

Одеський національний університет імені І. І. Мечникова
Факультет математики, фізики та інформаційних технологій
Кафедра Оптимального Керування і Економічної Кібернетики

Кваліфікаційна робота

на здобуття ступеня вищої освіти «бакалавр»

«Застосування методів кластерного аналізу для
просторового та часового розподілу сонячних спалахів
у різних циклах»

«Application of cluster analysis methods to spatial and
temporal distribution of solar flares in different cycles»

Виконав: здобувач денної форми навчання
спеціальності 113 Прикладна математика
Освітня програма «Прикладна математика»
Діденко Микола Олександрович

Керівник: канд. фіз.-мат. наук, доц. Страхов Є.М.,
канд. фіз.-мат. наук, с.н.с. Рябов М.І. (РІ НАНУ)

Рецензент: кандидат фіз.-мат. наук, доц. Вербіцький В.В.

Рекомендовано до захисту:
Протокол засідання кафедри
№ ____ від _____ 2025 р.
Завідувач кафедри

Захищено на засіданні ЕК № _____
Протокол № ____ від _____ 2025 р.
Оцінка _____ / _____ / _____
Голова ЕК

Одеса — 2025 р.

ЗМІСТ

Вступ	3
1 Фізична природа космічної погоди	4
1.1 Сонячна активність і сонячний цикл	4
1.2 Сонячні спалахи	7
1.3 Корональні викиди маси (СМЕ)	8
1.4 Протонні спалахи	10
1.5 Класифікація сонячних спалахів	11
2 Методи кластеризації	14
2.1 Метод k -середніх (K-Means)	14
2.2 Метод DBSCAN	15
2.3 Метод BIRCH	16
2.4 Метрика якості кластеризації	18
3 Обробка даних та їх аналіз	20
3.1 Формування бази даних	20
3.2 Візуалізація даних	21
3.3 Кластерний аналіз сонячних спалахів	24
3.4 Результати	26
3.4.1 Візуалізації	26
3.4.2 Кластеризації	29
Висновки	35
Список літератури	37
ДОДАТКИ	38

ВСТУП

Сонячна активність є одним із ключових чинників, що впливають на космічну погоду та умови існування в навколоземному просторі. Одним із найяскравіших проявів цієї активності є сонячні спалахи — короточасні, але потужні викиди енергії, які супроводжуються випромінюванням у широкому спектрі та можуть мати значні наслідки для техногенних систем і здоров'я людей.

Мета роботи: провести комплексне дослідження просторово-часового розподілу сонячних спалахів і виявити особливості 25-го циклу сонячної активності.

Об'єкт дослідження: дані каталогу потужних сонячних спалахів класу *X-ray* M1–X за 23–25 цикли сонячної активності.

Предметом дослідження є вивчення просторово-часових закономірностей виникнення сонячних спалахів у межах 23–25 циклів сонячної активності. У центрі уваги — аналіз змін координат спалахів, їхньої інтенсивності та динаміки впродовж цих циклів. Основним джерелом інформації слугує каталог потужних сонячних спалахів, який містить відомості про клас спалаху, геліографічні координати, час початку, максимуму та завершення кожної події.

Методи дослідження: алгоритми кластерного аналізу (K-Means, DBSCAN, BIRCH) з використанням силуетного коефіцієнта для оцінки якості кластеризації.

Робота виконувалась у тісній співпраці з колегами з Інституту радіоастрономії Національної академії наук України, що забезпечило високу наукову якість аналізу, актуальність задач та доступ до унікальних джерел даних. Таке партнерство дало змогу синхронізувати теоретичні підходи з практичними спостереженнями та зробити вагомий внесок у сучасне вивчення сонячної активності.

РОЗДІЛ 1

ФІЗИЧНА ПРИРОДА КОСМІЧНОЇ ПОГОДИ

1.1 Сонячна активність і сонячний цикл

Вперше прояви сонячної активності були зафіксовані у 1610 році італійським вченим Галілео Галілеєм. Використовуючи створений ним телескоп, він виявив темні плями на поверхні Сонця та спостерігав їх переміщення по сонячному диску. Це дозволило встановити період обертання Сонця, який, як з'ясувалося, залежить від широти розташування груп плям. Для середніх і приєкваторіальних широт цей період становив приблизно 27 діб.

Після відкриття Галілея настав тривалий період зниження сонячної активності — так званий мінімум Маундера (1645–1715 роки), протягом якого сонячні плями спостерігалися надзвичайно рідко. Однак у подальшому плями знову почали з'являтися на поверхні Сонця, і їх регулярність помітив німецький астроном Генріх Швабе. Саме він встановив, що частота появи сонячних плям має періодичність близько 10 років.

Подальший систематичний внесок у дослідження сонячної активності зробив швейцарський астроном Рудольф Вольф, який з 1847 року розпочав власні спостереження сонячних плям. Він запропонував чисельний індекс активності — так зване число Вольфа (W)

Означення 1.1. Число Вольфа — числовий показник рівня сонячної активності, що базується на кількості сонячних плям.

Число Вольфа W розраховується за формулою:

$$W = k \cdot (10 \cdot g + f),$$

де:

- g — кількість груп сонячних плям;
- s — загальна кількість окремих плям;
- k — поправковий коефіцієнт, що враховує умови спостереження (якість

інструментів, атмосферні умови, досвід спостерігача тощо).

Індекс Вольфа став одним з основних індикаторів рівня сонячної активності та досі використовується у сучасній геліофізиці.

У подальшому виявилось, що середня тривалість одного циклу сонячної активності становить приблизно 11 років. Саме тому ці періоди отримали назву «*11-річні цикли*». Проте на практиці тривалість окремих циклів може значно варіюватися — від 7 до 17 років.

Починаючи з 1755 року, коли розпочався перший умовно виділений цикл, всі подальші сонячні цикли прийнято нумерувати. Наприклад, 24-й цикл сонячної активності тривав орієнтовно з січня 2008 року (за іншими оцінками — з грудня 2008 або січня 2009) до жовтня-листопада 2019 року. Після його завершення розпочався 25-й цикл, який, за прогнозами, триватиме приблизно до 2030 року.

Характерною особливістю сонячного циклу є зміна широти появи сонячних плям. На початку кожного циклу плями виникають переважно на високих геліографічних широтах. У міру розвитку циклу зона утворення плям поступово зміщується до екватора. Цей процес триває до завершення циклу. Така закономірність відома як **закон Шперера**, за іменем німецького астронома Густава Шперера, який першим описав це явище.

Означення 1.2. закон Шперера - емпіричний закон у фізиці Сонця, який описує зміну геліографічних широт, на яких утворюються сонячні плями та інші активні області протягом сонячного циклу

Багаторічні спостереження показали, що глобальна конфігурація магнітного поля Сонця змінюється квазіперіодично із середнім періодом приблизно 22 роки. Цей період відображає повний *магнітний цикл* Сонця, також відомий як **цикл Гейла**. Він складається з двох послідовних 11-річних циклів сонячних плям, упродовж яких спостерігається зміна полярності магнітного поля.

Упродовж одного 11-річного циклу в біполярних групах плям ведучі (головні) плями в північній півкулі мають однакову полярність, яка є протилежною полярності ведучих плям у південній півкулі. У хвостових плям полярність завжди протилежна до полярності ведучих. У наступному циклі полярності змінюються на протилежні — як у ведучих, так і в хвостових плям. Одночасно з цим змінюється і полярність загального магнітного поля Сонця, магнітні полюси якого розташовані поблизу полюсів обертання.

Таким чином, магнітний цикл Сонця триває 22 роки, що є більш фізично обґрунтованою одиницею циклічності порівняно з традиційним 11-річним циклом за числом плям.

На початку кожного 11-річного циклу нові сонячні плями з'являються на широтах приблизно 30° , а протягом циклу зона їх утворення поступово зміщується до екватора. Якщо нанести на графік залежність широти появи плям від часу, отримаємо характерну структуру, яка нагадує крила метелика. Така візуалізація отримала назву «**діаграма метелика**» **Маундера** (англ. *Butterfly diagram*), на честь британського астронома Едварда Маундера.

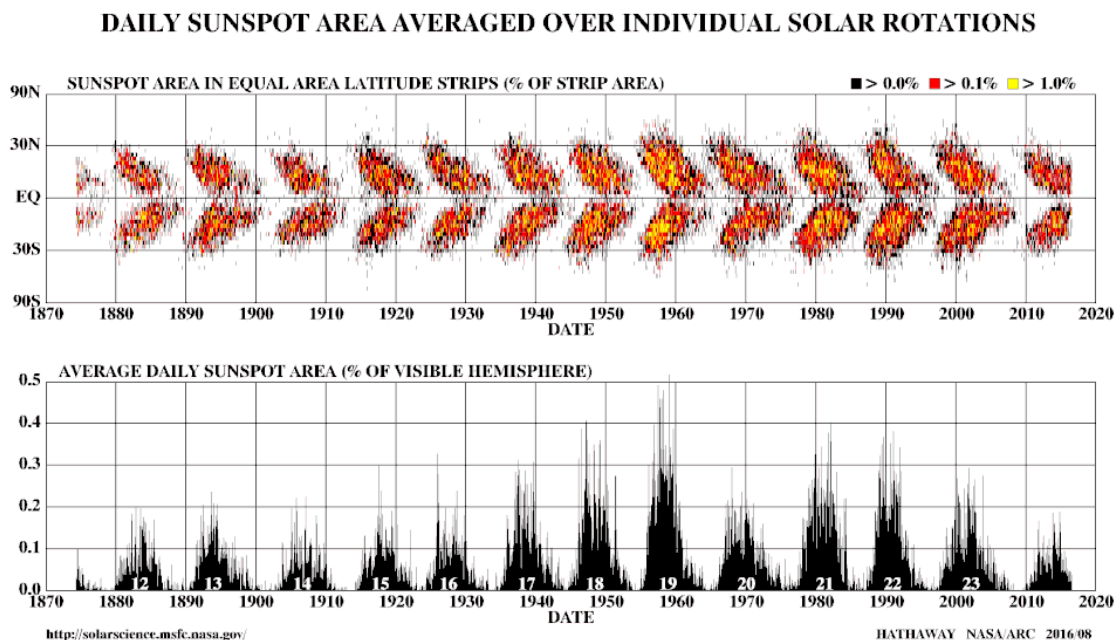


Рис. 1.1. Діаграма метелика та розподіл спалахів

1.2 Сонячні спалахи

Означення 1.3. Сонячний спалах – це раптове і потужне вивільнення енергії на поверхні Сонця або в його атмосфері.

Це надзвичайно потужні явища, що відбуваються в атмосфері Сонця. Вони є одними з основних джерел формування екстремальних станів космічної погоди та часто стають причиною магнітних бур на Землі.

Спалахи виникають у регіонах груп сонячних плям, зазвичай уздовж *нейтральної лінії* — межі між областями з протилежно спрямованими магнітними полями. З фізичної точки зору спалах є реакцією сонячної атмосфери на раптовий і швидкий процес вивільнення енергії, найімовірніше магнітного походження. Цей процес в основному зачіпає хромосферу та корону Сонця.

Вивільнення енергії викликає локалізоване тимчасове нагрівання (так звана *теплова спалахова компонента*), а також прискорення заряджених частинок — електронів, протонів і важчих іонів. У результаті температура в області хромосфери досягає приблизно 10^4 К (так звані хромосферні або низькотемпературні спалахи), а в короні може перевищувати 10^7 К (високотемпературні спалахи).

Прискорені частинки, що реєструються на навколосезній орбіті, мають енергії в діапазоні від ~ 20 кеВ (для електронів) до ≥ 10 ГеВ (для протонів). Повна енергія, яка вивільняється під час найпотужніших спалахів, може сягати 10^{25} Дж (або 10^{32} ерг), що робить їх одними з найенергійніших явищ у Сонячній системі.

Сонячний спалах супроводжується короткочасним, але інтенсивним електромагнітним випромінюванням у широкому діапазоні довжин хвиль — від жорсткого рентгенівського випромінювання (довжина хвилі $\lambda \sim 10^{-9}$ см) до радіохвиль з довжиною $\lambda \sim 10^6$ см. Основна частина цього випромінювання має *теплову природу*.

Проте на дуже коротких довжинах хвиль, таких як жорстке рентгенівське випромінювання ($\lambda < 1$ Å) та гамма-випромінювання ($\lambda < 0,01$ Å), реєструються *нетеплові імпульсні сплески*. Вони генеруються прискорени-

ми частинками в атмосфері Сонця — зокрема:

- нетеплове гальмівне та синхротронне випромінювання електронів;
- гамма-випромінювання, що виникає при взаємодії прискорених іонів з ядрами атомів сонячної атмосфери.

У діапазоні наддовгих хвиль (радіодіапазон) подібні сплески пов'язані з поширенням ударних хвиль у плазмі корони та сонячного вітру. Такі процеси є важливим індикатором динамічної взаємодії між магнітними структурами Сонця та його навколишнім середовищем.

1.3 Корональні викиди маси (СМЕ)

Під час сонячного спалаху плазма у відповідній ділянці може розігріватися до температур близько 10^7 К. Такий інтенсивний нагрів супроводжується потужним випромінюванням в ультрафіолетовому, рентгенівському та видимому діапазонах спектра.

Окрім випромінювання, сонячні спалахи часто супроводжуються викидом значної кількості плазми з корони Сонця у міжпланетний простір зі швидкостями порядку 1000 км/с (іноді може перевищувати 3000 км/с.) Ці явища отримали назву **корональні викиди маси** (англ. *Coronal Mass Ejections*, СМЕ).

Корональні викиди маси є надзвичайно динамічними явищами, під час яких плазма, що утримується у замкнених корональних магнітних структурах, раптово викидається в міжпланетний простір. Відрив подібних на вигляд стабільних і магнітно замкнених структур від Сонця ставить перед наукою низку фундаментальних питань, пов'язаних із закономірностями магнітогідродинаміки (МНД).

Частота СМЕ значно залежить від фази сонячного циклу. В період мінімуму сонячної активності СМЕ спостерігаються приблизно 1 раз на 1–2 доби (іноді — не частіше ніж раз на тиждень). У максимумі циклу кількість подій зростає до 2–5 викидів на добу. Якщо усереднити частоту по всьому 11-річному сонячному циклу, то в середньому спостерігається близько 1 СМЕ на добу.

Примітно, що хоча СМЕ часто асоціюються з потужними сонячними спалахами, між ними немає однозначного зв'язку. За статистичними даними:

- близько 90% сонячних спалахів різних класів не супроводжуються корональними викидами;
- водночас приблизно 60% СМЕ відбуваються без супутніх спалахів.

Через це в науковому середовищі досі не існує єдиної думки щодо механізму утворення СМЕ, зокрема — чи є вони самостійними подіями, чи частиною єдиного процесу «спалах–СМЕ». Також дискутується питання про відносну роль СМЕ порівняно із самими спалахами у контексті сонячно-земних зв'язків.

Наразі всі встановлені зв'язки між СМЕ та іншими проявами сонячної активності на диску Сонця мають суто статистичний характер. Чітко визначених фізичних механізмів взаємодії між цими явищами поки що не встановлено — це залишається предметом активних досліджень сучасної геліофізики.

Корональні викиди маси (СМЕ) були вперше виявлені 14 грудня 1971 року за результатами спостережень із борту супутника *OSO-7* (7-й орбітальний сонячний обсерваторій, англ. *Orbiting Solar Observatory*). Подальший значний прогрес у спостереженнях СМЕ був досягнутий після запуску космічної обсерваторії *SOHO* (Solar and Heliospheric Observatory) у грудні 1995 року.

На борту апарата *SOHO* встановлено два ключових інструменти:

- **LASCO** (Large Angle Spectrometric Coronagraph Experiment) — забезпечує зображення сонячної корони в білому світлі на відстані до 30 сонячних радіусів
- **EIT** (Extreme Ultraviolet Imaging Telescope) — отримує зображення сонячного диска та нижньої корони в ультрафіолетовому діапазоні.

Кожен окремий СМЕ може виносити в міжпланетний простір від 2×10^{14} до 4×10^{16} г сонячної речовини. За окремими оцінками, під час найпотужніших подій маса СМЕ може досягати до 2×10^{17} г. Така величезна маса була зафіксована, зокрема, під час події 28 жовтня 2003 року.

Згідно з сучасними напівемпіричними моделями сонячної корони, в межах до 10 сонячних радіусів вона містить приблизно $10^{18} - 10^{19}$ г речовини. Таким чином, поодинокий СМЕ може містити до 1 – 10% маси всієї корони. Зазвичай більша частина цієї речовини походить з нижньої корони або навіть із ще нижчих шарів сонячної атмосфери

1.4 Протонні спалахи

Протонні спалахи (англ. *Solar Proton Events*, SPE або SEP — *Solar Energetic Particles*) є важливим проявом сонячної активності, що супроводжує найпотужніші сонячні спалахи або корональні викиди маси (СМЕ). Під час таких подій у зовнішніх шарах сонячної атмосфери відбувається різке прискорення заряджених частинок, передусім протонів, а також електронів та важчих іонів. Унаслідок цього високоенергетичні частинки викидаються в міжпланетний простір і можуть досягати навколоземного середовища за лічені хвилини або години.

Протонні події є одними з найбільш небезпечних форм космічної погоди, оскільки:

- створюють серйозну радіаційну загрозу для астронавтів;
- можуть виводити з ладу електроніку супутників;
- спричиняють порушення в роботі навігаційних та радіозв'язкових систем;
- впливають на авіаційні маршрути у полярних широтах.

За класифікацією NOAA, до **протонних подій** відносять ті, при яких інтенсивність протонів з енергією > 10 MeV перевищує поріг 10 частинок/(см²·с·стер). Такі спалахи фіксуються переважно супутниками серії GOES та іншими космічними обсерваторіями.

Протонні події поділяються на:

- *спалахи, асоційовані зі спалахами типу X або M*, де джерелом прискорення частинок є область енергетичного вивільнення;
- *події, пов'язані з СМЕ*, коли основну роль відіграє ударна хвиля, що поширюється у міжпланетному середовищі.

У реальних умовах ці два механізми можуть діяти одночасно, а конкретна інтенсивність і тривалість протонної події залежать від положення джерела на Сонці (ефективність значно вища для заходних довгот), швидкості СМЕ, геометрії магнітних силових ліній та інших факторів.

1.5 Класифікація сонячних спалахів

Різноманітність проявів сонячних спалахів, а також специфіка методів їхнього спостереження зумовили необхідність розробки системи класифікації за потужністю та тривалістю.

До 1 січня 1966 року яскравість спалахів у лінії водню $H\alpha$ оцінювалась за 4-бальною шкалою: 1⁻ (субспалах), 1, 2 та 3. Найпотужнішим спалахам іноді призначали оцінку 3⁺. Згодом було запроваджено 5-бальну шкалу:

$$S \text{ (субспалах), } 1, 2, 3, 4,$$

де спалахи з оцінкою ≥ 2 вважалися сильними.

З початком позаатмосферних спостережень, зокрема з орбіти супутників, спалахи почали класифікувати також за рівнем рентгенівського випромінювання в діапазоні довжин хвиль 1–8 Å. Відповідно до стандарту NOAA/GOES, для класифікації використовуються літерні індекси:

- A ($> 10^{-8}$ Вт/м²),
- B ($> 10^{-7}$ Вт/м²),
- C ($> 10^{-6}$ Вт/м²),
- M ($> 10^{-5}$ Вт/м²),
- X ($> 10^{-4}$ Вт/м²).

Після літери вказується число від 1.0 до 9.9, що уточнює величину потоку. Наприклад, спалах класу X5 має потужність 5×10^{-4} Вт/м² рентгенівського випромінювання для спалахів класу X перевищує 10^{-1} ерг/(см²·с) у діапазоні 0.1–0.8 нм. Для порівняння, спалахи класу M є у 10 разів слабшими, а класу C — у 100 разів слабшими за типову спалахову подію класу X.

Така шкала широко використовується у сучасній геліофізиці, зокрема супутниковими місіями NOAA/GOES, для моніторингу рівнів сонячної активності.

Починаючи з 1977 року, сонячні спалахи умовно поділяють на два типи — *імпульсні* та *поступові* — залежно від тривалості сплеску м'якого рентгенівського випромінювання (англ. *Soft X-Ray*, SXR). Якщо спалах триває менше однієї години, він вважається імпульсним; якщо понад годину — поступовим.

Сучасна систематизація цих типів має ґрунтовну спостережну базу і представлена, зокрема, у класифікації Kallenrode (2001). Основні відмінності між імпульсними та поступовими спалахами подано у таблиці 3.1

Наразі безперервний моніторинг рентгенівських спалахів здійснюється за допомогою супутникової системи **GOES** (NOAA), яка забезпечує оперативну класифікацію спалахів, прогнозування космічної погоди та виявлення потенційно небезпечних для Землі подій.

Потужне рентгенівське випромінювання досягає Землі приблизно через 8 хвилин після початку спалаху, що може спричинити радіаційні бурі в іоносфері та порушення короткохвильового (КХ) радіозв'язку.

Якщо спалах супроводжується корональним викидом маси (СМЕ), спрямованим у бік Землі, то приблизно через 2–3 доби виникає геомагнітна буря. Вона може тривати від кількох годин до кількох діб, залежно від швидкості та маси викиду. Найпотужніші спалахи супроводжуються також викидом високоенергетичних протонів, які досягають навколоземного простору вже за кілька годин.

Показовим прикладом стала подія 14 липня 2000 року, відома як *Bastille Day Event (BDE)*. У день національного свята Франції стався потужний спалах, який спричинив стрімкий потік сонячних енергетичних частинок. Через 30 хвилин після спалаху ці частинки досягли навколоземного простору, викликавши «радіаційну бурю» на борту супутника *SOHO*. Через бомбардування детекторів частинками високої енергії на зображеннях камери *EIT* з'явився характерний «сніг».

Спалах також супроводжувався надзвичайно потужним СМЕ, який

викинув мільярди тонн плазми зі швидкістю близько 1800 км/с — майже вдвічі швидше за середній рівень.

Тривалість оптичного спалаху може змінюватися від кількох хвилин до кількох годин. Початкова фаза, що характеризується стрімким збільшенням яскравості та площі активної області, називається *флеш-фазою* (або імпульсною фазою) і зазвичай триває до 15 хвилин. У межах флеш-фази спостерігається короткий (до 1 хв) імпульсний сплеск яскравості на обмеженій ділянці, що супроводжується одночасними сплесками мікрохвильового та жорсткого рентгенівського випромінювання. Мікрохвилі з частотами в діапазоні ГГц є результатом синхротронного випромінювання електронів з енергіями 10–100 кеВ у магнітних полях.

РОЗДІЛ 2

МЕТОДИ КЛАСТЕРІЗАЦІЇ

2.1 Метод k -середніх (K-Means)

Метод k -середніх (англ. *K-Means*) — це один з найбільш використовуваних методів кластерного аналізу, який належить до групи ітеративних евристичних алгоритмів для розв’язання задачі класифікації без учителя.

Мета методу. Розбити множину n точок $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ у d -вимірному просторі на k неперетинних кластерів $\{C_1, C_2, \dots, C_k\}$ так, щоб усередині кожного кластера об’єкти були максимально подібні, а між кластерами — максимально відмінні.

Цільова функція. Метою є мінімізація суми квадратів евклідових відстаней від кожного об’єкта до центроїда відповідного кластера:

$$\min_{\{C_j\}_{j=1}^k} \sum_{j=1}^k \sum_{\mathbf{x}_i \in C_j} \|\mathbf{x}_i - \boldsymbol{\mu}_j\|^2,$$

де $\boldsymbol{\mu}_j = \frac{1}{|C_j|} \sum_{\mathbf{x}_i \in C_j} \mathbf{x}_i$ — центр маси кластера C_j .

Алгоритм.

- 1) Вибір k початкових центрів (випадково або методом *k-means++*).
- 2) Призначення кожної точки до найближчого центру.
- 3) Перерахунок центрів як середніх значень точок у кожному кластері.
- 4) Повторення кроків 2–3 до збіжності (тобто поки призначення кластерів не змінюється).

Особливості методу:

- Метод чутливий до початкової ініціалізації, тому рекомендується використовувати багатократне запускання з різними початковими точками.

- Кількість кластерів k має бути задана наперед.
- K-means ефективно працює з опуклими, сферичними кластерами з приблизно однаковим об'ємом.
- Не придатний для виявлення кластерів складної форми або з різною щільністю.

Вибір k : Для вибору кількості кластерів використовуються методи:

- Метод «лікоть» (elbow method) — точка зламу на графіку залежності сумарної похибки кластеризації від k ;
- Коефіцієнт силуету (silhouette score);

2.2 Метод DBSCAN

Метод DBSCAN (Density-Based Spatial Clustering of Applications with Noise) є одним із найпопулярніших алгоритмів кластерного аналізу, який базується на щільності розподілу об'єктів у просторі ознак. Його головна перевага — здатність виявляти кластери довільної форми та виявляти викиди (аномалії), що особливо важливо при аналізі просторового розподілу сонячних спалахів, де існують неоднорідності у щільності їх концентрації.

Основна ідея. DBSCAN групує точки, що лежать у щільних регіонах простору, та ігнорує точки, що знаходяться у розріджених регіонах (які вважаються шумом).

Параметри:

- ε (eps) — радіус околу для пошуку сусідів;
- **MinPts** — мінімальна кількість точок в околі ε , щоб точка вважалася *ядровою*.

Типи точок:

- **Ядрові точки (core points)** — мають не менше MinPts сусідів у межах ε ;

- **Прикордонні точки (border points)** — мають менше ніж MinPts сусідів, але лежать у ε -околі ядрової точки;
- **Шум (noise/outliers)** — не є ані ядровими, ані прикордонними.

Алгоритм:

- 1) Для кожної точки x визначається ε -окол.
- 2) Якщо кількість точок у цьому околі MinPts — x є ядром, формується новий кластер.
- 3) До кластера приєднуються всі точки, щільно зв'язані з x (прямо або опосередковано через інші ядрові точки).
- 4) Повторити для всіх ще не класифікованих точок.

Переваги DBSCAN:

- Не потребує заданої наперед кількості кластерів;
- Виявляє кластери довільної форми;
- Стійкий до викидів і шуму.

Недоліки:

- Важко підібрати оптимальні значення ε та MinPts , особливо для даних з неоднорідною щільністю;
- Не працює добре, якщо щільності кластерів істотно різняться.

2.3 Метод BIRCH

Метод **BIRCH** (Balanced Iterative Reducing and Clustering using Hierarchies) був запропонований для ефективної кластеризації *дуже великих* наборів даних, що постійно поповнюються. Алгоритм будує компактну ієрархічну структуру — *CF-дерево* (Cluster Feature Tree), яка інкрементно узагальнює інформацію про вхідні об'єкти й дає змогу кластеризувати дані практично за один прохід.

Cluster Feature (CF). Для кожного підкластера з N точок $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ у d -вимірному просторі зберігається трійка

$$\text{CF} = (N, \mathbf{LS}, \mathbf{SS}), \quad \mathbf{LS} = \sum_{i=1}^N \mathbf{x}_i, \quad \mathbf{SS} = \sum_{i=1}^N \mathbf{x}_i^2,$$

де $\mathbf{LS}$ — векторна сума, $\mathbf{SS}$ — сума квадратів координат. За CF легко обчислюються центр $\boldsymbol{\mu} = \mathbf{LS}/N$ та радіус/діаметр підкластера без потреби перевантажувати пам'ять усіма точками.

Ключові параметри.

- **Branching Factor B** — максимальна кількість дочірніх вузлів кожного внутрішнього вузла дерева.
- **Threshold T** — верхня межа радіуса (або діаметра) підкластера в листі CF-дерева.
- **Leaf Size L** — максимальна кількість записів у листовому вузлі.

Чотири фази алгоритму.

- 1) **Побудова CF-дерева.** Поступове внесення кожної точки у дерево з оновленням CF-записів; за потреби виконується перерозбиття листів (split).
- 2) **Конденсація.** За бажанням із CF-дерева видаляються «шумові» вузли, а близькі вузли консоліднуються з жорсткішим порогом T' .
- 3) **Глобальна кластеризація.** До компактного набору листових підкластерів застосовується зовнішній алгоритм (наприклад, k -середніх або ієрархічний), зменшуючи витрати з $O(n)$ до $O(|\text{листів}|)$.
- 4) **За потреби — рефінмент.** Додаткове уточнення границь кластерів.

Переваги для аналізу сонячних спалахів.

- *Автоматичне виявлення шуму:* рідкісні або аномальні спалахи потрапляють у малі CF-вузли й можуть бути відфільтровані у фазі конденсації.

Обмеження.

- Параметр T необхідно добирати емпірично; надто мале значення призводить до надмірного розгалуження дерева.

2.4 Метрика якості кластеризації

Після побудови кластерної моделі необхідно перевірити, наскільки вдало об'єкти згруповано. Однією з найуживаніших метрик для такого **внутрішнього** (intrinsic) оцінювання є *Silhouette Score* (силуетна оцінка), запропонована Р. J. Rousseeuw

Формула. Для кожної точки i обчислюємо:

$$a(i) = \frac{1}{|C_i| - 1} \sum_{\mathbf{x}_j \in C_i, j \neq i} d(\mathbf{x}_i, \mathbf{x}_j), \text{ (середня відстань до точок власного кластера),}$$

$$b(i) = \min_{C \neq C_i} \frac{1}{|C|} \sum_{\mathbf{x}_j \in C} d(\mathbf{x}_i, \mathbf{x}_j), \text{ (найменша середня відстань до чужого кластера).}$$

Тоді силуетний коефіцієнт окремої точки:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \in [-1, 1].$$

Інтерпретація.

- $s(i) \approx 1$ — об'єкт добре «вписався» у свій кластер і далеко від інших.
- $s(i) \approx 0$ — об'єкт лежить на межі двох кластерів (перекриття).
- $s(i) < 0$ — ймовірно, об'єкт віднесено до «чужого» кластера (помилка).

Середнє значення $\bar{s} = \frac{1}{n} \sum_{i=1}^n s(i)$ називають **Silhouette Score** усієї кластеризації.

Переваги.

- Не потребує «зовнішнього» еталону — підходить для ненаглядного аналізу.

- Дає зрозумілу шкалу $[-1,1]$.
- Дозволяє емпірично добирати кількість кластерів k : максимізуємо $\bar{s}(k)$.

Обмеження.

- Витрати $O(n^2)$ на відстані; для великих наборів потрібні вибірко-ві/наближенні методи.
- Чутлива до вибору метрики $d(\cdot, \cdot)$; класично — евклідова.
- Може віддавати перевагу кластерам схожій густини та розміру, «карально» оцінюючи дрібні, проте природні групи.

РОЗДІЛ 3

ОБРОБКА ДАНИХ ТА ЇХ АНАЛІЗ

Результати аналізу були реалізовані у вигляді веб застосунку. (див. у дод. Б)

3.1 Формування бази даних

Перший етап практичної частини роботи полягав у формуванні бази даних потужних сонячних спалахів з урахуванням їхніх просторових та часових характеристик. В якості джерела даних був використаний відкритий «Каталог сонячних спалахів класу X-ray M1–X», який нам передали наші колеги

Каталог представлений у форматі PDF-файлу та містить табличні дані про кожен зареєстрований спалах, включаючи:

- дату та час початку, максимуму й завершення спалаху;
- клас спалаху в рентгенівському діапазоні (X-ray) та оптичну важливість;
- координати на сонячному диску (широта та довгота);
- потік випромінювання (інтегральний потік у J/m^2);
- наявність корональних викидів маси (CME);
- інформацію про протонні події та інші параметри.

Для автоматизації вилучення інформації було розроблено веб-додаток з використанням фреймворку **Streamlit**. На сторінці **File Processing** реалізовано інтерфейс для завантаження PDF-файлу каталогу, парсингу його вмісту та попереднього перегляду даних. Застосовано алгоритм розбиття рядків за фіксованими позиціями, що відповідають структурі таблиці в оригінальному документі. Після обробки дані зберігаються у форматі таблиці **DataFrame** та можуть бути експортовані у форматі Excel для подальшого аналізу.

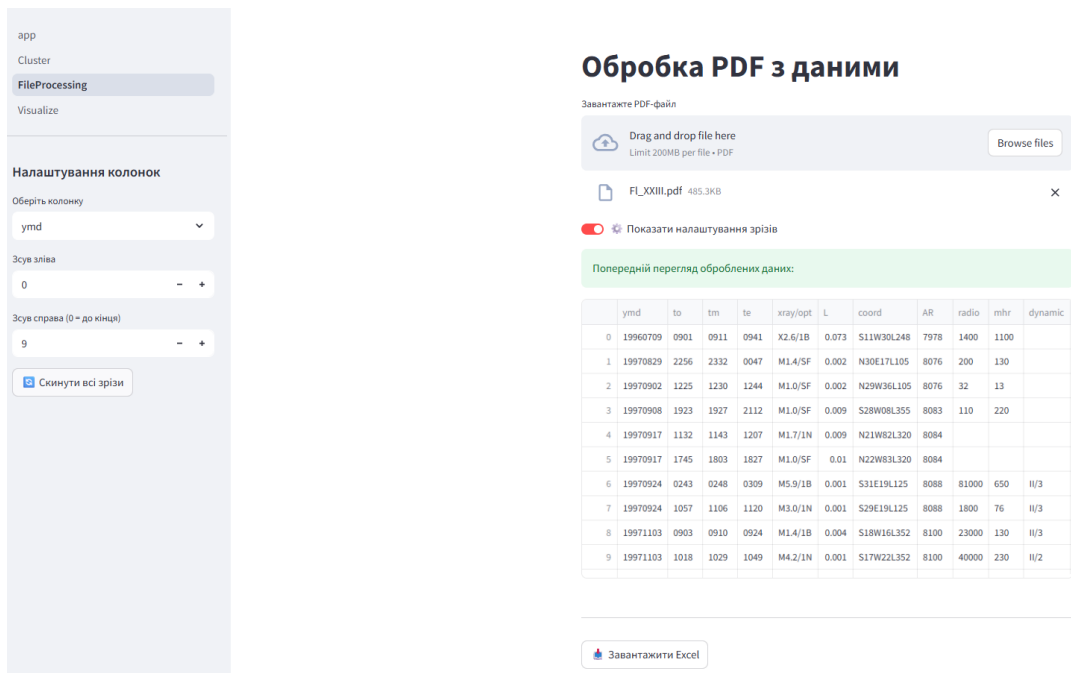


Рис. 3.1. Інтерфейс обробки PDF-файлу на етапі формування бази даних

Цей етап забезпечив коректне та ефективне перенесення інформації з паперового формату до структурованої бази даних (Excel), придатної для подальшого аналізу.

3.2 Візуалізація даних

Після формування структурованої бази даних сонячних спалахів було реалізовано етап візуального аналізу з метою дослідження просторово-часового розподілу подій. Цей етап дозволив виявити закономірності активності Сонця в межах кожного циклу та простежити характер розміщення потужних спалахів у північній та південній півкулях.

Візуалізація була реалізована у модулі `Visualize.py`. Вона включала такі графічні інструменти:

- **Стовпчикова діаграма частоти спалахів** — демонструє кількість зареєстрованих спалахів у кожен день, дозволяючи виявити пікові періоди активності.
- **Точковий графік “широта – час”** — відображає просторове положення спалахів за широтою відносно екватора Сонця в часовій шкалі. Для кожної події додатково вказано рентгенівський клас.

- **Графіки по кожному циклу окремо** — дають змогу детально дослідити розподіл спалахів у межах одного циклу, розділити події за класами (М/Х), за наявністю СМЕ та за півкулями.
- **Лінії тренду для окремих півкуль** — побудовано з урахуванням приналежності події до північної чи південної півкулі, що дозволяє дослідити зсув активності у часі між півкулями.

Перед побудовою графіків дані було попередньо очищено: виконано перетворення типів, обробку дати та числових значень, відфільтровано неповні записи. Усі графіки створено за допомогою бібліотеки Plotly, що забезпечує інтерактивність і гнучке налаштування візуалізації.

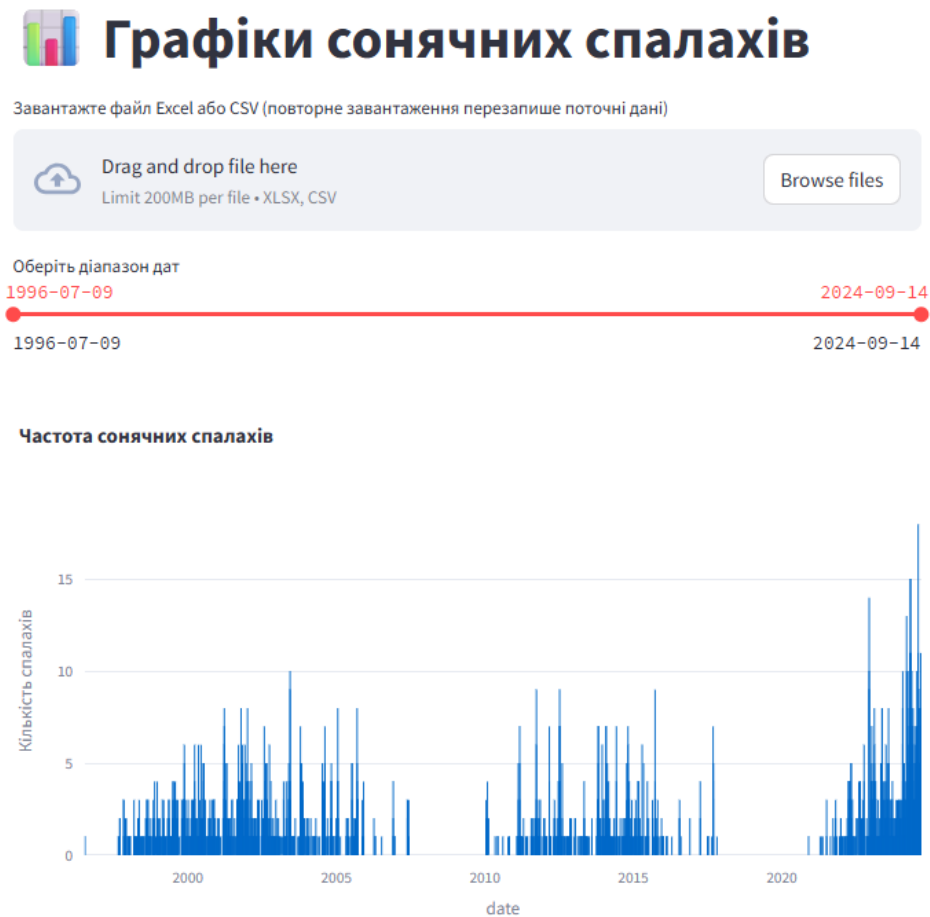


Рис. 3.2. Розподіл сонячних спалахів по дням

Розподіл сонячних спалахів за широтою у часі

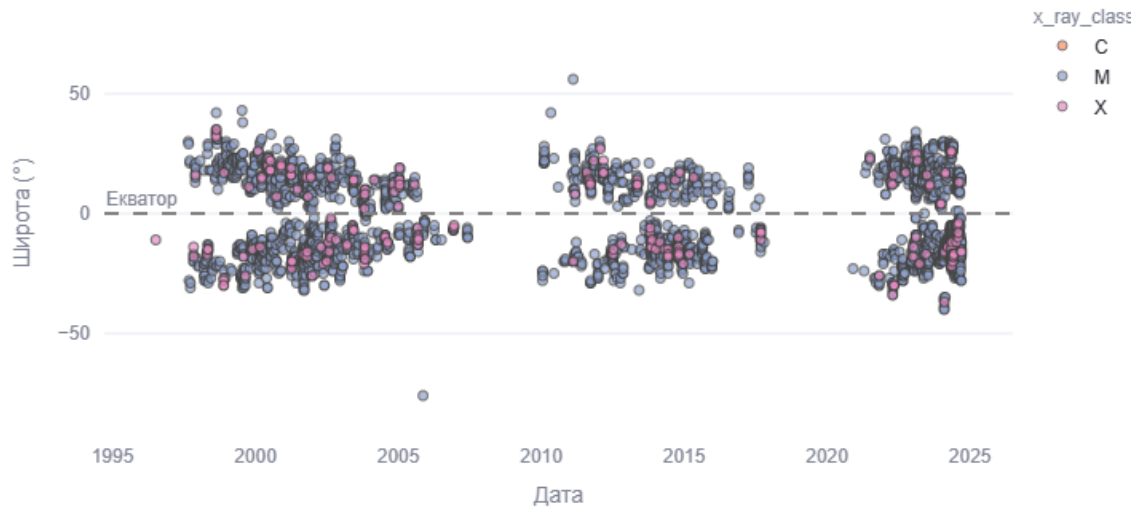


Рис. 3.3. Розподіл сонячних спалахів за широтою у часі

Цикл 25

Статистика по циклу 25:

- Загальна кількість спалахів: 1306
- 3 CME: 428 (32.8%)
- Протонних: 43 (3.3%)

Широти всіх спалахів у циклі 25

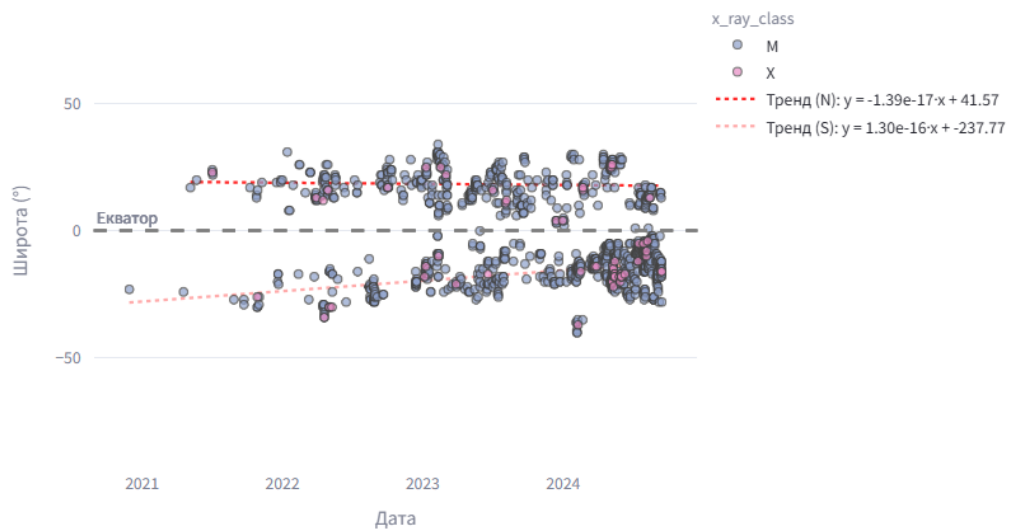


Рис. 3.4. Розподіл сонячних спалахів за широтою у часі 25 циклу

Таким чином, другий етап дозволив глибше зрозуміти динаміку сонячної активності у просторово-часовому вимірі

3.3 Кластерний аналіз сонячних спалахів

На третьому етапі було здійснено кластерний аналіз з метою виявлення структурованих груп сонячних спалахів за просторовими, енергетичними та часовими характеристиками. Для цього було реалізовано інтерактивний інтерфейс у веб-додатку, що дозволяє обирати ознаки, налаштовувати параметри кластеризації та переглядати результати.

У якості ознак для кластеризації було використано такі параметри кожного спалаху:

- числове значення рентгенівського класу (`x_ray_class_num`);
- пік потужності випромінювання (`peak_flux`);
- широта та довгота (`lat`, `lon`);
- тривалість (`duration_minutes`);
- енергетичний інтеграл (`L`).

Було реалізовано підтримку кількох методів кластеризації:

- **K-Means** — класичний алгоритм з фіксованою кількістю кластерів k ;
- **DBSCAN** — алгоритм на основі щільності, що дозволяє виявляти аномалії (кластер -1);
- **Birch** — ефективний для великих обсягів даних, з підтримкою налаштування порогу.

Перед кластеризацією:

- 1) були видалені записи з пропущеними значеннями у вибраних ознаках;
- 2) усі числові значення були нормалізовані за допомогою `StandardScaler` для забезпечення однакового масштабу.

Для аналізу результатів побудовано точковий графік з розфарбуванням точок за кластерною належністю. За віссю x використано дату події, за віссю y — широту.



Рис. 3.5. Інтерфейс розділу кластеризації

У бічній панелі додатку відображається значення метрики **Silhouette Score**, що дозволяє оцінити якість кластеризації (чим ближче до 1 — тим краще відокремлені кластери). У випадках, коли виявлено лише один кластер виводиться попередження.

На завершення формується підсумкова таблиця з такими показниками для кожного кластеру:

- середні значення ознак;
- стандартне відхилення;
- кількість об'єктів у кластері.

Та кнопка завантаження таблиці з визначеними кластерами

Статистика по кластерах

Клас спалаху - "A": 0, "B": 1, "C": 2, "M": 3, "X": 4

cluster	x_ray_class_num			peak_flux			L		
	mean	std	count	mean	std	count	mean	std	count
0	3	0	2798	1.7002	0.7064	2798	0.018	0.0216	2798
1	4	0	235	2.5766	2.7291	235	0.2671	0.3425	235
2	3	0	533	5.8482	1.649	533	0.0642	0.0651	533

Silhouette Score: 0.702

Завантажити Excel

Рис. 3.6. Підсумкова статистика кластерів

Таким чином, третій етап дозволив виявити групи сонячних спалахів з подібними характеристиками та виявити потенційні закономірності у їх розподілі в часі та просторі.

3.4 Результати

3.4.1 Візуалізації

Порівняння трендових ліній для півкуль 23 циклу



Рис. 3.7. 23 цикл сонячної активності

Тренди широти сонячних спалахів у 23-му циклі описуються рівняннями

$$y_N = -6.59 \times 10^{-17} x + 81.05, \quad y_S = 4.36 \times 10^{-17} x + 59.83,$$

де y — геліографічна широта (градуси), x — час (умовні одиниці).

Оскільки абсолютне значення кутового коефіцієнта для північної півкулі

$$|k_N| = 6.59 \times 10^{-17}$$

на приблизно 50% більше, ніж для південної

$$|k_S| = 4.36 \times 10^{-17}, \quad \frac{|k_N|}{|k_S|} \simeq 1.5,$$

північна активна смуга мігрує до екватора істотно швидше.

Таким чином, у 23-му циклі спостерігається помітна асиметрія закону Шпеллера: північні регіони активності “спускаються” до 0° широти динамічніше, тоді як південні роблять це плавніше.

Трендові лінії для 24-го циклу

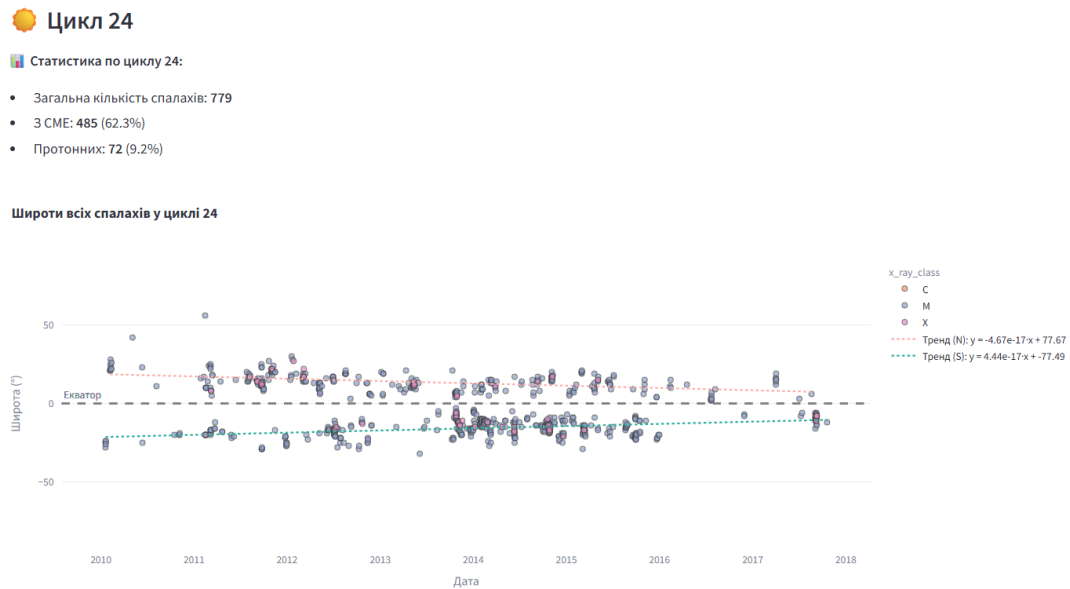


Рис. 3.8. 24 цикл сонячної активності

$$y_N = -4.67 \times 10^{-17} x + 77.67, \quad y_S = 4.44 \times 10^{-17} x - 77.49,$$

де y — геліографічна широта, x — час.

Абсолютні значення кутових коефіцієнтів

$$|k_N| = 4.67 \times 10^{-17}, \quad |k_S| = 4.44 \times 10^{-17}$$

майже збігаються:

$$\frac{|k_N|}{|k_S|} \simeq 1.05.$$

Висновок. У 24-му циклі міграція активних зон до екватора є практично симетричною: північний пояс спускається лише на $\approx 5\%$ швидше за південний.

Трендові лінії для 25-го циклу

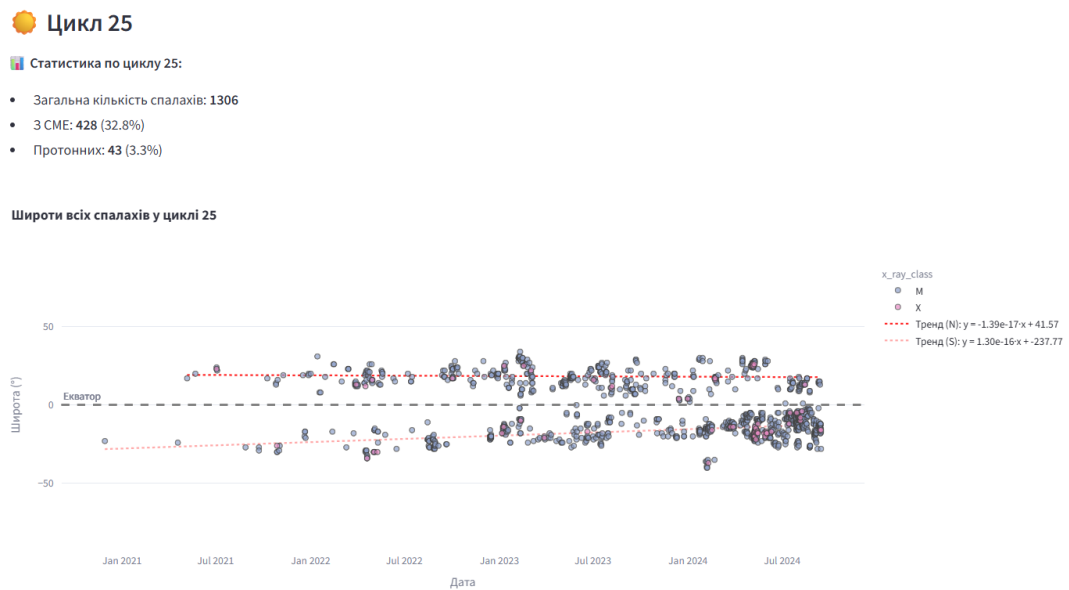


Рис. 3.9. 24 цикл сонячної активності

$$y_N = -1.39 \times 10^{-17} x + 41.57, \quad y_S = 1.30 \times 10^{-16} x - 237.77.$$

Кутові коефіцієнти мають абсолютні значення

$$|k_N| = 1.39 \times 10^{-17}, \quad |k_S| = 1.30 \times 10^{-16}.$$

Оскільки

$$\frac{|k_S|}{|k_N|} \simeq 9.4,$$

південна активна смуга «піднімається» до екватора майже в десять разів швидше, ніж північна.

Таким чином, на поточній фазі 25-го циклу спостерігається помітна асиметрія: міграція спалахів у південній півкулі є значно динамічнішою, тоді як північна залишається відносно стабільною.

3.4.2 Кластеризації

Вибір кількості кластерів k (метод «лікоть»)

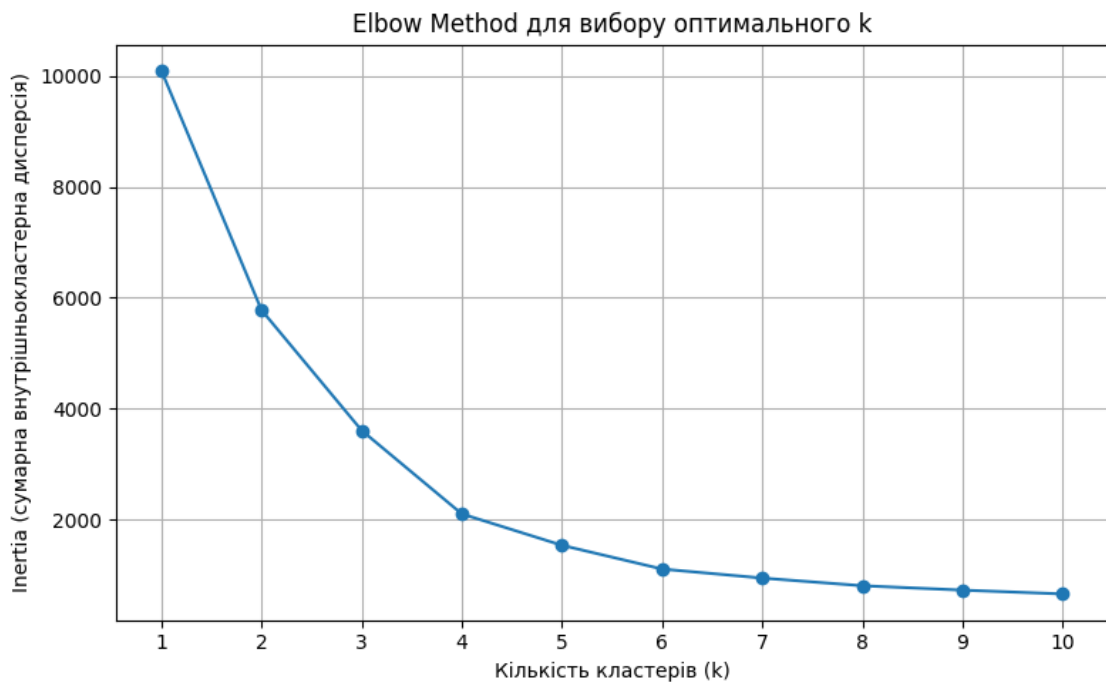


Рис. 3.10. Elbow Method

На рис. 3.14 зображено залежність інерції (WCSS) від числа кластерів k . Крива швидко спадає при переході від $k = 1$ до $k = 4$, після чого темп зменшення різко сповільнюється; подальший приріст k дає лише маргінальне покращення:

- $k = 1 \rightarrow 2$: інерція знижується майже удвічі;
- $k = 2 \rightarrow 3$: ще одна суттєва економія WCSS;
- $k = 3 \rightarrow 4$: останнє «помітне» покращення;
- починаючи з $k \geq 5$ кривина майже вирівнюється.

Отже, *точка перегину* («лікоть») розташована біля $k = 4-5$.

Щоб уникнути суб'єктивності, результати «ліктя» було зіставлено з силуетною оцінкою, де $\bar{s}_{\max} = 0.782$ спостерігалось при $k = 3$. Беручи до уваги обидва критерії та геофізичну інтерпретацію кластерів, прийнято компромісне рішення:

$$k_{\text{опт}} = 3-4$$

Алгоритм k -means

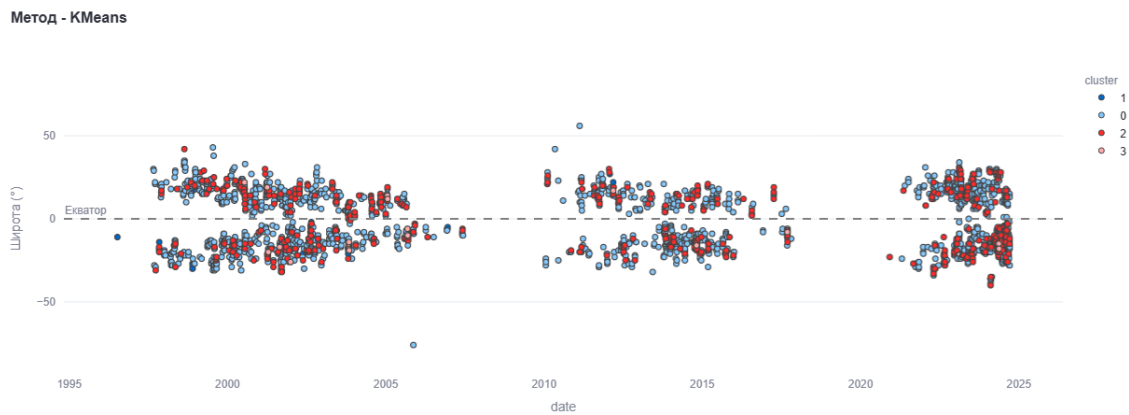


Рис. 3.11. Розподіл кластерів по широті за часом

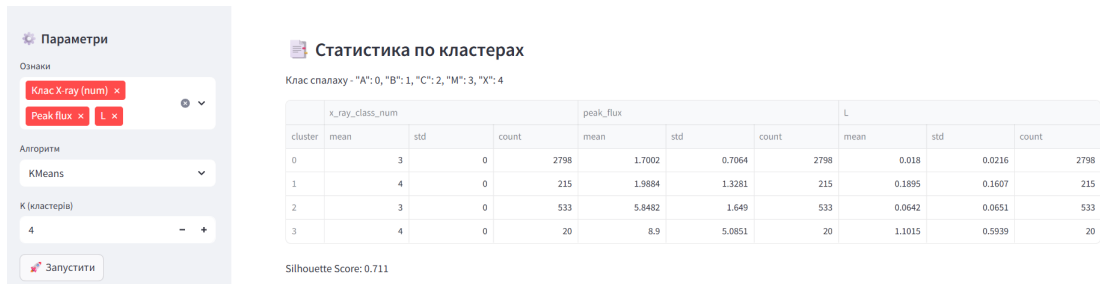


Рис. 3.12. Статистика кластерів

Алгоритм k -means з $k = 4$ розбив сукупність $n = 3\,566$ спалахів на чотири чітко відокремлені групи; середня силуетна оцінка $\bar{s} = 0.711$ вказує на високу внутрішньокластерну однорідність і добру міжкластерну роздільність.

Отримані групи

- 1) **Кластер 0** (2798 подій) — переважно спалахи класу **M** з помірним піковим потоком $F_{\text{peak}} \simeq 1.7$ та дуже малим інтегральним потоком $L \simeq 0.02$.
Це «типові» події середньої енергії.
- 2) **Кластер 1** (215 подій) — спалахи класу **X** середньої потужності $F_{\text{peak}} \simeq 2.0$, інтегральний потік зростає на порядок ($L \simeq 0.19$).
Представляє перехідну групу до більш енергійних екстремальних подій.
- 3) **Кластер 2** (533 події) — «сильні» події класу **M** з високим піковим потоком $F_{\text{peak}} \simeq 5.85$ та підвищеним інтегральним потоком $L \simeq 0.06$.
За енергетичними характеристиками займає проміжну позицію між 0 та 1.
- 4) **Кластер 3** (20 подій) — *екстремальні* спалахи класу **X** ($F_{\text{peak}} \simeq 8.9$) з максимальним інтегральним потоком ($L \simeq 1.10$).
Це рідкісні, але найбільш небезпечні події з погляду радіаційних ризиків.

Висновок.

- Модель виокремила два рівні потужності всередині класу *M* (кластер 0 vs 2) та два підкласи серед *X*-спалахів (1 vs 3), що різняться переважно піковим потоком і тривалістю.
- Екстремальні події ($\approx 0.6\%$ спостережень) зосереджені в компактному кластері 3, що спрощує їхню оперативну ідентифікацію для прогнозів космічної погоди.
- Високе значення $\bar{s}$ підтверджує адекватність $k = 4$ для даного набору ознак, тоді як подальше дроблення погіршувало б силует і не давало б фізично чистих груп.

Алгоритм DBSCAN

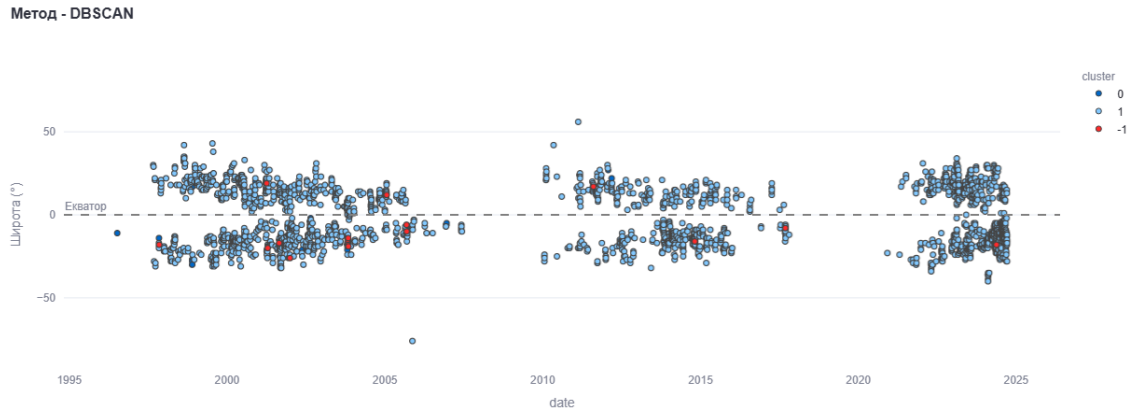


Рис. 3.13. Розподіл кластерів по широті за часом

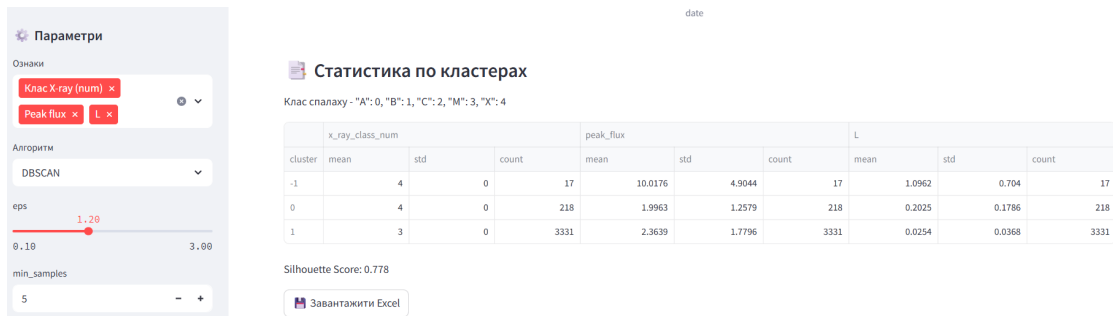


Рис. 3.14. Статистика кластерів

Параметри алгоритму: $\varepsilon = 1.2$, MinPts= 5. Силуетна оцінка $\bar{s} = 0.778$ перевищує аналогічний результат для k -means ($\bar{s} = 0.711$), що свідчить про кращу внутрішню однорідність кластерів.

Отримані групи

- **Кластер 1** ($n = 3331$) — переважно спалахи класу **М** із середнім піковим потоком $F_{\text{peak}} \simeq 2.36$ та дуже низьким інтегральним потоком $L \simeq 0.026 \text{ J m}^{-2}$.

Це «фоновий» щільний осередок середньопотужних подій.

- **Кластер 0** ($n = 218$) — спалахи класу **Х** із помірним піковим потоком $F_{\text{peak}} \simeq 2.00$ та інтегральним потоком $L \simeq 0.20$.

Умовно «середньоекстремальна» група підвищеної енергії, достатньо щільна, аби утворити власний осередок.

- **Шум / викиди** ($\text{cluster} = -1$, $n = 17$) — рідкісні події класу **X** з екстремальними характеристиками: $F_{\text{peak}} \simeq 10.0$, $L \simeq 1.10$. DBSCAN позначив їх як *outliers*, що не належать жодному щільному регіону простору ознак.

Висновок. DBSCAN виділяє лише два «справжні» щільні кластери (типові *M*-події та підвищено-енергетичні *X*-події) і автоматично маркує найпотужніші спалахи як шум. Такий підхід важливий з прикладної точки зору: саме ці 17 викидів найбільш імовірно пов’язані з небезпечними радіаційними та геомагнітними явищами.

Алгоритм BIRCH

Параметри моделі: $\text{threshold}=0.50$, задано 3 фінальні кластери.

Силуетна оцінка $\bar{s} = 0.782$ — найвища серед протестованих алгоритмів.

Отримані групи

- 1) **Кластер 2** ($n = 3331$). Базовий «фон»: спалахи класу **M** з середнім піковим потоком $F_{\text{peak}} \approx 2.36$ та дуже малим інтегральним потоком $L \approx 0.025 \text{ J m}^{-2}$.
Відповідає типово енергетичним подіям, що становлять більшість каталогу.
- 2) **Кластер 1** ($n = 228$). Помірно потужні спалахи класу **X**:
 $F_{\text{peak}} \approx 2.28$, $L \approx 0.221 \text{ J m}^{-2}$.
Формують окремий щільний осередок підвищеної енергії.
- 3) **Кластер 0** ($n = 7$). Рідкісні *екстремальні* події класу **X** з рекордними значеннями $F_{\text{peak}} \approx 12.27$, $L \approx 1.79 \text{ J m}^{-2}$.
BIRCH виокремив їх у самотійний кластер, що полегшує подальший моніторинг найбільш небезпечних спалахів.

Висновок. BIRCH компактно стиснув початкову вибірку до трьох інформативних груп, при цьому зберігши точне виділення екстремальних *X*-подій. Завдяки інкрементальній природі алгоритму та найкращому силуетному

показнику ($\bar{s} = 0.782$) цей підхід доцільно застосовувати для потокового оновлення бази спалахів у майбутніх циклах сонячної активності.

ВИСНОВКИ

У результаті проведеного дослідження досягнуто поставлену мету — проаналізовано просторово-часовий розподіл сонячних спалахів за допомогою методів кластерного аналізу в межах 23–25 циклів сонячної активності. Основні результати можна сформулювати так:

- 1) Побудовано структуровану базу даних сонячних спалахів на основі PDF-каталогу, що містить ключові характеристики кожної події: координати, тривалість, енергетичні параметри, наявність супутніх CME та протонних подій.
- 2) Реалізовано візуалізацію активності Сонця, зокрема:
 - діаграми щоденної частоти спалахів,
 - розподіли за широтою у часі,
 - трендові лінії для північної та південної півкуль.

Виявлено характерну асиметрію у 23-му та 25-му циклах, де північна або південна активна зона зміщується до екватора значно швидше.

- 3) Застосовано кілька методів кластерного аналізу:
 - **K-Means** — оптимальне число кластерів $k = 4$, що дозволило виявити 2 підкласи М-спалахів і 2 підкласи Х-спалахів, з середнім значенням Silhouette Score $s = 0.711$;
 - **DBSCAN** — успішно виділив 2 щільні кластери та групу аномальних подій (outliers), з вищим значенням $s = 0.778$;
 - **BIRCH** — показав найкраще значення Silhouette Score
- 4) Кластеризація дозволила виокремити рідкісні екстремальні події ($\approx 0.6\%$ вибірки), що можуть мати найбільший вплив на космічну погоду, та групи типових середньоенергетичних спалахів.
- 5) Продемонстровано, що використання кластерного аналізу є ефективним інструментом для виявлення закономірностей розподілу сонячних спалахів у просторі й часі. Такі підходи можуть бути корисними для автоматизованого моніторингу й прогнозування небезпечних сонячних подій.

Отримані результати мають не лише наукову, а й прикладну цінність — зокрема, для подальшого прогнозування космічної погоди, оцінки потенційних загроз для технічної інфраструктури, а також подальших досліджень сонячної активності.

Робота пройшла апробацію та була представлена на міжнародному семінарі BLU 2025 — Workshop of the Bulgaria–Latvia–Ukraine Initiative for Space Weather Investigations (2–6 червня 2025 р., Приморсько, Болгарія).

СПИСОК ЛІТЕРАТУРИ

1. Hudson H.S., Bougeret J.-L., Burkepile J.T. Coronal mass ejections: Overview and issues // *Solar Physics*. – 2006. – Vol. 234. – P. 109–126.
2. Ester M., Kriegel H.-P., Sander J., Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise // *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD'96)*. – Portland, OR, 1996. – P. 226–231.
3. Zhang T., Ramakrishnan R., Livny M. BIRCH: An efficient data clustering method for very large databases // *Proceedings of the ACM SIGMOD International Conference on Management of Data*. – Montreal, 1996. – P. 103–114.
4. M.I.Ryabov, A.L.Sukharev, ON THE ANOMALOUS RISE OF CYCLE 25, Odessa observatory URAN-4 Radio Astronomical Institutes NAS Ukraine
5. T. Ataç, Atila Özgüç, Flare Index of Solar Cycle 22, Kluwer Academic Publishers, 1998.
6. Яровий А. Т., Страхов Є. М. *Багатовимірний статистичний аналіз: навчально-методичний посібник для студентів математичних та економічних спеціальностей*. — Одеса: ОНУ ім. І. І. Мечникова, 2016. — 148 с.

ДОДАТКИ

Додаток А

Табл. 3.1. Класифікація сонячних спалахів за Kallenrode (2001)

Параметр	Імпульсні спалахи	Поступові спалахи
Тривалість м'якого рентгенівського сплеску (SXR)	< 1 години	> 1 години
Час спаду SXR	< 10 хвилин	> 10 хвилин
Висота над фотосферою	$\leq 10^4$ км	$\sim 5 \times 10^4$ км
Об'єм плазми	$10^{26} - 10^{27}$ см ³	$10^{28} - 10^{29}$ см ³
Питома енергія	Висока	Низька
Розмір області в лінії H α	Малий	Великий
Тривалість жорсткого рентгенівського випромінювання (HXR)	< 10 хвилин	> 10 хвилин
Тривалість мікрохвильового радіосплеску	< 5 хвилин	> 5 хвилин
Тип радіосплеску в метровому діапазоні	II, III	II, (III), IV
Наявність CME	Рідко	Часто

Додаток Б

Посилання на веб додаток:

`urlhttps://solar-analysis.streamlit.app/Visualize`

Посилання на код у GitHub:

`https://github.com/MykolaONU/streamlitApp`