Nonlinear Causal Link Estimation under Hidden Confounding with an Application to Time Series Anomaly Detection

Violeta Teodora Trifunov[1,2][0000−0002−9987−2356], Maha Shadaydeh[1][0000−0001−6455−2400], Jakob Runge[2][0000−0002−0629−1772], Veronika Eyring[4,5][0000−0002−6887−4885], Markus Reichstein[3,6][0000−0001−5736−1112], and Joachim Denzler[1,3][0000−0002−3193−3300]

[1]Computer Vision Group, Friedrich Schiller University Jena, Jena, Germany
[2] Climate Informatics Group, Deutsches Zentrum für Luft- und Raumfahrt e.V. (DLR), Institute for Data Science, Jena, Germany
[3] Michael Stifel Center Jena for Data-Driven and Simulation Science, Jena, Germany
[4] Deutsches Zentrum für Luft- und Raumfahrt e.V. (DLR), Institut für Physik der Atmosphäre, Oberpfaffenhofen, Germany
[5] University of Bremen, Institute of Environmental Physics (IUP), Bremen, Germany
[6] Max Planck Institute for Biogeochemistry, Jena, Germany, Maha Shadaydeh

Abstract. Causality analysis represents one of the most important tasks when examining dynamical systems such as ecological time series. We propose to mitigate the problem of inferring nonlinear cause-effect dependencies in the presence of a hidden confounder by using deep learning with domain knowledge integration. Moreover, we suggest a time series anomaly detection approach using causal link intensity increase as an indicator of the anomaly. Our proposed method is based on the Causal Effect Variational Autoencoder (CEVAE) which we extend and apply to anomaly detection in time series. We evaluate our method on synthetic data having properties of ecological time series and compare to the vector autoregressive Granger causality (VAR-GC) baseline.

1 Introduction

Causality analysis represents one of the most important tasks when examining dynamical systems such as ecological time series. Its principal difficulties are hidden causes of the observed phenomena, in addition to the often-found nonlinearities in the data. We propose to mitigate the problem of inferring nonlinear cause-effect dependencies in the presence of a hidden confounder by using deep learning together with domain knowledge. Moreover, we suggest a time series anomaly detection approach using causal link intensity increase as an indicator of the anomaly.

In ecosystems for instance, considering the problem of confounding is important when trying to determine a causal link between gross primary production (GPP) and the ecosystem respiration (R_{eco}). Since both of these variables are

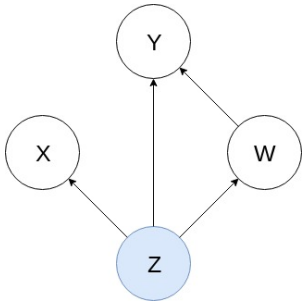


Fig. 1: Causal graphical model portraying hidden confounding with one proxy. Variable Y denotes an outcome, W an intervention variable, Z an unobserved confounder and X denotes a proxy variable providing noisy views on the hidden confounder Z.

influenced by the global radiation (R_g), one cannot be certain that the causal link between them is not influenced by this variable. Therefore, not considering a confounder may lead to incorrect conclusions. Two variables, W and Y, are said to be *confounded* if there exists another variable Z that causes both W and Y. In order to verify whether the confounder is influencing the link between W and Y, we need to intervene on W in the sense of do-Calculus [15] and thereby remove any influence of Z on W. If the intervention on W does not affect the outcome, it is clear that the causal link between W and Y is influenced exclusively by the hidden confounder Z itself.

When a confounder is observed, the usual approach for accounting for its effect is to "control" for it, for instance by covariate-adjusted regression or propensity score regression [12]. However, if a confounder is hidden or unmeasured, it is impossible to estimate the effect of the intervention on the outcome without further assumptions [15]. This is, nevertheless, a problem of utmost importance in observational studies, Simpson's paradox [23] being a good example of the type of bias that may occur in causal inference if unmeasured confounding is not properly dealt with. One way to tackle this issue is by using a proxy to the hidden confounder instead of the confounder itself. In the previously described ecosystem example, the air temperature (T) can be utilized as a proxy to the confounder R_g. Figure 1 depicts a version of this problem in the form of a causal graphical model when there is only one proxy variable, as suggested by [13]. For more general proxy models, as well as conditions under which they could be identified, see [14].

In ecological time series, variables often contain trends or periodic components such as diurnal and seasonal cycles. These act as an unobserved confounder, concealing the true causal effect between the affected variables. It was recently shown in [21] that time domain causality analysis of ecosystem variables based on vector autoregressive Granger Causality (VAR-GC) [7], may result in spurious causal links due to the diurnal or seasonal cycle. To tackle this issue,

the authors in [21] proposed to use the parametric frequency domain representation of VAR-GC. It was further shown in [20], [19] that anomalous events can be detected as those events where the causal intensities between the variables in certain frequency bands differ considerably from the average causal intensities. The application of our method to anomaly detection builds upon these findings. Namely, we estimate the causal link intensity between confounded variables and by observing an increase of this estimation, we are able to detect anomalies. Moreover, our work extends to a setting where seasonal cycles or trends are acting as the unobserved confounder. Additionally, we are able to perform causal inference of not only linear, but also the nonlinear inter-variable relationships, which are difficult to consider using GC methods. Our suggested method is based on the causal effect variational autoencoder (CEVAE) [13], a deep graphical model designed to estimate the unknown latent space summarizing the confounders and the causal effect by relying on a noisy proxy of the hidden confounder, as seen in Figure 1. It is required that the causal graphical model used by the CEVAE satisfies the back-door criterion [15] in order for it to be possible to use the do-Calculus and calculate the desired causal effect. We extend CEVAE for ecological time series and use it to infer a nonlinear causal link between variables confounded by the periodic component such as the seasonal or diurnal cycle. We apply our proposed method in this setting to estimate the intensity of the previously mentioned causal link. By being able to do so, we use its increase to detect anomalies. Furthermore, we are, to the best of our knowledge, first to use this deep graphical model for anomaly detection. In summary, our method which builds upon the CEVAE is in line with the trend to apply deep learning in Earth system analysis for describing the spatio-temporal dependency of ecosystems on climate and the interacting geo-factors as recognized by [17].

In Section 2, we discuss methods of causality analysis and anomaly detection on time series, whereas we devote Section 3 to outlining the CEVAE method along with our adaptation of it to ecological data. In Section 4 we describe synthetic data used to evaluate our method, followed by experimental results and the comparison to the VAR-GC method. Finally, Section 5 concludes our paper.

2 Related work

The analysis of causal dependencies in time series has become a focal point of study in various fields such as engineering, finance, the physical and life sciences [18]. The main assumption of this probabilistic concept of causality is that causes always come before their effects in time. This means that if one time series causes another series, knowing the former series should be helpful for predicting future values of the latter series after influences of all other variables have been considered. A standard method used for this purpose is Granger Causality (GC) applied in the setting of no hidden confounding [4] and only when causal links are linear. These limitations persist in anomaly detection methods relying on GC for time series [16]. Seeking to improve the conventional way of causality

analysis, several deep learning approaches have been suggested. One such approach is introduced for inferring interactions between variables while learning the dynamics in an unsupervised manner [10]. Furthermore, Causal Effect Network (CEN) [11] has been proposed for assessing causal relationships of time series, as well as their time delays between different processes. However, these methods cannot be applied when causal inter-variable relations are nonlinear, nor when they are driven by a hidden confounder. In [24], causality between the global radiative forcing and the annual global mean surface temperature anomalies (GMTA) is measured as the time rate of information flowing from one time series to another. A different branch of research that deals with modelling the latent variable space using deep graphical models was introduced in the recent years, specifically by the introduction of a Variational Autoencoder (VAE) in [9]. It is a deep learning method combined with a directed probabilistic graphical model for efficient inference in the presence of continuous latent variables with intractable posterior distributions. Moreover, it represents a crucial building block of a CEVAE [13], which allows for estimation of the unknown latent space and inference of the causal links between the confounded variables. Our work extends the capabilities of a CEVAE to time series, as well as to its novel application to anomaly detection using an increase of the causal link intensity.

An unsupervised method for discovering anomalies considering intervals of multivariate time series is proposed by [1]. It proposes that instead of regarding one point at the time, it is beneficial to compare probability distribution of samples within an interval to that of the rest of the data. A recently developed method for anomaly detection of time series using a Variational Recurrent Autoencoder (VRAE) [5] is proposed in [2]. It applies a latent-space detection approach which considers the variability of the latent representations, as well as their expectation and computes the anomaly score using the median Wasserstein distance [25] between a test sample and other samples within the test set of latent representations.

Our method differs substantially from other anomaly detection approaches as we rely on causal link intensity changes in the presence of an unobserved confounder for detecting anomalies in ecological time series which has, to the best of our knowledge, not been done so far.

3 Methodology

3.1 Causal effect variational autoencoder

Based on a VAE [9] and a TARnet [22] generative model structure, CEVAE [13] is a deep learning method dealing with hidden confounding as it estimates the latent space and summarizes the causal effect of discrete or continuous, non-sequential variables, using a noisy proxy related to the confounder, as shown in Figure 1. One of its original applications was to medical data, so that in Figure 1 W denotes treatment, Y an outcome of the treatment, whereas a hidden confounder Z represents the socio-economic status of each patient and its proxy X represents patient's income for the previous year and a place of residence. The

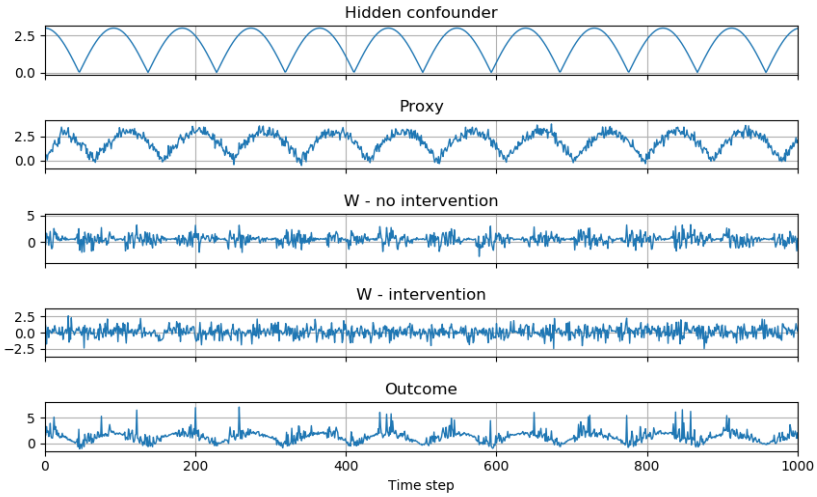


Fig. 2: Synthetic data. The first row shows the hidden confounder Z with parameters $b = 5$ and $f = 36.5$; the second row shows the noisy proxy X for $s = 500$ and $\beta = 0.3$; the third row shows variable W for $\mu_w = 0.55$ and $e = 0.4$ without intervention; the fourth row shows variable W with intervention; the fifth row shows the outcome Y for $g = 0.8$ and $\alpha = 3$.

main objective was, therefore, recovering the Individual Treatment Effect (ITE) and the Average Treatment Effect (ATE) defined in (1) and (2), respectively:

$$ITE(x) := \mathbb{E}(Y|X = x, do(W = w^1)) - \mathbb{E}(Y|X = x, do(W = w^0)) \quad (1)$$

$$ATE := \mathbb{E}(ITE(x)) \quad (2)$$

These metrics are defined for each value x of variable X, and by w^1 we denote applied treatment, while values of W when no treatment is applied are denoted by w^0. ATE is easily calculated once we recover the ITE, and to do that we need to recover the joint probability $p(Z, X, W, Y)$, as shown by Theorem 1 in [13]. Obtaining this joint distribution is done through a model network of a CEVAE by estimating the true posterior over Z which depends on X, W and Y, whereas Z itself is modelled by the standard normal distribution. The estimate of the posterior is then inferred via TARnet [22] by splitting it for each intervention group in W. It is then possible to construct a single objective for the inference and model networks, i.e. the *variational lower bound*

$$\begin{aligned}\mathcal{L} = \sum_{i=1}^{N} \mathbb{E}_{q(z_i|x_i,w_i,y_i)} (\log p(z_i) - \log q(z_i|x_i, w_i, y_i) \\ + \log p(x_i, w_i|z_i) + \log p(y_i|w_i, z_i))\end{aligned} \quad (3)$$

of the causal graphical model from Figure 1. By x_i we denoted an input data point, w_i corresponds to the treatment assignment, y_i to the outcome of the specific treatment, z_i corresponds to the latent hidden confounder and by q we denote estimation of the probability distribution with the same arguments. Finally, since it is necessary to know the intervention assignment w together with its outcome y before inferring the posterior distribution over Z, two auxiliary distributions are introduced, helping to predict w_i and y_i for new samples, so the variational lower bound becomes

$$\mathcal{F}_{CEVAE} = \mathcal{L} + \sum_{i=1}^{N} (\log q(w_i = w_i^* | x_i^*) + \log q(y_i = y_i^* | x_i^*, w_i^*)), \tag{4}$$

where x_i^*, w_i^*, y_i^* are the observed values for the input, intervention and outcome variables in the training set.

3.2 CEVAE for ecological time series

When analysing ecological time series, one often encounters variables having periodic components such as diurnal and seasonal cycles. This can make it difficult to infer inter-variable causal dependencies as the underlying cycle may be influencing them as well. Synthetic data we use for the evaluation of our method, shown in Figure 2, is generated such that these periodic components act as the hidden confounder. Our task is to infer the causal link intensity between W and Y in the presence of this confounder and detect anomalies induced by the causal link intensity's increase. In contrast to the conventional CEVAE setting, our intervention variable W is a time series. This means we needed to find a different way of intervening on W in order to estimate the desired cause-effect relations. To further put the CEVAE into our context, we adjust several required probability distributions. Namely, we model a conditional distribution of W given Z as defined in (5).

$$p(W|Z) = \mathcal{N}(\mu_w, \sigma_w^2), \quad [\ \mu_w, \sigma_w] \ = f_1(Z) \tag{5}$$

Estimation of this distribution is obtained through the use of the proxy X:

$$q(W|X) = \mathcal{N}(\mu_{\hat{w}}, \sigma_{\hat{w}}^2), \quad [\ \mu_{\hat{w}}, \sigma_{\hat{w}}] \ = f_2(X) \tag{6}$$

Functions f_1 and f_2 are feedforward neural networks with three layers. To measure the intervention effect of W to Y we extend ITE to the case of a sequential intervention and define the Interval Intervention Effect (IIE) and the Average Intervention Effect (AIE):

$$\begin{aligned} IIE(x) := \mathbb{E}(Y | x_i \leq X \leq x_{i+1}, do(W = w^1)) - \\ \mathbb{E}(Y | x_i \leq X \leq x_{i+1}, do(W = w^0)) \end{aligned} \tag{7}$$

$$AIE := \ \mathbb{E}(IIE(x)) \tag{8}$$

In (7) and (8), x_0 and x_i are the interval limits for $i \in \{1, \dots m\}$ and $m = 256$, as we regularly quantize variable X into an 8-bit word, whereas w^1 and w^0 denote intervention or no intervention on W, respectively.

For detecting anomalies in the intervention variable W under hidden confounding, we utilize the increase of the intensity of the causal link from W to Y regardless of its linearity and deploy a sliding window approach documenting the estimated AIE for each window.

3.3 Vector autoregressive Granger Causality

The main assumption of Granger causality (GC) [7] is that causes precede their effects and can be used for their prediction. Let $u_i, i = 1, \cdots, N$ be the time series of N ecological variables. Each time series $u_i(t), t = 1, \cdots, k$ is a realization of length k real valued discrete stationary stochastic process $U_i, i = 1, \cdots, N$. These N time series can be represented by a pth order vector autoregressive model (VAR(p)) of the form

$$\begin{bmatrix} u_1(t) \\ \vdots \\ u_N(t) \end{bmatrix} = \sum_{r=1}^{p} A_r \begin{bmatrix} u_1(t-r) \\ \vdots \\ u_N(t-r) \end{bmatrix} + \begin{bmatrix} \epsilon_1(t) \\ \vdots \\ \epsilon_N(t) \end{bmatrix}. \tag{9}$$

The residuals $\epsilon_i, i = 1, \cdots, N$ form a white noise stationary process with covariance matrix Σ. The model parameters at time lags $r = 1, \cdots, p$ comprise the matrix $A_r = [a_{ij}(r)]_{N \times N}$. Let Σ_j be the covariance matrix of the residual ϵ_j associated to u_j using the model in (9), and let Σ_j^{i-} denote the covariance matrix of this residual after excluding the ith raw and column in A_r. The time domain VAR-GC of u_i on u_j conditioned on all other variables is defined by [6]

$$\gamma_{i \to j} = \ln \frac{|\Sigma_j^{i-}|}{|\Sigma_j|}. \tag{10}$$

4 Experiments

By experiments on synthetic data we first demonstrate that our method is sensitive to the increase of the nonlinear causal link's intensity between the confounded variables, which we then exploit to achieve the second goal of this work, i.e. to apply CEVAE for detecting anomalies. In regard of the neural network architecture, we closely followed [13]. We used feedforward neural networks, namely f_1 and f_2 having 3 hidden layers, with ELU [3] nonlinearity and 200 neurons in each layer. We note, however, that more hidden layers can be used. We modelled variable Z as normally distributed with 20 dimensions, due to its latency. We used a small weight decay term for all parameters, with $\lambda = 0.0001$. For optimization, Adamax [8] was utilized with a learning rate of 0.01. Furthermore, early stopping according to the lower bound on a validation set was performed. For obtaining the outcomes $p(y|x_i \leq X \leq x_{i+1}, do(W = w^1))$ and $p(y|x_i \leq X \leq x_{i+1}, do(W = w^0))$ we averaged over 100 samples from the approximate posterior $q(Z|X) = \sum_w \int q(Z|w, y, X) q(y|w, X) q(w|X) dy$.

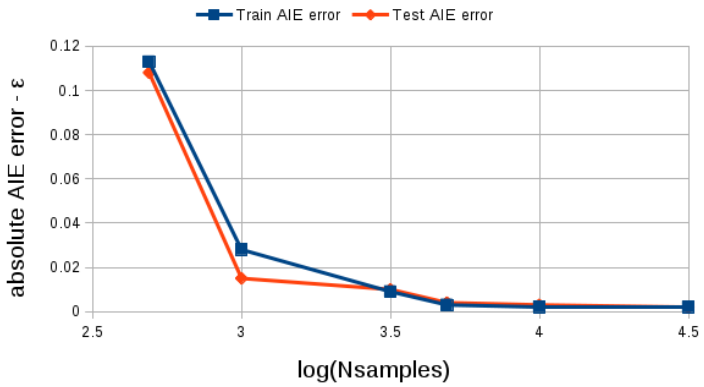


Fig. 3: Absolute error of AIE estimation on synthetic data samples for $\alpha = 1$, $\beta = 0.3$ and quantization level $m = 2^8$. We note that absolute AIE error ϵ is already quite small for the sample size $N = 1000$.

4.1 Synthetic data

The synthetic data was created according to causal relationships of the graphical model in Figure 1, which we consider to be the ground truth. In real data, these causal relationships are extracted from the prior expert knowledge. We create a hidden confounder Z as:

$$Z_t = |b \cdot \cos(\frac{\pi}{2} \cdot \frac{t}{f})|, \text{ for } b, f \in \mathbb{R} \tag{11}$$

where $t \in \{0, \dots N\}$ and N denotes the sample size. It is defined to resemble a periodic component such as daily or seasonal cycle often encountered in ecological time series. Noisy proxy X is defined through shifting Z by a constant $s \in \mathbb{N}$ and the noise level $\beta \in (0, 1)$:

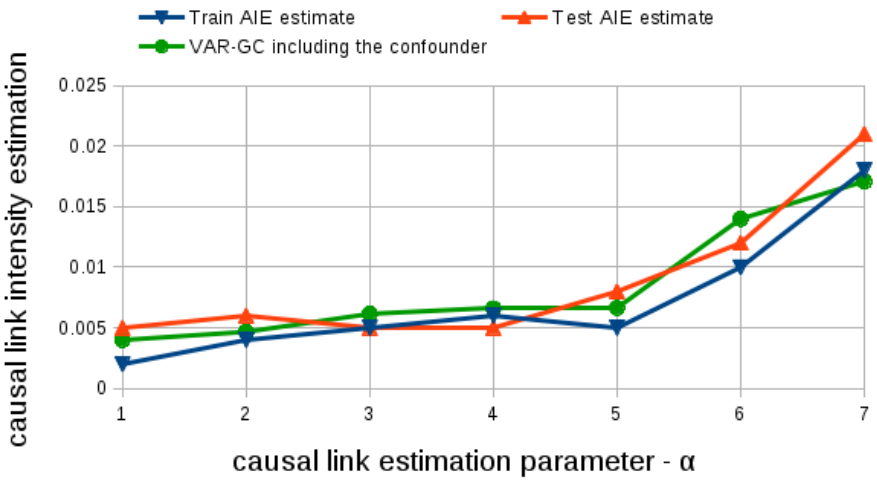
$$X_t = Z_{t-s} + \beta \cdot \epsilon_X, \text{ for } \epsilon_X \sim \mathcal{N}(0, 1). \tag{12}$$

The intervention variable W is modelled to be influenced by the unobserved confounder as follows:

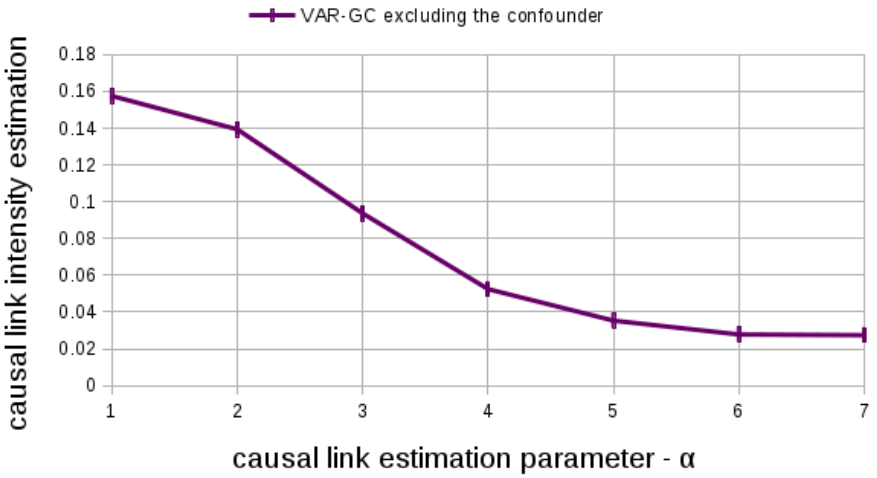
$$W_t \sim \mathcal{N}(\mu_w, e \cdot Z_t), \ e \in (0, 1) \tag{13}$$

As an intervention, we consider values of the intervention variable where values of the proxy's periodic component are less than its half. This type of the intervention was chosen in order for it to satisfy the properties of *do*-calculus. Moreover, it allows for an almost straightforward application of our method to real data. Namely, after intervention, causal link between W and its parent Z should either be removed or so small, that it can be neglected.
The outcome Y is modelled to be influenced by both the hidden confounder Z and the intervention variable W with added Gaussian noise as:

$$Y_t = 0.7 \cdot Z_t + g^{-(\alpha \cdot (W_t - \mu_w) + \mu_w)} + \epsilon_Y, \text{ for } \epsilon_Y \sim \mathcal{N}(0, 0.1), \ g \in (0, 1) \tag{14}$$



(a)



(b)

Fig. 4: (a) Causal link estimation results of our method in comparison to the vector autoregressive Granger causality. Blue and orange curves show our method's estimation of the AIE during training and test, respectively, for $\beta = 0.3$, $m = 2^8$, $s = 500$ and sample size $N = 1000$. Green curve shows results of the VAR-GC when the confounder Z is included. (b) Causal link estimation results of the VAR-GC method when the confounder is hidden. We note that our method performs better than VAR-GC baseline when the confounder is excluded and comparatively well to the VAR-GC with the included confounder.

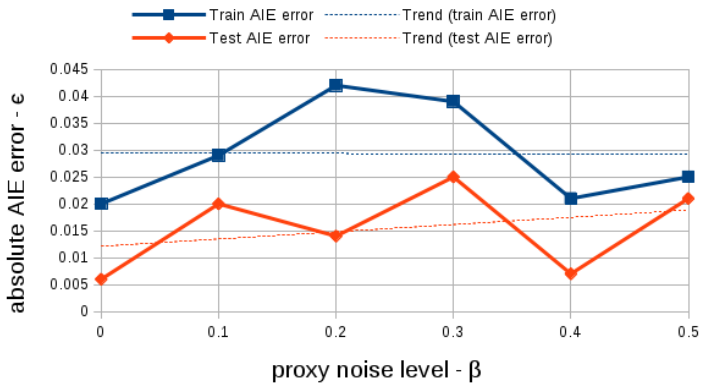


Fig. 5: Proxy noise level vs absolute AIE error ϵ for $\alpha = 2$, $m = 2^8$, $s = 500$ and sample size $N = 1000$. We observe that trend lines of the absolute AIE error for different values of the proxy noise level β are very close to constant.

4.2 Causal link intensity estimation

In the case when $\alpha \in (0, 1)$, perturbations in the variance of W are very small and we consider them neither as the intervention nor as the anomaly. We rather focus on causal link intensity estimation for $\alpha \geq 1$. A relationship between α, the parameter of a function proportional to the causal link's intensity, and the AIE metric's estimation is shown in Figure 4(a). This illustrates our method's sensitivity to the increase of the causal link strength between the confounded variables, predicting its estimation accordingly. To choose the most suitable sample size, we conducted an experiment in which we ran our method for nonlinear causal link intensity estimation with different values of N to choose the most suitable sample size. We used the absolute AIE error

$$\epsilon = |AIE^* - AIE| \tag{15}$$

for measuring our method's accuracy. This was done for each sample size $N \in \{500, 1000, 3000, 5000, 10000, 30000\}$ as seen in Figure 3. Here AIE* denotes the predicted average intervention effect, while its ground truth value is denoted by AIE. Finally, we have chosen $N = 1000$ for our sliding window size.

As the baseline, we applied the VAR-GC method to all four variables $u_1 = W$, $u_2 = Y$, $u_3 = X$ and $u_4 = Z$ from Figure 2 and compared it to our method during training and testing, as shown in Figure 4(a). We note that our method behaves comparatively well to this baseline when the confounding variable Z was included. To determine if VAR-GC can detect the increase of the nonlinear causal link's strength between W and Y without the use of the confounder Z, we excluded it and observed that it is not the case, as shown in Figure 4(b).

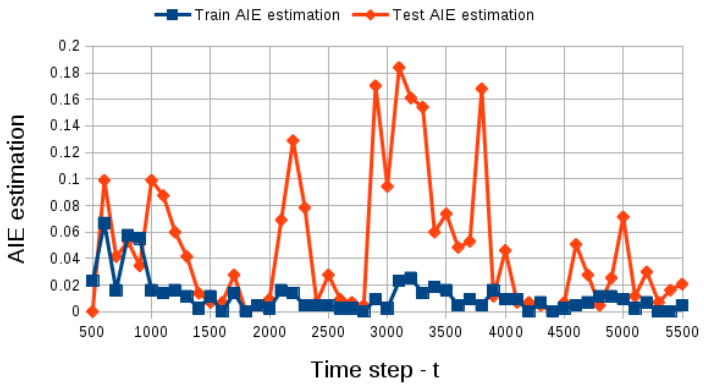


Fig. 6: Anomaly detection for $\alpha = 3$, $\beta = 0.3$, $m = 2^8$, $s = 500$, $N = 6000$ and ground truth anomaly with $\alpha = 7$ positioned at $t \in \{2500, \ldots, 4499\}$. Marked samples on the X-axis denote window centres for each sliding window.

4.3 Proxy noise levels

Since the confounder is unobserved, we wanted to ensure that the proxy we are using is not influencing the causal link between W and Y. To this end, we have performed a proxy noise level experiment as seen in Figure 5. We observe that, on average, changes in the absolute AIE error, as defined in (15), for different values of the proxy noise level β are constant. Therefore, we conclude that the proxy variable X is not influencing the causal link between W and Y. This means that the hidden confounder Z is not causing the link between the confounded variables, but that W is the actual cause of the outcome Y.

4.4 Anomaly detection in synthetic data

Using a sliding window approach and estimating the AIE of W on Y, we propose to detect anomalies in cause-effect relationship intensity between those two variables. We create the anomaly as an increased value of α in (14) by a certain value $a \in \mathbb{N}$ in a particular time interval. More precisely, in this interval the outcome variable becomes:

$$Y_t = 0.7 \cdot Z_t + g^{-((\alpha+a)\cdot(W_t-\mu_w)+\mu_w)} + \epsilon_Y, \text{ for } \epsilon_Y \sim \mathcal{N}(0, 0.1),\ g \in (0, 1) \quad (16)$$

Specifically, in our approach a window consisting of 1000 samples is shifted by 100 in each iteration. We train the adapted CEVAE on time series data, as described in the beginning of this section, using data with an intervention on W i.e. where values of the proxy's periodic component are less than its half, and test on data containing the anomaly. Figure 6 shows our anomaly detection results for the nonlinear coupling between W and Y defined in (14) for $g = 0.8$, $\alpha = 3$ and $\alpha = 7$ for the interval containing the anomaly. The AIE estimation in Figure 6 is

depicted after training and testing, where each value on the x-axis corresponds to each window's centre. Significantly higher AIE intensity from 2200 to about 4000 samples indicates considerable increase in the causal intensity and thus a possible anomaly.

5 Conclusion

In this paper, we have extended CEVAE to ecological time series in order to tackle the problem of nonlinear causal inference of variables in the presence of an unobserved confounder. Furthermore, we have shown that the proxy variable is not influencing the causal link between the confounded variables, meaning that the confounder itself is not the cause of the said link. After successfully establishing our method's sensitivity to increase of the causal link intensity on synthetic data, we utilized its estimates to detect anomalies induced by its increase. We used the VAR-GC method as a baseline and obtained better results when the confounder was hidden. To strengthen our method, we intend to incorporate time-delay embeddings as well as recurrent neural networks for time series anomaly detection.

References

1. Barz, B., Rodner, E., Guanche, Y., Denzler, J.: Detecting regions of maximal divergence for spatio-temporal anomaly detection. In: IEEE Transactions on Pattern Analysis and Machine Intelligence. vol. 41, pp. 1088–1101 (2019)
2. Cardoso Pereira, J.P.: Unsupervised anomaly detection in time series data using deep learning. Master's thesis, Instituto Superior Tecnico Lisboa (2018)
3. Clevert, D.A., Unterthiner, T., Hochreiter, S.: Fast and accurate deep network learning by exponential linear units (elus). arXiv:1511.07289 [cs.LG] (2016)
4. Eichler, M.: Causal inference in time series analysis, pp. 327–354. Wiley Series in Probability and Statistics, John Wiley & Sons, United States (2012). https://doi.org/10.1002/9781119945710.ch22
5. Fabius, O., van Amersfoort, J.R.: Variational recurrent auto-encoders. arXiv:1412.6581v6 [stat.ML] (2014)
6. Geweke, J.: Measurement of linear dependence and feedback between multiple time series. In: Journal of the American statistical association. vol. 77, pp. 304–313 (1982)
7. Granger, C.W.J.: Investigating causal relations by econometric models and cross-spectral methods. Econometrica - Journal of the Econometric Society **37**(3), 424–438 (1969)
8. Kingma, D.P., Adam, J.B.: A method for stochastic optimization. International Conference on Learning Representations (ICLR) (2015)
9. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: Proceedings of the 2nd International Conference on Learning Representations (ICLR), arXiv: 1312.6114 [stat. ML] (2014)
10. Kipf, T., Fetaya, E., Wang, K.C., Welling, M., Zemel, R.: Neural relational inference for interacting systems. In: Internvational Conference on Machine Learning 2018 (ICML), arXiv:1802.04687v2 [stat.ML] (2018)

11. Kretschmer, M., Coumou, D., Donges, J.F., Runge, J.: Using causal effect networks to analyze different arctic drivers of midlatitude winter circulation. Journal of Climate **29**, 4069–4081 (2016)
12. Li, L., Kleinman, K., Gillman, M.W.: A comparison of confounding adjustment methods with an application to early life determinants of childhood obesity. Journal of developmental origins of health and disease **5**(6), 435–447 (2014)
13. Louizos, C., Shalit, U., Mooij, J., Sontag, D., Z., R., Welling, M.: Causal effect inference with deep latent-variable models. In: Advances in Neural Information Processing Systems 30. pp. 6446–6456 (2017)
14. Miao, W., Geng, Z., Tchetgen Tchetgen, E.: Identifying causal effects with proxy variables of an unmeasured confounder. In: arXiv preprint arXiv:1609.08816 (2016)
15. Pearl, J.: Causality. Cambridge University Press (2009)
16. Qiu, H., Liu, Y., Subrahmanya, N.A., Li, W.: Granger causality for time-series anomaly detection. In: 2012 IEEE 12th International Conference on Data Mining. pp. 1074–1079 (Dec 2012). https://doi.org/10.1109/ICDM.2012.73
17. Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., Prabhat: Deep learning and process understanding for data-driven earth system science. Nature **566**, 195–204 (2019)
18. Runge, J.: Causal network reconstruction from time series: From theoretical assumptions to practical estimation. Chaos: An Interdisciplinary Journal of Nonlinear Science **28**(7), 075310 (2018)
19. Shadaydeh, M., Denzler, J., Guanche, Y., Mahecha, M.: Time-frequency causal inference uncovers anomalous events in environmental systems. GCPR (2019)
20. Shadaydeh, M., Guanche, Y., Mahecha, M., Denzler, J.: Baci deliverable 5.4: Methods for attribution scheme and near real-time baci. Tech. rep. (2018), available online at: `http://baci-h2020.eu/index.php/Outreach/Deliverables`
21. Shadaydeh, M., Guanche, Y., Mahecha, M., Reichstein, M., Denzler, J.: Causality analysis of ecological time series: a time-frequency approach. In: Climate Informatics Workshop 2018 (2018)
22. Shalit, U., Johansson, F., Sontag, D.: Estimating individual treatment effect: generalization bounds and algorithms. In: arXiv:1606.03976v5 [stat.ML] (2016)
23. Simpson, E.H.: The interpretation of interaction in contingency tables. pp. 238–241. No. 13 in B (1951)
24. Stips, A., Macias, D., Coughlan, C., Gracia-Gorriz, E., Liang, X.S.: On the causal structure between co_2 and global temperature. Sci. Rep. **6**(21691) (2016). https://doi.org/10.1038/srep21691
25. Villani, C.: The wasserstein distances. Springer Berlin Heidelberg pp. 93–111 (2009)