundancy; Synergy; Information Structure Decomposition; Air Pollution

1. Introduction

Ecological environment pollution [1] and industrial failure [2] directly affect the production and life of modern society, which have attracted extensive attention from relevant departments and researchers. Data-driven approaches have become a common tool for analyzing complex causality, among which Bayesian network [3], transfer entropy [4], Granger causality [5] are commonly used.

Granger causality analysis method is used to infer the potential causal mechanism between different time series. Fatemeh gholamzadeh [6] used a VAR model to predict the air pollution concentration in Tehran on the next day and explored the causal relationship between the variables using GC test, but none of them considered the effect of air quality in the surrounding cities on the central city. Zuofangzheng [7] used GC test to explore the time lag relationship of PM 2.5 in Beijing Tianjin Hebei region, using standard deviation, skewness and kurtosis coefficient to analyze air pollution. These statistics render difficult to estimate the influence of causality on air pollution in central cities. This paper considers using Granger causality and time series to solve this problem.

When the information is overlapping, the traditional GC based on vector autoregressive model has some limitations, which may lead to complex or even wrong experimental calculation. The interaction between variables can be summarized as redundancy and synergy. Redundant action means that the variables jointly transmit more information than the sum of the information transmitted separately, while synergy is the opposite. Redundancy and synergy can affect the performance of the model [8], so how best to decompose confounding information remains an open question. PaulL. Williams [9] proposed a definition of partial information decomposition, which exhaustively decomposes Shannon information in a multivariate system. FEIWANG [10] defined a redundancy index and stratified variables for fault diagnosis.

The main contributions of this paper can be summarized as follows: 1) the definition of partial information atom is introduced to divide information structure and reduce data granularity. 2) Based on the mean value, the average Granger causality index (AGCI) is proposed to express the concentration trend of time series data. 3) The mixed relation index (MRI) is put forward to reveal the potential causal relationship between variables intuitively. Experimental results show that this method overcomes the shortcomings of traditional models and is effective.

The paper is organized as follows: The second section introduces the concept of the GC and information theory. The

third section proposes AGCI and MRI In Section 4, this method is illustrated with the application to Beijing's air pollution data. And summarized in the last part.

2. Methods Introduction

In 1969, C.W. Granger [11] proposed the definition of the Granger causality: if the past values of X and Y can be used to predict the future of process y better than the past value of X, then X Granger-causes Y. And making the following two assumptions [12]:

I. The future can be affected by the past and present, otherwise.

II. The cause set does not contain redundant information.

Similar to GC, Shannon mutual information [13] can be used to quantify the information transfer between variables and capture pairwise information between variables. The purpose of introducing mutual information theory in this paper is to use the definition of partial information atoms, decompose R_i's information in a multivariate information system thoroughly, and determine the information about the time series S provided by each PI-atom, which may be uniquely, redundantly and synergistically. In the following of this paper, the expression is simplified by using $\{i\}$ to represent $\{R_i\}$, and the redundant atoms are considered as $\Pi_{\mathbf{R}}(S;\{i\}\{j\})$, synergistic information atom as $\Pi_{\mathbf{R}}(S;\{ij\})$.

In the following, the AGCI and MRI are defined to reduce the complex calculation caused by information overlap.

3. Mixed relation index

This paper mainly discusses information structure decomposition for more than four variables (including target variables). To simplify the trivariate relationship, consider determining the set of relationships between two variables, and then calculating the mixed relationship with the third variable.

Suppose that $I(S;R)$ represents the information of S provided by R, $\Pi_{\mathbf{R}}(S;\{i\})$ represents the unique information provided by R_i. $\Pi_{\mathbf{R}}(S;\{ij\}), i \neq j, \Pi_{\mathbf{R}}(S;\{i\}\{j\}), i \neq j$, indicates the synergistic or redundant information between R_i and R_j, when i or j is redundant or cooperative information atom, they represents mixed information.

Quantitatively, assume that the target variable is $\alpha = \{\alpha_1, \alpha_2, \cdots, \alpha_n\}$. Considering the time series $\{x_\alpha(t)\}_{\alpha=1,\cdots,n}$, the lag state vectors are denoted as $X_\alpha(t) = (x_\alpha(t-m), \cdots, x_\alpha(t-1))$, where m is the order of the model.

Based on the mean value, the average Granger causality index is proposed to indicate the central tendency of causality among multiple variables. Let B be a subset of $X = \{x_1, x_2, \cdots, x_m\}$. For the target variables $\alpha = \{\alpha_1, \alpha_2, \cdots, \alpha_j\}$ and variable B, the AGCI is defined as follows:

$$F_X(B \to \alpha) = \sum_B \sum_\alpha p(\alpha_j, B) log \frac{\varepsilon(x_\alpha | X \setminus B)}{\varepsilon(x_\alpha | X)} \quad (1)$$

Where $\varepsilon(x_\alpha|X)$ is the mean square error used to predict x_α based on all vectors X, $\varepsilon(x_\alpha|X \setminus B)$ is the mean square error based on all variables except B in vector X. Different variables are input with probability $p(B)$, and the output probability of the target variable is $p(\alpha_j)$. For different B and α_j, expressed by the joint probability $p(\alpha_j, B)$. For multiple possible outcomes, the traditional approach is to calculate Granger causality indices for the variables and the target variable in each case, leading to complex calculations for subsequent experiments. The AGCI proposed in this paper solves this problem. The joint probability of B and α_j is introduced and the expectation is calculated to obtain the AGCI, which is used to measure the overall relationship between the two variables.

In order to avoid a large number of calculations caused by changing subsets, formula (1) is rewritten into a unnormalized form:

$$F_X^U(B \to \alpha) = \sum_B \sum_\alpha p(\alpha_j, B)[\varepsilon(x_\alpha | X \setminus B) - \varepsilon(x_\alpha | X)] \quad (2)$$

Since the AGCI takes an overall perspective and treats the problem in an average sense, the AGCI must be positive.

If $\{x_\beta\}, \beta \in B$ in Formula (2) are statistically independent, then its contributions to α are additive, i.e

$$F_X^U(B \to \alpha) = \sum_{\beta \in B} F_{X\backslash B,\beta}^U(\beta \to \alpha) \tag{3}$$

When the contribution of B to α is greater than the sum of the contributions of all variables in subset B to α, i.e. $F_X^U(B \to \alpha) > \sum_{\beta \in B} F_{X\backslash B,\beta}^U(\beta \to \alpha)$, it means that produced the synergy between the variables. Conversely, if $F_X^U(B \to \alpha) < \sum_{\beta \in B} F_{X\backslash B,\beta}^U(\beta \to \alpha)$, illustrate the redundancy.

Define the index of the mixed relationship between i and j as:

$$\begin{aligned} \boldsymbol{\Psi}_{\alpha(i,j)} &= F_{X\backslash j}^U(i \to \alpha) - F_X^U(i \to \alpha) \\ &= F_X^U(i, j \to \alpha) - F_X^U(i \to \alpha) - F_X^U(j \to \alpha) \end{aligned} \tag{4}$$

Where, i and j represent the different atoms decomposed by the information structure in Section 3. This definition provides a quantitative measure of the mixed relationship between a pair of atoms. $\boldsymbol{\Psi}_{\alpha(i,j)} = 0$ means that the information atom is independent of the target variable. If $\boldsymbol{\Psi}_{\alpha(i,j)} > 0$, corresponding to redundancy; $\boldsymbol{\Psi}_{\alpha(i,j)} < 0$, it indicates synergy.

Property: $\boldsymbol{\Psi}_{\alpha(i,j)}$ has symmetry.

Proof: According to Equation (2), $\boldsymbol{\Psi}_{\alpha(j,i)} = F_X^U(j, i \to \alpha) - F_X^U(j \to \alpha) - F_X^U(i \to \alpha)$, Since there is no ordering relationship between i and j, $\boldsymbol{\Psi}_{\alpha(j,i)} = F_X^U(i, j \to \alpha) - F_X^U(i \to \alpha) - F_X^U(j \to \alpha) = \boldsymbol{\Psi}_{\alpha(i,j)}$.

In order to limit the MRI within [-1,1], this paper performs the following normalization.

$$\Psi'_{(i,j)} = \frac{\Psi(i,j)}{\max(|\Psi|)} \tag{5}$$

Where, Ψ represents all the MRI values in the system. In the subsequent experiments, the normalized results are presented.

4. Theoretical analysis

This paper uses the air pollution data of Beijing and its surrounding cities to verify the effectiveness of the method. Hourly concentration data of PM2.5, CO were selected from December 16 to December 31 in 2020. First, deal with the outliers and missing values of the original data. Then, stationarity test of time series was verified by ADF test.

Based on the initial definition of GC, the significance level was set at 0.05, and p-values were calculated as shown in Table 1. When $p \leq 0.05$, indicates that the city is the Granger cause of Beijing for a certain pollutant indicator. Table 1 shows:

1)Tianjin, Tangshan, Zhangjiakou and Chengde are granger causes of PM2.5 in Beijing.

2)Tianjin, Tangshan, Zhangjiakou and Chengde are granger causes of Beijing CO.

Table 1 P values of GCA

	tianjin	tangshan	baoding	zhangjiakou	chengde	langfang
PM2.5	0.040	0.050	0.521	0.002	0.002	0.169
CO	0.007	0.001	0.077	0.000	0.068	0.046

Table 2 Delay to choose

	tianjin	tangshan	baoding	zhangjiakou	chengde	langfang
PM2.5	2	2	-	3	4	-
CO	2	3	-	3	-	2

There is a time lag in air pollution in Beijing and surrounding cities. The experimental results show that the time lag is selected as shown in Table 2.

For $p(\alpha_j, B)$ in Formula (1), the thresholds are set separately to count the duration of the concentration exceeding the

threshold. Using the frequencies of pollutant exceedances in the central city and surrounding cities, the AGCI was calculated by replacing probabilities with frequencies. The thresholds are set as PM2.5: 75 $\mu g/m^3$, CO: 75 $\mu g/m^3$.

Figs.1-2 were obtained by calculating the MRI, which show the inter-city MRI with Granger effects on PM2.5, CO in Beijing.

In Fig.3-4, the values on the diagonal represent the unique information index provided by the city (or group of cities), which are all positive values. The lower triangle represents information provided by a city or group of cities. Red represents the synergy information index ($\Psi'_{(i,j)} < 0$), and blue represents the redundancy information index ($\Psi'_{(i,j)} > 0$). When considering the relationship between a city and a city group, a subordinate relationship indicates full redundancy, i.e., the relationship between the two is equivalent to the relationship within the city group.

In PM2.5, Tangshan uniquely provides the most information, indicating that Beijing is vulnerable if Tangshan PM2.5 is polluted. Tianjin and Tangshan produces the greatest redundancy, indicating that redundant information is generated between Tianjin to Beijing and Tangshan to Beijing. The experiment shows that Tangshan is the Granger cause of Tianjin, i.e., Tianjin is susceptible to Tangshan PM2.5, which is a redundant effect. Similarly, Zhangjiakou, Chengde and Tangshan produces a clear synergy. The conclusions can be verified by the degree of approximation of pollutant concentrations in the city (Fig. 3-4).

In CO, Tangshan provides the most unique information, Tangshan Zhangjiakou and Tianjin Zhangjiakou produce the largest redundancy, Tianjin Zhangjiakou and Tangshan produce synergy.

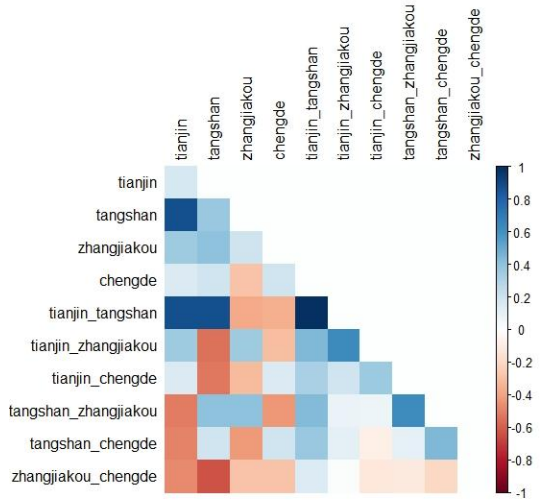

Figure 1 MRI of PM2.5

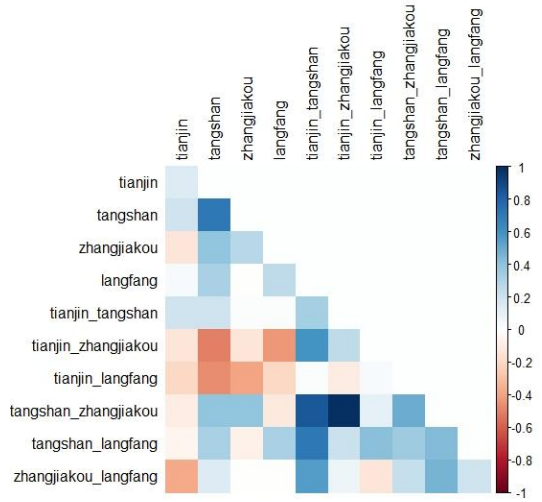

Figure 2 MRI of CO

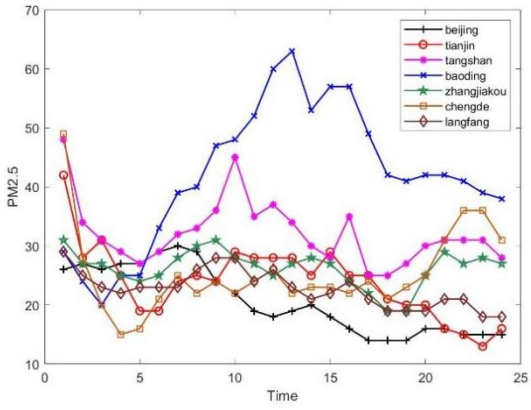

Figure 3 The concentration of PM2.5

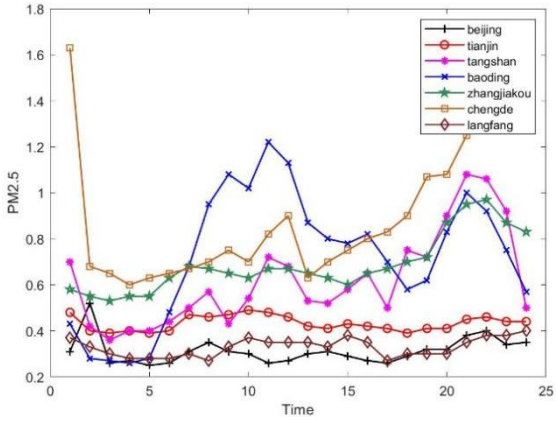

Figure 4 The concentration of CO

5. Conclusion

In this paper, information is divided into unique, redundant, cooperative and mixed information atoms by decomposing information structure to reduce data granularity. Based on the probabilities assigned to the predictor variables, the AGCI is proposed to capture their concentrated trends influenced by multiple variables. By nested synergies and redundant relationships, the MRI is proposed in a multivariate system to obtain the influence of each variable on the target variable. The air pollution data of cities around Beijing are used to explore the sources that can provide the most information for Beijing's air pollution, which proves the effectiveness of the MRI.

References

[1] Petrowski K, Bührer S, Strau B, et al. Examining air pollution (PM10), mental health and well-being in a representative German sample[J]. Scientific Reports, 2021, 11(1):18436.

[2] Luo B, Wang H, Liu H, et al. Early Fault Detection of Machine Tools Based on Deep Learning and Dynamic Identification[J]. IEEE Transactions on Industrial Electronics, 2018.

[3] Weidl G, Madsen AL, Israelson S. Applications of object-oriented Bayesian networks for condition monitoring, root cause analysis and decision support on operation of complex continuous processes[J]. Computers & Chemical Engineering, 2005, 29(9):1996-2009.

[4] Bauer M, Cox JW, Caveness MH, et al. Finding the Direction of Disturbance Propagation in a Chemical Process Using Transfer Entropy[J]. IEEE Transactions on Control Systems Technology, 2006, 15(1):12-21.

[5] Yuan, T., and Qin, SJ. (2014). Root cause diagnosis of plant-wide oscillations using Granger causality. Journal of Process Control, 24, 450-459.

[6] Gholamzadeh F, Bourbour S. Air pollution forecasting for Tehran city using Vector Auto Regression[C]// 2020 6th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS). 2020.

[7] Zheng Z, Xu G, Yang Y, et al. Statistical characteristics and the urban spillover effect of haze pollution in the circum-Beijing region[J]. Atmospheric Pollution Research, 2018, 1062-1071.

[8] Che JX. Optimal sub-models selection algorithm for combination forecasting model[J]. Neurocomputing, 2015, 151:364-375.

[9] Williams PL, Beer RD. Nonnegative Decomposition of Multivariate Information[J]. 2010.

[10] Wang F, et al. "An Improved Granger Causal Analysis Framework Based on Redundancy Index." 2020 IEEE 9th Data Driven Control and Learning Systems Conference (DDCLS) IEEE, 2020.

[11] Granger C, Granger C, Granger C, et al. Investigationg causal relations by econometric models: cross spectral methods. 1969.

[12] Granger, CWJ. (1980). Testing for causality: A personal viewpoint. Journal of Economic Dynamics and Control, 2, 329-352.

[13] Shannon CE and Weaver W, The Mathematical Theory of Communication (Univ of Illinois Press, 1949).