УДК 004.048:519.7 https://doi.org/10.15407/pp2022.03-04.183

ВІД ТЕМПОРАЛЬНИХ ДАНИХ ДО ДИНАМІЧНИХ КАУЗАЛЬНИХ МОДЕЛЕЙ

Олександр Балабанов

Доповідь присвячена огляду проблем виведення динамічних каузальних моделей з емпіричних даних, з акцентом на моделі векторних авторегресійних процесів. Типізовано і охарактеризовано схеми збору, архітектуру і форми репрезентації темпоральних даних та часових рядів даних. Показано, що вимоги до архітектури темпоральних даних диктуються характером динамічного процесу і потребами виведення адекватної моделі. Виділено основні типи динамічних процесів, зокрема, рекурентні процеси, «запрограмовані» процеси, потоки стохастичних подій, лічильні процеси тощо. Для характеристики довжини темпоральних даних запропоновано кілька часових горизонтів, визначених на основі структури моделі, лагу каузальних зв'язків процесу, довжини шляхів впливу, довжини зворотних зв'язків тощо. Частота вимірювання даних критично важлива для адекватності виведеної моделі і визначається тривалістю елементарних впливів між компонентами векторного процесу і швидкістю дії зворотних зв'язків. Відображено роль припущень у виведенні динамічної моделі з даних, зокрема, припущень стаціонарності та регулярності структури. Виділено особливості виведення динамічних каузальних моделей (у порівнянні із статичними моделями), окреслено тактику врахування темпорального порядку змінних. Проаналізовано проблеми, пов'язані з невідомим лагом післядії та існуванням прихованих автокорельованих часових рядів. Представлено концепцію каузальності за Грейнджером і вказано на її недосконаліть в реальних умовах неповноти інформації. Проведено порівняльний аналіз критерію каузальності за Грейнджером та правил орієнтації ребер в апараті каузальних мереж з точки зору їх спроможності виявляти каузальні відношення.

Ключові слова: каузальні моделі, динамічні процеси, часові ряди даних, умовна незалежність, каузальність за Грейнджером, каузальні зв'язки.

We present a brief review of dynamic causal model inference from data. A vector autoregressive models is of our prime interest. The architecture, representation and schemes of measurement of temporal data and time series data are outlined. We argue that requirement to data characteristics should come from the nature of dynamic process at hand and goals of model inference. To describe and evaluate temporal data one may use terms of longitude, measurement frequency etc. Data measurement frequency is crucial factor in order to an inferred model be adequate. Data longitude and observation session duration may be expressed via several temporal horizons, such as closest horizon, 2-step horizon, influence attainability horizon, oscillatory horizon, and evolutionary horizon. To justify a dynamic causal model inference from data, analyst needs to assume the dynamic process is stationary or at least obeys structural regularity. The main specificity of task of dynamic causal model inference is known temporal order of variables and certain structural regularity. If maximal lag of influence is unknown, inference of dynamic causal model faces additional problems. We examine the Granger's causality concept and outline its deficiency in real circumstances. It is argued that Granger causality is incorrect as practical tool of causal discovery. In contrast, certain rules of edge orientation (included in known constraint-based algorithms of model inference) can reveal unconfounded causal relationship.

Keywords: causal model, dynamic process, time series, conditional independence, Granger causality, causal relationship.

1. Емпіричні каузальні моделі. Динамічні моделі

Каузальні моделі створюються з метою адекватно відображати і пояснювати впливи та взаємодії між підпроцесами (або подіями) в середовищі. На відміну від суто описової (феноменологічної) моделі, каузальна модель має бути не тільки узгоджена із зібраними емпіричними даними, але й адекватна в розумінні здатності прогнозувати наслідки управління об'єктами (середовищем), тобто спроможна прогнозувати ефекти втручань в об'єкт. Традиційно каузальні моделі створювалися зусиллями прикладних експертів та математиків. Водночас каузальний ефект визначали на основі активних рандомізованих експериментів. З настанням епохи Великих Даних аналітики змінили постановки задач, і тепер визначальну роль відіграє аналіз емпіричних даних спостережень (а експертні знання та ненадійні припущення залучаються обережно) [1, 2]. Протягом останніх трьох десятиліть увага дослідників зсунулася до моделей типу каузальних мереж. Емпірична каузальна мережа має структурно описувати систему зв'язків та каузальних відношень на такому рівні деталізації й точності, який відповідає точності, повноті та обсягу використаних даних. Каузальні мережі є багатоцільовими предиктивними та генеративними моделями [2–7]. Вони дозволяють оперативно розв'язувати аналітичні задачі у бажаному форматі (обираючи цільову змінну без потреби повторно виводити модель). Динамічні каузальні мережі розширюють можливості стандартних каузальних мереж. Доповідь доповнює огляд аналітики Великих Даних (поданий в [2]) в аспекті темпоральних даних та виведення динамічних моделей.

Каузальна мережа («статична») задається як пара (G,Θ), де G – орграф, Θ – параметри (прив'язані до G). Вершини графу відповідають змінним, а ребра – безпосереднім статистичним (каузальним) зв'язкам. Зв'язок $X \to Y$ поєднує причину X та наслідок Y. Параметри Θ кількісно характеризують зв'язки. Зазвичай орграф G – ациклонний, тобто не містить строго орієнтованих циклів. Тоді набір каузальних зв'язків моделі накладає відповідні обмеження на темпоральний порядок змінних (відносну послідовність змінних у часі). Наприклад, якщо в моделі є орієнтований шлях $X \to Y \to \cdots \to Z$, то Z не може стояти в порядку раніше, ніж X. Можна задати темпоральний порядок змінних як знання (апріорні, предметні). Темпоральний порядок може бути частковим або повним. Якщо на вході не задано темпорального порядку змінних, то

модель, виведена з даних, складається переважно з неорієнтованих та неповністю орієнтованих ребер. Тоді темпоральний порядок змінних залишається майже невизначеним.

Каузальна мережа з темпоральним порядком змінних ще не є динамічною каузальною моделлю. Динамічні каузальні мережі характеризуються поглибленою часовою прив'язкою змінних моделі. За стандартної репрезентації каузальна мережа («статична») утворюється з різних змінних, пов'язаних статистичними залежностями. Зазвичай кожна змінна відображає окремий показник (підпроцес, компоненту), виміряний в «характерний момент» розвитку процесу, причому інтервали часу між виміряними змінними можуть залишатися не визначеними. Перший крок до динамічного відображення – це приписування тривалості інтервалам часу між змінними (подіями). Щоб модель набула повноцінного динамічного характеру, замість кожної окремої («великої», «титульної») змінної в моделі представляють серію «моментальних» змінних з мітками часу, і кожна серія репрезентує послідовні стани (фази) відповідного підпроцесу (компоненти). Тобто змінна «розщеплюється» на фази (розгортається).

Дані, з яких виводиться стандартна («статична») каузальна мережа, складаються з незалежних випадків (прецедентів). Для того, щоб стандартна модель могла описати процес переходу між фазами одного і того самого підпроцесу (між однойменними змінними в різні моменти), значення цих змінних мають міститися в межах кожного запису даних (в одному й тому випадку). Залежності між записами даних навіть не розглядаються в ході виведення «статичних» каузальних мереж. Натомість дані, призначені для виведення моделі динамічного процесу, організовані так, що послідовні записи описують той самий процес на послідовних фазах розвитку. Тоді розбиття даних на записи стає несуттєвим, технічним питанням програміста. В ході виведення моделі аналізуються залежності між змінними з кількох записів даних одночасно. Завдяки часовому виміру динамічні моделі отримують можливість оперувати таким поняттями, як лаг післядії, швидкість зміни, періодичність, спектр, форма сигналу і т.д. З'являються підстави розділяти «корисний сигнал» та «гамір» у єдиному потоці даних. Саме динамічні каузальні мережі здатні коректно описати функціонування системи із зворотними зв'язками. У науковій літературі сформовано кілька математичних репрезентацій для опису поведінки динамічних об'єктів. Найбільш відомі репрезентації: марковський ланцюг; дискретний процес у неперервному часі; неперервний процес у неперервному часі; дифузний процес; стохастичний точковий процес; системи диференційних рівнянь (звичайних та стохастичних) тощо [8–12].

2. Динамічні процеси та архітектура темпоральних даних

Дані для виведення «статичної» каузальної мережі вимірюються таким чином, щоб зафіксувати потрібні характеристики в «характерні моменти» процесу. Такі «характерні моменти» обирає фахівець (або на етапі збору даних, або на етапі попередньої обробки). Зазвичай обрати такі моменти легко, наприклад, коли потрібні характеристики змінюються дуже повільно або швидко досягають стабільного стану (після завершення перехідного процесу). Образно кажучи, дані збираються «ощадливо» через підхожий «трафарет». Натомість коли йдеться про виведення моделей динамічних процесів, доводиться збирати дані більш інтенсивно, регулярно й систематично. Потрібно враховувати характер поведінки процесів, для яких хочемо вивести модель.

Для того, щоб надати уявлення про різноманіття динамічних процесів, доцільно виділити кілька архетипів об'єктів моделювання, які потребують принципово різних підходів до аналізу. Можна класифікувати темпоральні дані згідно природи генераторів даних. В [2] виокремлено наступні класи генераторів: природні процеси; ергатичні системи (або цілеспрямовані об'єкти); бібліотеки «паттернів», послідовностей, сигналів. В контексті даної доповіді доцільно виділити декілька кластерів динамічних процесів з точки зору парадигми їх розвитку (генерації). Тоді, по-перше, виділяємо «затемнені», чи «запрограмовані», процеси, програма яких невідома. «Затемнені» процеси повністю керуються невідомими й незрозумілими зовнішніми (латентними) драйверами. Більш того, дані часто являють собою суміш «затемнених» процесів (які «запрограмовані» по різному). Легко зрозуміти, що спостереження за такими процесами дають недостатньо інформації для виділення закономірностей, пояснення механізму їх розвитку чи вироблення прогнозів. По-друге, можна уявити деякі повністю автономні процеси, ізольовані від зовнішнього світу. Динаміка кожного такого процесу є результатом саморозвитку. Поведінка різних ізольованих процесів може варіювати від дуже простої й прогнозованої до хаотичної. Можливості вивести адекватну модель подібного процесу – доволі обмежені, але кращі, ніж для «затемнених» процесів. В центрі уваги доповіді знаходяться «типові» процеси, які розвиваються згідно стабільної структури зв'язків між компонентами процесу та згідно «інерційних» залежностей. Припускаємо, що типові процеси спостерігаються достатньо систематично (хоча й не повністю), і зазнають впливу зовнішніх факторів (більшість з яких спостерігається). Отже, «типові» процеси розвиваються закономірно, так що можливо емпірично вивести доволі адекватну модель генерації даних. Виведення моделей вказаних процесів спирається на аналіз післядії минулих станів і ефектів впливу спостережуваних зовнішніх змінних (факторів). Для детального аналізу оберемо звужений клас процесів, а саме – рекурентні процеси з дійсними змінними. Для таких процесів переважно використовуються моделі авторегресійного типу. Для опису і аналізу таких процесів релевантним апаратом є марковські властивості та умовні незалежності. До цих питань повернемося у наступному розділі.

Початкове уявлення про архітектуру темпоральних даних дає схематизація, відображена на рис. 1. Дотримуємося загальноприйнятого принципу: одна модель – один об'єкт (середовище). Виділяємо два роди об'єктів моделювання: 1) один екземпляр (складної) системи; 2) популяція багатьох екземплярів,

які різняться несуттєво і тому описуються спільною моделлю. Множина екземплярів популяції породжує множину спостережень (записів) даних для аналізу. Повторюваність спостережень необхідна не тільки для статистичних методів аналізу й виведення моделі, але також і для евристичних методів, що відносяться до самонавчання алгоритмів [2]. Повторення спостережень може розгортатися у часі або у «просторі». Повторюваність спостережень у просторі – це багато екземплярів популяції. (Зазвичай прийнято, що екземпляри популяції є автономними і в певному сенсі взаємонезалежними. В разі, якщо доступний тільки один екземпляр об'єкту, потрібний обсяг даних створюється багатократними повтореннями спостережень за цим екземпляром (у часі). Як сурогатний спосіб повторень у просторі можна застосувати повторення у часі у спеціальному режимі, з багатократним перезапуском функціонування об'єкту від стартових умов. Згідно схеми повторюваності спостережень отримуємо початкову класифікацію архітектури даних, як показано на рис. 1. З чотирьох схем (типів архітектури), відображених на рис. 1, практично типовими є дві: а) багато (коротких) рядів даних; б) один довгий (тривалий) ряд даних. Вважається, що в обох цих схемах дані генеруються єдиною моделлю. Втім, у ситуації «б» можливий випадок, коли схема генерації даних повільно змінюється. З іншого боку, в деяких випадках ситуації «а» дані навряд чи доцільно називати темпоральними. Зазначимо, що на рис. 1 наочні зображення наводять на думку про регулярний формат даних. Насправді дані можуть збиратися в нерегулярних форматах. Наприклад, дані можуть бути виміряні «спорадично», згідно «трафарету», за схемою «вікно з маскою». Тоді кожна змінна вимірюється у свій відповідний момент (зазвичай зсунутий відносно вимірювання інших змінних).

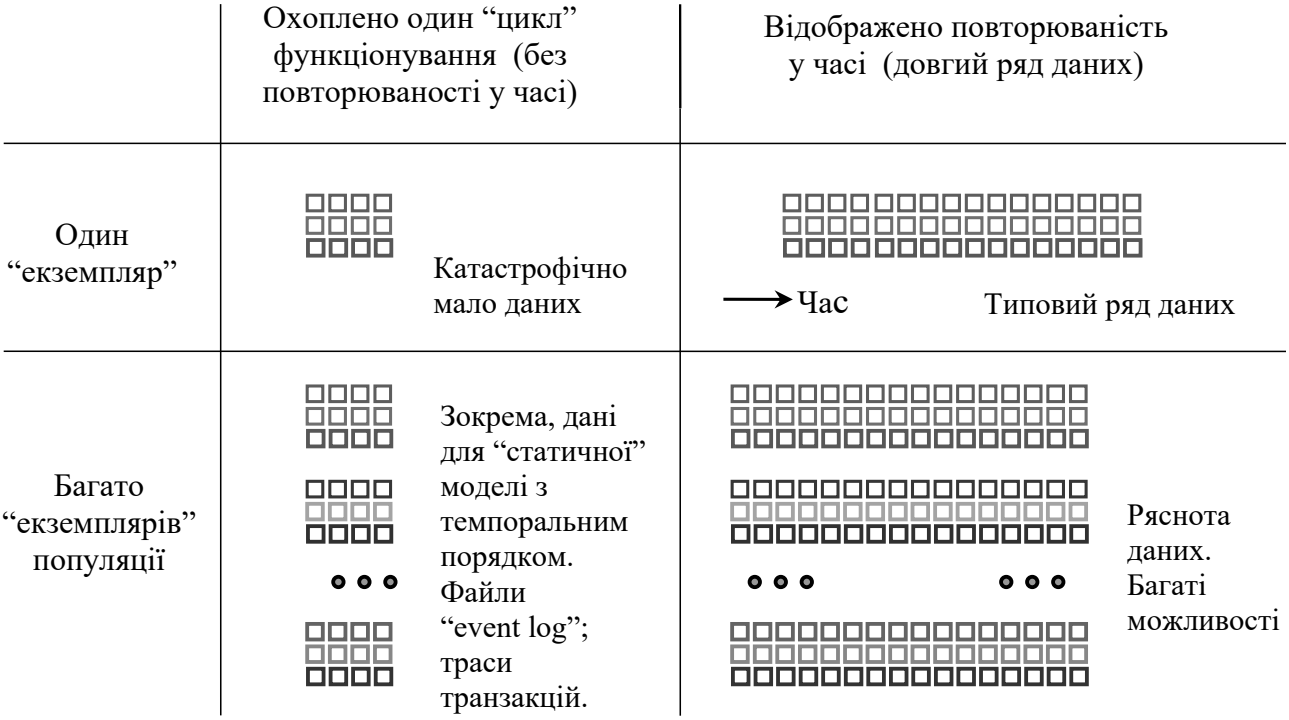

Рис. 1. Архітектура темпоральних даних (повторюваність у часі та у просторі)

Якщо для повторюваності спостережень обрано багатократне виконання циклу функціонування одного об'єкту, причому це досягається перезапуском з початкових умов, то виведена модель буде адекватною саме для цього режиму (повторення циклу функціонування об'єкту з перезапуском). Відзначимо також, що не завжди можна й чітко розмежувати схеми «один екземпляр – багато екземплярів». Об'єкт моделювання може радикально еволюціонувати в часі, так що важко говорити, що на початку та в кінці еволюції маємо той самий «екземпляр». Наприклад, уявімо, що поставлена задача вивести модель поведінки комахи, життєвий цикл якої включає гіперметаморфоз. Навряд чи розумно виводити одну модель для фази личинки та для фази імаго. Втім, на практиці зазвичай мають справу з менш радикальними змінами характеру поведінки процесу. Зокрема, моделюються зміни поведінки, які називають переключенням режиму функціонування (коли додаються або зникають окремі зв'язки в моделі та змінюються параметри) [13].

Тепер можна уточнити поняття довжини часового ряду даних (тривалості спостереження за динамічним процесом). Зрозуміло, що немає абсолютної одиниці (мірки) часу для визначення довжини ряду. Все залежить від динаміки та швидкості процесів. Наприклад, в деяких медичних та соціальних задачах кожний запис даних може охоплювати ціле життя людини. При цьому дані можуть формально вважатися «короткими», і доречно виводити «статичну» каузальну модель. (Така ситуація відповідає лівому долішньому квадранту на рис. 1). Можна виділити декілька горизонтів (масштабів) довжини рядів даних. «Найближчий» (мінімально-необхідний) горизонт – коли «сеанс» спостереження процесу охоплює всі безпосередні (елементарні) зв'язки між змінними (в тому числі – з найдовшим лагом).

Протягом «найближчого» горизонту каузальні «сигнали» встигають добігти від причини до безпосередніх наслідків. Другий («2-кроковий») горизонт довжини потребує, щоб протягом сеансу спостереження встигали подіяти (розповсюдитися) кожні два послідовні безпосередніх впливи. (Це важливо для ідентифікації каузального характеру зв'язків.) Третій горизонт довжини даних – «горизонт досяжності впливу», – потребує, щоб протягом сеансу спостереження встигали подіяти (розповсюдитися) всі опосередковані впливи між компонентами ряду. Циклічний або «осциляторний» горизонт довжини передбачає, що охоплено такий період часу, протягом якого встигають спрацювати зворотні зв'язки. Тобто якщо в об'єкті «сигнал» від змінної X поширюється через інші змінні ($Z, W..,$) і потім повертається назад до змінної X, то тривалість обігу сигналу по цьому колу має цілком вкладатися в сеанс спостереження процесу. Найближчий та 2-кроковий горизонти довжини даних актуальні, коли маємо дані за схемою «а», тобто багато коротких рядів даних (на рис. 1 – внизу зліва).

Ще один («еволюційний») горизонт довжини ряду даних претендує не те, щоб сеанс спостереження процесу охоплював період часу, протягом якого встигає відбутися зміна характеру поведінки процесу (еволюція), яку ми хочемо відобразити в моделі. Для виведення моделі з адекватним відображенням процесу розвитку (еволюції) потрібні дані, віднесені до четвертого квадранту на рис. 1 (багато екземплярів з довгими рядами даних для кожного екземпляру). В принципі, можна вивести модель розвитку з одного ряду даних, але тоді вимоги до довжини ряду багатократно посилюються. Повнота даних в часовому вимірі також потребує достатньої частоти (темпу) збору даних. Тільки пара «довжина + частота» даних спільно можуть забезпечити достатню інформацію про поведінку динамічного процесу у часі. Виконання вимог до цієї пари характеристик робить можливим виведення адекватної моделі процесу. Питання частоти збору даних роз'яснюється після конкретизації «фізики» процесів у часі та типів змінних (показників).

Виділимо два принципових типи процесів: рекурентний процес та процес у неперервному часі. Відповідно отримуємо два типи даних: 1) дані рекурентних процесів; 2) дані процесів у неперервному часі (рис. 2). Ситуацію рекурентного процесу («1») маємо, коли розглядається послідовність «активних фаз (станів)» процесу в певні моменти часу (короткі інтервали), а в проміжках часу між моментами «активного стану» процес або не визначений, або недоступний в принципі. (Наприклад, марковський ланцюг визначений тільки на ліченій множині моментів, а «пасивних станів» просто немає.) Можлива ситуація, що характеристики процесу в «пасивному стані» також доступні вимірюванню; в такому разі розгляд даних як для рекурентної моделі можна пояснити тим, що вимірювання «пасивних станів» не дадуть додаткової інформації для створення моделі. Рекурентний процес розвивається чітко визначеними (регулярними) дискретними кроками. Образно кажучи, такий динамічний процес стрибає у часі. Ясно, що дані рекурентного процесу мають збиратися в моменти «активного стану». Отже, темп (частота) збору даних має бути прив'язаний до ритму рекурентного процесу. Якщо частота збору даних є меншою, ніж частота активних станів, то такі дані будуть незадовільними (у літературі це називають терміном undersampled data).

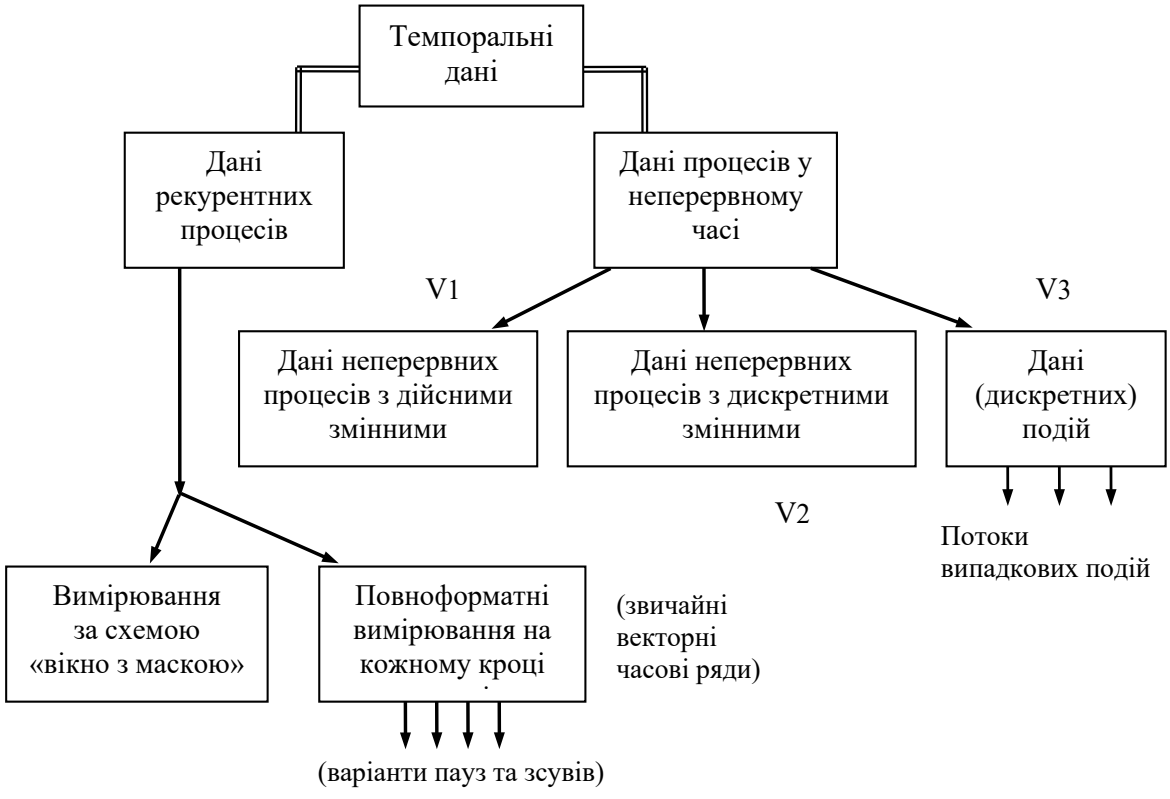

Рис. 2. Архітектура темпоральних даних (шкали часу та типи змінних)

Натомість процес у неперервному часі («2») розвивається і розглядається безперервно (континуально); відтак, аналітик й користувач потребують моделі, де час входить у рівняння в явному вигляді. В такому разі немає наперед визначених моментів, коли потрібно вимірювати дані. Це залежить від обставин і технічних можливостей. (Також можна виокремити проміжну проблемну ситуацію, коли процес розвивається як неперервний, але інтерес аналітика й користувача зосереджений виключно на певних моментах часу. Тоді до якого з двох вказаних типів процесів віднести обраний процес, залежить від вибору аналітика: яку модель він хоче виводити; чи буде дійсна змінна часу входити у рівняння в явному вигляді, чи ні.) Аналітик може обрати модель рекурентного процесу як апроксимацію процесу у неперервному часі, і тоді точність апроксимації буде критично залежати від частоти збору даних.

Для подальшої класифікації процесів у неперервному часі та даних для них потрібно розглянути наступну ознаку архітектури даних – простір значень змінних (рис. 2). Виділяємо три типи даних: дані неперервних процесів з дійсними змінними (V1); дані процесів з дискретними змінними (V2); дані потоків (дискретних) подій (V3). Випадок «V1» зазвичай відображає неперервний довготривалий процес. Випадок «V2» відображає дискретний процес у неперервному часі. Дані неперервних процесів у неперервному часі утворюються за допомогою вимірювання значень змінних у заплановані моменти часу. Дані спостережень за потоками подій («V3») утворюються за допомогою фіксації часу, коли відбулась подія. Отже, «прилад» спостереження має постійно перебувати у готовності, бо момент події (і генерації запису первинних даних) є непередбачуваний і визначається самою «природою». Якщо зарезервовано кілька класів (типів) подій, які фіксуються одним «приладом», то фіксується також клас (ідентифікатор) події. Дані процесів із дискретними змінними можна трактувати аналогічно неперервним змінним або аналогічно подіям. Дані типу «V2» відрізняється від даних типу «V3» («події») тим, що дискретна змінна існує у неперервному часі (навіть коли вона залишається постійною). Якщо поглянути на багатовимірний дискретний процес як на послідовність подій, то такі події об'єктивно груповані відповідно до окремих підпроцесів. Не завжди обов'язково фіксувати кожний перехід дискретної змінної від одного значення до іншого (такий перехід може бути менш важливим, ніж справжня «подія»).

У науковій літературі створено апарат, який дозволяє потік стохастичних подій репрезентувати як дискретний процес у неперервному часі. Для цього використовуються лічильні процеси, або процеси підрахунку (counting processes) [8, 11]. Особливість лічильних процесів полягає в тому, що значення кожної змінної зростає на одиницю у випадкові моменти часу. У дискретних процесах, що описують обслуговування у чергах, зміни стану відбуваються в обох напрямках (+1 або –1). Згідно теорії стохастичних процесів, за дискретним процесом стоїть дійсна характеристика, «інтенсивність», яка визначає ймовірність зміни стану дискретного процесу, але яка безпосередньо не спостерігається.

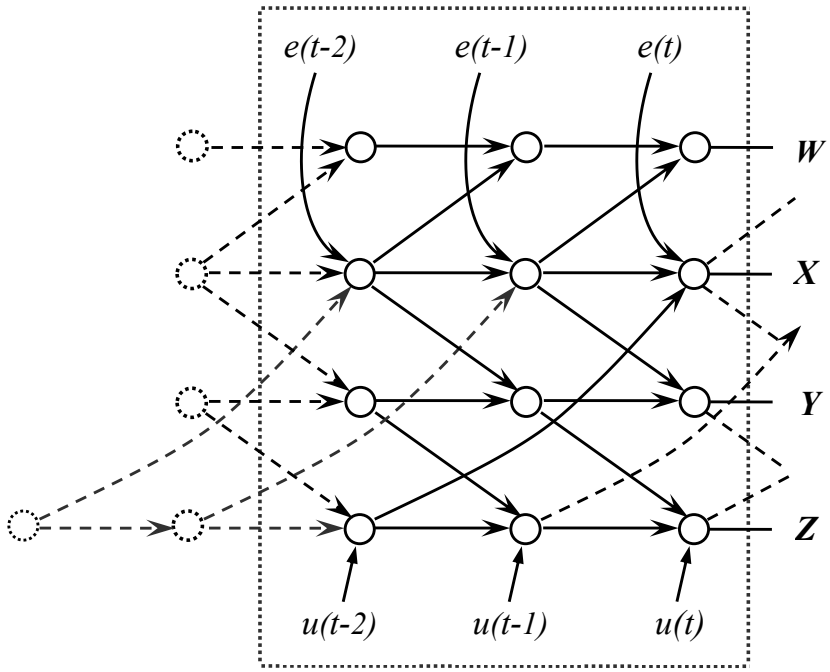

Рис. 3. Структура авторегресійного процесу

Нехай маємо неперервний процес з дійсними змінними; якщо похідні змінних у часі dx/dt мають великі значення, то вимоги до частоти збору даних різко підсилюються. На якісному рівні можна сформулювати наступну, не жорстку, але важливу, вимогу до частоти збору даних. Нехай в системі існують зворотні зв'язки, які функціонують як поширення «сигналу» вздовж відповідних шляхів у структурі моделі. Тоді частота (темп) збору даних має бути такою, щоб зафіксувати проміжні фази поширення сигналу вздовж шляху кожного зворотного зв'язку. Тоді зворотні зв'язки будуть відображені в моделі у вигляді спіралі (а не у вигляді циклону).

Проілюструємо вищеозначені горизонти (масштаби) довжини рядів даних. На рис. 3 зображено схему рекурентного динамічного процесу, який генерується згідно моделі векторної авторегресії. В цій моделі найдовший безпосередній каузальний зв'язок має лаг 2; це зв'язок $Z_{t-2} \to X_t$. «Вікно» спостереження (позначене пунктирним квадратом) охоплює три послідовні значення кожної змінної (три кроки). В даному разі таке «вікно» «захоплює» всі безпосередні каузальні зв'язки моделі. Тому, якщо дані спостережень зібрані у форматі такого «вікна», то найближчий горизонт (масштаб) довжини рядів даних забезпечено. Але таке «вікно» не забезпечує 2-крокового горизонт довжини, бо є ланцюг зв'язків $Z_{t-2} \to X_t \to W_{t+1}$, і для його явної репрезентації потрібні випадки (записи) даних довжиною 4 кроки. Вказане «вікно» довжиною три кроки також не покриває «осциляторного» горизонту довжини для цього часового ряду, бо в системі функціонує зворотний зв'язок, що складається із наступної послідовності безпосередніх каузальних зв'язків: $X_{t-2} \to Y_{t-1} \to Z_t \to X_{t+2}$. Цей зворотний зв'язок здійснюється за чотири кроки, і щоб наочно побачити його, треба репрезентувати ряд даних довжиною 5 кроків кожної змінної. Зрозуміло, якщо дані зібрано за схемою «довгі ряди даних» (на рис. 1 – справа), то вказані горизонти довжини забезпечуються автоматично. Натомість проблема частоти (темпу) збору даних – більш критична. Якщо частота збору даних є меншою, ніж справжня періодичність вказаного рекурентного динамічного процесу, можливості вивести адекватну модель звузяться. Наприклад, якщо в ряду даних значення змінних фіксуються тільки для одного із п'яти кроків вказаного процесу (рис. 3), то не вдасться відтворити спіраль зворотного зв'язку $X_{t-2} \to Y_{t-1} \to Z_t \to X_{t+2}$. З'являться моментальні (ізохронні) зв'язки.

Як правило, за умовчанням вважається, що вимірювання значень даних відбувається «миттєво». Якщо маємо швидкий неперервний процес з дійсними змінними (коли dx/dt сягає великих значень), а вимірювальний прилад фіксує якесь «середнє» значення змінної x за кінцевий інтервал часу Δt, то такі дані можуть спричинити викривлення моделі.

3. Каузальність та марковські властивості у векторних часових рядах

Серед відомих підходів та методів аналізу часових рядів даних, мабуть, найбільшого поширення набула техніка векторної авторегресії [9, 10]. З економетрики ці методи розповсюдилися на інші сфери, зокрема, експериментальну біологію. В моделях авторегресії, як і в каузальних мережах, поведінка системи описується за допомогою прямих зв'язків між компонентами. Такі моделі формально відповідають рекурентному процесу.

Одна з перших концепцій каузальності була запропонована К. Грейнджером (Granger C.W.J.) [16–18]. (Строго кажучи, первинно була сформульована концепція «не-каузальності», і вже як комплементарне з'явилося поняття каузальності.) Концепція каузальності за Грейнджером походить від традиційних уявлень, які були поширені в соціології та економетриці першої половини 20-го століття. Початкова ідея проста: для того, щоб відповісти на питання, чи є змінна X причиною для Y, треба знайти всі змінні, які (можливо) впливають на Y. Тоді «фіксація» значень всіх цих змінних, разом із змінною X, мусить призвести до того, що Y обернеться на константу. Змінюємо умови цього уявного експерименту: фіксуємо значення всіх вказаних змінних (причин), окрім самої X. Тоді, якщо варіювання X супроводжується варіюванням Y (причому змінна X йде у часі раніше, ніж Y), то змінна X є причиною для Y. Але накреслену ідею важко провести у житті. Практично неможливо зафіксувати абсолютно всі фактори, які можуть впливати на обрану зміну Y. Тому для практичного виявлення каузальних відношень обирають критерій з менш жорсткими умовами експерименту. Треба зафіксувати тільки ті фактори, які впливають на X та Y паралельно (тобто в умови включають тільки спільні причини змінних X та Y). Тоді критерієм каузального відношення буде не альтернатива «константність – неконстантність» ефекту Y, а альтернатива «незалежність – залежність» Y від X за обраних умов. Залежність за вказаних умов свідчить, що X є причиною для Y. І цю залежність можна використати для прогнозування.

Відштовхнемося від поняття «не-каузальності» за Грейнджером. Грейнджер розглядає задачу прогнозування показника (ефекту) Y й аналізує гіпотетичну причину X з точки зору її інформативності для прогнозу. Задача полягає у прогнозуванні значення показника Y_{t+1} (ефекту, наслідку) на основі наявної інформації. Ясно, що для прогнозу аналітик чи комп'ютер може знати і врахувати тільки інформацію про минуле (включно із найостаннішим, «свіжим»). (Такі умови відрізняють задачу прогнозування від задачі заповнення пропуску в базі даних, яку іноді теж називають «предикція».) Нехай Ω_t – вся інформація у світі на момент часу t (принаймні інформація, релевантна одночасно для X та Y), а X_t – значення змінної X в момент часу t. Позначимо через $\Omega_t \setminus X_t$ інформацію у світі на момент часу t, за виключенням значення X_t. Вище аргументовано: якщо X не є причиною для Y, то з врахуванням інформації $\Omega_t \setminus X_t$ значення Y_{t+1} не залежить від X_t. Відтак, прогноз розподілення змінної Y_{t+1} на основі інформації $\Omega_t \setminus X_t$ не відрізняється від прогнозу на основі інформації Ω_t. «Не-причина» не покращує прогнозу наслідку. Формально це виражається як рівність умовних розподілень: $p(Y_{t+1} \mid \Omega_t \setminus X_t) = p(Y_{t+1} \mid \Omega_t)$. У більш стандартній формі це записується як

$$p(Y_{t+1} \mid \Omega_t \setminus X_t) = p(Y_{t+1} \mid \Omega_t \setminus X_t, X_t). \quad (1)$$

Отже, «не-каузальність» за Грейнджером може бути формалізована як умовна незалежність прогнозу від «не-причини» за умови врахування причин (релевантної інформації). Далі, поняття каузальності (за Грейнджером) формально з'являється як констатація не виконання рівності (1), тобто як констатація порушення цієї умовної незалежності. Змістовно, каузальність змінної X для наслідку Y (за Грейнджером) означає, що X несе «незамінну» інформацію для прогнозу наслідку Y. Проблема в тому, що факт цієї «незамінності» важко встановити. Вся релевантна інформація Ω_t у світі практично недоступна, і навіть не можна бути впевненим, що вдалося знайти всі можливі спільні причини для X та Y. Тому неможливо надійно перевірити чинність рівняння

(1). Припустимо, аналітик здійснив перевірку сурогату рівняння (1) з використанням неповної інформації Ω_t^*. Тоді, якщо встановлено, що сурогат рівності (1) виконується, то висновок про відсутність каузального зв'язку між X та Y буде коректним. Натомість якщо встановлено, що сурогат рівності (1) **не** виконується, то висновок про каузальний зв'язок між X та Y правдоподібно може бути *некоректним*. Отже, концепції каузальності та не-каузальності (за Грейнджером) – несиметричні. Повернемося до цього пункту згодом. А зараз коротко розглянемо роль апарату умовної незалежності в динамічних і каузальних моделях.

Класичне визначення марковської властивості встановлено для марковських ланцюгів. Але апарат марковських ланцюгів не є продуктивним для задачі виведення структур моделей (через абстрактність поняття стану). Важливою віхою досліджень в потрібному напрямку стала робота [12], яка адаптувала класичну теорію марковських процесів до реалістичної проблематики каузального моделювання. Замість «унітарного» марковського процесу аналізується система підпроцесів, що взаємодіють.

Нехай структура авторегресійного процесу невідома. Для спрощення приймемо, що всі зовнішні фактори (тобто змінні поза набором $\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{W}$) є взаємно незалежними, і кожний з них впливає на одну спостережувану змінну. Тоді наївний спосіб застосувати концепцію каузальності (за Грейнджером) для тестування гіпотетичного каузального зв'язку $W_{t-1} \to X_t$ зведеться до перевірки виконання формули

$$p(X_t \mid \Omega\_{t}) = p(X_t \mid \Omega\_{t}, W_{t-1}), \tag{2}$$

де $\Omega\_{t} = \{W_0, W_1, .., W_{t-2}, X_0, X_1, .., X_{t-2}, X_{t-1}, Y_0, Y_1, .., Y_{t-2}, Y_{t-1}, Z_0, Z_1, .., Z_{t-2}, Z_{t-1}\}$. Набір умов Ω_{t} охоплює все минуле рядів $\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{W}$, за виключенням W_{t-1} (попереднього стану підпроцесу $\mathbf{W}$). Зрозуміло, що виконати перевірку рівності у подібному форматі практично неможливо. Для надійної оцінки умовної ймовірності з такою складною умовою потрібно мати величезний обсяг даних, а власне обчислення буде дуже важким. Потрібен акуратний й ефективний апарат аналізу даних, а також апарат аналізу графової структури темпоральної моделі. Принципова вразливість концепції каузальності Грейнджера полягає в тому, що вона залишає невизначеним, який прогноз мається на увазі – пасивний чи активний (ефект втручання).

В класичних («статичних») каузальних мережах успішно застосовується поняття умовної незалежності та відповідний конструктивний критерій *d*-сепарації [3, 4, 7]. Як перенести цей апарат в контекст векторних часових рядів, де маємо довгі паралельні серії змінних? Приміром, якщо поглянути на векторний авторегресійний процес, зображений на рис. 3, то зрозуміло, що всі ряди основних змінних ($\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{W}$) є взаємозалежними. Але ці ряди відіграють різну роль в системі каузальних впливів. Зокрема, очевидно, що ряд $\mathbf{W}$ не впливає на інші ряди, але безпосередньо залежить тільки від ряду $\mathbf{X}$. Щоб аналізувати структуру каузальних зв'язків між компонентами процесу, необхідно розглядати умовну незалежність у відповідному часовому форматі. Для векторних часових рядів функція сепарації полягає у тому, щоб акуратно «перерізати» паралельні серії змінних у відповідних фазах. Тобто конструктивний темпоральний критерій сепарації має залучати змінні, що знаходяться на коректно обраному «часовому фронті». Умовну незалежність «фронтового» типу в динамічних моделях називають «локальна незалежність».

Припустимо, що в рекурентному процесі немає моментальних (ізохронних) зв'язків (вигляду $V_t \to R_t$), і що максимальний лаг автокореляційних зв'язків (вигляду $V_{t-j} \to V_t$) дорівнює m. Нехай треба перевірити факт, що ряд змінних $\mathbf{W}$ не впливає на ряд $\mathbf{X}$. Нехай відомо, що немає жодного «третього» ряду, який впливає паралельно на ряди $\mathbf{X}$ та $\mathbf{W}$. Спочатку хочемо перевірити, що ряд $\mathbf{W}$ не впливає на ряд $\mathbf{X}$ з лагом 1. Для того, щоб отримати остаточну позитивну або негативну відповідь на це питання, треба аналізувати умовну залежність між X_t та W_{t-1} за кондиціонування змінних $X_{t-1}, .., X_{t-m-1}, W_{t-2}, .., W_{t-m-2}$. Якщо ця умовна залежність встановлена, то зв'язок $W_{t-1} \to X_t$ присутній, інакше – відсутній. Наступним кроком можна тестувати факт, що ряд $\mathbf{W}$ не впливає на ряд $\mathbf{X}$ з лагом 2. Для цього треба тестувати умовну залежність між X_t та W_{t-2}, з використанням в умові відповідних минулих значень цих рядів. Якщо відомо, що зв'язок $W_{t-1} \to X_t$ існує, то змінну W_{t-1} треба включати в умову вказаного тесту. Якщо цього не зробити, то умовна залежність між X_t та W_{t-2} може існувати завдяки шляху $W_{t-2} \to W_{t-1} \to X_t$, а не завдяки безпосередньому зв'язку $W_{t-2} \to X_t$ з лагом 2. Тестування безпосередніх зв'язків з великим лагом відкриває все більше можливостей для «опосередкованих» впливів. Отже, в умови тестів безпосереднього зв'язку необхідно включати не тільки спільне минуле, але й можливих посередників. Але в процесі аналізу структури динамічного процесу доцільно уникати складних тестів. Процес аналізу має просуватися за принципом «від простих тестів до складніших».

Конкретно для структури рекурентного процесу, показаного на рис. 3, той факт, що ряд $\mathbf{W}$ не впливає на ряд $\mathbf{X}$ (послідовно з лагами 1, 2, 3, ..), підтверджується тестами умовної незалежності між X_t, з одного боку, та (послідовно) $W_{t-1}, W_{t-2}, W_{t-3}, ..$ – з іншого, причому всі ці тести використовують як умову одну змінну X_{t-1}. Таким чином, для конкретної вказаної структури відсутність зв'язків впливу ряду $\mathbf{W}$ на ряд $\mathbf{X}$ (з різними лагами) ідентифікується тестами першого рангу. Звертаємо увагу на наступну особливість відношень в динамічних процесах. За кондиціонування змінної X_{t-1} змінна X_t умовно *незалежна* від W_{t-1}. Натомість за кондиціонування змінної W_{t-1} змінна W_t умовно *залежна* від X_{t-1}. (Ця залежність забезпечується зв'язком $X_{t-1} \to W_t$.) Маючи на увазі несиметричність вказаної конструкції, в літературі іноді кажуть, що локальна незалежність є несиметричною (на відміну від звичайної незалежності). Насправді вказана несиметричність не є властивістю відношення незалежності. Просто факти незалежності в часових рядах даних завжди виражаються через змінні з часовими прив'язками (індексами). Якщо маємо незалежність між X_k та Y_m, то незалежність не обов'язково збережеться після обміну індексами. Обмін індексами у змінних означає перехід до змістовно іншого відношення. Вказана несиметричність є проявом несиметричності структури зв'язків між часовими рядами, а також проявом спрямованості процесів в один бік (від минулого до майбутнього).

Щоб формалізувати техніку сепарації для динамічних моделей, необхідно обрати форму графової репрезентації динамічного процесу. Один варіант репрезентації фактично використано на рис. 3. Зручна і компактна форма графової репрезентації векторного авторегресійного процесу («базова динамічна мережа») створюється на основі «найближчого» горизонту довжини векторного ряду даних. Базова динамічна мережа виглядає як фрагмент багатовимірного ряду з набором безпосередніх зв'язків. Модель компактно презентує увесь довгий ряд (в згорнутому вигляді). Розмір моделі (її «часова глибина») визначається довжиною (глибиною) безпосереднього зв'язку з найбільшим часовим лагом. В лівій («ранній») частині моделі структура має додаткові «неявні» зв'язки, які є наслідком залежностей з минулого. (На рис. 3 вони показані пунктиром.) Це порушує візуальну регулярність структури, але не заважає аналізу. Просто треба пам'ятати, що в процесі аналізу зв'язків завжди треба розглядати праву (пізнішу) частину такої структури. Якщо для аналізу потрібна динамічна мережа з більшим горизонтом часу, то структура «нарощується», тобто з правого боку моделі додаються копії змінних для наступного такту з усіма копіями зв'язків (відображених суцільними лініями).

У вищенаведеній аргументації та викладках суттєвими вважалися тільки зсуви часу (різниця між індексами), а абсолютні значення індексів часу розглядалися як умовні. Тобто за умовчанням вважалося, що проведений аналіз стосується всього векторного часового ряду цілком. Насправді це не завжди вірно, і необхідно розглянути питання стаціонарності та стабільності процесів.

4. Типи стаціонарності процесів і припущення для аналізу векторних часових рядів

Виведення каузальної моделі (як й будь-якої статистичної моделі), а також її верифікація на основі даних, спирається на обґрунтування концептуально чіткої схеми обчислення статистик з даних. Для обчислення статистик треба виділити з даних «випадки» і гарантувати певні вимоги регулярності генерації даних. Найбільш відома форма регулярності генерації даних – схема I.I.D, – традиційно була сформульована для «статичних» даних (лівий долішній квадрант на рис. 1). Схема I.I.D може бути адекватною також у ситуації, коли маємо багато екземплярів популяції з довгим рядом даних для кожного екземпляру (четвертий квадрант на рис. 1). Маючи багато екземплярів довгих рядів даних, можна аналізувати навіть нестаціонарний динамічний процес, зокрема, процес, структура якого змінюється у часі. В такому разі для кожного такту (кроку) j процесу статистика збирається «вертикально», по одному запису (для такту j) з кожного ряду даних. Проблеми обґрунтувань і припущень для статистичного аналізу виникають для типової ситуації, коли маємо один довгий (тривалий) ряд даних (правий верхній квадрант на рис. 1). Тоді доводиться узагальнювати дані з різних тактів одного процесу, а відтак, постають питання щодо стандартних припущень про статистичну вибірку даних (незалежність, ідентичність, однорідність, стаціонарність, стабільність тощо). Увесь часовий ряд даних відображає історію функціонування об'єкту (середовища). В реальності об'єкт може змінюватися й еволюціонувати. Для того, щоб відомі методи були спроможні вивести адекватну модель з даних, мають виконуватися обмеження на характер та швидкість змін поведінки об'єкту. Типова риса динамічного процесу – це міцні «короткі» автокореляційні зв'язки $X_t \to X_{t+1}$. Втім, такі зв'язків не обов'язково присутні.

Усталені в економетриці методи аналізу часових рядів (а також визначення стаціонарності) за умовчанням виходять з припущення про рівномірність повторення вимірювань даних у часі, а також про одночасність вимірювань компонентів вектору характеристик з одним індексом часу. Нехай векторний часовий ряд включає l паралельних рядів, тобто складається з набору l змінних $(X,Y,Z,...)$. Одночасність вимірювань компонентів вектору характеристик означає, що значення кожної компоненти вимірюють в моменти часу з єдиного переліку $t = 0,1,2,...,\infty$. Отже, природним чином визначається вектор-стовпчик $(X_t,Y_t,Z_t,...)$ значень, виміряних в один і той самий момент t. Можна позначити вектор-стовпчик $(X_t,Y_t,Z_t,...)$ через $\mathbf{S}_t$ й назвати його станом процесу у момент t. Але, в принципі, не є обов'язковим, щоб всі змінні з одного інтервалу часу (кроку) вимірювались строго в один і той самий момент. Можлива ситуація, коли існують деякі «зсуви» між моментами вимірювань «квазі-одночасних» змінних. Проте такі зсуви мають бути фіксованими, інакше дані з нерегулярною схемою вимірювань будуть породжувати помилки, зміщення і викривлення. Про зсув між моментами вимірювання різних змінних має бути відомо аналітику, бо ця інформація важлива для каузального виведення (це темпоральний порядок, який треба врахувати). Також за умовчанням в усталених методах аналізу часових рядів прийнято, що всі інтервали часу між моментами вимірювання всіх змінних – однакові. Але, взагалі кажучи, не обов'язково виконувати всі вимірювання через фіксовані інтервали часу. Проте варіативність інтервалів часу призведе, зокрема, до розмивання функції автоковаріації, і взагалі, буде породжувати ілюзію нестаціонарності. Стаціонарність – традиційне припущення в аналізі часових рядів. Багато методів аналізу спираються на це припущення. Реальні обставини й потреби розширити сфери застосувань методів змушують аналітиків послабити вимоги стаціонарності, але в таких межах, щоб не втратити спроможності вивести адекватну модель з даних.

Переглянемо відомі форми статистичної стаціонарності динамічного процесу та часових рядів даних. *Строга стаціонарність* (векторного) часового ряду означає, що імовірнісна поведінка значень часового ряду не змінюється з просуванням у часі, тобто сумісний розподіл ймовірностей значень часового ряду є незмінним (ідентичним) для довільного зсуву індексів часу [9–11]. Внаслідок строгої стаціонарності, зокрема, автокореляційна функція залежить тільки від різниці індексів часу (від часового лагу) і не залежить від абсолютного часу. *Слаба стаціонарність* (векторного) часового ряду означає, що: 1) математичні очікування значень ряду є константні, тобто не залежать від часу; 2) кожна коваріація залежить тільки від різниці індексів часу (від часового лагу) і не залежить від абсолютного часу.

Назвемо деякі нетрадиційні форми стаціонарності динамічного процесу. «Перехідна» стаціонарність означає, що *умовний* розподіл ймовірностей значень часового ряду є незмінним (стабільним) у часі, тобто $p(\mathbf{S}_t \mid \mathbf{S}_{t-1}, \mathbf{S}_{t-2}, ..)$ не залежить від t. В аналізі часових рядів (за умовчанням) прийнято припущення про регулярну структуру зв'язків протягом всього часу спостереження. (Це забезпечує обчислення коректних статистик.) «Структурна» стаціонарність чи регулярність означає незмінність набору зв'язків у часі. Найпростіший варіант структурної регулярності – елементарна регулярність, повторюваність всіх зв'язків на кожному такті (кроці) всього процесу. Тобто якщо є зв'язок $X_i \to Y_{i+1}$, то він присутній для всіх тактів $i = 0,1,2,...,\infty$. Можливі й інші трактовки структурної регулярності (наприклад, повторення зв'язків з періодом два чи більше тактів). Зазначимо, що часовий ряд може мати тренд навіть за умови, що чинні «перехідна» стаціонарність й структурна стаціонарність. Наявність тренду створює проблеми для багатьох методів виведення моделі [19]. Один із способів обійти або послабити проблему полягає у трансформації даних. Аналізуються не самі оригінальні дані, а відповідні різниці. Наприклад, переходять до змінних $\Delta X_t = X_t - X_{t-1}$, завдяки чому умовний розподіл ймовірностей значень $p(\Delta X_t \mid Y_{t-1}, ..)$ або $p(\Delta X_t \mid \Delta Y_{t-1}, ..)$ може виявитися стабільним. Нарешті, зауважимо, що коли динамічний процес об'єктивно генерується системою диференціальних рівнянь, поведінка процесу може бути квазі-хаотичною (для таких процесів рекурсивні моделі не придатні).

Якщо всі інтервали часу між моментами вимірювання даних – однакові, то час можна відображати цілими індексами $i = 0,1,2,...,N$. Стан процесу в такті i відображається вектором-стовпчиком $\mathbf{S}_i = (X_i, Y_i, Z_i, ...)$. Для утворення статистик, необхідних для виведення моделі, треба спиратися на певну регулярність структури. Нехай маємо найпростіший варіант структурної регулярності, коли період повторення зв'язків – одиниця. Тоді елементарний випадок статистичної вибірки – це зріз векторного ряду даних для одного з індексів часу i. Він включає вектор-стовпчик $\mathbf{S}_i$ та кілька векторів-стовпчиків з меншими значеннями індексів часу, відповідно до розміру лагу. Тобто елементарний «випадок» вибірки виглядає як $<\mathbf{S}_i, \mathbf{S}_{i-1}, \mathbf{S}_{i-2}, .., \mathbf{S}_{i-\tau}>$. Величина τ має дорівнювати найбільшому часовому лагу безпосередніх зв'язків процесу. Якщо довжина ряду даних дорівнює N, то отримаємо $N - \tau$ повноформатних «випадків» вибірки.

5. Виведення каузальних динамічних моделей і роль припущень

Коротко розглянемо особливості виведення динамічної моделі з векторних часових рядів даних методами, основаними на незалежності [14, 15, 19–22]. Вибір цих методів дозволяє чітко висвітлити суть проблем і роль припущень. Базовими операціями цих методів є тестування умовної незалежності. Методи, основані на незалежності, не є ізольованими від традиційних методів. Зокрема, можна провести паралелі між цими методами та методами регресійного аналізу. Дійсно (за певних стандартних умов) є еквівалентність між тестуванням умовної незалежності та тестуванням значущості змінної як регресора. Якщо, приміром, змінна X умовно незалежна від Y за умови на Z, то регресія Y на (X, Z) має дати незначущий коефіцієнт для регресора X.

Задача виведення динамічної каузальної моделі відрізняється від аналогічної «статичної» задачі низкою особливостей. Перша – це заданий темпоральний порядок змінних. Друга – це зазвичай збільшена кількість змінних, а відтак, і зв'язків (автокореляційних та крос-кореляційних). Але зростання складності стримується іншою особливістю задачі – стандартними припущеннями про регулярність системи зв'язків. Припущення про регулярність зв'язків доповнюється припущенням про стабільність кількісних характеристик (сили зв'язків). Знання темпорального порядку змінних дозволяє впорядкувати виведення динамічних каузальних мереж (порівняно з звичайними мережами), спрощує пошук сепараторів та допомагає визначенню орієнтацій ребер. Розумно спочатку ідентифікувати «короткі» зв'язки (з малим лагом); це спростить пошук довгих зв'язків. Відкриваються можливості оптимізувати тактику пошуку сепараторів, тобто умов для тестів незалежності.

Якщо найбільший лаг післядії (довжина зв'язків) – невідомий, виникає проблема, якої немає при виведенні «статичної» каузальної мережі. По-суті, в такому разі невідомим стає розмір («часова глибина») моделі. В такій ситуації треба вирішувати, коли припинити аналіз вглиб і зафіксувати розмір моделі. Просте практичне рішення – перед запуском алгоритму виведення директивно встановити максимальний лаг (вказати цифру). Більш адаптивне рішення (яке має певне теоретичне виправдання) полягає в тому, що при досягненні певних ознак «насичення» моделі робиться висновок, що знайдено справжній розмір моделі. Відтак, нарощування моделі треба припинити. Такою ознакою «насичення» моделі є той факт, що після аналізу зв'язків чергової довжини k (найбільшої на цей момент) з'ясувалося, що зв'язки такої глибини відсутні. Вказана ознака в багатьох випадках виправдана, але не завжди гарантує досягнення повної глибини автентичної моделі. Така ознака дійсно свідчить про відсутність зв'язки з лагом більше k тільки за припущення регулярності структури та за умов, що всі ряди мають достатньо сильні короткі автокореляційні зв'язки (вигляду $X_{i-1} \to X_i$). Короткі автокореляційні зв'язки присутні у більшості реальних процесів, але не можна гарантувати, що вони є скрізь.

Припущення елементарної структурної регулярності може не виконуватися. Регулярність зв'язків може мати іншу форму – як періодичне повторення зв'язків. Тобто якщо є зв'язок $X_{i-1} \to X_i$, то також є зв'язок $X_{i-1+\omega} \to Y_{i+\omega}$. Згорнута репрезентація такої моделі має «довжину» $\tau + \omega$, де τ – довжина зв'язку з найбільшим лагом, а ω – довжина періоду повторення зв'язків. Якщо розмір періоду повторення зв'язків невідомий, задача виведення моделі радикально ускладнюється. Треба багатократно переглядати схему обчислення статистик.

Зв'язки з великим часовим лагом є джерелом проблем. Давайте поглянемо на існування таких зв'язків з точки зору фізичного сенсу і здорового глузду. Пояснення, що далеке минуле безпосередньо впливає на сучасне – непереконливе й неправдоподібне. Адже, строго кажучи, минулого вже не існує. Втім, цей парадокс легко

розв'язується. Треба припустити, що існують ланцюги не спостережуваних змінних, через посередництво яких минуле передало інформацію і змогло здійснити вплив на змінні поточного часу.

Ідентифікація ребра моделі – це виявлення безпосереднього («лаг-специфічного») зв'язку. Для ідентифікація кожного ребра доводиться виконати багато тестів умовної незалежності (знайти сепаратор). В ході пошуку сепаратора для пари Y_i, X_{i-k} треба відразу відкинути всі змінні V_j з індексами часу $j > i$. Першочергові кандидати у сепараторний набор – це змінні, можливо-суміжні до Y_i або X_{i-k}, й найближчі до них. Назвемо кілька запропонованих алгоритмів виведення моделі з багатовимірних часових рядів даних. Алгоритми SVAR-FCI та SVAR-GFCI [22] є результатом адаптації методу FCI до часових рядів з прихованими змінними. Ці алгоритми припускають існування ізохронних зв'язків, але без утворення циклонів. (Циклони не утворюються, якщо впливи протягом одного інтервалу вимірювання не встигають оббігти цикл й повернутися до стартової змінної.) Алгоритм Regime-PCMCI відрізняється додатковими можливостями; він автоматично шукає фрагменти векторного ряду, які відповідають різним режимам функціонування процесу [13].

Характер (орієнтацію) деяких ребер розпізнати неможливо, особливо якщо припускаються приховані спільні причини. В таких обставинах алгоритм виведення моделі оперує кількома формами опису ребер. Ребро вигляду $V \to Z$ є каузальним. Біорієнтоване ребро $Q \leftrightarrow W$ показує, що існує деяка прихована змінна U, яка впливає рівночасно (паралельно) на Q та W. Ребро вигляду $X \circ\!\!-\!\!\circ Y$ відображає ситуацію, коли каузальний характер цього зв'язку зовсім не визначений. Ребро з частково визначеною орієнтацією $V \circ\!\!\to Z$ резервує два варіанти (біорієнтоване, каузальне). Після ідентифікації присутності ребер моделі алгоритм переходить до етапу орієнтації ребер. Якщо змінна X йде у часі раніше, ніж Y, позначаємо це $X \succ Y$. Чинне правило «фізичного сенсу»: якщо відомо, що $X \succ Y$, і маємо ребро $X \circ\!\!-\!\!\circ Y$, то орієнтуй $X \circ\!\!\to Y$. Втім, замість того, щоб вставляти в алгоритм подібне правило, простіше вже на етапі виявлення безпосередніх зв'язків відразу ставити ребра вигляду $X \circ\!\!\to Y$. Отже, при виведенні моделі з даних часових рядів всі ребра відразу отримують вістря на передньому кінці. Тому деякі стандартні правила орієнтації [6, 7], розроблені для статичних каузальних мереж, стають непотрібними. Повний набір правил працює для орієнтації ізохронних ребер. Але тут розглядаємо випадок, коли ізохронних ребра немає. Розглянемо тільки перші два правила орієнтації ребер – $\Re 0$ та $\Re 1$. Оскільки завдяки відомому темпоральному порядку змінних всі ребра відразу отримують вістря ($A \circ\!\!\to B$), створюються умови для застосування правила орієнтації ребер $\Re 1$. Тому на перший погляд може здатися, що немає необхідності застосовувати правило орієнтації ребер $\Re 0$. Проте це хибна думка. Правило $\Re 0$ забезпечує також розпізнавання «вістря» на задньому кінці ребер. Тим самим правило $\Re 0$ розпізнає біорієнтовані ребра, а отже, виключає можливість, що цей зв'язок є каузальним.

Для ситуації відомого темпорального порядку змінних правило орієнтації ребер $\Re 1$ формулюється наступним чином

$$X^* \to Y \circ\!\!\to Q \;\&\; Y \in \mathrm{Sep}_{\min}(X,Q) \;\Rightarrow\; X^* \to Y \to Q, \tag{3}$$

де $\mathrm{Sep}_{\min}(X,Q)$ – це мінімальний сепаратор для пари змінних X, Q. Позначка кінця ребра «*» означає, що цей кінець ребра залишається не специфікованим і може бути довільним.

Для ілюстрації виведення в умовах спільних прихованих причин розглянемо генеративну модель, показану на рис. 4,а. В цій структурі приховані причини e_i – взаємонезалежні. (У наведеній структурі кожна e_i впливає відразу на дві спостережувані змінні. Це є відхиленням від стандартних припущень.) Маючи достатньо даних, типові алгоритми виведуть модель, показану на рис. 4,в. Біорієнтовані ребра виводяться завдяки правилу $\Re 0$. Модель адекватна. Проблем не виникло, оскільки приховані спільні причини e_i не займають позицій конфаундерів для каузальних зв'язків [7].

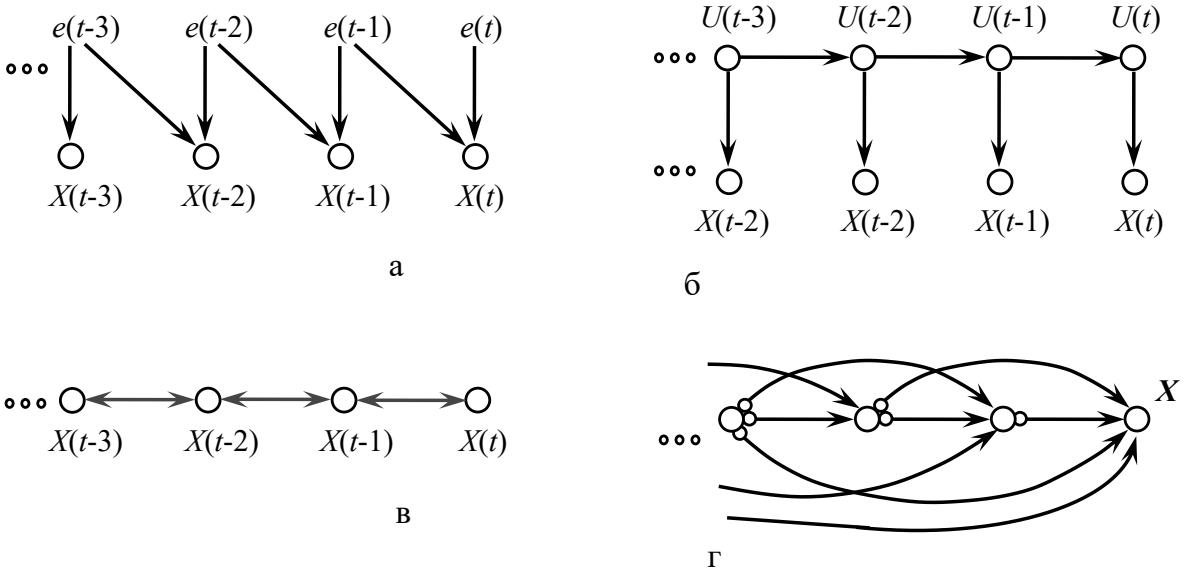

Рис. 4. Ілюстрація виведення каузальних моделей в умовах прихованих змінних та прихованих часових рядів; (а) та (б) – генеративні моделі; (в) та (г) – виведені моделі.

Повертаємося до фізичного пояснення впливів з великим часовим лагом. Ясно, що існують механізми, які передають вплив минулого через посередників (приховані змінні). Змінні, що є першоджерелом впливу, можуть бути організовані як динамічний автокорельований процес. Тому можна припустити, що і посередники впливу також організовані як автокорельований процес. Отже, можна очікувати, що існують приховані автокорельовані часові ряди, які взаємодіють з іншими підпроцесами. Такі часові ряди можуть створити ілюзію зв'язків з нескінченно-довгим часовим лагом. Це підриває тестування марковських властивостей як принцип виведення моделі.

Покажемо можливі наслідки існування прихованих часових рядів. Розглянемо екстремально спрощену структуру моделі (рис. 4,б) з наступними особливостями. Прихований часовий ряд U_i – автономний, тобто не зазнає впливів від спостережуваної частини моделі. Немає безпосередніх зв'язків між спостережуваними змінними X_i. Спочатку уявімо ідеалізовану постановку задачі виведення моделі, коли всі факти незалежності строго випливають з фактів d-сепарації. Тоді всі змінні ряду X_i – взаємозалежні, і немає жодної умовні незалежності в рамках спостережуваної моделі. Як наслідок, алгоритм виведе повнозв'язану (насичену) структуру, частково показану на рис. 4,г. Всі ребра мають вигляд $X_i \circ\!\!\rightarrow X_j$. Глибина моделі буде визначена технічними обмеженнями алгоритму. Тепер уявімо виведення такої моделі з реалістичних даних. В ході виведення можуть відбутися різні сценарії, пов'язані з тим, що певні непрямі залежності (через ланцюги) поведуть себе як слабкі. Якщо будуть виявлені факти безумовної незалежності, то (завдяки правилу $\Re 0$) в моделі з'являться біорієнтовані ребра, зокрема, $X_{i-1} \leftrightarrow X_i$. Якщо алгоритм знайде факти умовної незалежності, то правило $\Re 1$ виведе каузальні ребра (і це буде грубою помилкою).

6. Виявлення каузальних відношень в даних і каузальність за Грейнджером

Навіть в умовах прихованих причин апарат каузальних мереж (за певних відомих обставин) забезпечує виведення каузальних ребер, які свідчать про каузальні відношення [3, 4, 6, 7]. Натомість концепція каузальності К.Грейнджера розрахована на ситуацію повної спостережуваності світу (приховані змінні, які впливають на цільові змінні, не припускаються). В реальних ситуаціях доводиться застосовувати критерій Грейнджера в сурогатному режимі, тобто в умовах доступності лише неповної інформації Ω_t^*. Порівняємо можливості цих двох підходів. Спочатку відзначимо другорядну особливість, що відрізняє «причину» (за К. Грейнджером) від «причини» в апараті каузальних мереж. В каузальній мережі змінна X вважається причиною для Y, якщо є строго орієнтований шлях від X до Y. Тобто в перелік причин включено також і непрямі (опосередковані) причини. Натомість за К. Грейнджером непрямі причини X будуть визнані не-причинами, якщо доступний достатній набір посередників, через які здійснюється вплив X на Y.

Суттєвим є те, що сурогатний критерій К. Грейнджера може призводити до систематичних помилок у визначенні каузальності. Можна описати роботу критерія К. Грейнджера так: формується набір предикторів $\Im_{t-1}$, який мінімізує помилку прогнозу змінної Y_t, а потім з набору $\Im_{t-1}$ видаляються всі «надлишкові» предиктори. Виключення надлишкових предикторів не погіршує прогноз. Всі предиктори, що не були видалені як надлишкові, визнаються «причинами» для Y_t за Грейнджером. Легко бачити, якщо, припустимо, для X_{t-k} та Y_t алгоритм виведення моделі не знайшов сепаратора, то X_{t-k} буде визнана причиною для Y_t за Грейнджером. Отже, критерій К. Грейнджера включає в перелік «причин» для Y_t всі змінні, які поєднані з Y_t ребром і які йдуть у часі раніше за Y_t. Зокрема, якщо є ребро $Z_{t-k} \leftrightarrow Y_t$, то змінна Z_{t-k} також буде визнана причиною для Y_t, що є помилкою. Не-причина може бути необхідною для покращення прогнозу, бо несе «незамінну» інформацію в умовах недоступності справжніх причин. Наприклад, якщо маємо генеративну модель, показану на рис. 4,б, то, згідно Грейнджеру, причинами змінної X будуть визнані всі минулі значення цієї змінної.

Більш того, неадекватність критерія Грейнджера простягається ще далі. В певних ситуаціях цей критерій може включити в набір «причин» для Y_t змінну, яка є безумовно незалежна від Y_t. Так станеться, якщо в моделі є ребро $Z_{t-k} \leftrightarrow Y_t$ та ребро $Z_{t-k} \leftrightarrow Q_{t-l}$. Це може здатися парадоксальним, але включення змінної Q_{t-l} в набір предикторів зменшує помилку прогнозу змінної Y_t (за умови, що на змінну Z_{t-k} не здійснюється втручання). Отже, в реалістичних умовах критерій К. Грейнджера є некоректним для задачі виявлення причин.

Звернемося до правил орієнтації ребер в методах виведення каузальних мереж. Як показано в [7], якщо можливі конфаундери, то серед всіх відомих правил тільки правило $\Re 1$ виводить каузальне ребро. Це правило є коректним засобом виявлення причинних зв'язків. Але для того, щоб правило орієнтації ребер $\Re 1$ спрацювало, має виконуватися відповідна умовна незалежність. А для цього треба, щоб алгоритм «дістався» до відповідного зв'язку з найбільшим лагом, і щоб не було прихованого конфаундера. Якщо жодного непустого сепаратора не знайдено, то тоді не буде виведено жодного каузального зв'язку, і всі зв'язки залишаться у невизначеному статусі (в формі ребра $X \circ\!\!\rightarrow Y$).

Висновки

Динамічні каузальні моделі є потужним засобом аналізу й дослідження різноманітних процесів та об'єктів і забезпечує відображення таких рис, як швидкість змін, періодичність, лаг післядії, зворотні зв'язки тощо. Серед різноманітних типів динамічних процесів (які включають, зокрема, рекурентні процеси, «запрограмовані» процеси, потоки стохастичних подій, лічильні процеси і так далі) виділено векторні авторегресійні процеси. (Відомо, як для них можна виводити каузальні моделі.) Вимоги до архітектури темпоральних даних, способів та частоти їх збору диктуються характером динамічного процесу. Частота вимірювання даних має критичне значення для адекватності виведеної моделі і визначається тривалістю елементарних впливів між

компонентами векторного процесу. Для виведення динамічним моделей з даних важливу роль відіграють припущення стаціонарності та регулярності структури. Особливостями виведення динамічних каузальних моделей є врахування темпорального порядку змінних та проблеми, пов'язані з невідомим лагом післядії та існуванням прихованих автокорельованих часових рядів.

Проведено порівняльний аналіз критерію каузальності за Грейнджером та правил орієнтації ребер в апараті каузальних мереж з точки зору їх спроможності виявляти каузальні відношення. Концепція каузальності за Грейнджером в реальних умовах неповноти інформації непридатна для виявлення каузальних відношень в даних. За критерієм Грейнджера, в набір предикторів для прогнозу наслідку Y будуть включені всі змінні, корисні для покращення пасивного прогнозу. Взагалі, коректність критерію залежить від формату керування. Оскільки деякі справжні причини X для наслідку Y є недоступними, то корисними (для певних форматів втручання в об'єкт) можуть бути не-причини Z, які передають «незамінну» (в даному контексті) інформацію для прогнозу наслідку Y. (Зокрема, це змінні, пов'язані будь-яким ребром з Y, які йдуть у часі раніше Y.) Коректним засобом виявлення каузальних відношень в даних є правило орієнтації ребер $\mathfrak{R}1$ (розроблене в апараті каузальних мереж). Але це правило результативне тільки за певних умов (каузальне ребро не знаходиться під дією прихованого конфаундери і не оточене повно зв'язаною структурою).

Бібліографія

1. Handbook of Big Data. P. Bühlmann, P. Drineas, M. Kane, M. van der Laan (Eds.). Boca Raton, FL.: CRC Press. Taylor & Francis Group, 2016. 452p.
2. Балабанов О.С. Задачі та методи аналізу великих даних (огляд). (Оновлено) Jan 2020. – [Electronic resource] URL: https://www.researchgate.net/publication/338714383_Tasks_and_methods_of_Big_Data_analysis_a_survey_revised.
3. Pearl J. Causality: models, reasoning, and inference. Cambridge: Cambridge Univ. Press, 2000. 526 p.
4. Spirtes P., Glymour C. and Scheines R. Causation, prediction and search. New York: MIT Press, 2001. 543 p.
5. Spirtes P., Zhang K. Causal discovery and inference: concepts and recent methodological advances. *Applied Informatics*. (2016). V.3: 3. 28 p.
6. Zhang J. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*. (2008). V. 172. P. 1873–1896.
7. Балабанов О.С. Логіка каузального виведення з даних в умовах прихованих спільних причин. *Кибернетика и системный анализ*. 2022. № 2. С. 10–28.
8. Aalen O.O., Borgan O., Gjessing H.K. Survival and Event History Analysis. A Process Point of View. Springer, New York, 2008. 539 p.
9. Lütkepohl H. New introduction to multiple time series analysis. Springer-Verlag, 2005. 764 p
10. Shumway R.H., Stoffer D.S. Time series analysis and its applications with R examples. Springer, 2011, 596 p.
11. Kulkarni V.G. Introduction to modeling and analysis of stochastic systems. (2nd Ed.) Springer, 2011. 313 p.
12. Schweder T. Composable Markov processes. *J. Appl. Probab*. (1970). V. 7. P. 400–410.
13. Reconstructing regime-dependent causal relationship from observational time series / E. Saggioro, J. de Wiljes, M. Kretschmer, J. Runge – *Chaos*. (2020). V. 30 (n.11). 113115. –22p. ISSN 1089-7682.
14. Gong M., Zhang K., Schölkopf B., Tao D. and Geiger P. Discovering temporal causal relations from subsampled data. Proc. of the 32nd Intern. Conf. on Machine Learning, 2015. P. 1898–1906.
15. Plis S., Danks D., Freeman C., and Calhoun V. Rate-agnostic (causal) structure learning. In: Advances in Neural Information Processing Systems. 2015. P. 3303–3311.
16. Granger C.V.J. Testing for causality. A personal viewpoint. *Journal of Economic Dynamics and Control*. (1980). V. 2. Issue 1, P. 329–352.
17. Granger C.V.J. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*. (1969). V. 37. P. 424–459.
18. Swanson N.R. and Granger C.W.J. Impulse response functions based on a causal approach to residual orthogonalization in vector autoregressions. *J. of the American Statistical Association*. (1997). V. 92. N 437. P. 357–367.
19. Malinsky D. and Spirtes P. Learning the structure of a nonstationary vector autoregression. The 22nd Intern. Conf. on Artificial Intelligence and Statistics. – Proc. of Machine Learning Research, PMLR, 2019. V. 89. P. 2986–2994.
20. Entner D. and Hoyer P.O. On causal discovery from time series data using FCI. *Proc. of the 5th European Workshop on Probabilistic graphical models*. 2010, Helsinki, Finland. P. 121–128.
21. Runge J. Causal network reconstruction from timeseries: From theoretical assumptions to practical estimation. *Chaos*. (2018). V. 28, paper 075310. 20 p.
22. Malinsky D. and Spirtes P. Causal structure learning from multivariate time series in settings with unmeasured confounding. *Proc. of 2018 ACM SIGKDD Workshop on Causal Discovery*, 2018, London, UK. – *PMLR*, V. 92. P. 23–47.

References

1. Handbook of Big Data. P. Bühlmann, P. Drineas, M. Kane, M. van der Laan (Eds.). Boca Raton, FL.: CRC Press. Taylor & Francis Group, 2016. 452p.
2. Balabanov O.S. (2020) Tasks and methods of Big Data analysis (a survey). (Revised). (preprint at ResearchGate) DOI: 10.13140/RG.2.2.18586.39367 [in Ukrainian]
3. Pearl J. Causality: models, reasoning, and inference. Cambridge: Cambridge Univ. Press, 2000. 526 p.
4. Spirtes P., Glymour C. and Scheines R. Causation, prediction and search. New York: MIT Press, 2001. 543 p.
5. Spirtes P., Zhang K. (2016) Causal discovery and inference: concepts and recent methodological advances. *Applied Informatics*. V.3: 3. 28 p.
6. Zhang J. (2008) On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*. V. 172. P. 1873–1896.
7. Balabanov O.S. (2022). Logic of causal inference from data under presence of latent confounders. *Cybernetics and Systems Analysis*. V. 58. No. 2, P. 171–185. DOI 10.1007/s10559-022-00448-z
8. Aalen O.O., Borgan O., Gjessing H.K. Survival and Event History Analysis. A Process Point of View. Springer, New York, 2008, 539 p.
9. Lütkepohl H. New introduction to multiple time series analysis. Springer-Verlag, 2005, 764 p.
10. Shumway R.H., Stoffer D.S. Time series analysis and its applications with R examples. Springer, 2011, 596 p.
11. Kulkarni V.G. Introduction to modeling and analysis of stochastic systems (2nd Ed.) Springer 2011. 313 p.
12. Schweder T. (1970) Composable Markov processes. *J. Appl. Probab*. V. 7. P. 400–410.
13. Reconstructing regime-dependent causal relationship from observational time series / E. Saggioro, J. de Wiljes, M. Kretschmer, J. Runge – *Chaos*. (2020). V. 30 (n.11). 113115. –22p. ISSN 1089-7682.
14. Gong M., Zhang K., Schölkopf B., Tao D. and Geiger P. (2015) Discovering temporal causal relations from subsampled data. *Proc. of the 32nd Intern. Conf. on Machine Learning*, P. 1898–1906.

15. Plis S., Danks D., Freeman C., and Calhoun V. Rate-agnostic (causal) structure learning. In: Advances in Neural Information Processing Systems. 2015. P. 3303–3311.
16. Granger C.V.J. (1980) Testing for causality. A personal viewpoint. *Journal of Economic Dynamics and Control*. V. 2. issue 1, P. 329–352.
17. Granger C.V.J. (1969) Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*. V. 37. P. 424–459.
18. Swanson N.R. and Granger C.W.J. (1997) Impulse response functions based on a causal approach to residual orthogonalization in vector autoregressions. *J. of the American Statistical Association*. V. 92. N 437. P. 357–367.
19. Malinsky D. and Spirtes P. (2019) Learning the structure of a nonstationary vector autoregression. The 22nd Intern. Conf. on Artificial Intelligence and Statistics. – Proc. of Machine Learning Research, *PMLR*. V. 89. P. 2986–2994.
20. Entner D. and Hoyer P.O. (2010) On causal discovery from time series data using FCI. *Proc. of the 5th European Workshop on Probabilistic graphical models*. Helsinki, Finland. P. 121–128.
21. Runge J. (2018) Causal network reconstruction from timeseries: From theoretical assumptions to practical estimation. *Chaos*. V. 28, paper 075310. 20 p.
22. Malinsky D. and Spirtes P. (2018) Causal structure learning from multivariate time series in settings with unmeasured confounding. *Proc. of 2018 ACM SIGKDD Workshop on Causal Discovery*, London, UK. – *PMLR*, V. 92. P. 23–47.

Одержано 19.08.2022

***Про автора*:**

Балабанов Олександр Степанович,
доктор фіз.-мат. наук,
провідний науковий співробітник.
Кількість наукових публікації в українських виданнях – 62.
Кількість наукових публікації в іноземних індексованих виданнях – 13.
Індекс Хірша – 7.
http://orcid.org/0000-0001-9141-9074

Місце роботи автора:

Інститут програмних систем НАН України,
03187, м. Київ-187,
проспект Академіка Глушкова, 40
Тел.: (38)(044) 526-34-20
e-mail: bas@isofts.kiev.ua

Прізвища та ініціали авторів і назва доповіді англійською мовою:
Balabanov O. S.
From temporal data to dynamic causal models

Прізвища та ініціали авторів і назва доповіді українською мовою:
Балабанов О.С.
Від темпоральних даних до динамічних каузальних моделей

Контакти для редактора: м. тел. 095-5405228