香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

Final Report of DDA4210

Causal Inference of Hong Kong Sea Water Quality Parameters

Wang Ping 120090853

Han Ruobing 120090213

Liu Yuxuan 120090737

Yu Lihang 119010409

School of Data Science

May 6, 2023

1 Significance

The main goal of our study is to investigate the causal linkages of water pollution features by using different causality models. In order to gauge the causal transmission patterns we employ many classical methods of: (i) Causal Discovery Toolbox, (ii) Toda and Yamamoto level VAR (Toda & Yamamoto, 1995), (iii) Dowhy respectively. The study illustrates how such methods may be appropriately augmented or combined in a compatible fashion to unearth previously unfounded linkage properties inherent amongst a system of water pollution features.

Our major aim in this study is to make an attempt to better incorporate different models to draw to a strong conclusion as well as illustrate the characteristics between different methods. Our results would inspire future researchers to better the models, adopting their good points and avoiding their shortcomings, building up a more solid causal relation inference.

2 Data Collection and Preprocessing

Our raw data is collected from the governmental website of Hong Kong (https://data.gov.hk/tc-data/dataset/hk-epd-marineteam-marine-water-quality-historical-data-tc), using the data from 2019-2021. Data preprocessing would be divided into 3 steps.

In the first step, we use $DataFrame.info()$ function in python library Pandas to check out NaN values in the table, as shown in Figure 2.1. According to the result, attributes Total Kjeldahl Nitrogen (mg/L), Total Nitrogen (mg/L), and Total Phosphorus (mg/L) are removed from the table due to their outstanding number of NaN values included. Value including string "<" are eliminated since they are considered to be not significant.

In the second step, we construct different graphs to visualize the data, in order to have a clear understanding of the distributions of the data, as well as their intrinsic correlation. The visualization is achieved using functions from python library seaborn. The 4 types of graphs are histogram by $sns.distplot()$, box graph by $sns.boxplot()$, heat map by $sns.heatmap()$, and trend graph by $sns.lmplot()$. The graphs can be found in Appendix (Figure 2, 3, 4, 5).

After gaining a basic view on the graphs, the last step before we move on to implement the models is picking target attributes according to domain knowledge of water pollution. According to Huang et al.'s study in 2018, we lock on attributes including Temperature (°C) and Dissolved oxygen (mg/L) as potential causal factors or receptor.

3 Methodology

3.1 Causal Discovery Toolbox

Different models will fit well in some specific situations. Before using CDT, we first tried to fit one classical causal discovery method, called the PC algorithm discovery. This method constructed the set of vertices in graph C and did the iteration to do the casual discovery. However, when we fit the data into this tool and choose four features that we took them

as highly correlated factors, the result always shew in undirected way (see Figure 6). Considering the reason might be the large scale difference in four factors, we also standardized the data using the original data subtracting the mean and dividing the standard error. The change also did not change the result.

We know that the the PC algorithm discovery was not suitable to do the causal discovery here. Then, we tried to use a tool that is not a specific method to get the development. Here, we chose the Causal Discovery Tool. This method is based on the graphics of the neural network : CGNN. The usage of the package in Python is open-source and under the MIT license. Causal Discovery Toolbox (CDT) is an open-source Python package observing causal findings aims to learn causal maps and associated causal mechanisms from a sample of the joint probability distribution of the data. CDT includes many state-of-the-art causal modeling algorithms (some of which are imported from R), with support for GPU hardware acceleration and automatic hardware detection.

Download and import the package in python, then fit all the full dataset into the tool. Then, we get the whole causal discovery for all the features. we can discover an existing causal relationship pointing from temperature to dissolved oxygen. Beside our targets, we also discover two other features that connects with target features at the same time, Fecal coliform and NTU (see Figure 7). Then we chose the 4 features and ran the CDT again. The result shew the direct relationship from temperature to dissolved oxygen is gone (see Figure 8). It seems that the confounders are contributing to their relationship.

One possible explanation is that the causal relationships are complex and may depend on the presence of other variables. When only include the four features, these other variables may not be present, leading CDT to identify a different causal structure. Another possibility is that the results may be affected by noise in the data. And for 4 features, the data may be less noisy and the hyperparameter may contribute to the variation.

3.2 DoWhy

Given we' ve discovered some causal relations with Causal Discovery Toolbox, we can now take a step further to quantify and verify such causal relation and even do statistical tests on whether these relations are indeed true. Two variables we noticed are Temperature(°C) and Dissolved Oxygen (mg/L). By commonsense, a rise in temperature would cause a drop in solubility of oxygen in water, and we've pinned down this relationship through the NRDC research. But this don't necessarily mean such causal relationship is observable in real-world data since we cannot guarantee water to be saturated with oxygen. The dataset contains a large group of variables to choose from and the causal relationship between temperature and dissolved oxygen (measured in mg per liter) may be affected by them to the point where the effects of temperature on it has become minimal. And DoWhy is introduced to help identify the numerical, causal effect of temperature on dissolved oxygen.

Based on previous CDT discovery, we construct the causal graph as shown in Figure 9."Water Control Zone" variable is not considered here because we deemed it not necessary to do so, as Figure 5 (in the appendix) would suggest that a general dropping trend can

be observed and thus quantified. The graph is constructed by picking the variables that we consider most relevant to temperature and dissolved oxygen, and then taken from the CDT discovered causality graph. With these we use dowhy.CausalModel.identify_effect to identify an estimand for the effect.

And to check the causality is indeed true, we further use some several off-the-shelf methods in DoWhy to check. Here we choose 3 from all that are available. The first is "*random_common_cause*", which adds a randomly generated confounding variable to both the treatment (temperature) and effect (dissolved oxygen). A high p-value in this case is indicator of high confidence in the null hypothesis (treatment and effect having the estimated causal effect). The second is "$data\_subset\_refuter$", where the package would check whether the same effect would exist in a subset of all data. Again, a high p-value here would indicate high confidence in the null hypothesis being true. And the results below shows that the causal effect has indeed been verified. (Outputs of terminal) The estimated effect sits at around -0.086, meaning a 1°C increase in temperature causes dissolved oxygen (measured in mg/liter) to drop by 0.086. The terminal output is shown in Figure 10 and Figure 11 in the Appendix.

3.3 Toda and Yamamoto level VAR

Toda and Yamamoto (1995) proposed a method based on the estimation of augmented VAR model $(k + d_{max})$ where k is the optimal time lag on the first VAR model and d_{max} is the maximum integrated order on system's variables (VAR model). The current model is modified to corporate the multivariate case following the procedures and equations below:

• To find the maximum integration order for each series (d_{max}),we developed an authomated method. More specifically, the order of integration is tested by using the function ndiffs from the Rob J. Hyndman's forecast package. The function ndiffs uses a unit root test to determine the number of differences required for time series x to be made stationary. The test we use is Kwiatkowski–Phillips–Schmidt–Shin (KPSS);

• We create a VAR model on series levels regardless of integration order that we found, the order of VAR model (k) from lag length is determined by final prediction error (FPE), AIC, SC, HQ criteria, which is 1, 2 or 6 in this case, illustrated in Figure 9 in Appendix. We use the library vars to fit a VAR model augmented of equation lags.

• We test if VAR $(k+d_{max})$ (adjusted VAR model) is correctly specified or stable. Model with $k = 6$ is less likely to be serially correlated. Thus model with $k = 6$ is selected.

• We apply Granger causality test using pairwise equations and modified Wald test (MWald) for the significance of parameters on examined equations on number time lags $(k + d_{max})$, which follows Chi-square distribution asymptotically and the degrees of freedom are equal to the number of time lags $(k+d_{max})$. The wald test is achieved by the aod library in R.

• Rejection of null hypothesis entails the rejection of Granger causality.

• Finally, we check if there is cointegration on VAR model. If two or more series are cointegrated, then there is one causal relationship (unidirectional or bilateral) but not vice versa.

The Toda-Yamamoto approach overcomes some issues that can arise when testing nonstationary series, a condition leading to the risk of spurious causality (He, Maekawa, 2001)

and thus is applicable to more dataset than Granger test.
The results of tests is summarized in Table 1 (part of the results is shown due to the space limit) and causality graph with linkages over 99% significant level are shown in Figure 13. The p-value of temperature and dissolved oxygen is 0.003191, also indicating the strong causality between the two features. The current approach also finds that the strong causal effect of temperature on pH, which has a p-value equal to 0.002740616. The causality between temperature, pH and dissolved oxygen is inaccordance with Huang et al.'s study.

4 Conclusion

This paper focuses on patterns of causal linkages among water quality factors. In dealing with such an issue, a comparative analysis was made of three approaches in conducting causal inferences in systems containing possibly integrated as well as cointegrated processes. At a more substantive level, this study provides further evidence of significant causality between Temperature and Dissolved Oxygen (Huang et al., 2018). While we do not present any evidence to over-ride this finding, comparative analysis of current results with other results from the same field studies suggest that, causality inference between water quality parameters tend to be more similar among area that have similar land use practices and industrial activities. For example, AI-Ansari et al. (2015), Wang et al. (2017) and Shen et al. (2019) all concluded the strong causality effects of temperature on dissolved oxygen while the regions they studied, including Huang et al.'s study, all emphasized heavily on agricultural land use. Thus the current research can be extended to explore the causality of certain water quality parameters and land use practice.

Limitations of this analysis should not be ignored. These include the aggregated nature of this study (monthly rather than a higher frequency). While we test for three classical approaches and give a blended and common conclusion, a useful extension of the paper would be to adjust for applying more classical or augmented model in the interest of robustness of statistical results. A more detailed study focusing on the field and using data observed at daily or weekly intervals as well as adopting more approches could provide a positive and practical step for future research.

References

[1] Al-Ansari, N., Knutsson, S., & Ali, A. (2015). Causal relationships among water quality parameters in the Euphrates river basin. *Water*, *7*(11), 6446-6465. doi: 10.3390/w7116446.

[2] Diviyan Kalainathan. (2020). Causal Discovery Toolbox: Uncovering causal relationships in Python. *Journal of Machine Learning Research*, *21*, 1-5.

[3] He, Z., & Maekawa, K. (2001). On spurious Granger causality. *Economics Letters*, *73*(3), 307-313.

[4] Huang, G., Chen, S., & Yang, M. (2018). Causal relationships between water quality parameters in river systems. *Water Science and Engineering*, *11*(2), 95-104. doi: 10.1016/j.wse.2018.06.003.

[5] Malinsky, D., & Danks, D. (2018). Causal discovery algorithms: A practical guide. *Philosophy Compass*, *13*(1), e12470. doi: 10.1111/phc3.12470.

[6] Spirtes, Peter, & Clark Glymour (1991). An algorithm for fast recovery of sparse causal graphs. *Social science computer review*, 62-72.

[7] Shen, Y., Liang, J., Zhang, X., & Chen, L. (2019). Causal relationships among water quality parameters in the Chesapeake Bay. *Water Research*, *157*, 81-91. doi: 10.1016/j.watres.2019.03.020.

[8] Toda, H. Y., & Yamamoto, T. (1995). Statistical inference in vector autoregressions with possibly integrated processes. *Journal of econometrics*, *66*(1-2), 225-250.

[9] Wang, J., Chen, B., & Yang, S. (2017). Causal relationships among water quality parameters in the Yangtze river basin. *Environmental Science and Pollution Research*, *24*(25), 20560-20574. doi:10.1007s11356-017-9552-9.

5 Appendix

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2708 entries, 0 to 2707
Data columns (total 29 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   水質管制區              2708 non-null   object
 1   監測站                2708 non-null   object
 2   日期                 2708 non-null   object
 3   樣本編號               2708 non-null   int64
 4   深度                 2708 non-null   object
 5   五天生化需氧量 (毫克/升)     2708 non-null   object
 6   氨氮 (毫克/升)           2708 non-null   float64
 7   葉綠素-a (微克/升)        2708 non-null   object
 8   溶解氧飽和百分率 (百分率)     2708 non-null   int64
 9   溶解氧 (毫克/升)          2708 non-null   float64
 10  大腸桿菌 (菌落數/100毫升)   2708 non-null   object
 11  糞大腸菌群 (菌落數/100毫升)  2708 non-null   object
 12  硝酸鹽氮 (毫克/升)         2708 non-null   object
 13  亞硝酸鹽氮 (毫克/升)        2708 non-null   object
 14  正磷酸鹽磷 (毫克/升)        2708 non-null   object
 15  酸鹼值                2708 non-null   float64
 16  脫鎂色素 (微克/升)         2708 non-null   object
 17  鹽度 (psu)           2708 non-null   float64
 18  透明度 (米)            2708 non-null   float64
 19  硅 (二氧化硅) (毫克/升)     2708 non-null   object
 20  懸浮固體 (毫克/升)         2708 non-null   object
 21  溫度 (攝氏)            2708 non-null   float64
 22  無機氮 (毫克/升)          2708 non-null   float64
 23  凱氏氮 (毫克/升)          1214 non-null   object
 24  總氮 (毫克/升)           1214 non-null   float64
 25  總磷 (毫克/升)           1214 non-null   object
 26  混濁度 (NTU)          2705 non-null   float64
 27  非離子氨氮 (毫克/升)        2708 non-null   object
 28  揮發性固體總量 (毫克/升)     2708 non-null   object
dtypes: float64(9), int64(2), object(18)
memory usage: 613.7+ KB
```

Figure 1: Data attributes information

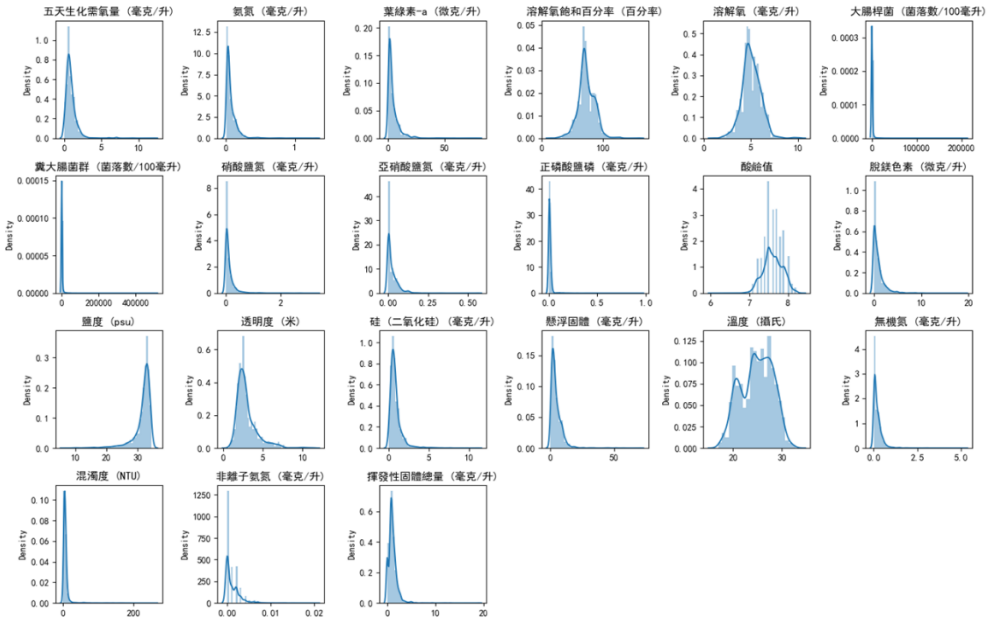

Figure 2: Histograms of the attributes

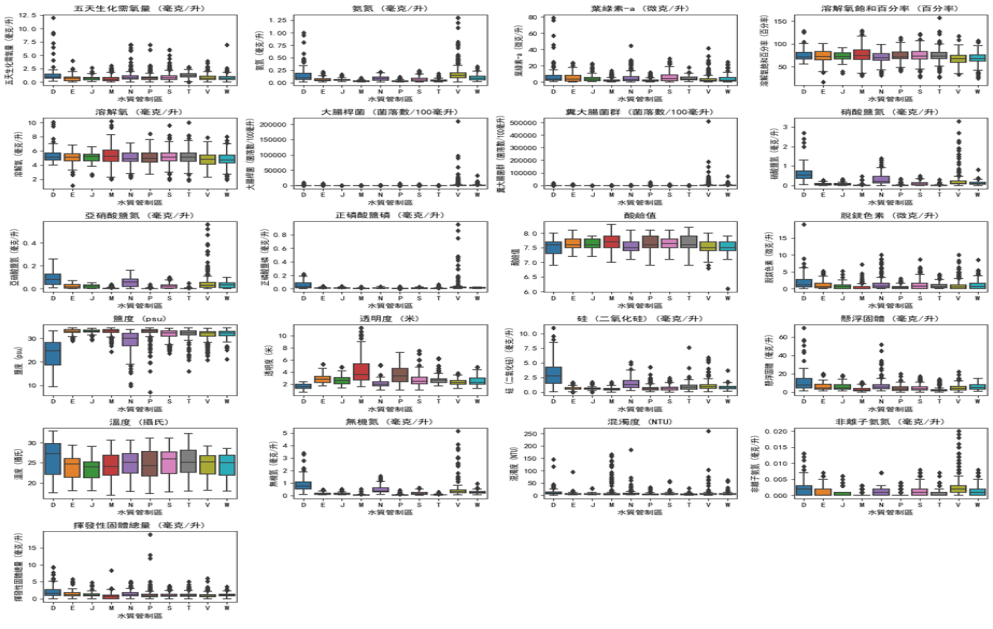

Figure 3: Box graphs of the attributes

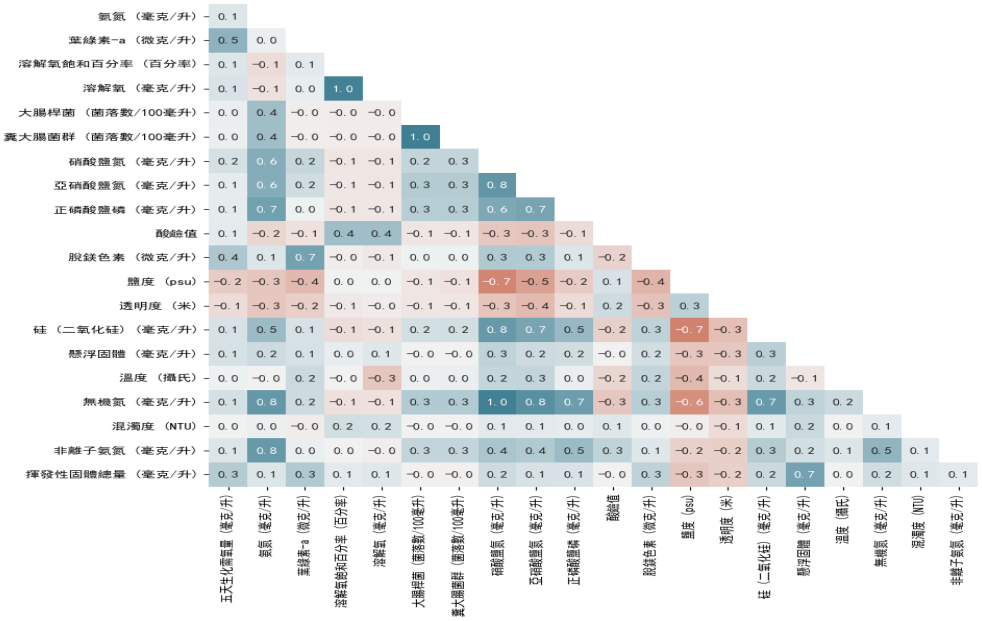

Figure 4: Heat map of the attributes

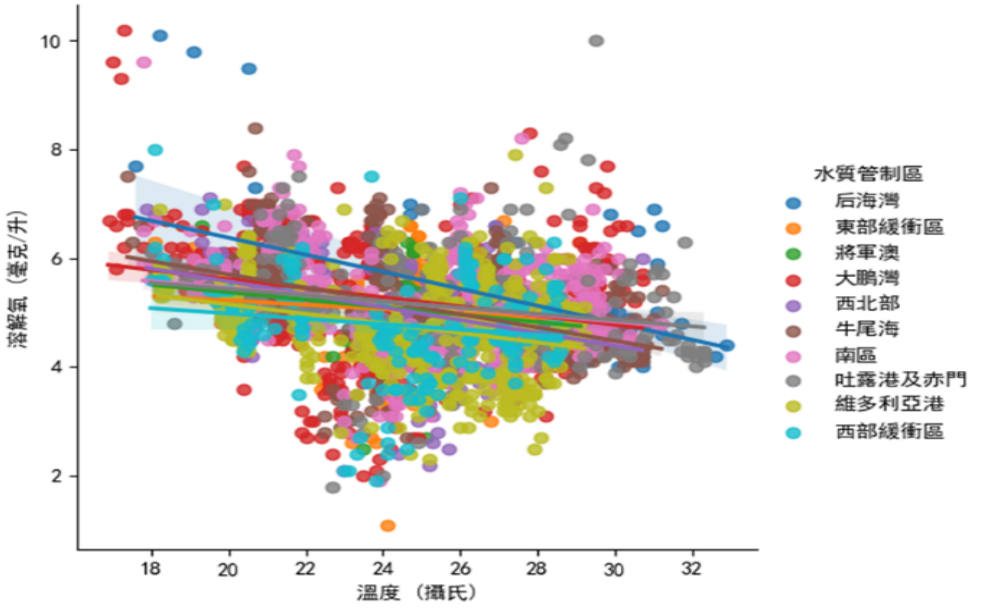

Figure 5: Trend graph of Temperature and Dissolved oxygen

Green: undirected; Blue: directed; Red: bi-directed

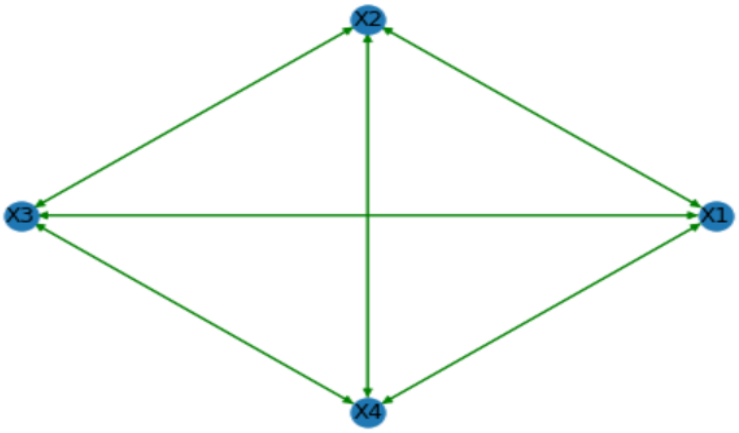

Figure 6: Graph constructed from CDT

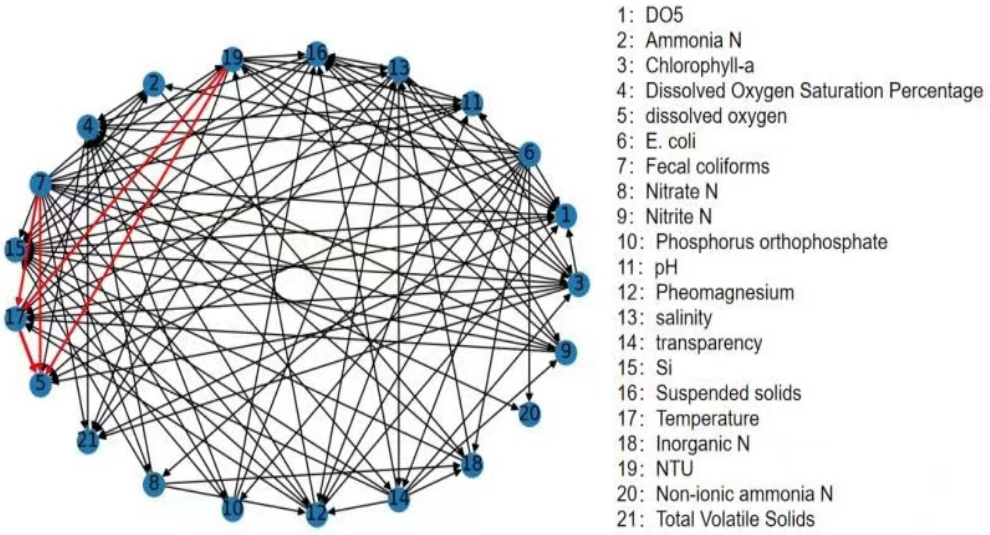

Figure 7: Graph constructed from CDT

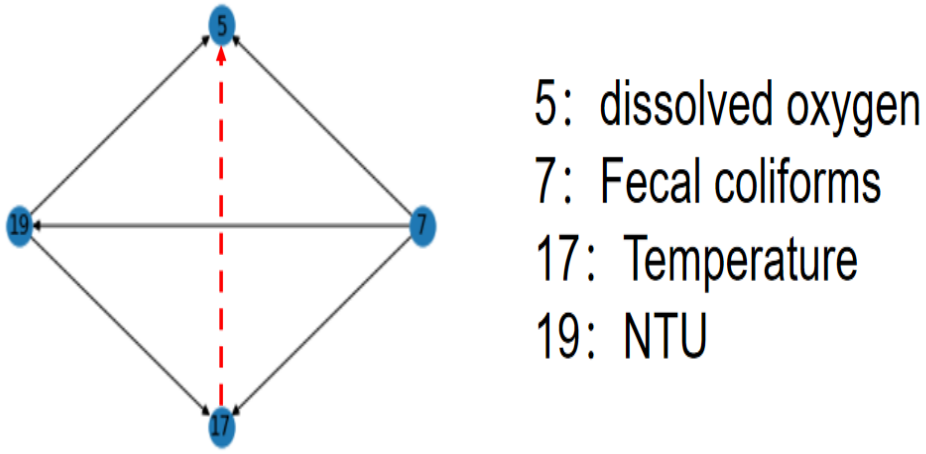

Figure 8: Graph constructed from CDT

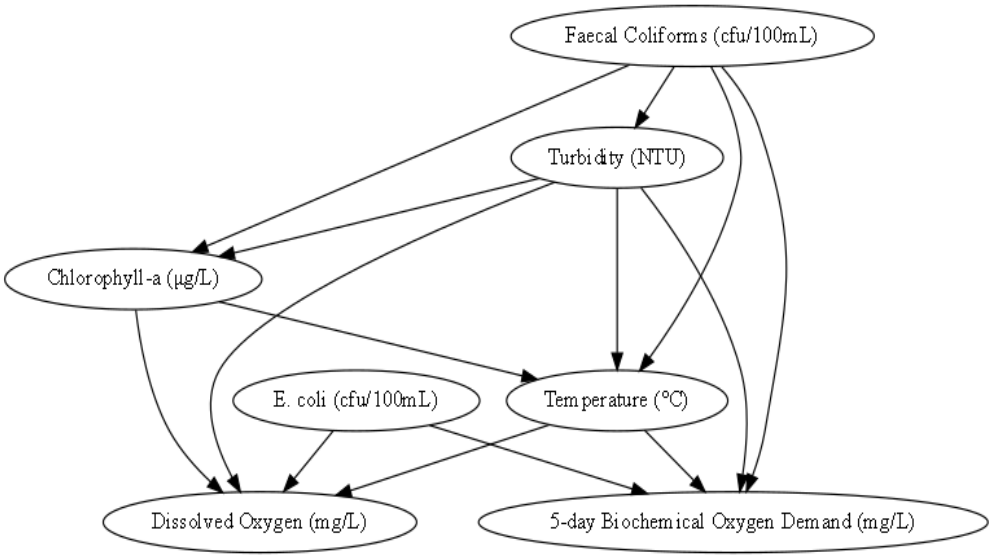

Figure 9: Graph constructed from CDT

Figure 10: Terminal Output of Dowhy 1

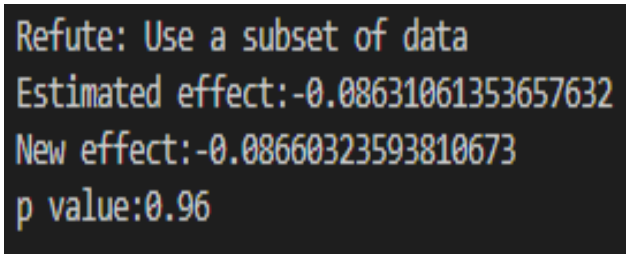

Figure 11: Terminal Output of Dowhy 2

	Dissolved Oxygen	Nitrite Nitrogen	pH	Salinity	Temperature
Dissolved Oxygen	1	0.376	0.447	0.809	0.708
Nitrite Nitrogen	0.114	1	0.468	0.764	0.785
pH	0.371	0.067	1	0.358	0.014
Salinity	0.681	0.053	0.279	1	0.003
Temperature	0.003	0.108	0.003	0.344	1

Table 1: Summary of part of the causality results based on Toda–Yamamoto [k+d_{max}]th-order level VAR procedure.Reported above are significance levels associated with asymptotic Wald statistic. The row indicates the cause and the column indicates the effect. For example, the significance level between Salinity and Nitrite Nitrogen is $0.053 < 0.1$, then Salinity causes Nitrite Nitrogen at a 90% significance level

```
$selection
AIC(n)  HQ(n)  SC(n) FPE(n)
     6      2      1      6

$criteria
                1           2           3           4           5           6
AIC(n) -5.673499951 -6.241343667 -6.508296420 -6.608281583 -6.73654177 -6.77588200
HQ(n)  -5.220508050 -5.355950406 -5.190501800 -4.858085602 -4.55394443 -4.16088330
SC(n)  -4.436383248 -3.823342837 -2.909411465 -1.828512501 -0.77588856  0.36565533
FPE(n)  0.003435873  0.001947467  0.001491599  0.001350382  0.00118885  0.00114445
                7           8           9          10
AIC(n) -6.746407553 -6.687660463 -6.677286078 -6.661316860
HQ(n)  -3.699007493 -3.207859043 -2.765083297 -2.316712720
SC(n)   1.576013907  2.815645123  4.006903635  5.203756979
FPE(n)  0.001180785  0.001255201  0.001272155  0.001297557
```

Figure 12: integration order determined by different criteria

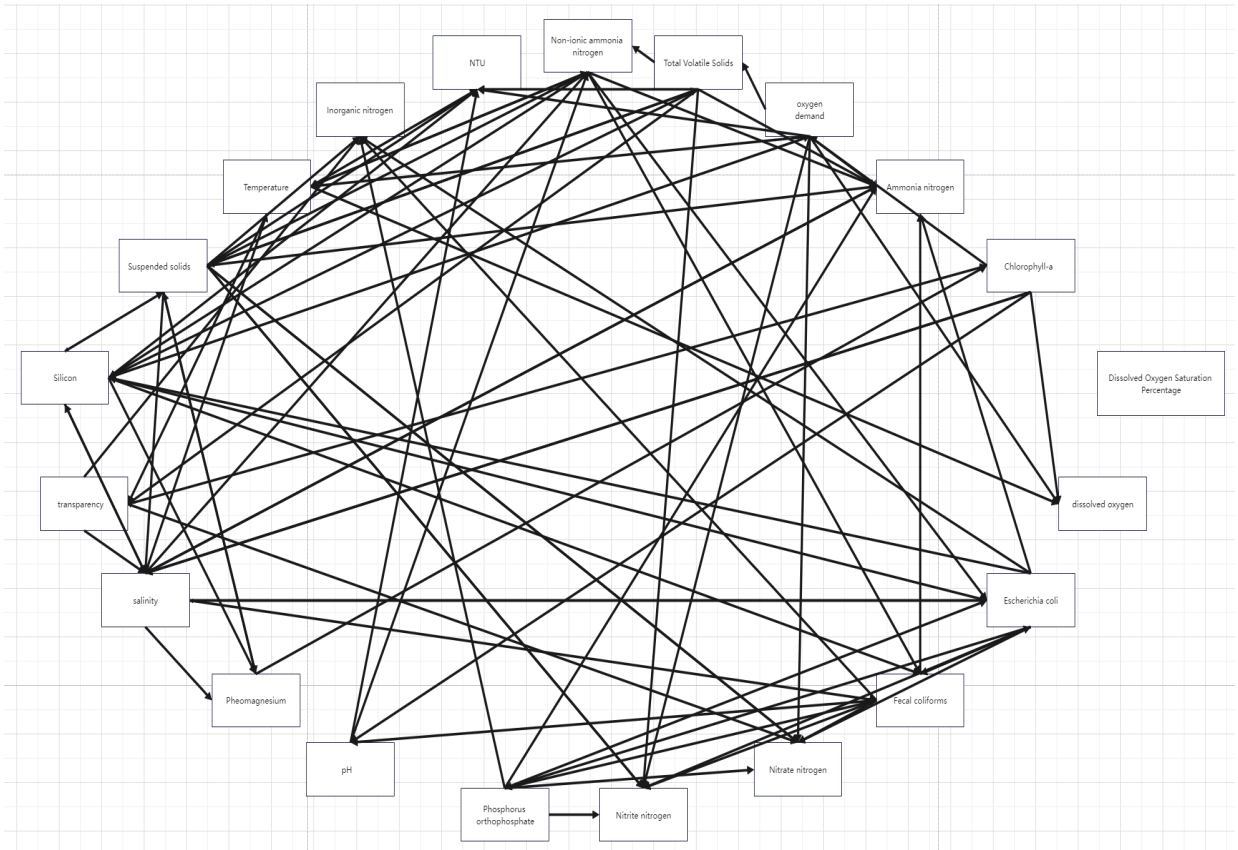

Figure 13: Causality Graph of linkages with 99% significance level