Valve Status Verification and Sensor Error Detection via Causal Inference from Sensor Data

Dirk Vries[1,*], Joost van Summeren[1]

[1] KWR Watercycle Research Institute, Groningenhaven 7, 3433PE Nieuwegein, The Netherlands

[*] *dirk.vries@kwrwater.nl*

ABSTRACT

Recent developments in (near) real-time sensor applications have the potential to provide operators and managers with useful information on drinking water distribution supply and need of its maintenance. A systematic methodology based on causal inference from observational data is proposed to increase knowledge of water supply distribution systems equipped with sensor networks. This methodology can be used to help identify deviations from expected operation of water supply and sensor infrastructure, using only observational data. We outline the first steps of two distinct procedures that use data from a sensor network, to infer a map of a causal dependence structure. These procedures are applied to scenario studies where an unexpected change in operation occurs, i.e. a valve status is different and a sensor bias is introduced. A draft outline of future steps is given that could improve and validate the methodology.

Keywords: graph modelling, sensor networks, distribution model

1 BACKGROUND

Recent developments in (near) real-time sensor applications have the potential to provide operators and managers with useful information on drinking water distribution supply and need of maintenance. Although implementation of sensor networks is not yet common practice, numerous numerical studies have demonstrated potential benefits of a sensor network, such as real time event detection of water quality contaminations (e.g. [1]) or leakage and pipe burst detection and localization (e.g. [2]). We also foresee that sensor networks will provide operational benefits such as improving distribution network models and the effectiveness of sensor networks. Automated monitoring and control of water supply services using sensor data and models imply a strong reliability on sensor data and network models. This reliance poses a challenge because knowledge of distribution networks is not always correct, comprehensive, and up-to-date and sensors are known to be imperfect (false positives and negatives, drift and failure, etc.). A systematic methodology to increase actual knowledge of the systems can help identify deviations from expected operation of the drinking water distribution and sensor infrastructure.

A novel method investigated in this work is aimed at quantifying operational benefits using a heuristic approach and testing the methodology with a laboratory scale model of a real-life distribution network. In this paper, we focus on the development of a graph theoretical procedure aimed at improving the quality of system information and models that rely on such information. We outline the first steps of two distinct procedures to use only observational data, i.e. data from a sensor network, to infer a map of a causal dependence structure. We test these steps and give a draft outline of the remaining steps that could improve the methodology in the identification of deviations from expected operation in water supply networks. In order to provide tangible and

quantified benefits of a sensor network we narrow the work down to two practical applications, i.e. (C1) detection of changed valve statuses and (C2) detection of erroneous sensor measurements.

2 METHODS

The information flow of a sensor network provides the basis of our procedure to infer a graph model for a baseline situation, i.e. where the operation of the drinking water distribution network (DWDN) is assumed normal. It is our hypothesis that any deviations with respect to this baseline case (BC) occurring from sensor faults, leakage or changed valve status values, etc. is revealed as a change in the newly estimated graph. Hence the estimated information flow is presented as a graph, i.e. each node represents a (sensor) variable such as flow, pressure, or electrical conductivity, and each edge represents a correlation (undirected graph) or a direct cause between nodes (directed acyclic graph, DAG). It is assumed that there are no feedback loops, hence acyclic, or hidden variables. We implement and evaluate two notions of causality inference from synthetic DWDN data by the following steps:

1. a framework for graph theoretical analysis is set up and written in the Python programming language to determine if there are causal relationships in the sensor network (Figure 1). The framework consists of calls to the R statistical software package (via python module rpy2), calls to EPANET (epanettools), calls to statistical tests (stattools) and methods to construct and visualise the sensor and DWDN via the python module networkx.

   The framework enables testing and evaluating two procedures (Figure 2):

   - a graph theoretic methodology that uses the Peter and Clark (PC)-algorithm directly onto sensor data [3,4]. Conceptually, the algorithm starts with a complete, undirected graph between each node (here: each sensor signal) and recursively deletes edges based on Markov conditional independence tests until a minimal set of connected nodes is reached. The R package pcalg provides pre-defined functions to perform these independence tests on Gaussian, discrete or binary data and discover directed (acyclic) graphs. The PC algorithm is evaluated on data which is pre-processed to remove time lags. These time lags are (manually) estimated on the basis of delays in step response signals. This procedure will be referred to as PPC;
   - a Monte Carlo (MC) analysis of Granger causality tests and detected time lags for every sensor couple in the set of sensors (nodes) using the sensor data. Granger causality means that 'X Granger-causes Y, if Y can be better predicted using the histories of both X and Y than it can by using the history of Y alone'. The same data as in the PPC procedure is used, but now the lag estimation is part of the causality tests. Based on a set of estimated lags and causal relations between nodes, a subset of most likely occurring lags (and thus Granger-causal relations) are selected to draw a directed graph. This will be further called the PMC procedure.

2. a DWDN model (EPANET) was constructed with a layout and sensor network depicted in Figure 2. The DWDN represents an experimental scale model of a real-life supply zone [5]. It includes two main water supply sources with different water quality in the West (Noard-Burgum, NB) and South (Wirdum) and the transport and large distribution mains (of real-life diameters of >300 mm). Demands are represented by 31 demand nodes. A tank is included, but is not actively taken into account in the calculations. Three sensors (X28, X42,

X32) are placed along a North-side transport main of eastward water flow from the water supply source in the West, one sensor (X280) is placed near the centre of the DWDN, one (X335) near the water supply source in the South and at the Eastside (X113). All sensors measure water quality, i.e. the concentration of a chemical tracer.

3. We define and evaluate test conditions (variation in signals, number of sensors) and define the baseline scenario BC and case scenarios C1 and C2:
    - BC: 1 week and 4 weeks of tracer data is simulated by EPANET with a 15 minutes sampling frequency to check the sensitivity of the method to the amount of available data. Repeating weekly demand patterns are set. A chemical tracer is supplied at NB with an average concentration value of 3.0 and perturbed with Gaussian noise with standard deviation 0.1. Similarly, the water supply at Wirdum contains a tracer with an average concentration of 1.0 and a Gaussian noise perturbation with a standard deviation of 0.1.
    - C1: similar to BC with 4 weeks of data, except that one valve, i.e. valve V52 (Figure 2) is closed.
    - C2: similar to BC with 4 weeks of data, except that one sensor, i.e. sensor X42 (Figure 2) has a bias of +1.0 during a period of 168 hours from the start of the simulation.
4. Simulation tests of cases C1 and C2 for an EPANET model of a water supply distribution network (DWDN) in order to determine the performance and applicability of the procedures.

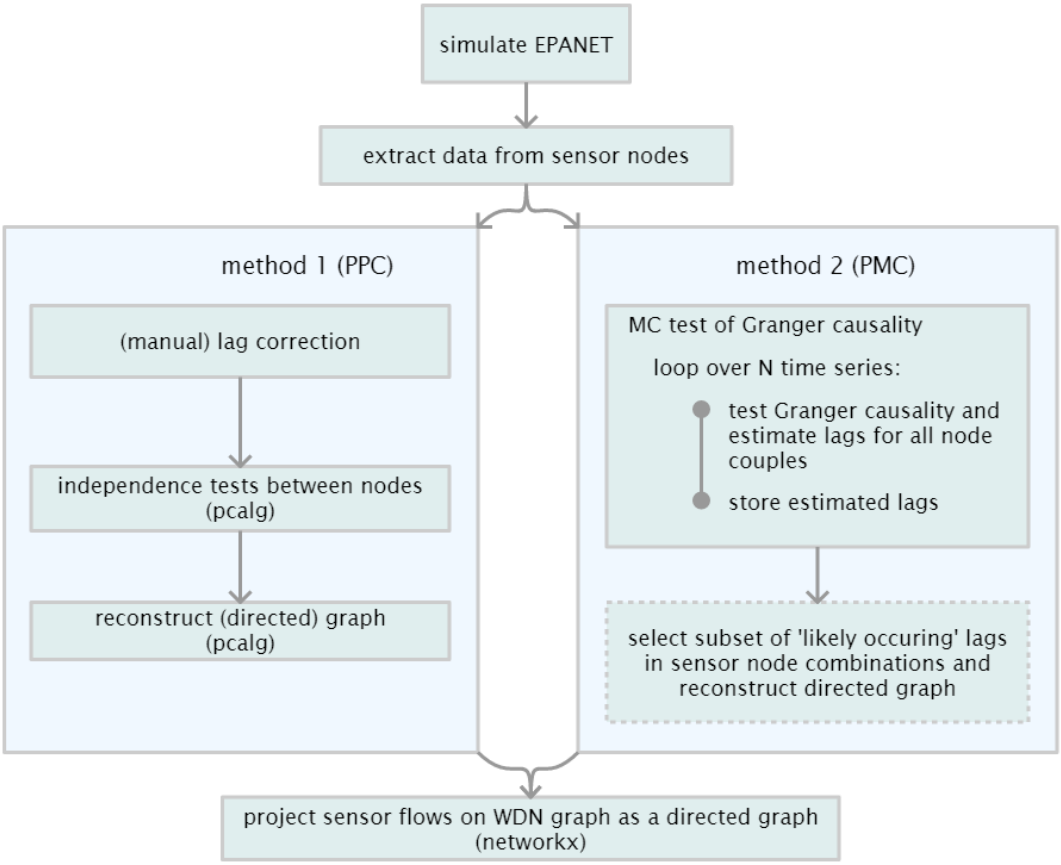

Figure 1. Scheme of followed procedures to estimate a causal structure between sensor nodes in a DWDS.

Figure 2. Layout of the EPANET scale model. Pipe widths are plotted on a gray scale, flow directions are shown by arrows, and the concentration of tracer chemicals are shown by the purple colors. Sensors are placed at nodes X<numeric value>. At Wirdum, the average tracer concentration is 1.0, at Noardburgum the average concentration value is 3.0. The valve used in scenario C1 is shown by the blue symbol labeled V52.

3 RESULTS

3.1 PMC Method

The results with the PMC method vary to a large extent. While the results were promising for a small, academic example with 4 nodes and Gaussian noise; the method does not work well with (sensor) data simulated with EPANET. Lags are only estimated between X42 pointing towards X28 (lag: 3 hours) and X32 pointing towards X28 (lag: 1.45h) and are off from the estimated lags deduced from step responses and the causal effect is exactly in the opposite direction (X28 to X42: 3.75 hours).

3.2 PPC Method

The PPC procedure relies on the R library 'pcalg' to infer causal maps (directed graphs). Results are shown in Figure 3. When no causal relation between two nodes is found, no edge is drawn. The conditional independence tests are run with an uncertainty threshold of 5% (orange lines). Results of the simulated baseline scenario (BC) are shown for the case where water quality data during a period of 1 week is available (Figure 3a) and a period of 4 weeks (Figure 3b). Based on the simulation in EPANET (Figure 2), we expect that information flows from Wirdum (no sensor present) via X335 to X113 and most of the time to X280, while information from reservoir

Noardburgum (NB) will flow via X28 towards X42, passing V52 towards X32 and possibly from X42 or X32 to X280.

The BC lag corrected case when using *1 week* of chemical tracer data is shown in Figure 3a. In this graph, the algorithm finds that X335 is correlated to X113, but no causality is found. (Partial) correlation is also resolved between nodes X28 – X42. No (causal) relation between X42 – X32 is found. When a *4 week* period of data is available, the graph looks different (Figure 3b). The edge X42 – X32 is now added, and the information flow is correctly resolved towards X32. Furthermore, X32 is apparently effected by X280. The direction in causality is remarkable, because from the EPANET simulations we know that X280 is lagging 8.75 hours w.r.t. the NB reservoir, while X32 has a delay of 7.75 hours. Note that no correlation or causality could be inferred from data of X335 and X113.

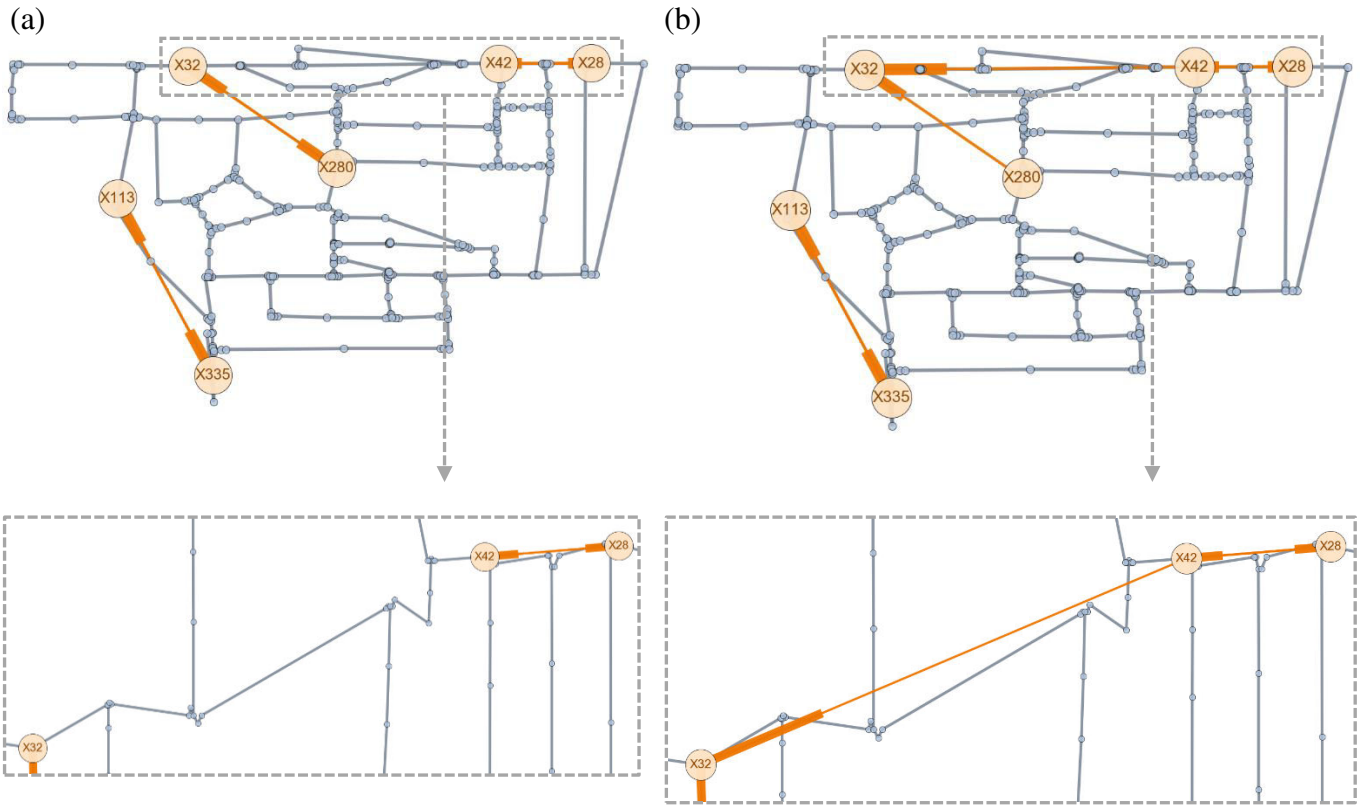

Figure 3: DAG representation of the sensor network observations (orange coloured edges, a stub represents an arrowhead) projected onto the scale model (blue) by PPC for the BC scenario. DAG reconstruction by (a) 1 week and (b) 4 weeks of data , respectively. The lower panels zoom in on the region of the Northern trajectory (X28 > X42 > X32).

Figure 4 shows the results of the PPC procedure applied to the practical cases of an unexpected valve status (C1) and sensor failure (C2). For the case C1, the graph shows that information is obstructed between the sensors in the direct vicinity of the valve (X42, X32, X280). Compared to Figure 2b, the procedure indicates an interrupted flow of information in the vicinity of the valve that was closed in this simulation (X42 – X32 and X32 – X280), while the other connections and some of the directions remain unchanged. The inferred graph of scenario C2 in Figure 4 also differs from the BC graph (Figure 3b), but now in the node set X28, X42, X32 and X280. Apparently, the introduced bias in sensor X42 leads to a re-shuffling of links: X42 now has a causal effect on X28

and X28 has an effect on X32. It should be noted that the perturbed signal of X42 is not Gaussian anymore, which violates an important assumption for the independence tests of the PC algorithm. These preliminary results must be substantiated with further tests, but suggest that the followed approach can be used to detect deviations from expected operation, even with a limited number of sensors.

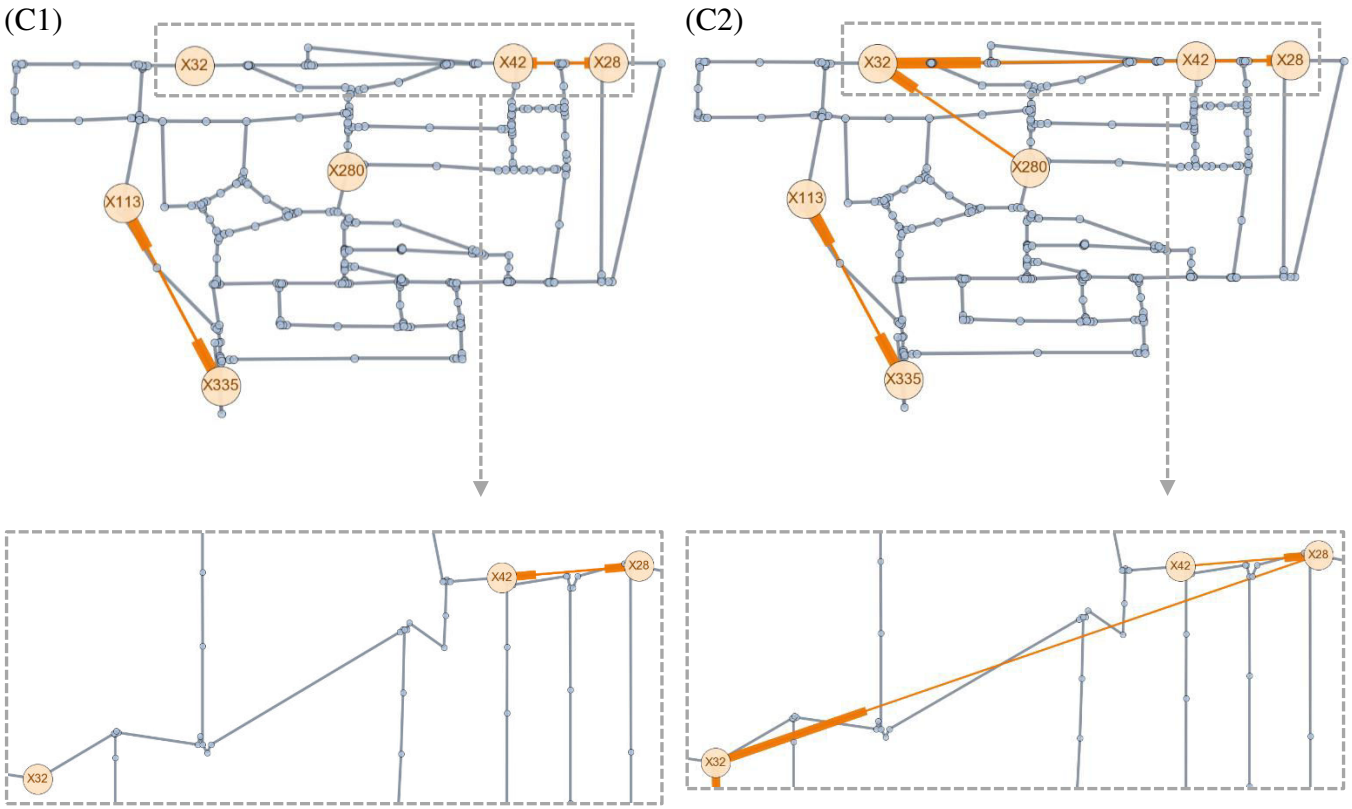

Figure 4: graph notation as Figure 3, for practical applications of an unexpected valve status (C1) and erroneous sensor measurements (C2). The lower panels show the zoomed in region of the Northern trajectory (X28 > X42 > X32).

4 DISCUSSION AND FUTURE STEPS

While the PC algorithm as introduced by Peter and Clark is - by definition - not equipped for lagged signals, a drawback in applying the PMC method to the presented application is that it relies on a heuristic approach that searches for a high probability in connectivity based on time lag information between node couples, but not for the whole network. As a consequence, there is a significant risk of introducing inconsistencies in the resulting graph with respect to the sum of estimated lags if interconnected nodes are present. Furthermore, it is known that Granger causality tests should be applied to data that is stationary, and, possible co-integration should be corrected first. However, in this work, stationarity is not guaranteed due to several reasons, e.g. possible mixing of water, changing flow directions during the course of a day or when bias occurs (as is the case for C2).

Based on experiences with both techniques, we propose to pursue the following steps to improve the method:

- Check whether the use of other data (flow, pressure or other water quality data) reduces the amount of data needed to capture the information flow.

- Extend the PMC procedure with (i) a check whether the data is stationary, semi-stationary over a time window, or has an order of integration with the augmented Dickey–Fuller test and proceed with (ii) the estimation of a VAR (vector autoregressive regression) using the procedure as outlined in [6] and (iii) check for Granger causality.
- For PPC, there are different options to cope with lagged sensor signals:
    i. With the assumption that the unlagged (source) signal is the only cause for the lagged signal and there is no co-integration or autoregression present, time delays can be estimated by minimisation of the squared difference of the unlagged (source) signal and lagged signal [7]. This rather straightforward approach resembles the work here, except that the time delays are now estimated from data.
    ii. As a first step, it is assumed that all sensor signals can be modelled by a structured VAR (SVAR) process, i.e. $A(L)Y_t = E_t$, where Y_t is a n x 1 column vector of n sensor variables at time step t; A is a n x n conformable matrix whose terms are polynomials in a fixed lag value L, and E is a column vector of errors at time t. The idea is to generate SVAR models by a Monte Carlo approach with N realisations of A, estimate the SVAR and use the residuals (filtered Y_t minus Y_t) as an input for the PC algorithm in each realisation. Note that N is typically in the order of 10^4 to 10^5. Then, the selected graph is compared to a reference graph ('PC true graph'), i.e. the graph that indexes the equivalence class to which the true graph belongs. Finally, every possible link is evaluated. See [8] for more details.

In addition, we plan to apply the technique to a real-life scale model of a distribution network from the Dutch water company Vitens and evaluate the results to relate this new information to tangible benefits. The use of a scale model (as opposed to numerical simulations) allows for testing the methodology under (close-to) real-life circumstances, including realistic imperfections of sensors, drinking water and pipes in a controlled environment. Another research subject is to address the optimal sensor placement problem based on maximising causal inference by the PMC or PPC method.

5 CONCLUDING REMARKS

- A Monte Carlo approach for testing Granger causality between any sensor pair followed by selection of the most frequently occurring lags did not yield satisfactory results. Pre-processing, lag consistency checks and further tests are needed to validate whether this approach holds promise.
- The PC algorithm seems promising to infer (changes in) causalities from sensor network data of a DWDN, at least in a simulated environment where source patterns of water quality are defined as signals with Gaussian noise and the observation output signals at the sensor node positions are corrected for time lags.
- The PC algorithm is sensitive to the 'information content' of the signal, or more generally speaking, whether the system excitation was sufficient. Part of the causality in the sensor network was not resolved when either the signal duration was too short or when the standard deviation in the noise sequence was relatively small.

Acknowledgements

This research was conducted within a project of the Joint Research Program (BTO) of the Dutch Water Companies. We acknowledge the valuable comments by Mirjam Blokker (KWR) on the manuscript, the support of Vitens to use the scale model of Leeuwarden for experiments and fruitful discussions with Peter van Thienen (KWR).

References

[1] Zhao, Y., Schwartz, R., Salomons, E., Ostfeld, A., Vincent Poor, H. New formulation and optimization methods for water sensor placement (2016). Environmental Modelling & Software, 76 Issue C, 128-136.

[2] Mounce, S. R., Mounce, R. B., Jackson, J., Boxall, J.B. Pattern matching and associative artificial neural networks for water distribution system time series data analysis (2014). Journal of Hydroinformatics, 16.3, 617-632.

[3] Spirtes, P. and G. Clark. An algorithm for fast recovery of sparse causal graphs. Social science computer review 9.1 (1991): 62-72.

[4] Kalisch, M. and P. Bühlmann. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. Journal of Machine Learning Research 8.Mar (2007): 613-636.

[5] Van Summeren, J., Meijering, S., Beverloo, H. and Van Thienen, P. Design of a distribution network scale model for monitoring drinking water quality (2017). Journal of Water Resources Plannning and Management, 143(9), 1-10.

[6] Toda, H.Y., and Yamamoto, T. Statistical inference in vector autoregressions with possibly integrated processes (1995). Journal of econometrics 66.1: 225-250.

[7] Giovanni, J and Scarano, G. Discrete time techniques for time delay estimation (1993). *IEEE Transactions on signal processing* 41.2: 525-533.

[8] Demiralp, S., and Hoover, K.D.. Searching for the causal structure of a vector autoregression (2003). *Oxford Bulletin of Economics and statistics* 65: 745-767.