Зимова школа із системного аналізу та штучного інтелекту

Національний технічний університет
ДНІПРОВСЬКА ПОЛІТЕХНІКА
1899

Параметричні моделі або автокореляційні залежності – знайти зайвого чи знайти компроміс?

К.т.н., доц. Желдак Тімур Анатолійович,

Завідувач кафедри системного аналізу та управління

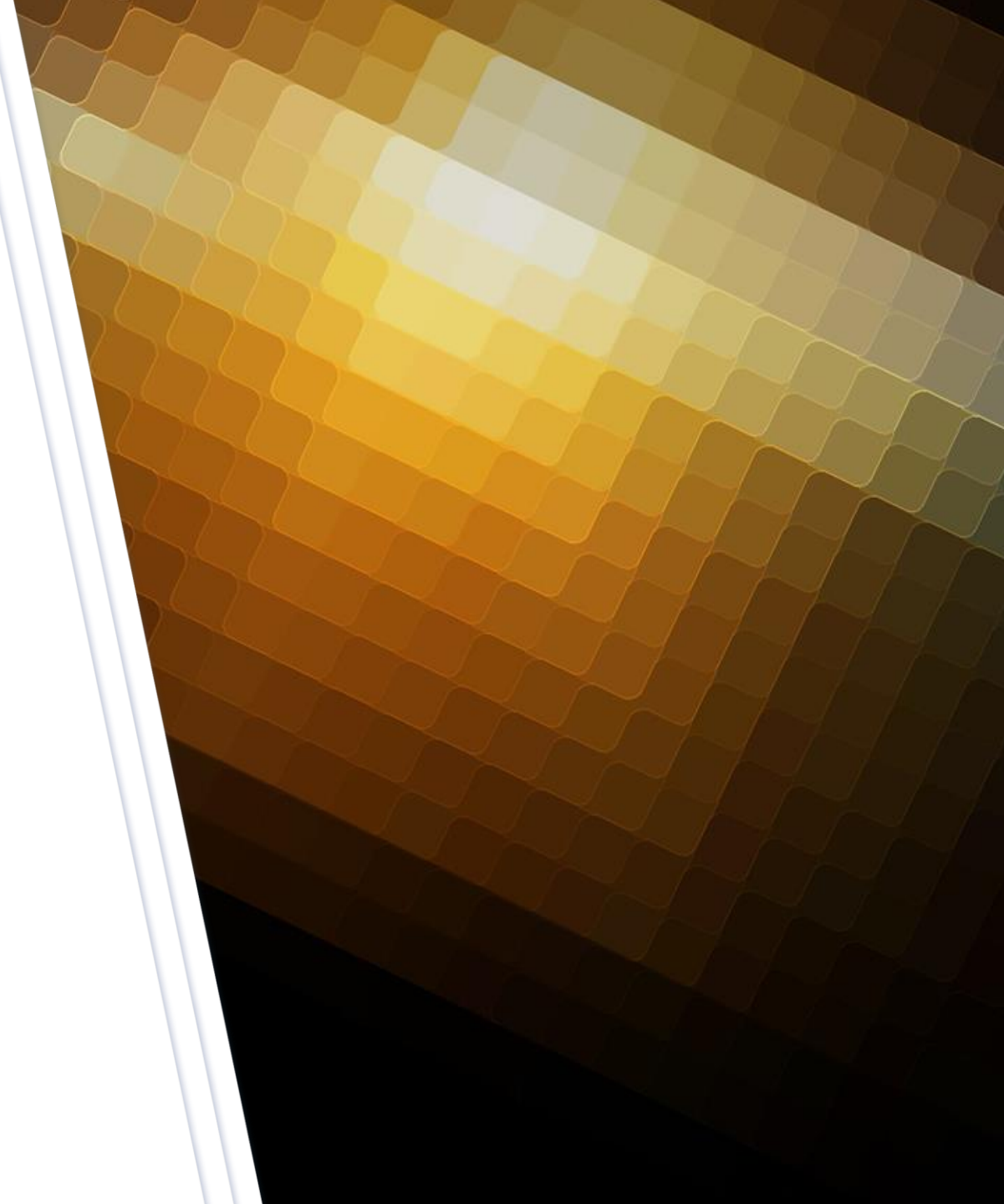
Гаранжа Дмитро Миколайович,

Старший викладач кафедри системного аналізу та управління

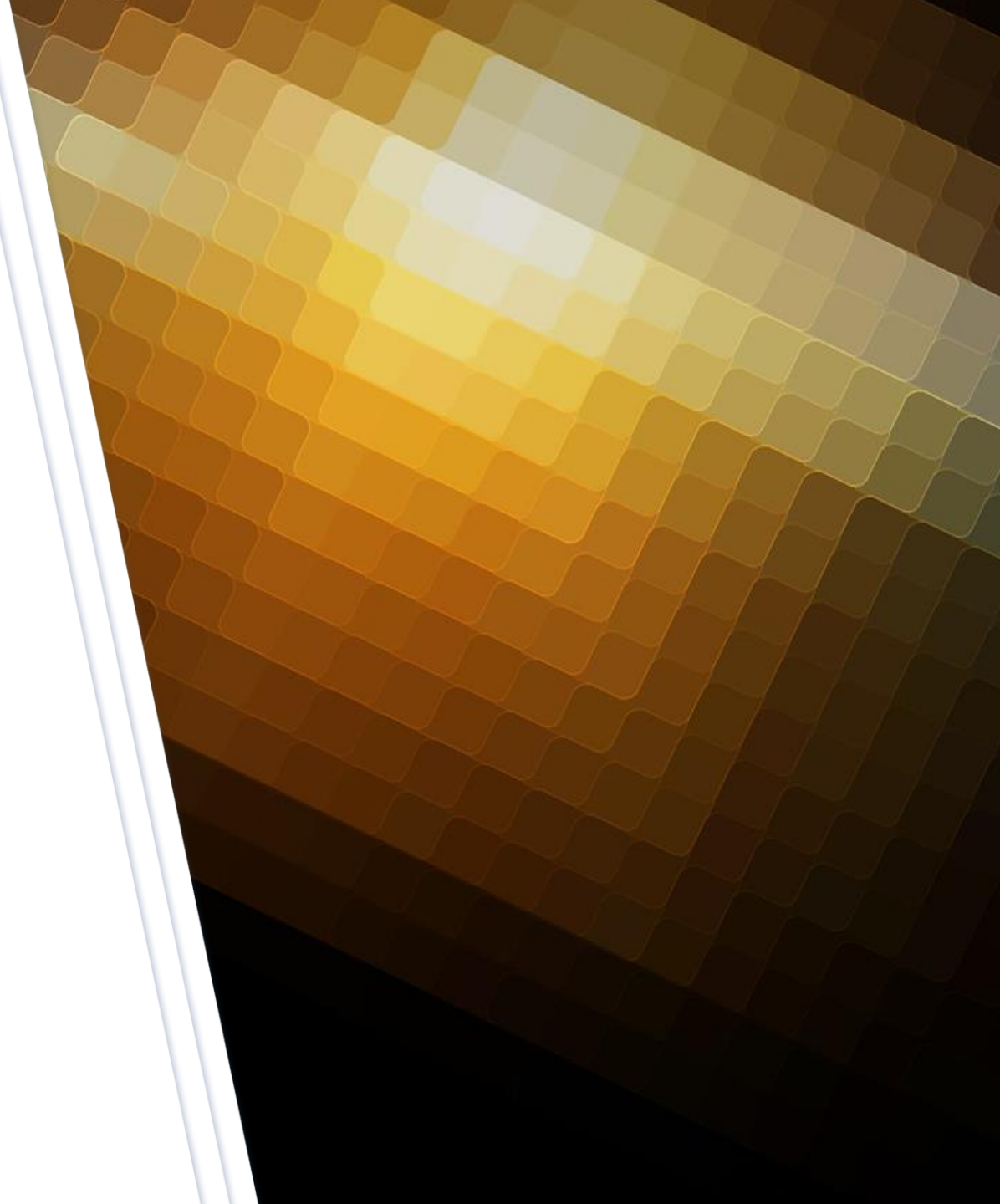
Співавтори:

Зміст

1. Класичний підхід до побудови параметричної моделі прогнозування
2. Класичний підхід до побудови моделі прогнозування часових рядів.
3. Що ми робимо, якщо ми системні аналітики.



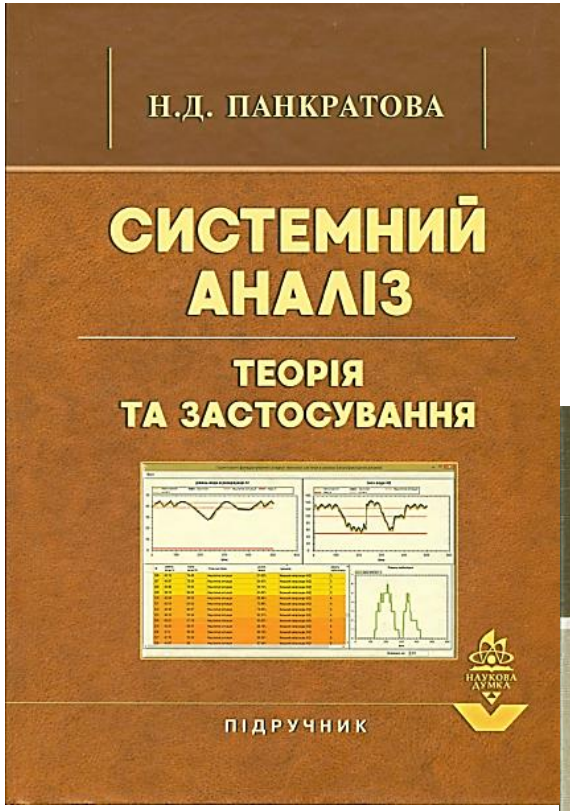
Розділ 1.
Враховуємо фактори, боремося з автокореляцією

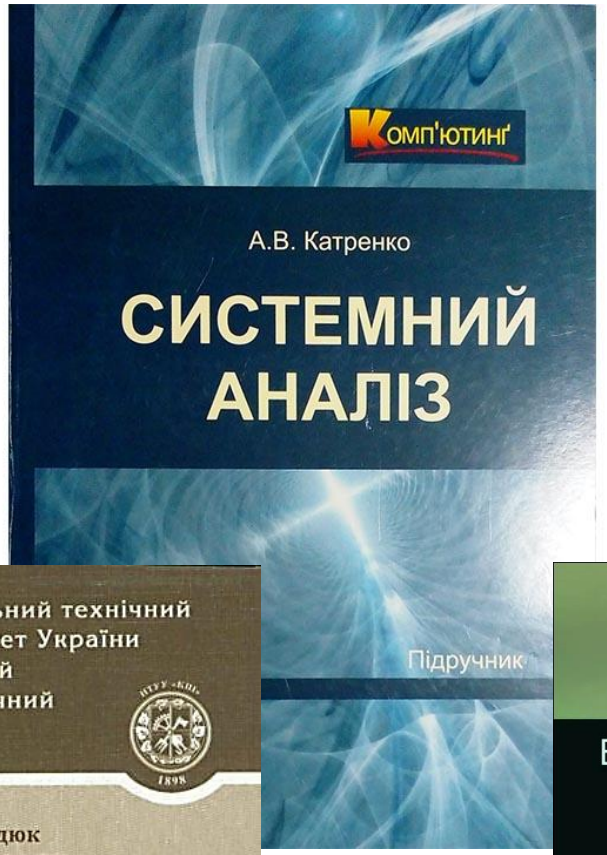


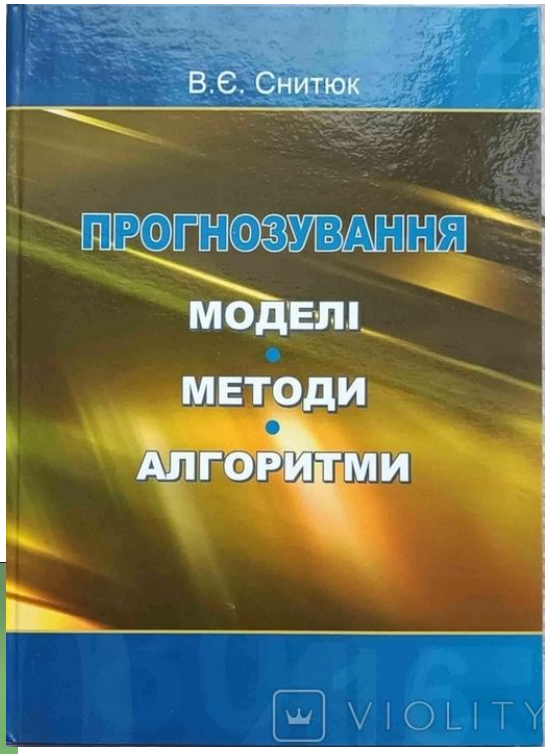
Типова задача – багато стовпчиків, серед яких є дата

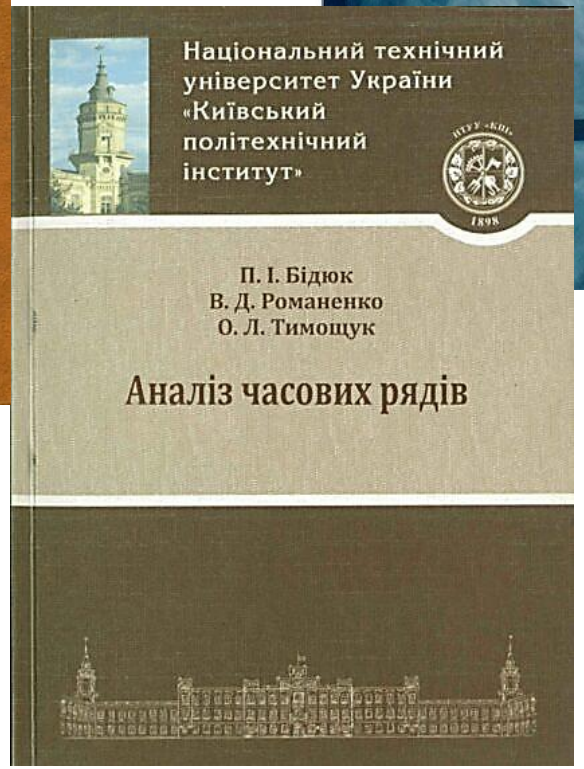
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Date	direct	bing_ad	bing_organic	google_ad	google_organic	email	Social	leads_profit	Impr. (Abs. Top) %	Impr. (Top) %	Desktop	Mobile	Google_Clicks	Google_CPL	Bing_Clicks	Bing_CPL	Conversion_rate
2	27.06.2019	71	57	7	293	37	14	19	921	0.2597	0.6464	203	278	3040	10.37542662	594	10.42105263	6.84
3	28.06.2019	153	49	9	338	31	23	12	916	0.2861	0.6591	241	350	3531	10.44674556	483	9.857142857	7.75
4	29.06.2019	106	50	5	357	19	16	10	854	0.2955	0.69	168	366	2946	8.25210084	509	10.18	6.09
5	30.06.2019	48	47	6	332	23	17	27	775	0.2597	0.6747	105	363	2691	8.105421687	839	17.85106383	7.61
6	01.07.2019	187	77	10	474	45	23	15	1299	0.304	0.6784	335	470	4560	9.620253165	883	11.46753247	8.01
7	02.07.2019	137	93	13	517	33	26	7	1346	0.3446	0.6806	289	507	4552	8.804642166	815	8.76344086	6.54
8	03.07.2019	95	71	28	377	40	23	10	989	0.305	0.6729	233	389	3284	8.710875332	511	7.197183099	7.48
9	04.07.2019	54	49	6	270	33	11	4	689	0.2511	0.6683	108	300	2259	8.366666667	736	15.02040816	6.82
10	05.07.2019	80	61	13	293	46	24	5	732	0.2071	0.6436	177	332	2327	7.941979522	810	13.27868852	6.83
11	06.07.2019	65	58	3	259	29	13	7	678	0.216	0.6625	104	309	2176	8.401544402	551	9.5	5.75
12	07.07.2019	64	51	8	301	29	16	9	761	0.2051	0.6367	116	338	2431	8.07641196	915	17.94117647	5.91
13	08.07.2019	106	90	15	583	68	40	8	1423	0.2999	0.6588	295	572	4930	8.45626072	938	10.42222222	6.04
14	09.07.2019	161	90	16	453	53	18	10	1225	0.2513	0.6447	332	444	3854	8.507726269	729	8.1	6.45
15	10.07.2019	116	73	13	418	55	42	11	997	0.237	0.6433	288	419	3606	8.626794258	593	8.123287671	6.91
16	11.07.2019	92	53	8	413	48	18	14	1100	0.2662	0.6592	207	410	3862	9.351089588	541	10.20754717	7.73
17	12.07.2019	199	57	13	415	50	33	9	967	0.2854	0.6616	320	439	3492	8.414457831	410	7.192982456	5.27
18	13.07.2019	106	30	3	265	32	11	4	714	0.2602	0.6684	134	298	2225	8.396226415	547	18.23333333	5.18
19	14.07.2019	56	41	6	346	35	13	8	892	0.2805	0.6652	120	363	3197	9.239884393	705	17.19512195	6.05
20	15.07.2019	127	67	12	522	62	22	10	1361	0.2589	0.6593	300	496	4528	8.674329502	669	9.985074627	6.25
21	16.07.2019	102	62	17	433	55	22	11	1070	0.2438	0.6545	261	421	3631	8.385681293	683	11.01612903	5.98
22	17.07.2019	113	68	29	445	57	25	12	1132	0.2743	0.6582	289	436	4005	9	629	9.25	7.16
23	18.07.2019	107	54	13	437	48	38	34	1092	0.2927	0.6682	239	459	3806	8.709382151	501	9.277777778	5.77
24	19.07.2019	93	51	9	506	44	32	20	959	0.2721	0.6554	312	421	3272	6.466403162	396	7.764705882	5.32
25	20.07.2019	186	48	7	445	34	16	5	834	0.2906	0.676	336	379	2990	6.719101124	436	9.083333333	5.4
26	21.07.2019	113	57	8	556	43	19	8	934	0.2587	0.6706	306	467	3183	5.724820144	688	12.07017544	5.57
27	22.07.2019	123	106	20	598	76	44	19	1191	0.2394	0.6449	415	529	3982	6.658862876	593	5.594339623	7.64
28	23.07.2019	91	85	13	445	63	21	13	1143	0.2433	0.6418	270	427	3694	8.301123596	540	6.352941176	5.34
29	24.07.2019	159	59	11	440	57	31	10	1026	0.2556	0.6551	294	443	3679	8.361363636	550	9.322033898	7.7
30	25.07.2019	108	58	9	452	45	14	11	1157	0.2635	0.6559	225	443	3964	8.769911504	485	8.362068966	7.61

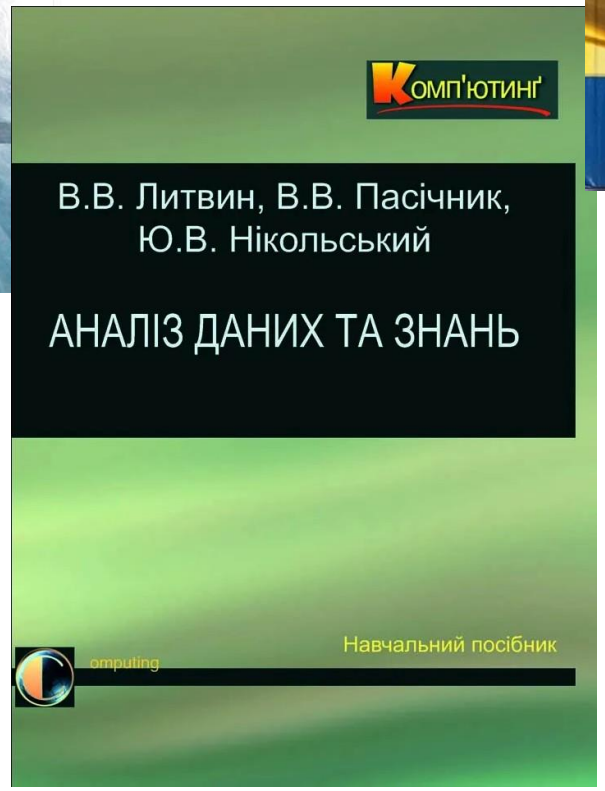
Що пишуть класики?











Три етапи аналізу даних (насправді 4…)

(не залежно, чи є там дата, час, ідентифікатор клієнта…)

Аналіз і попередня обробка даних – preprocessing

Синтез моделі –discovery («вільний пошук»)

Прогностичне моделювання – prediction / forecasting

… Аналіз виключень і відхилень - forensic analysis & deviation detection

Етап 1 – найголовніший: ПРЕПРОЦЕСИНГ

- Нормалізація і стандартизація
- Боротьба з пропусками і різкими відхиленнями
- Вибілювання (збільшення ентропії)
- Виключення мультиколінеарності та гетероскедастичності
- Пониження розмірності

Нормалізувати можна по-різному.

важливо не вихлюпнути дитину….

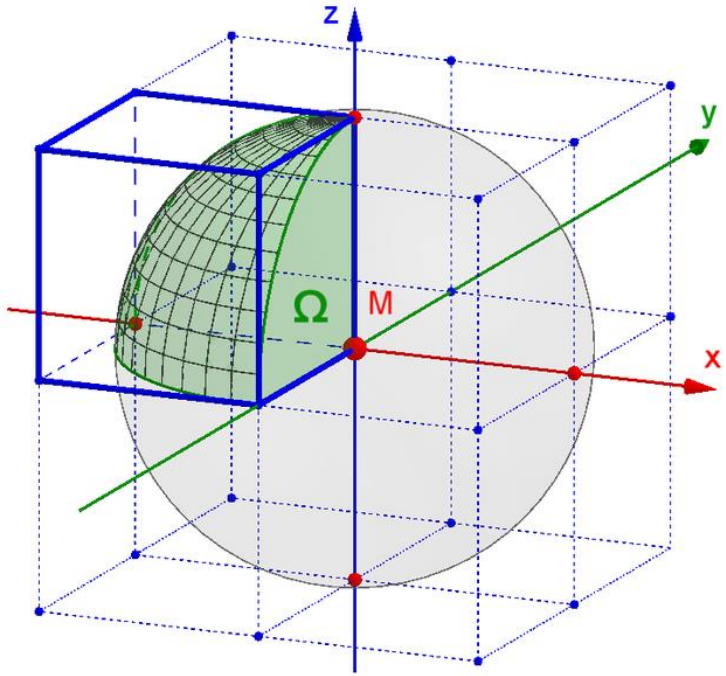
$$x_i' = \frac{x_i - min(x_i)}{max(x_i) - min(x_i)};$$

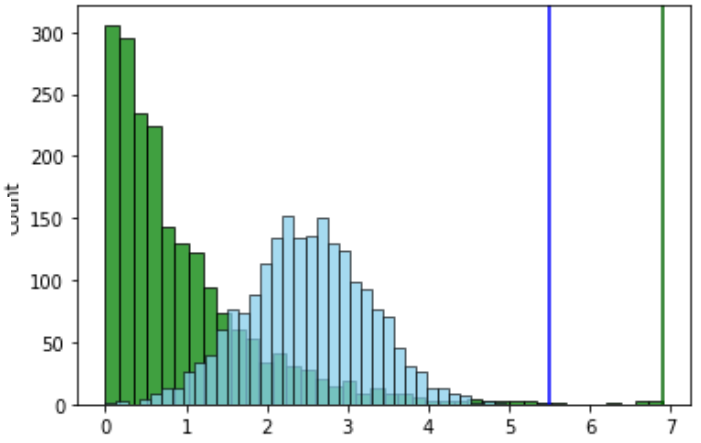
$$x_i' = \frac{max(x_i) - x_i}{max(x_i) - min(x_i)};$$

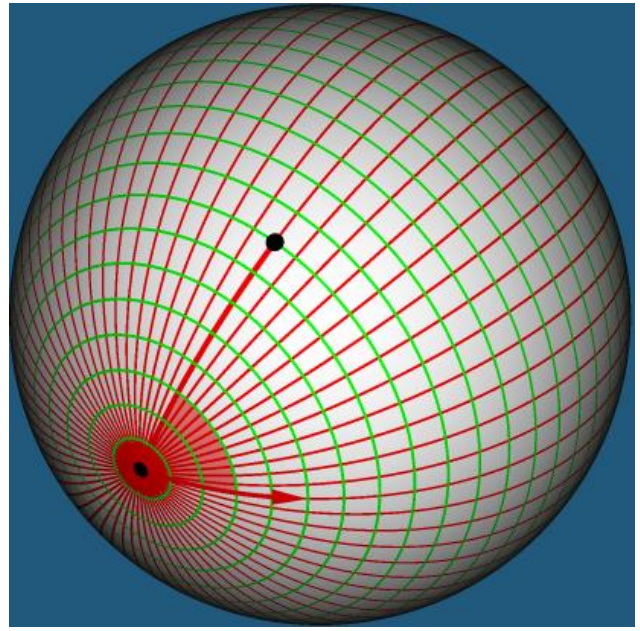
$$x_i' = \frac{x_i - \overline{x_i}}{\sigma_{x_i}};$$

$$x_i' = \frac{2(x_i - min(x_i))}{max(x_i) - min(x_i)} - 1;$$

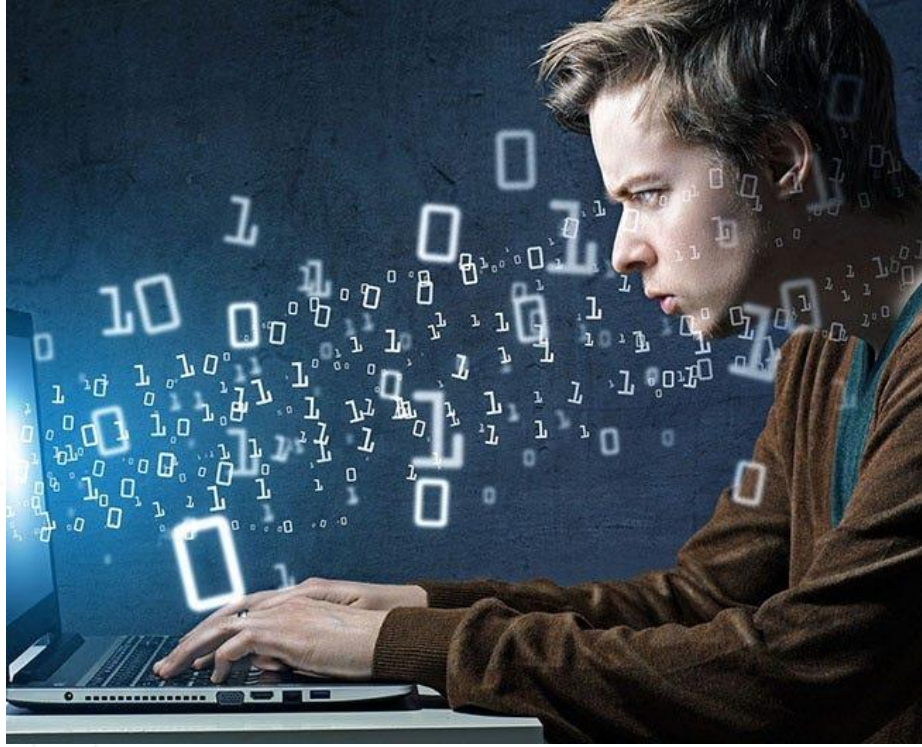
$$x_i' = \frac{1}{1 + e^{-x_i}}.$$







Чому дані можуть бути неінформативні?



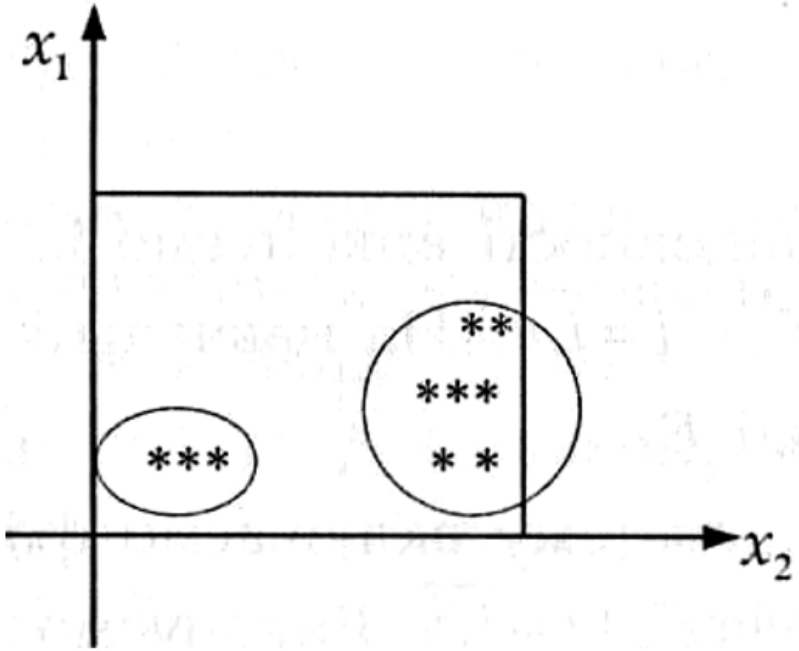


Рис. 1.1 - Неінформативні дані

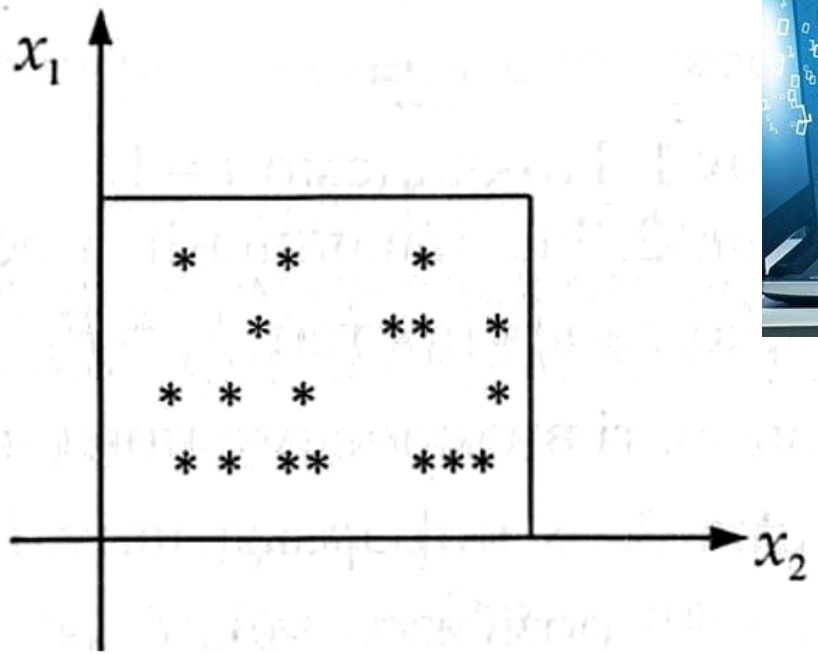


Рис. 1.2 - Інформативні дані

І як їх зробити інформативними? І до чого тут вибілювання

Крок 1. Для кожного вхідного фактору знайдемо його середнє значення

$$\bar{x}_j = \frac{1}{m}\sum_{i=1}^{m} x_{ij}\,, \qquad j = \overline{1,n}. \tag{10}$$

Крок 2. Обчислимо коваріаційну матрицю K, у якій

$$k_{ij} = \frac{1}{m-1}\sum_{l=1}^{m}(x_{li} - \bar{x}_i)\big(x_{lj} - \bar{x}_j\big), \qquad i,j = \overline{1,n}. \tag{11}$$

Крок 3. Визначимо лінійне перетворення, яке діагоналізуватиме коваріаційну матрицю: $KV = \lambda V$, де λ - власні числа матриці K, V - матриця з власних векторів матриці K.

Крок 4. Виконаємо перетворення

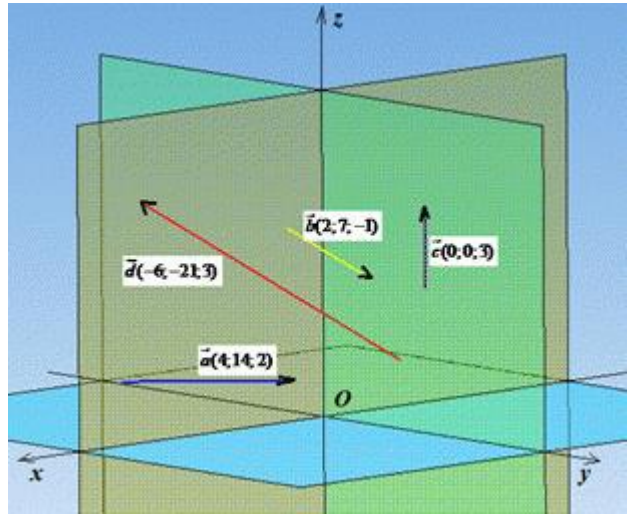
$$\tilde{X} = X_{norm} \cdot \frac{V}{\sqrt{\lambda}}, \tag{12}$$

де матрицю X_{norm} одержують з X відніманням від елементів кожного стовпчика його середнього значення за (10).

Крок 5. Закінчення алгоритму.

А якщо фактори пов’язані за природою?

- Одна змінна є лінійною комбінацією інших
- Ряд змінних утворюють повну суму
- Якась змінна є нелінійною функцією іншої (кількох інших)



1. З критерієм Пірсона визначаємо, чи є в таблиці мультиколінеарність

2. За критерієм Фішера – які змінні в таблиці мають колінеарні пари

Алгоритм Фаррара-Глобера

3. За критерієм Стьюдента – хто кому колінеарний (безпосередньо пара)

З пари колінеарних факторів можна робити згортку, але частіше (бо простіше) – викинути зайвий
Питання: як виявити «зайвого». Дивимось докладніше https://www.youtube.com/watch?v=BBksyLUOLC0

Як це роблять в Пайтоні https://stackoverflow.com/questions/48223443/how-to-run-a-multicollinearity-test-on-a-pandas-dataframe

Ще один крутий кейс https://www.datasklr.com/ols-least-squares-regression/multicollinearity

А якщо при цьому дані ще й залежать від часу?

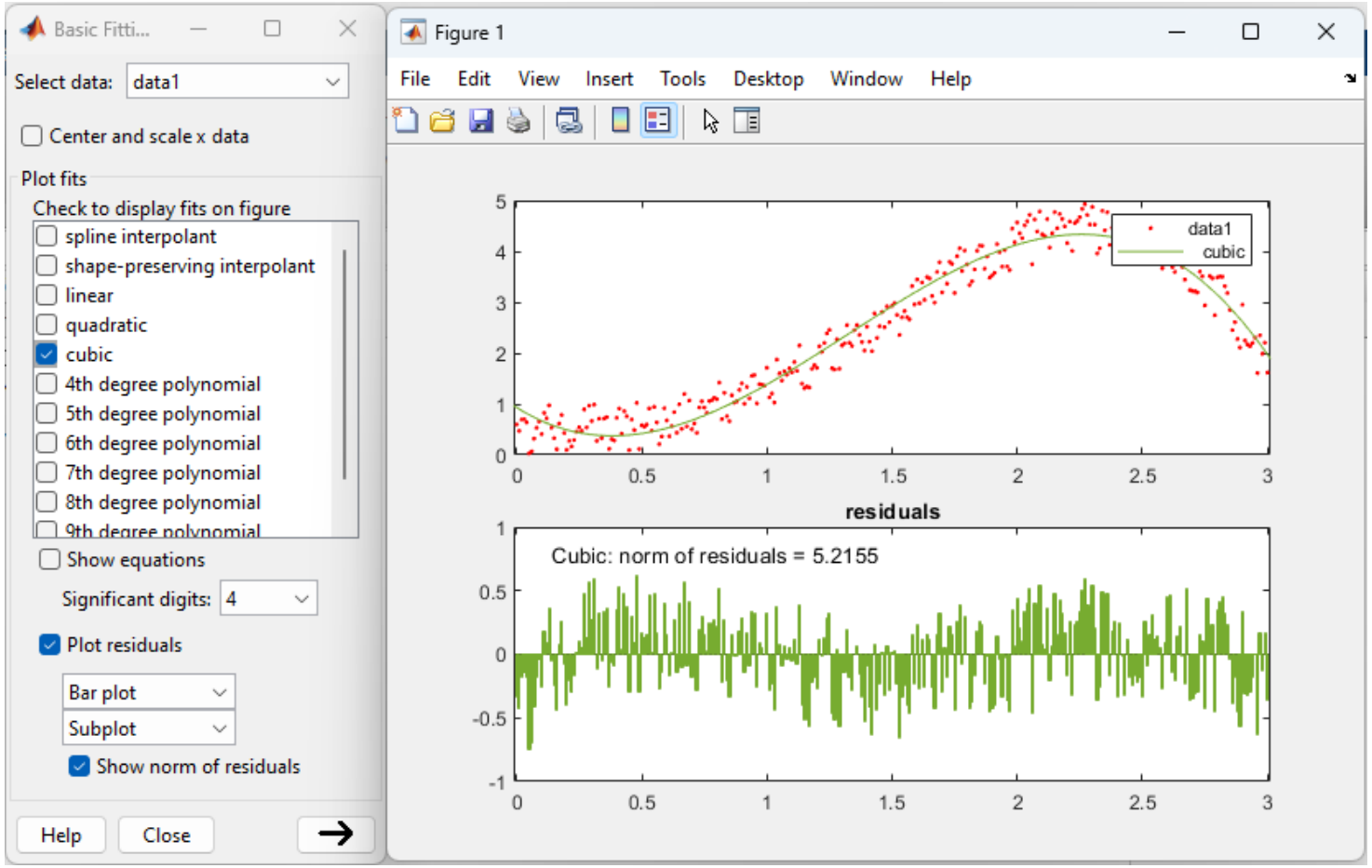
Найголовніше – залишки від застосування моделі до даних (відхилення)

Параметрична модель **некоректна**, якщо

1) Закон розподілу залишків міняється залежно від частини графіка (зліва в рази більше ніж праворуч або навпаки)

2) В залишках присутня автокореляція (має місце бодай якась періодичність)

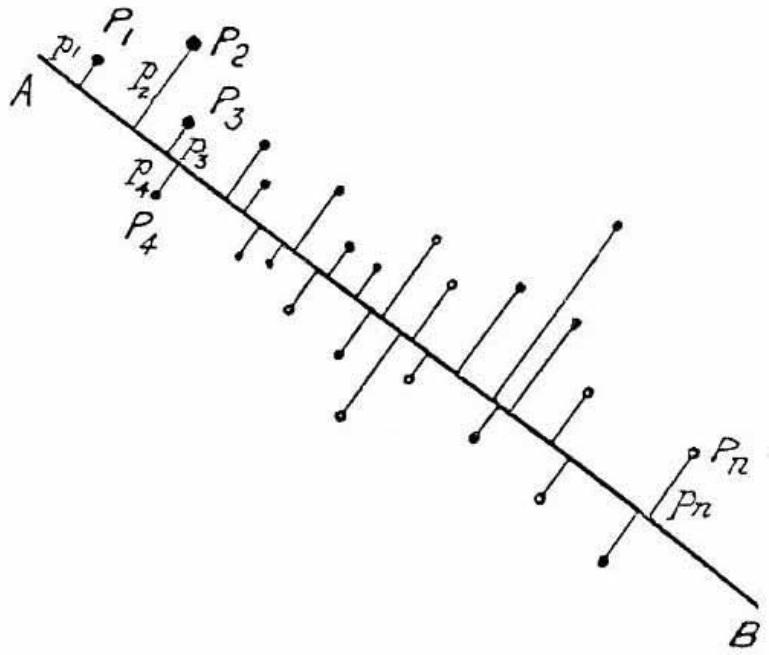
ВИХІД:
- Або розглядати як часовий ряд і відповідні методи
- Або тасувати дані для виключення часової складової



Дивимось докладніше https://www.youtube.com/watch?v=yOd1MG6xO00

Що робити, якщо фактори всі коректні, але їх забагато? (заскладні моделі, проблеми коефіцієнтів)

Метод головних компонентів (МГК, англ. principal component analysis, PCA) — метод факторного аналізу в статистиці, який використовує ортогональне перетворення множини спостережень з можливо пов'язаними змінними (сутностями, кожна з яких набуває різних числових значень) у множину змінних без лінійної кореляції, які називаються головними компонентами.



Метод головних компонентів — один з основних способів зменшити розмірність даних, втративши найменшу кількість інформації. Винайдений Карлом Пірсоном у 1901 році та доповнений і розширений Гарольдом Готелінґом в 1933 р. Застосовується в багатьох галузях, зокрема, в економетриці, біоінформатиці, обробці зображень, для стиснення даних, у суспільних науках.

Обчислення головних компонент може бути зведене до обчислення сингулярного розкладу матриці даних або до обчислення власних векторів і власних чисел коваріаційної матриці початкових даних.

А якщо без злої математики і ортонормальних базисів….

Вектори головних компонент можуть бути знайдені як розв'язки однотипних задач оптимізації:

1. Централізуються дані (відніманням середнього): $x_i := x_i - \overline{X}$. Тепер $\sum_{i=1}^{m} x_i = 0$;
2. Відшукується перша головна компонента як розв'язок задачі:

$$a_1 = \underset{\|a_1\|=1}{\operatorname{argmin}} \left(\sum_{i=1}^{m} \|x_i - a_1(a_1, x_i)\|^2 \right).$$

якщо розв'язок не єдиний, то вибирається один з них.

3. З даних віднімається проєкція на першу головну компоненту:

$$x_i := x_i - a_1 (a_1, x_i);$$

4. Відшукується друга головна компонента як розв'язок задачі:

$$a_2 = \underset{\|a_2\|=1}{\operatorname{argmin}} \left(\sum_{i=1}^{m} \|x_i - a_2(a_2, x_i)\|^2 \right).$$

Якщо розв'язок не єдиний, то вибирається один з них.

Далі процес триває, тобто на кроці $2k-1$ віднімається проєкція на $(k-1)$-у головну компоненту (до цього моменту проєкції на попередні $(k-2)$ головні компоненти вже відняті):

$$x_i := x_i - a_{k-1} (a_{k-1}, x_i);$$

і на кроці $2k$ визначається k-а головна компонента як розв'язок задачі:

$$a_k = \underset{\|a_k\|=1}{\operatorname{argmin}} \left(\sum_{i=1}^{m} \|x_i - a_k(a_k, x_i)\|^2 \right)$$ (якщо розв'язок не єдиний, то вибирається один з них).

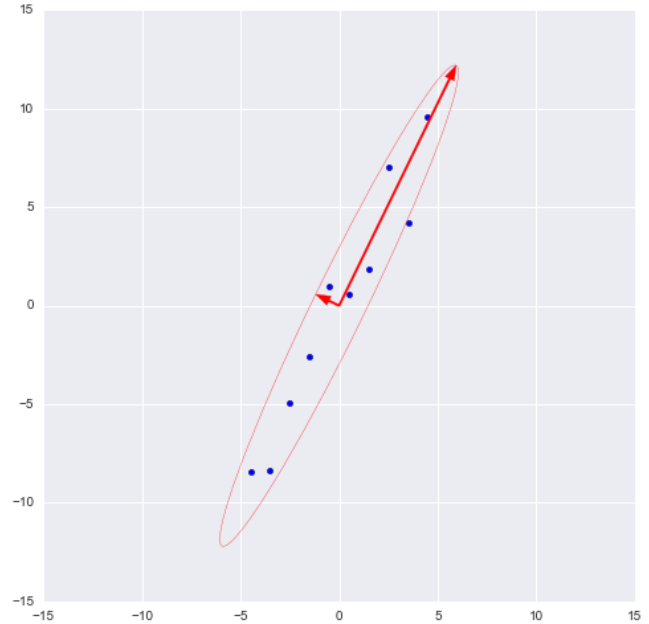
І що це дає на практиці?

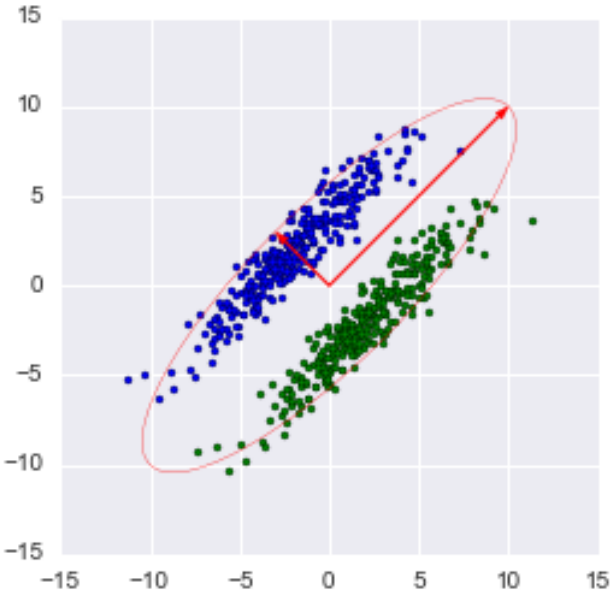
Значне підвищення наочності представлення вихідних даних, що досягається їх проектуванням на спеціальним чином визначений одно-, дво- чи тривимірний простір.

Зменшення на порядки складності досліджуваних моделей, що одночасно дозволить спростити розрахунки та інтерпретацію моделей.

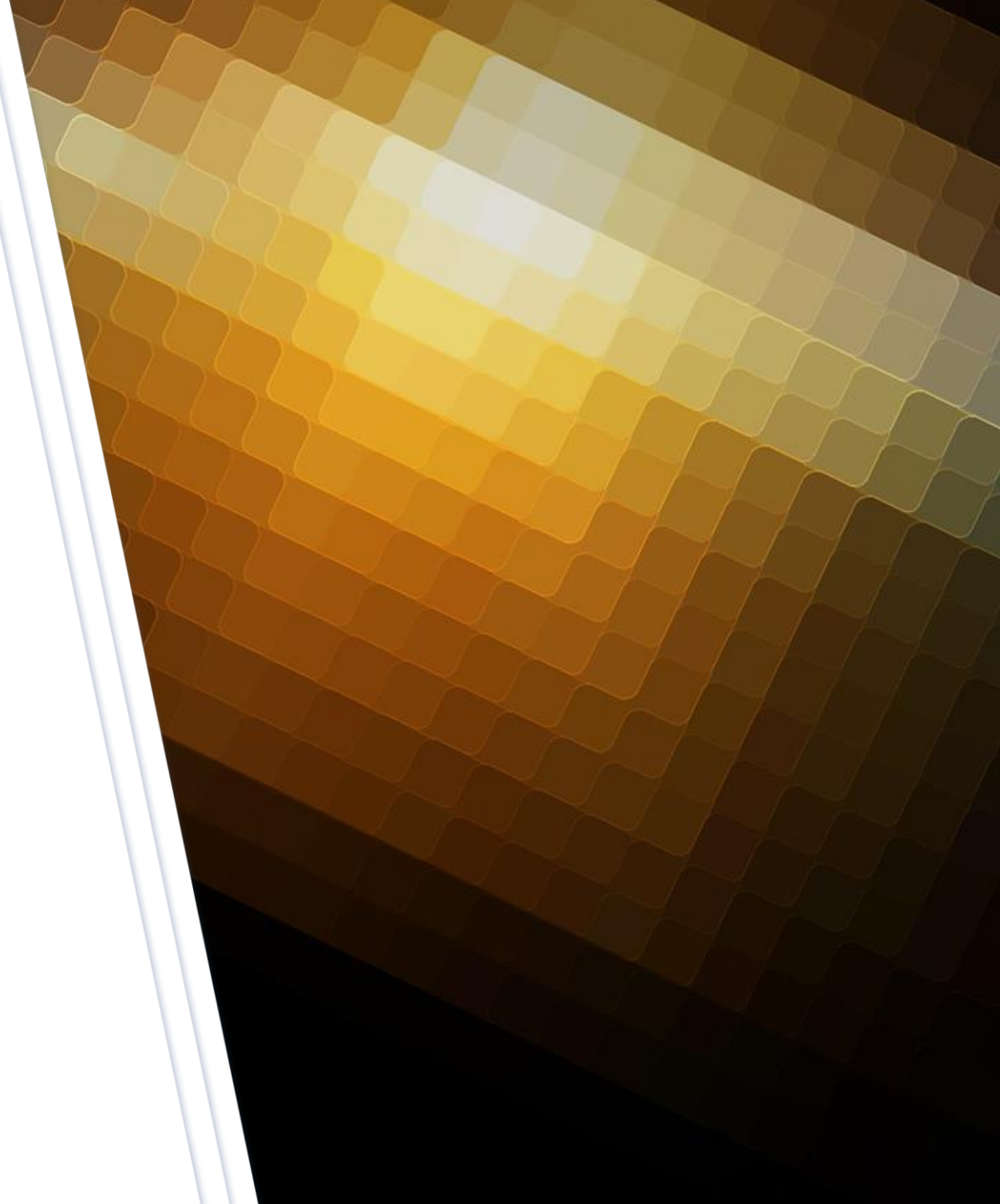
Ущільнення та стиснення обсягів статистичної інформації.

Для методу головних компонент існують стандартні функції
https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html





Розділ 2.
Вивчаємо історію, нехтуємо факторами



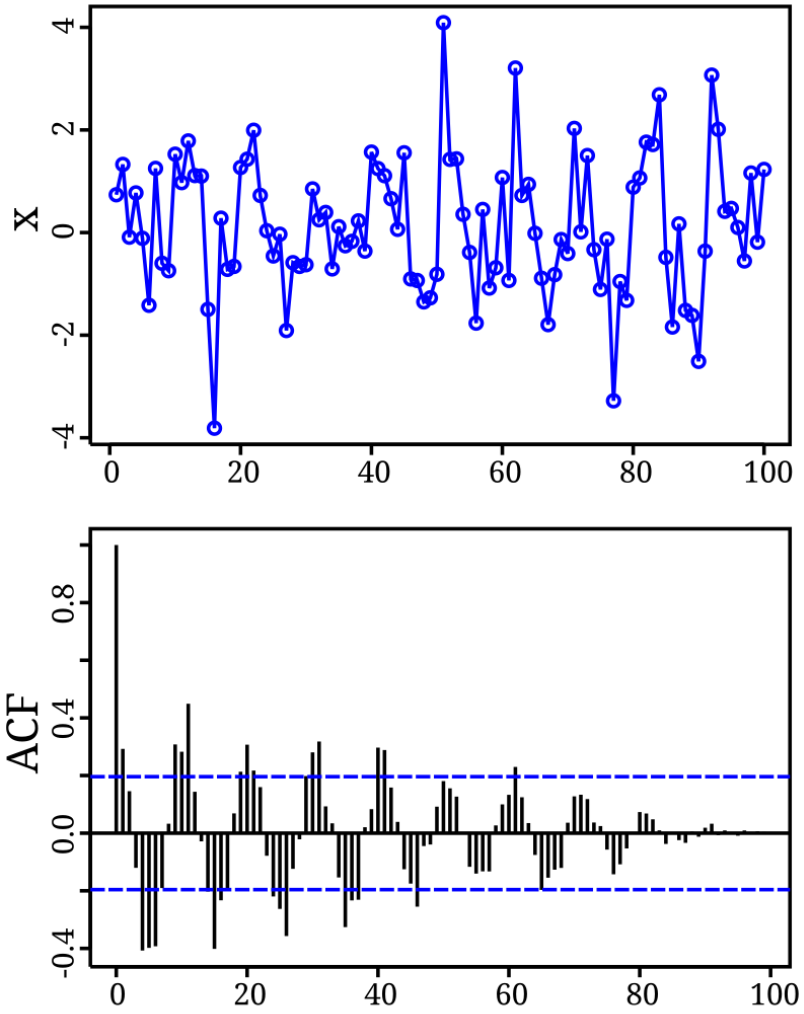
Якщо сьогодні визначається нашим вчора…

(насправді майже завжди)

Автокореляція (англ. autocorrelation), іноді відома як послідовна кореляція (англ. serial correlation), у випадку дискретного часу — це кореляція сигналу із затриманою копією самого себе як функція від затримки.

Неформально — це схожість між спостереженнями як функція від відставання в часі (англ. time lag) між ними.

Аналіз автокореляції — це математичний інструмент для пошуку повторюваних закономірностей, таких як наявність періодичного сигналу

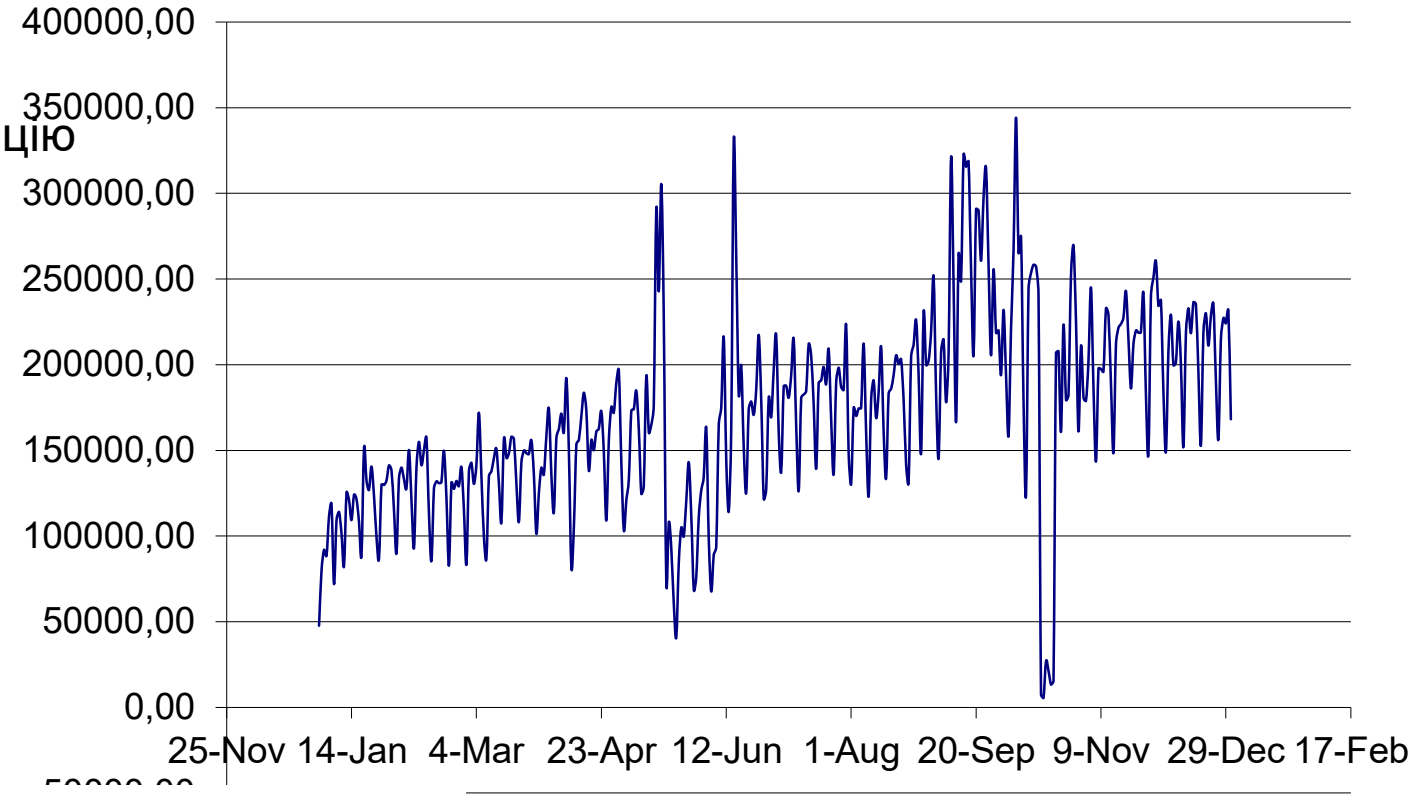


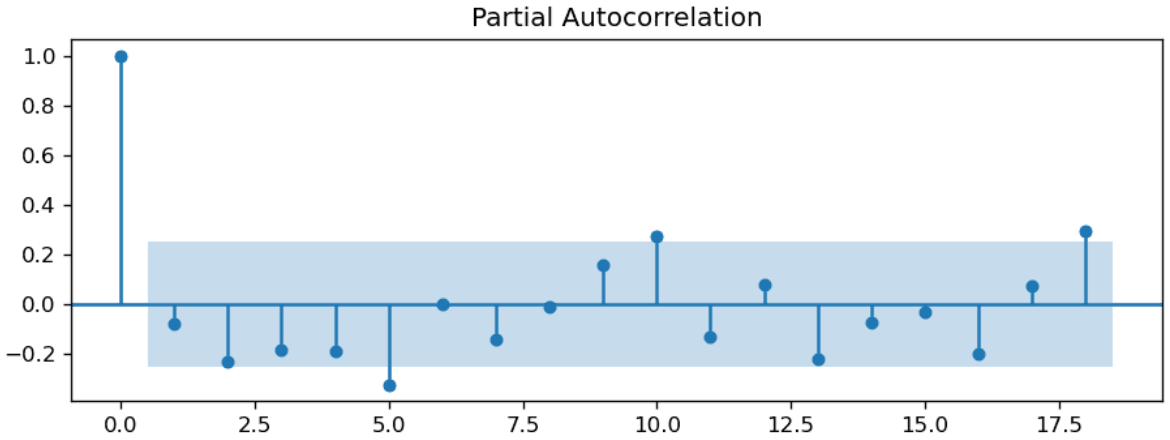
Як діяти?

Автокореляційна функція (ACF) показує кореляцію між точками часового ряду на різних лагах.

Допомагає визначити довжину сезонних компонентів.

Найчастіші випадки:
- Тижнева (5 або 7 днів);
- Добова (часи «пік» і спади);
- Місячна (дні зарплат, сплати податків,…)
- Річна (відпустки, посівна, сесії,…)

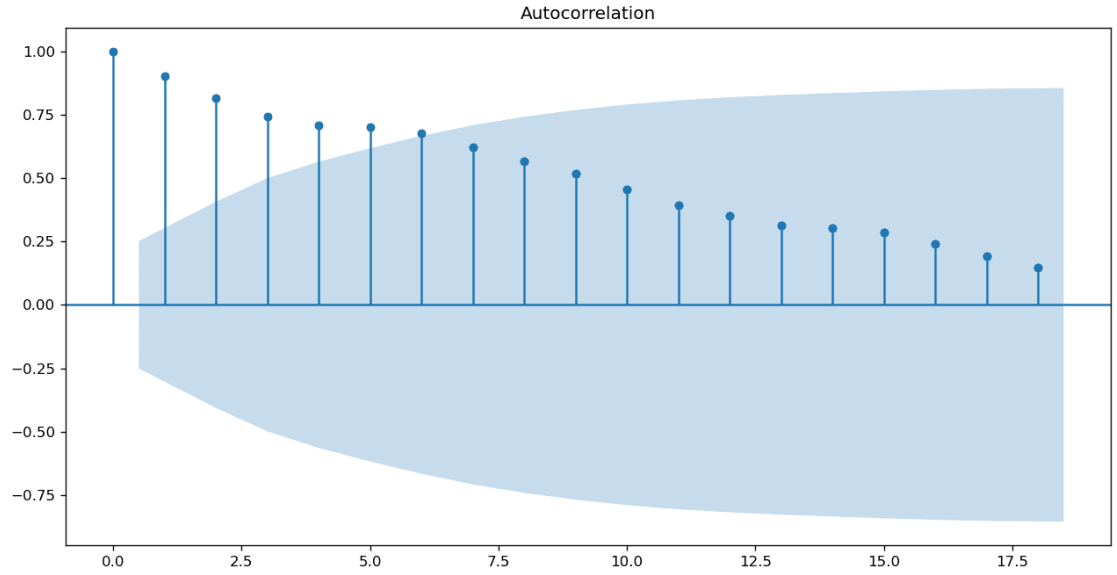


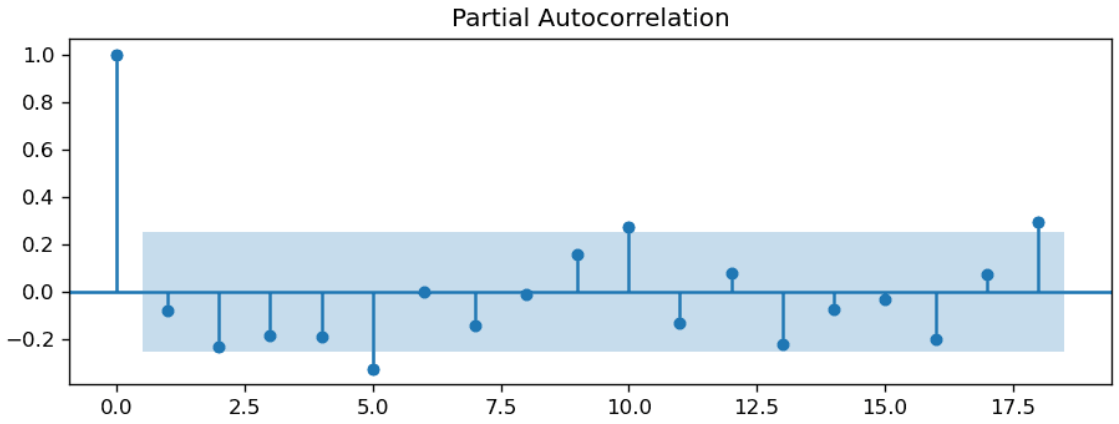


Автокореляція – повна і часткова

Можна спиратися лише на попередні значення, можна на попередні помилки прогнозів, а можна на все разом

	Теоретичні моделі ACF і PACF	
Тип моделі	Типова модель ACF	Типова модель PACF
AR (p)	Експоненційний розподіл або синусоїдна структура, або одночасно	Значні стрибки через лаги p
MA (q)	Значні стрибки через лаги q	Знижується в геометричній прогресії
ARMA (p,q)	Експоненційний розподіл	Експоненційний розподіл





ARMA, ARIMA, SARIMA,…

ARMA (Autoregressive moving average) - we математична модель, яка використовується для аналізу і прогнозування стаціонарних рядів в статистичних даних. ARMA об’єднує в собі прості моделі часових рядів: авторегресії (AR) і ковзне середнє (MA). Тобто, це стаціонарний процес виду:

$$y_t = c + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \ldots + \beta_p y_{t-p} + \varepsilon_t + a_1 \varepsilon_{t-1} + \ldots + a_q \varepsilon_{t-q}$$

Де ε – білі шуми, c — константа, і ми припускаємо, що сума p + q мінімальна можлива.

Позначається модель як ARMA (p, q), перший параметр p – це кількість лагів по авторегресії (AR), а q – це кількість лагів за змінним середнім (MA).

ARIMA (Auto Regressive Integrated Moving Average) – розвиток ARMA. Integrated – процес, який призводить часовий ряд до стаціонарності.
Кожен з компонентів вказаний в якості параметрів моделі. Використовується стандартне позначення моделювання ARIMA (p, d, q). Тобто прогнозоване значення = константа + лінійна комбінація лагів (p) + лінійна комбінація лагів прогнозованої помилки (q) + ступінь нестаціонарності (d).

Докладніше почитати:

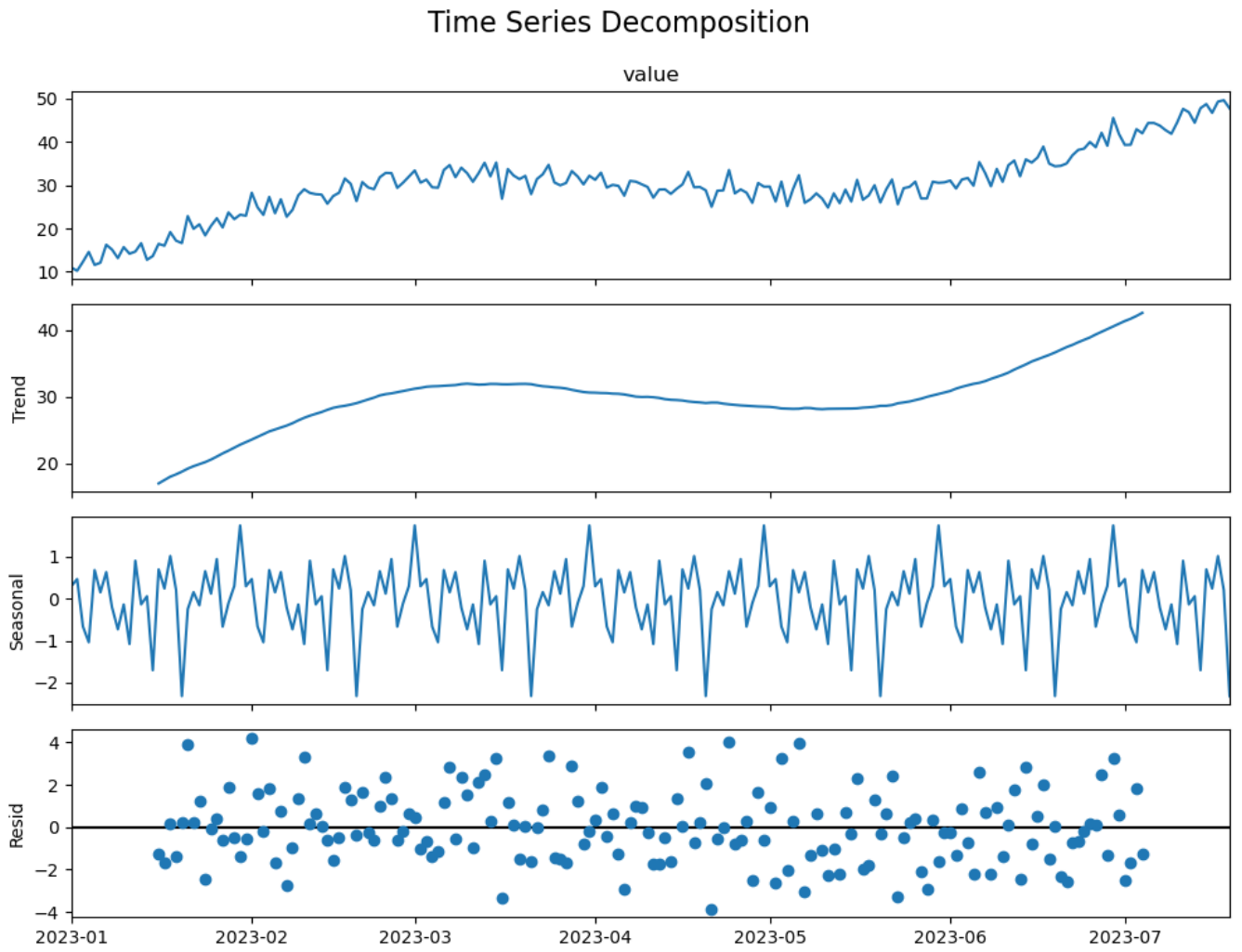
ARMA https://caseware.com.ua/arma-3/
ARIMA https://caseware.com.ua/arima-3/

Але ж є ще сезонність, і з нею буває складніше всього

SARIMA (Seasonal Autoregressive Integrated Moving Average) є універсальною та широко використовуваною моделлю прогнозування часових рядів.
Це розширення несезонної моделі ARIMA, розробленої для обробки даних із сезонними моделями.
SARIMA фіксує як короткострокові, так і довгострокові залежності в межах даних, що робить її надійним інструментом для прогнозування.

Ефективне застосування (правда з помилкою)
https://javascript.org.ua/analiz-chasovih-ryadiv-za-dopomogoyu-statsmodels-v-python/

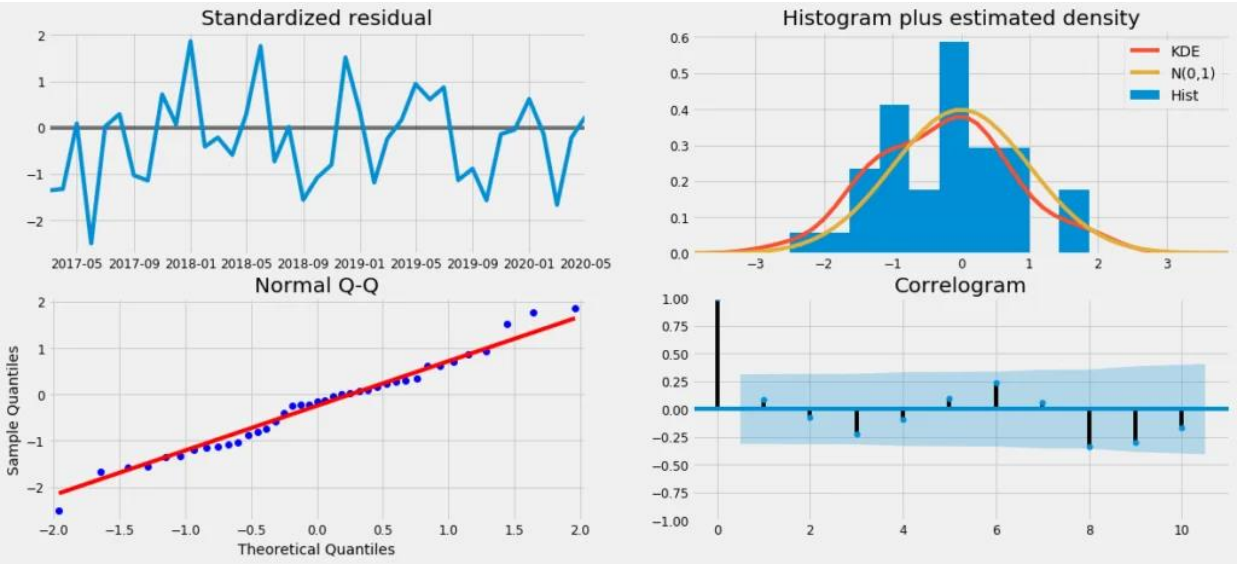
Докладно про повну модель
https://www.geeksforgeeks.org/sarima-seasonal-autoregressive-integrated-moving-average/



Автокореляційні моделі – це таки погано чи добре?

Переваги:
- Простота обчислень;
- Врахування трендів, сезонності, нестаціонарності та випадкових складових

Недоліки:
- Непридатність для сильно шумових даних.
- Труднощі в ідентифікації тренду (поліном, ряд, …).
- Висока залежність від правильного вибору параметрів.
- Не враховує зовнішні фактори



Яка модель найкраща, знає критерій Акаіке

$$AIC = 2k - n\,ln\left(\sum_{i=1}^{n}(\overline{y_i} - y_i)^2/n\right),$$

Всі методи автокореляційного аналізу реалізує бібліотека statsmodels.tsa
https://www.statsmodels.org/stable/tsa.html

Завдання для груп №1 (80 хвилин, хто готовий, розпочинає)

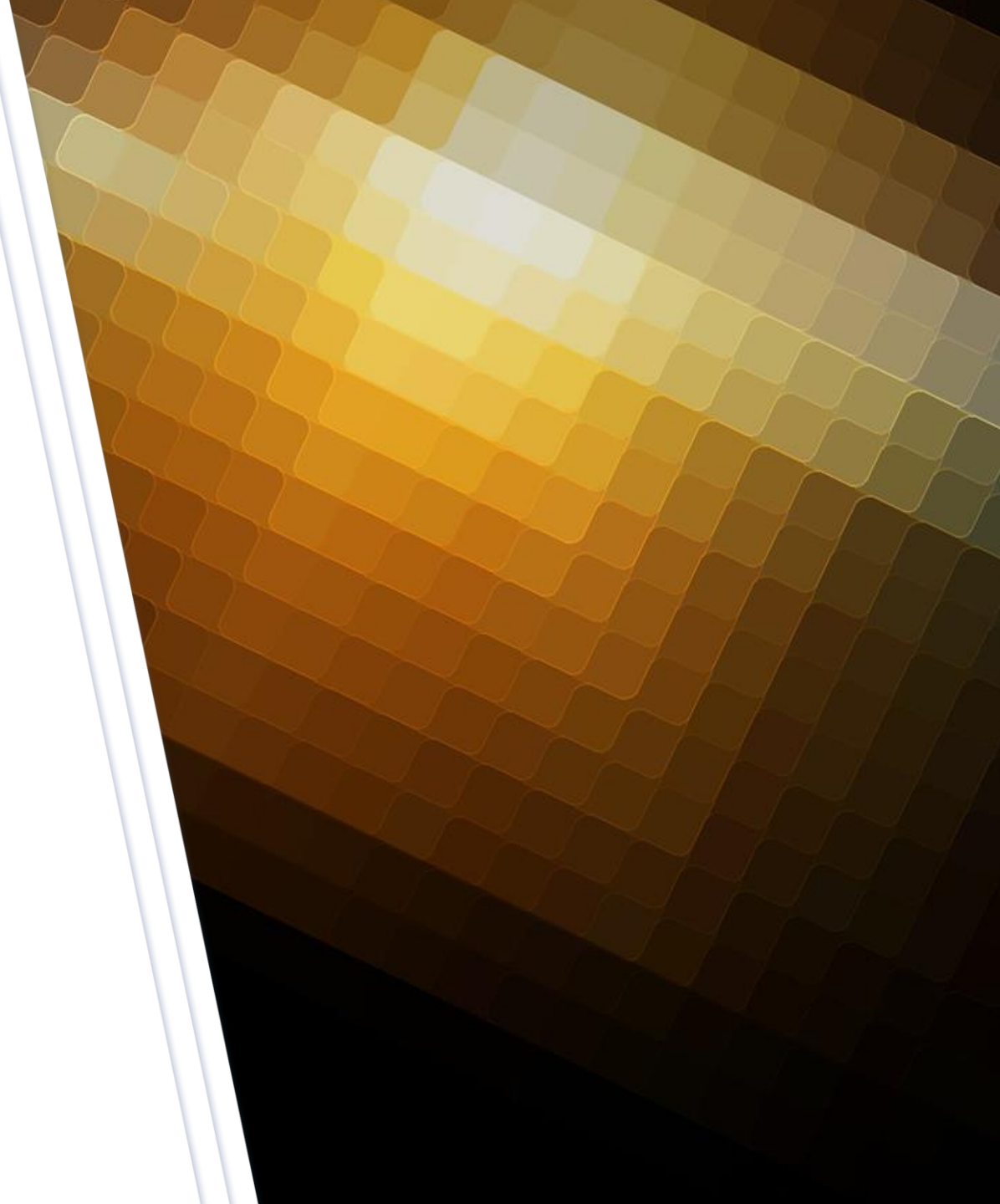
Дано:
- Файл з 14 показниками, доступними для реєстрації, які можуть бути факторами (а можуть і не бути!), що впливають на єдиний цільовий показник – Conversion_Rate
- Дані супроводжуються міткою дати, яка є реальною датою.

Необхідно:
- Побудувати довільну модель (параметричну чи автокореляційну) таку, що на 16 днях лютого 2021 року забезпечить найкращий прогноз
- - за максимальною відносною похибкою;
- - за середньою відносною похибкою.
- Модель повинна мати мінімальну кількість вільних коефіцієнтів.
- Всі коефіцієнти моделі мають бути значимими

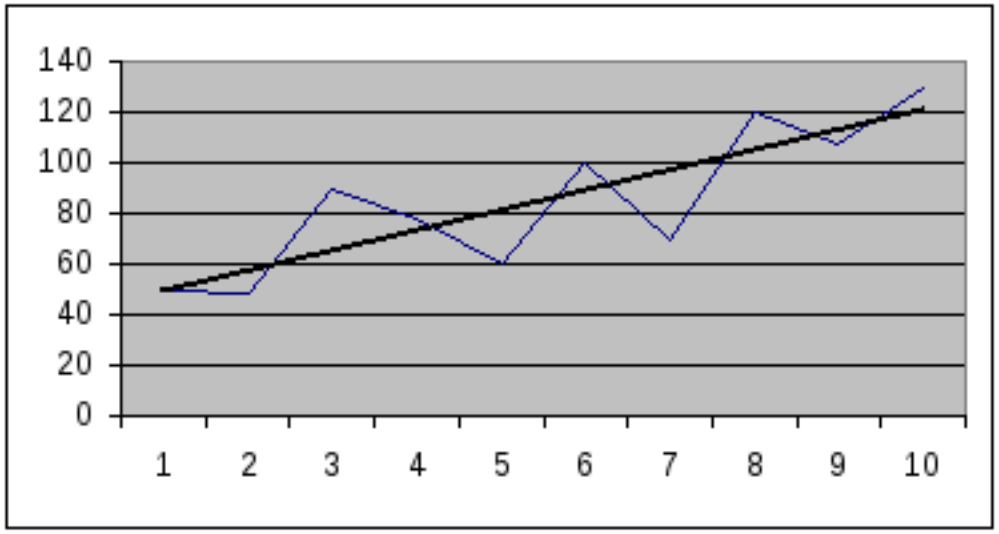
Обмежень по мові програмування, середовищу для реалізації, використаних (або не використаних) методах та відкинутих змінних немає.

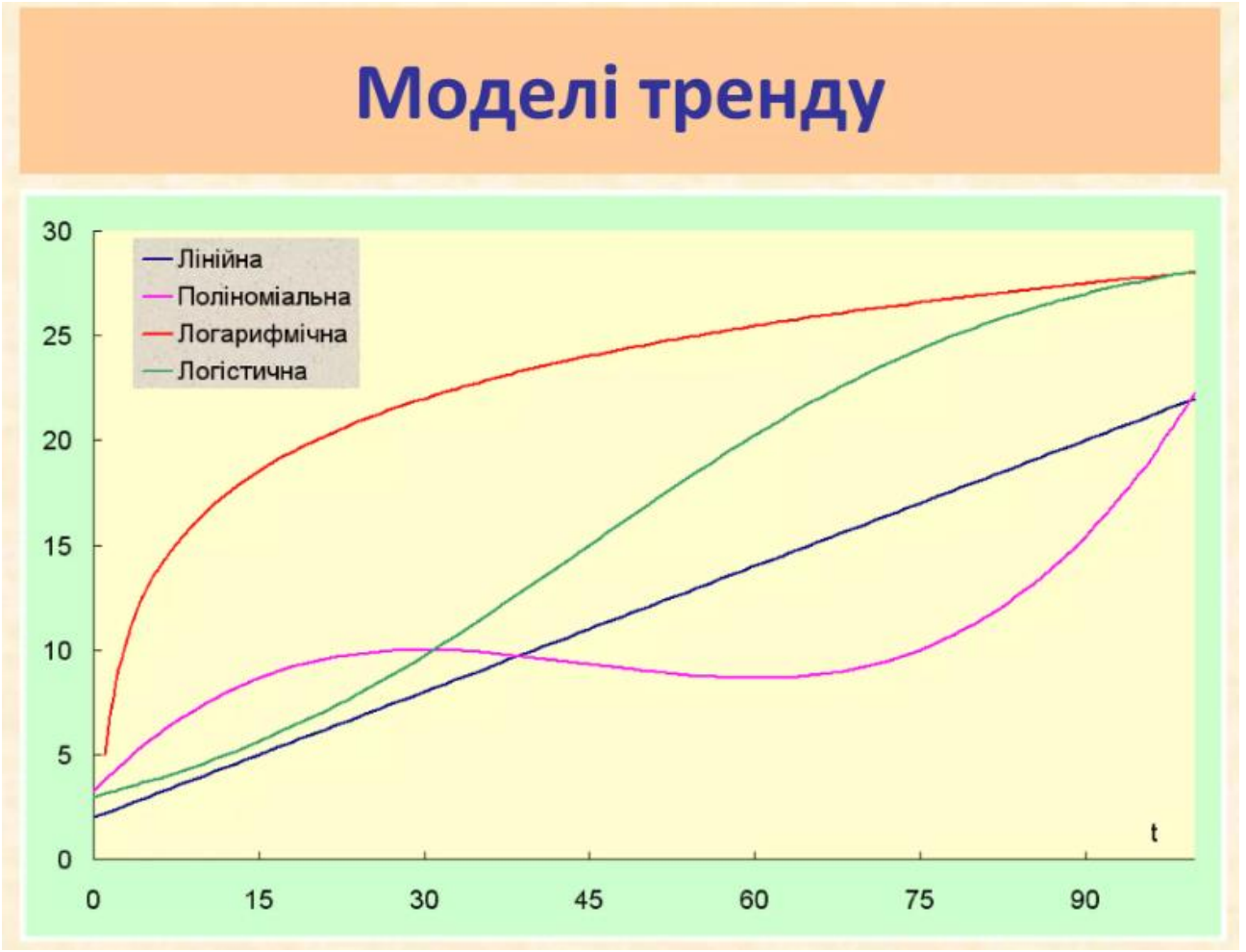
Оцінка складається з якості рішення (60%) і його обґрунтування (40%).

Розділ 3.
Тренди і згортки

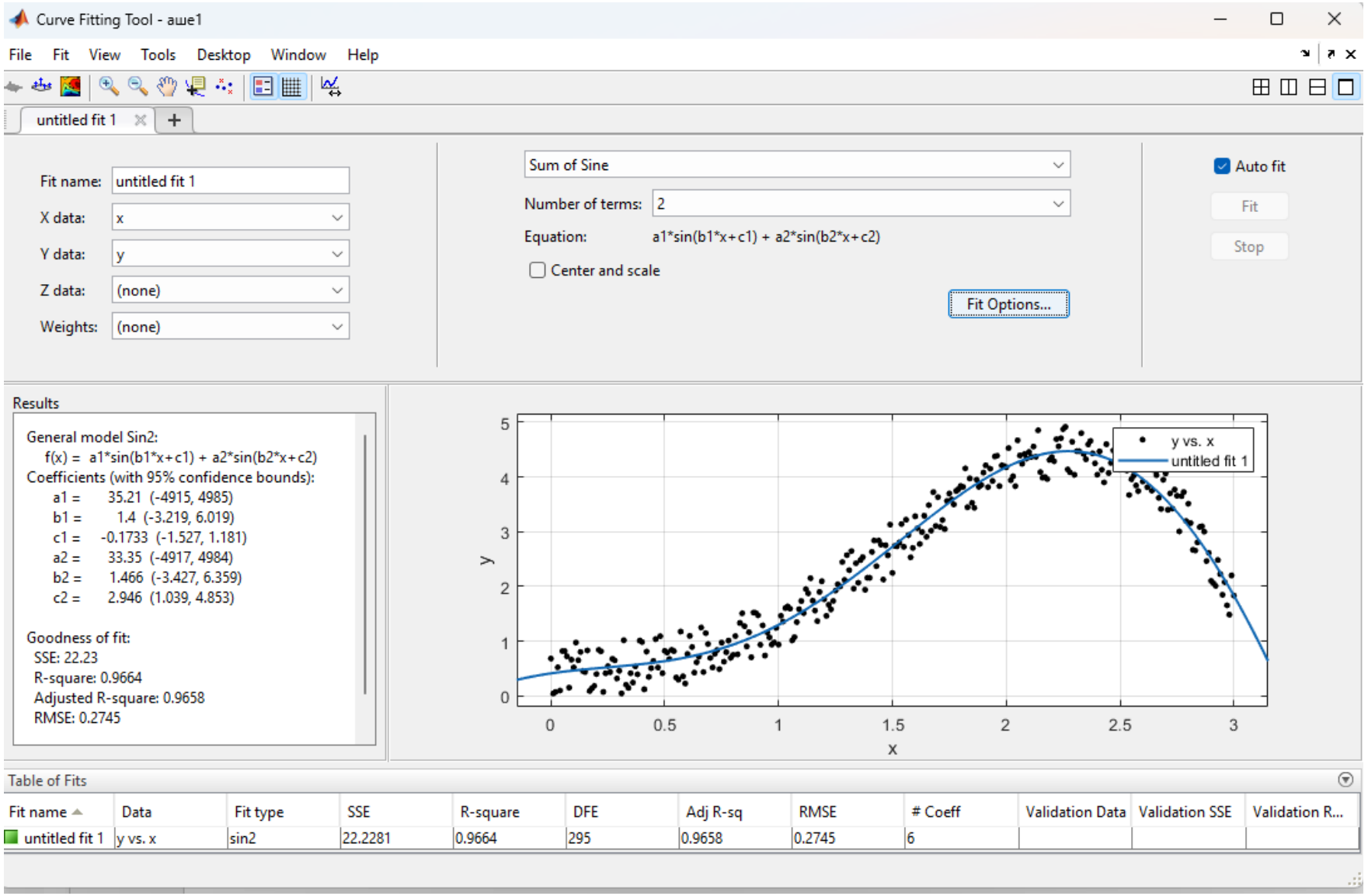


Тренд – яка тенденція у числового ряду?

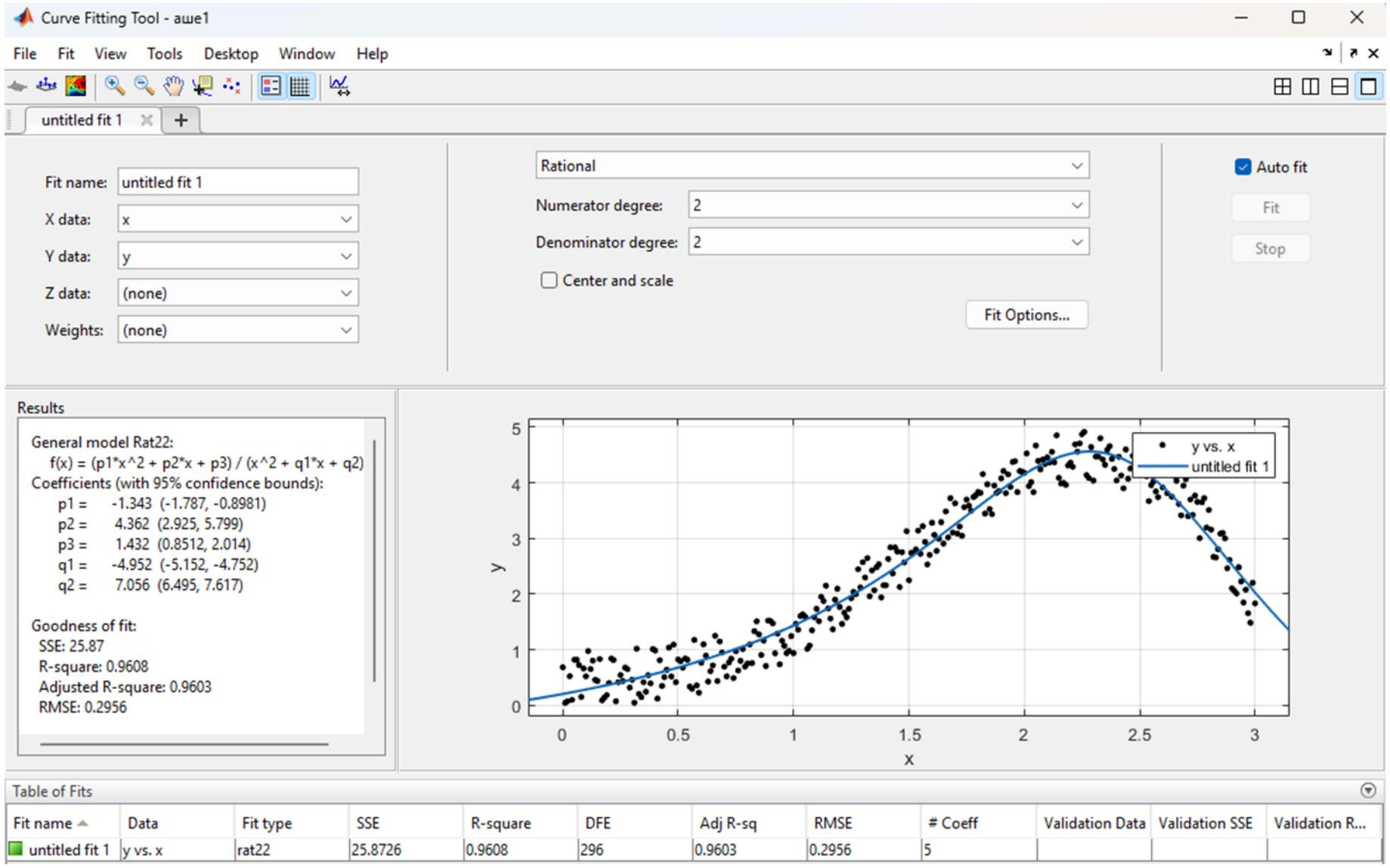




Якщо дані не тривіальні, може бути складніше



І який тренд кращий?

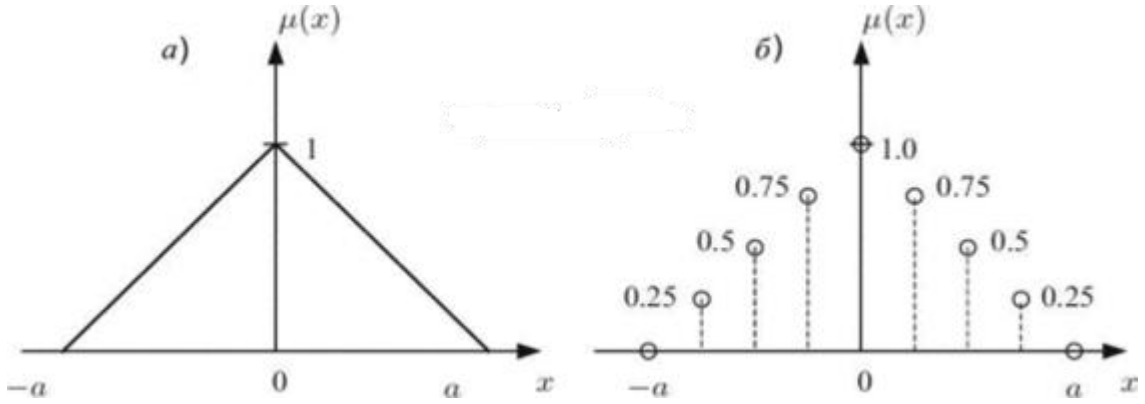


У найпростішому випадку – згладжування!

$$\tilde{y}_i = (2y_{i-2} + 4y_{i-2} + 8\, y_i + 4\, y_{i+1} + 2\, y_{i+2})/20$$

Також відомі фільтри:

- параболічний;
- Чебишева;
- Баттерворта;
- еліптичний

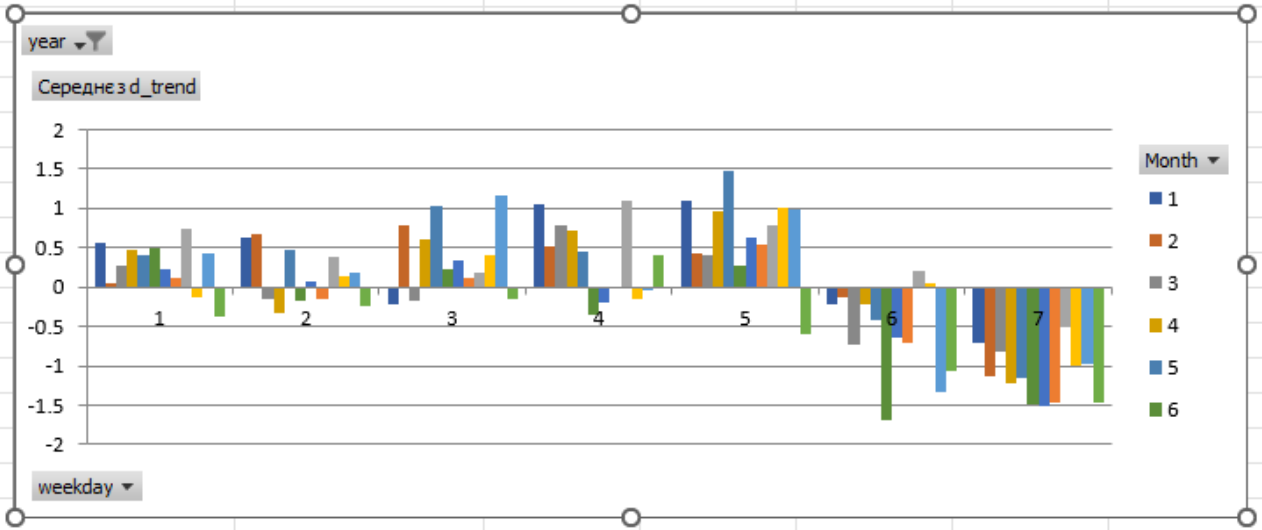


Якщо сезонність явна, її можна описати коефіцієнтами

	A	B	C	D	E	F
1	DATE JOINED	PERCENT	year	Month	week	weekday
2	01.01.2018	3.61	2018	1	1	1
3	02.01.2018	6.43	2018	1	1	2
4	03.01.2018	6.2	2018	1	1	3
5	04.01.2018	5.38	2018	1	1	4
6	05.01.2018	4.7	2018	1	1	5
7	06.01.2018	3.74	2018	1	1	6
8	07.01.2018	1.89	2018	1	2	7
9	08.01.2018	5.12	2018	1	2	1
10	09.01.2018	4.6	2018	1	2	2
11	10.01.2018	5.74	2018	1	2	3
12	11.01.2018	5.96	2018	1	2	4
13	12.01.2018	5.32	2018	1	2	5
14	13.01.2018	5.43	2018	1	2	6
15	14.01.2018	4.39	2018	1	3	7
16	15.01.2018	4.36	2018	1	3	1
17	16.01.2018	5.71	2018	1	3	2
18	17.01.2018	4.03	2018	1	3	3
19	18.01.2018	4.64	2018	1	3	4
20	19.01.2018	2.83	2018	1	3	5
21	20.01.2018	4.08	2018	1	3	6
22	21.01.2018	3.04	2018	1	4	7
23	22.01.2018	2.99	2018	1	4	1

1	mon	0.420454	0.273252	-0.13759	0.320489
2	tue	0.094751	0.110645	0.295901	0.119635
3	wed	0.49055	0.343945	-0.26897	0.370617
4	thu	0.518522	0.354211	-0.6086	0.356516
5	fri	0.679288	0.699179	0.429751	0.660302
6	sat	-0.77971	-0.55161	-1.17728	-0.75104
7	sun	-1.02368	-1.11994	-1.50305	-1.10049
		2021	2020	2019	week
		0.6	0.3	0.1	1

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	year	2020												
2														
3	**Середнє з d_trend**	**Позначки стовпців**												
4	**Позначки рядків**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**Загальний підсумок**
5	1	0.566541119	0.055222849	0.265609797	0.481043036	0.410984928	0.487923048	0.216861121	0.105655112	0.730074063	-0.126272551	0.418151403	-0.378853723	0.273251926
6	2	0.629110275	0.666127372	-0.163031043	-0.328316587	0.467442313	-0.167533698	0.071352463	-0.158262768	0.381206178	0.138294106	0.190598705	-0.241005324	0.110645463
7	3	-0.213046824	0.784208263	-0.180070576	0.611019132	1.036243249	0.225716927	0.348143357	0.124438263	0.193550019	0.407242111	1.164797636	-0.143910084	0.343944934
8	4	1.047718191	0.517002809	0.793240281	0.725785461	0.439885237	-0.358657225	-0.206130498	-0.022534565	1.097266184	-0.144129575	-0.048431632	0.397172913	0.354210695
9	5	1.107444008	0.437048382	0.41460498	0.953570989	1.476704614	0.272238653	0.630462586	0.540782682	0.794010946	1.011652209	0.98342687	-0.58918162	0.699179319
10	6	-0.209301966	-0.138915363	-0.725964235	-0.209284385	-0.426975671	-1.684068773	-0.631935317	-0.701044172	0.212574015	0.046819688	-1.324611969	-1.073242483	-0.551613809
11	7	-0.70959433	-1.138542441	-0.82721642	-1.214286687	-1.158790109	-1.482552616	-1.514233693	-1.457843511	-0.519565212	-0.993979423	-0.973450328	-1.46479352	-1.119936414
12	**Загальний підсумок**	**0.34895475**	**0.158265246**	**-0.077933759**	**0.180497614**	**0.286223131**	**-0.350244846**	**-0.11514337**	**-0.268660271**	**0.404374032**	**0.073317696**	**0.036220794**	**-0.450419286**	**0.01749078**



Завдання для груп №2 (80 хвилин)

Дано:
- відомо, що серед всіх факторів, розглянутих раніше, впливати реально можна лише на 2 – CPL_Google та Email;
- відомо, що ряд має значну тижневу сезонність і важкий тренд.

Необхідно:
- Знайти найкращу автокореляційну модель;
- Знайти найкращу трендову модель;
- Знайти найкращу параметричну модель;
- Знайти спосіб узгодження моделей для максимізації критеріїв регулярності.

Оцінками якості прогнозування вважати:
- Мінімум середньої відносної похибки прогнозування;
- Максимум коефіцієнта кореляції прогнозів і реального ряду.

Важливі і корисні посилання

1. Побудова множинної лінійної регресії в Пайтон з виключенням мультиколінеарності https://www.datasklr.com/ols-least-squares-regression/multicollinearity
2. Застосування автокореляційної прогностичної моделі SARIMA https://www.geeksforgeeks.org/sarima-seasonal-autoregressive-integrated-moving-average/
3. Метод головних компонентів у Пайтон https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

Зимова школа із системного аналізу та штучного інтелекту

Національний технічний університет
ДНІПРОВСЬКА ПОЛІТЕХНІКА
1899

Моделювання та оптимізація виконання скінченних послідовностей замовлень

К.т.н., доц. Желдак Тімур Анатолійович,

Завідувач кафедри системного аналізу та управління

Контакти: Zheldak.t.a@nmu.one +380676319926

Співавтори: