УДК 543.061+543.42+543.166

ПОШУК ЗАЛЕЖНОСТЕЙ В МАСИВАХ ДАНИХ ПРИРОДНЬОЇ ГАММА-АКТИВНОСТІ ЗРАЗКІВ ДОВКІЛЛЯ

[1]Стець М.В., [2]Матьовка О.М.

[1]Інститут електронної фізики НАН України, м. Ужгород
[2] Ужгородський національний університет, м. Ужгород

1. Проблематика.

1.1. Перед початком будь-якого наукового дослідження дослідником певного об'єкта - представника ("in-vivo": в живому; "in-vitro": в пробірці; in-situe: в полі) досліджуваної системи, чи досліджуваного процесу в системах, над системами, або між системами, обов'язково (явно, чи неявно), формується мета досліджень, будується план цих досліджень, та можлива модель, котра, в результаті, має бути підтверджена, певним чином зкоректована, або відкинута. На мові системного підходу можна стверджувати, що є досліджувана система, в якій існує ця досліджувана проблема, котра трансформується в аналітичну проблему. Якщо для цієї аналітичної проблеми шукається аналітичне рішення, і якщо воно знайдено, тоді це означає завершення дослідження.

1.2. Практика засвідчує, що можливості, та реалізація сучасних (не в останню чергу, гамма-спектрометричних) досліджень, у багатьох випадках є достатніми для можливості пошуку, та виявлення латентної інформації в масивах отриманих даних – інформації (нових аналітичних проблем), котра не була "цікавою" на попередніх етапах досліджень. Більш детальніше ця проблематика розглянута в [1].

1.3. Розробці методів пошуку, виявлення латентних залежностей, як носіїв певних закономірностей, присвячений цілий напрямок досліджень в галузі інформатики та штучного інтелекту, який отримав назву "Data Mining - видобуток, розкопування даних" [2] .

1.4. Прикладна ядерна гамма-спектрометрія (ПЯГС) зразків довкілля використовується в першу чергу, для задач радіаційної безпеки: оцінки рівнів гамма-активності (як правило, шляхом визначення питомих активностей Ап гамма-активних нуклідів (ГАН)) в цих зразках; і в подальшому, для оцінки загального рівня радіоактивності.

В першу чергу для цих проблем цікавими є техногенні ГАН – ГАН - уламки поділу ядерного палива U235, U238, та ГАН ядерної зброї – ізотопи Pu, Am.

1.5. Зараз (в після-чорнобильський період), зміст та об'єм даних ПЯГС природньої гамма-активності зразків у більшості випадків, визначається значеннями питомих активностей Ап менш "цікавих, рутинних" ГАН - ГАН, що входять в ряди (сімейства) Th232, U235, U238, а також K40, Cs137.

1.6. Нами було виконано виміри природньої (фонової) гамма-активності декількох типів ґрунтів, глин, та виробів з них- цегли, черепиці, посуди. Глина, та вироби з неї – будівельна, технічна та побутова кераміка залишаються суттєвим матеріальним (в тому числі і радіаційним) компонентом нашого довкілля.

Стартовими аналітичними проблемами (задачами) були саме ці задачі: визначення нуклідного складу; визначення питомих активностей Ап зразків.

1.7. В подальшому масив значень Ап ГАН був використаний для оцінки можливостей виявлення додаткової (латентної) інформації. Мета цих досліджень викладена в [6], і може бути коротко сформульована так:

– використання даних ПЯГС ГАН K40, Ac228 (ряд TH232), Pb214 (ряд U238) в цих зразках для розробки методів ідентифікації зразків в рамках можливостей програми Microsoft Excel 2000 , та її наступних модифікацій [3]. Розробка методів ідентифікації, в свою чергу, зводиться до розробки мір подібності та відмінності, аналогічних відстаням у евклідовому просторі, які невеликі у випадках подібності, і великі у випадках відмінності.

1.8. Встановлено, що ці математичні, статистичні, графічні, та інші можливості Excel дозволяють в певній мірі здійснювати систематизацію даних (значень Ап), і виявляти факти факторизації, тим самим вирішуючи задачі вибору мір подібності та відмінності.

Факторизація [4] – декомпозиція (розклад), або агрегація (об'єднання) досліджуваних множин (виборок) на фактор-множини (множини, однорідні по певному фактору - значенню ознаки), лежить в основі теорії вимірювань.

1.9. Можливість факторизації з використанням значень Ап "рутинних" ГАН K40, Ac228, Pb214, котра виявлена на відносно невеликих виборках зразків - це цікавий факт. Ця факторизація проявляється графічно у вигляді скупчень (кластерів) точок. Це - скупчення точок, які є фактор - множинами, об'єднаними відношеннями близькості (подібності).

1.10. Цікавим є і виявлення іншого типу фактор-множин значень – впорядкованої певним чином послідовності точок. Виявлення таких послідовностей, як трендів (регресійних залежностей між значеннямив термінах Excel) є процедурою виявлення цих фактор - множин, які об'єднані відношенням функціональної залежності.

Досить швидко було встановлено, що ці регресійні залежності, з достатньо високим значенням коефіцієнта детермінації $R^2 > 0,7$ [4], мають самостійне значення (див. пп. 1.2, 1.3).

Коефіцієнт детермінації R^2 в термінах Excel має назву коефіцієнта достовірності апроксимації. Інші тлумачення – див [5].

Викладемо зараз коротко схему експерименту, розрахунків, та деякі основні результати пошуку таких залежностей.

2. Аналіз даних.

2.1. Організація і рандомізація масивів даних (зразків). Для вирішення задачі стартово було підібрано (і згруповано) такі типи зразків-глина (11 зразків глин та ґрунтів), цегла (6 зразків), черепиця (3 зразки), посуд (16 зразків).

Кожен з цих типів розглядався, як окрема множина (вибірка). Разом з тим, деякі типи об'єднувались, і теж розглядались як самостійні типи (вибірки): – "цегла + черепиця", "глина+цегла". Зразки типу глини, навпаки, роз'єднувались на підмножини (підтипи).

Зразки посуду не руйнувались. Всі інші зразки подрібнювались, доводились (сушились) до постійної ваги, однак вимірювались і несушеними, з метою виявлення систематичних похибок, пов'язаних з цим етапом підготовки зразків.

Таким чином, ми розглядаємо два надтипи глиняних зразків - зразки, котрі проходили, по крайній мірі, один раз технологію переробки (будівельна кераміка та неглазурований посуд), два рази (глазурований посуд), і зразки, які не оброблялись (глина). Перший надтип (оброблені зразки) може розглядатись, як артефакт.

Окрім цих мір організації, як певної систематизації, були здійснені і "дезорганізаційні" міри (рандомізація): випадковий вибір зразків; відсутність даних про вік, місце виготовлення зразків; випадкова послідовність вимірів.

2.2. Виміри. Виміри природньої гамма – активності зразків здійснювались на спектрометричному комплексі, до складу котрого входить Ge(li) - детектор гамма–випромінювання ДГДК-100В, багато-канальний аналізатор, та програмне забезпечення SBS-40.

2.3. Питомі активності Ап. Результатом вимірів є апаратурні гамма-спектри (АГС), та протоколи (лістинги) обробки цих АГС. На основі даних, що містяться в цих документах, визначаються питомі активності Ап (розмірність Ап - Бекерель/кілограм), розраховані для кожної аналітичної лінії Eg ГАН: Ac228, Pb212,

Tl208 (ряд Th232); Ra226, Pb214, Bi214 (ряд U238); також K40 і Cs137. Цей великий масив значень Ап повинен бути використаний, але потім, якщо можливо, зменшений.

Обробка даних показала, що зменшення масиву значень величин, які використовуються в якості характерних ознак, можливе, і необхідне, шляхом застосування процедур підгонки та апроксимації, які в ядернофізичній термінології мають назву фітінгу (англ.- fitting). Ми будемо використовувати саме цей термін.

2. 4. Фітінг даних.

2.4.1. Оцінка радіоактивної рівноваги. Можливість зменшення масива даних грунтується на встановленому експериментально (для наших зразків, режимів пробовідбору і вимірів) факті виконання радіоактивної рівноваги (РАР) - рівності в межах статистичної похибки значень Ап для ГАН, надійно виділених програмою SBS-40, та пов'язаних генетично: в ряді Th232: Ап(Ac228)=Ап (Pb212) = Ап(Tl208); в ряді U238: Ап(Pb214) = Ап(Bi214).

Поскільки значення Ап, визначені для кожної аналітичної лінії Eg одного і того ж ГАН повинні бути, і були, в межах статистичної похибки, рівними, подальший розгляд здійснювався для усереднених по лініям значень Ап.

2.4.2. Оцінка фону вимірів. Для здійснення цих процедур був виконаний регресійний лінійний аналіз, в якому шукався тренд:

Апі(зразок)=f(Апj)=Апj(зразок) (1),

де Апі, Апj – значення питомих активностей для різних аналітичних ліній Eg i,j- одного і того ж ГАН, в різних зразках одного і того ж типу.

Для оцінки вкладу фону виміру, а саме значень Ап (зразок+фон), шукалась залежність:

Ап (зразок+фон)=Ап (зразок)+Ап (фон) (2)

Відмінність між виразами (1) і (2) очевидна. У випадку (1) пряма проходить через 0, що і повинно бути для значень Ап зразка без фону (Ап(зразок)). У випадку (2) пряма через 0 може не проходити, якщо рівень фона помітний (Ап(фон)>0). Зрозуміло, що більш правильними є значення Ап(зразок) із (1), які в подальшому і використовувались.

Для K40, у якого є тільки одна аналітична лінія Eg=1460 КеВ, методика оцінки вкладу фону буде іншою, і грунтується на регресійному аналізі значень Ап K40, визначених для різних мас одного і того ж зразка.

Точність фітінгу визначається значеннями коефіцієнта детермінації R^2, і може бути досить високою. Для досліджуваних задач задовільною вважали R^2 >0,7.

Після виконання названих вище процедур фітінгу, які повинні здійснюватись одночасно, та попереднього статистичного аналізу, масив значень Ап можна зменшити, без втрати інформації, розглядаючи апроксимовані значення Ап. Тому надалі розглядались тільки ГАН Ac228 (ряд Th232), Pb214 (ряд U238), K40.

Таким чином, вибором саме цих ГАН, ми виключаємо час як фактор.

Ra226 та Cs137 не розглядались, поскільки були надійно виділені програмою SBS-40 тільки у деяких вимірах.

2.5. Стандартизація і нормалізація. Значення Ап, які є характерними ознаками (факторами), використовувались: безпосередньо, а також після процедур стандартизації, та нормалізації..

Для стандартизованих (Аст) та нормалізованих (Ан) значень визначались середні значення для Ап (функція Excel СРЗНАЧ), стандартні відхилення σ (функція СТАНДОТКЛОН), а потім, відповідно, стандартизовані значення Аст =Ап/Ап, та нормалізовані значення Ан, визначені за допомогою функції НОРМАЛИЗАЦИЯ. Подальший розрахунок виконувався для цих трьох масивів (варіантів) значень - Ап, Аст, та Ан. При зміні типа зразка (в тому числі за рахунок створення нових типів - див. вище) здійснювався перерахунок Аст та Ан.

Крім значень Ап, визначались і використовувались в якості характерних ознак їх відношення: Ac228/K40, K40/Ac228, Ac228/Pb214, K40/Pb214, Pb214/K40. Для цих величин теж здійснювалась стандартизація та нормалізація.

Розрахунки, виконані в одиницях Ап, Аст, Ан, показали, що, як і очікувалось, їх графіка не відрізняється.

2.6. Діаграми "Апі-Апj-k". Отримані значення Ап, Аст, Ан, та їх відношень, дозволяють побудувати виразні графічні представлення – патерни (образи) - для кожного ГАН кожного типу зразка. Це - точкові та бульбашкові діаграми Excel [4, 5], які дозволяють здійснити мапінг - дво-, або тривимірне представлення даних в графічному вигляді. Приклади мапінгу - власне географічні мапи (карти), та томографії. В нашому випадку - це діаграми "Апі-Апj-k" питомих активностей Ап ГАН I, та ГАНj зразка k (k-імя, або, частіше, номер зразка в множині зразків одного типу).

Діаграми "Апі-Апj/Апl.-k". Характерними ознаками можуть бути також і відношення питомих активностей Апі/Апl, і відповідно, діаграми "Апі-Апj/Апl-k", котрі пов'язують трійку ГАН- K40, Ac228, Pb214 в різних комбінаціях { i, j, l }.

Діаграми "Апі/Апj-Апj/Апl-k". Порівняльний аналіз масивів отриманих діаграм "Апі-Апj-k", діаграм "Апі-Апj/Апl-k", та діаграм "Апі/Апj-Апj/Апl-k", засвідчує, що вони також є патернами типів і можуть бути використані для вирішення певних задач ідентифікації.

Важливо, що діаграми дозволяють виявляти в них різні структури - скупчення., окремі точки, впорядковані сукупності точок (тренди, залежності). Таким чином, їх використання в певній мірі вирішує задачу пошуку мір подібності та відмінності. Цими мірами, що цікаво, є кількісні величини значення питомих активностей, та похідні від них величини - відношення питомих активностей.

2.7. Регресійні залежності. Для виявлення, та оцінки залежностей Апі = F(Апj), Апі = F(Апj / Апl), Апі/Апj = F(Апі/Апl), використано можливості регресійного аналізу Excel.

Дані апроксимувались лінійними залежностями, інтерпретація яких є, як відомо, найбільш простою і зрозумілою.

2.8. Алгоритм пошуку залежностей. Алгоритм пошуку залежностей полягав в послідовному виключенні, та включенні точок множини (типу), поки не буде отримано лінійний тренд із значенням коефіцієнта детермінації $R^{2.} > R^{2}0$, де $R^{2}0$ – коефіцієнт детермінації для всієї множини точок.

Після зупинки, коли $R^2 \geq R^{2.}$зад, пошук залежності розпочинається в множині виключених точок. $R^{2.}$зад- значення коефіцієнта детермінації, що задовільняє умовам задачі.

Пошук завершується, коли знайдено всі можливі залежності.

2.9. Приклади пошуку залежностей. Деякі основні результати пошуку регресійних лінійних залежностей приведені на рис.1-рис.18, у вигляді діаграм "Апі – Апj - k", для пар ГАНі, ГАНj: Ac228-Pb214; Pb214-Ac228; Ac228-K40; Pb214-K40. Частина діаграм приведено в одиницях Ап, деякі в одиницях Аст.

Приводяться дані для типу зразків "глина" (рис.1- рис.6), типу зразків "посуд" (рис.7 - рис.12), та типу "цегла + черепиця" (рис.13 - рис.18).

Діаграми, які ілюструють можливості пошуку регресійних лінійних залежностей, приведені в порівняльному "режимі" - до регресійного аналізу (непарні номери рисунків); після регресійного аналізу (парні номери рисунків).

2.9.1.Тип зразків "глина" (рис.1-рис.6). Цей тип зразків (див. пункт 2.1), є, по суті, складним, що видно на рис 1, 3,. 5, та по невисоким значенням коефіцієнта детермінації R^2 , однак дозволяє знайти залежності

2.9.2. Тип зразків "посуд" (рис.7 - рис.12). На рис 7, 9, 11 видно факторизацію скупчень точок, а також факторизацію послідовностей (див. рис. 9). Регресійний аналіз підтверджує цю апріорну факторизацію з високим значеннм R^2 , виявляючи і інші тренди.

2.9.3. Тип "цегла + черепиця" (рис.13 - рис.18) створений, зрозуміло, штучно, шляхом об'єднаних розрахунків Аст, та Ан. На рис.13, 15, 17, зразки черепиці позначені світлими кружками, зразки цегли - темними.

Видно що для пари Ac228-Pb214 тренди цих типів майже ортогональні, а значення R^2 високі, тому факторизація (див. рис. 13, 14) знаходить ці зразки. У інших випадках, де значення R^2 для черепиці

невисокі, розпізнавання не відбувається: факторизація виділяє інші залежності (див., напр., рис.18), де чітко виділяються два паралельні тренди, які, можливо, свідчать про дві, відмінні одна від одної, технології виготовлення будівельної кераміки (технології виготовлення цегли і черепиці).

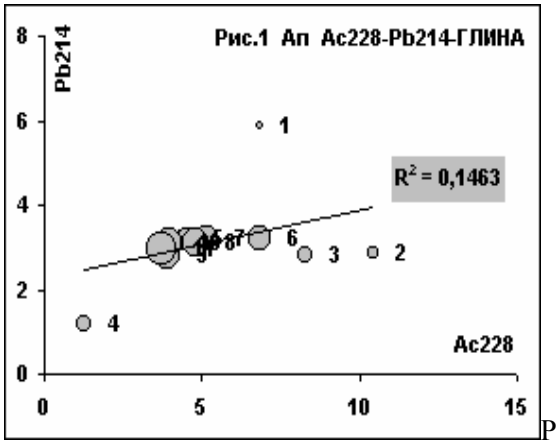

Рис. 1. Діаграма “Апі – Апj - k” Ac228, Pb214для зразків глини до регресійного аналізу.

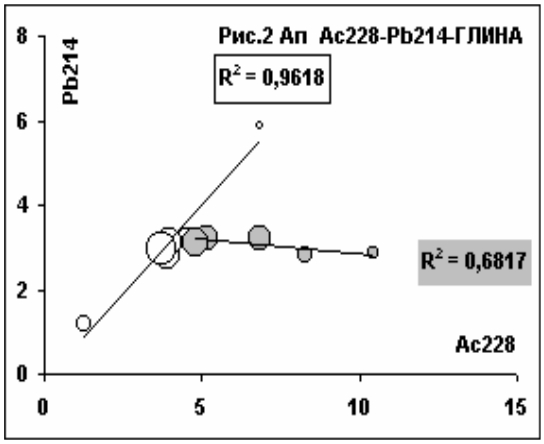

Рис.2. Діаграма “Апі – Апj - k” Ac228, Pb214 для зразків глини після регресійного аналізу.

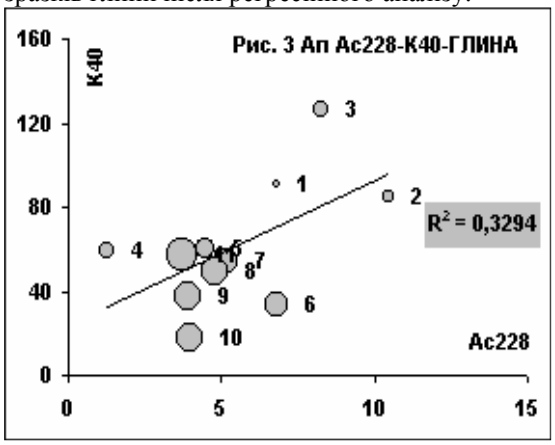

Рис.3. Діаграма “Апі – Апj - k” Ac228, K40 для зразків глини до регресійного аналізу.

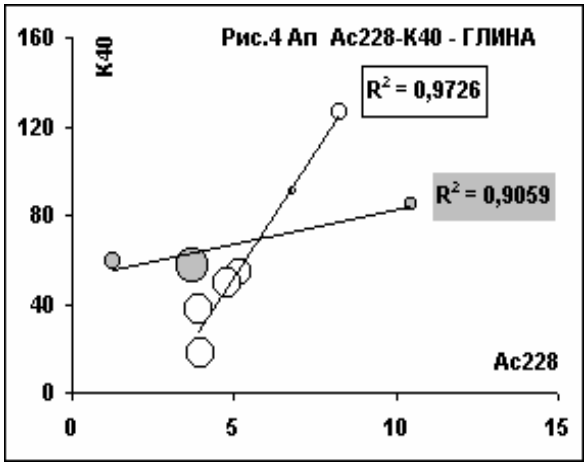

Рис. 4. Діаграма “Апі – Апj - k” Ac228, K40 для зразків глини після регресійного аналізу.

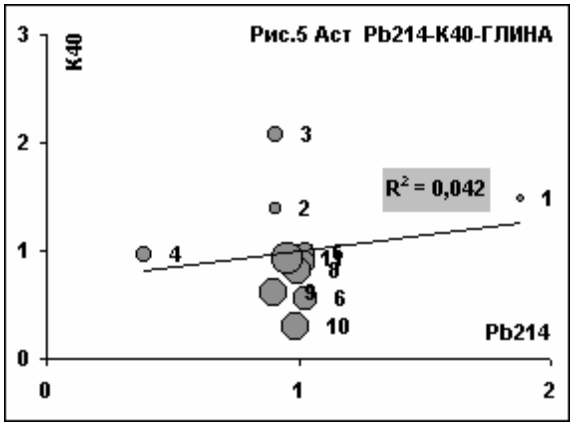

Рис.5. Діаграма “Асті – Астj - k” Pb214, K40 для зразків глини до регресійного аналізу.

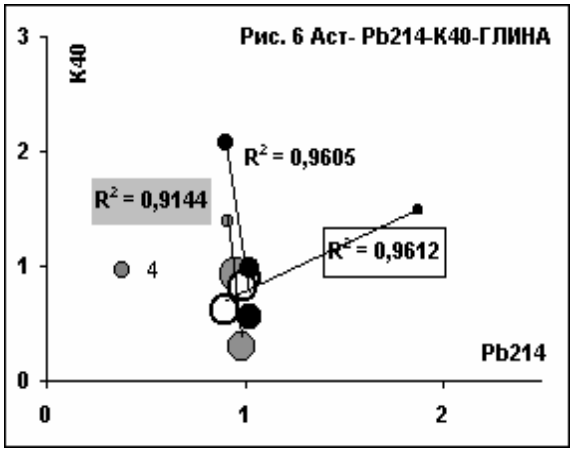

Рис.6. Діаграма “Асті – Астj - k” Pb214, K40 для зразків глини після регресійного аналізу.

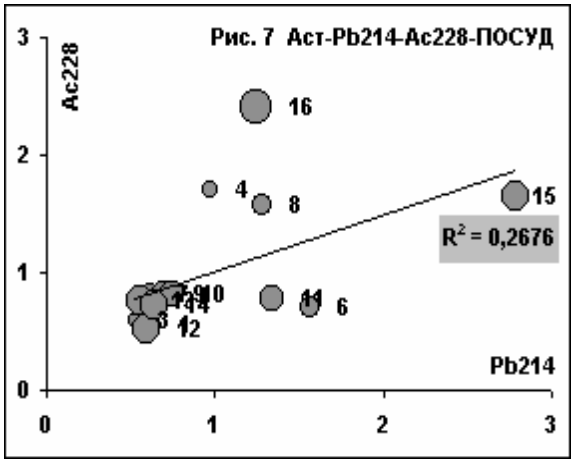

Рис.7. Діаграма “Асті – Астj - k” Pb214, Ac228 для зразків посуду до регресійного аналізу.

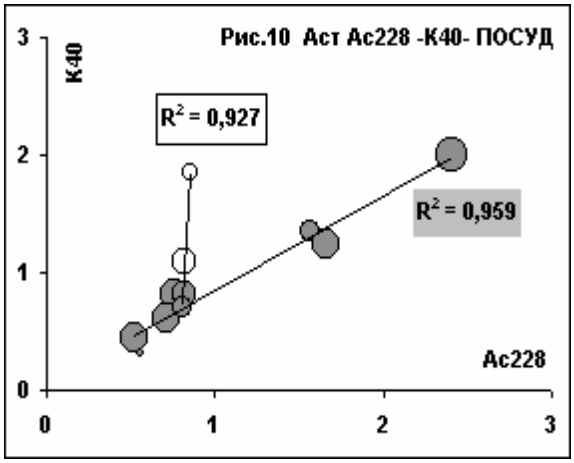

Рис.10. Діаграма “Апі – Апj - k” Ac228, K40 для зразків посуду після регресійного аналізу.

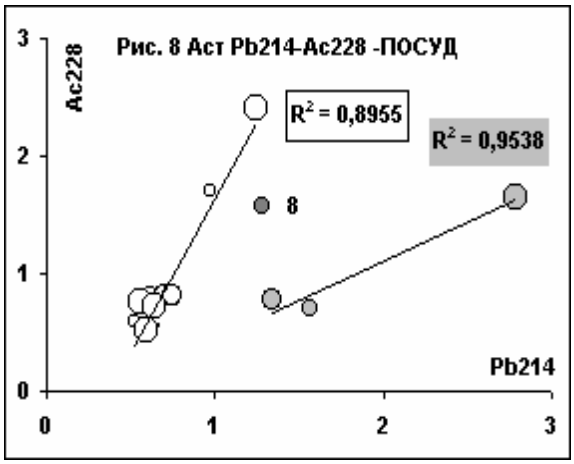

Рис.8. Діаграма “Асті – Астj - k” Pb214, Ac228 для зразків посуду до регресійного аналізу.

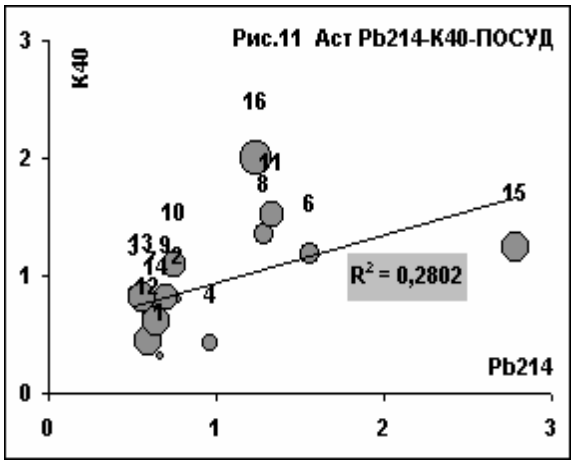

Рис.11. Діаграма “Асті – Астj - k” Pb214, K40 для зразків посуду до регресійного аналізу.

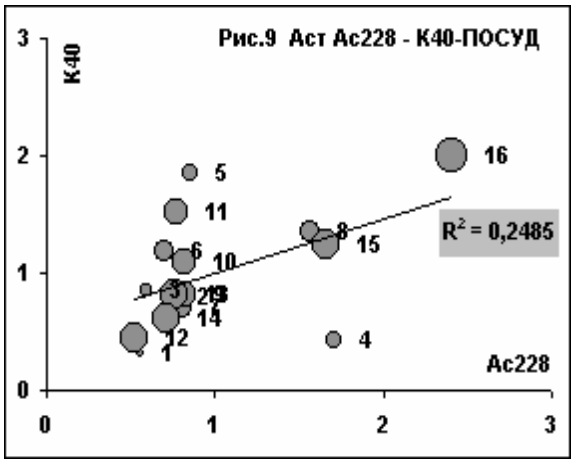

Рис.9. Діаграма “Асті – Астj - k” Ac228, K40 для зразків посуду до регресійного аналізу.

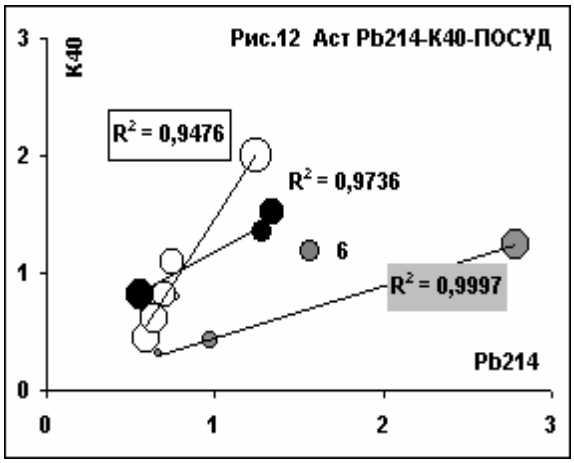

Рис.12. Діаграма “Асті – Астj - k” Pb214, K40 для зразків посуду після регресійного аналізу.

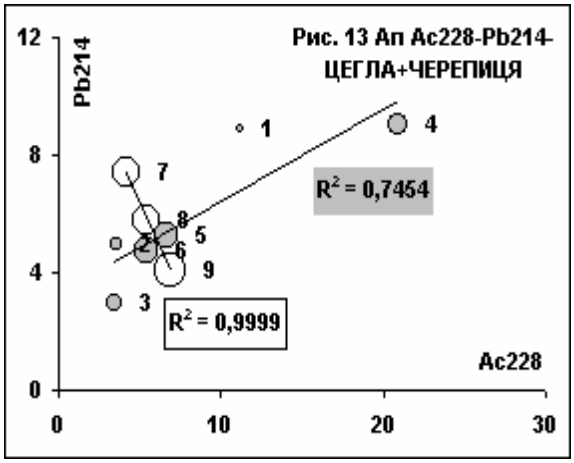

Рис.13. Діаграма “Апі – Апj - k” Ac228, Pb214 для зразків” цегла+черепиця” до регресійного аналізу.

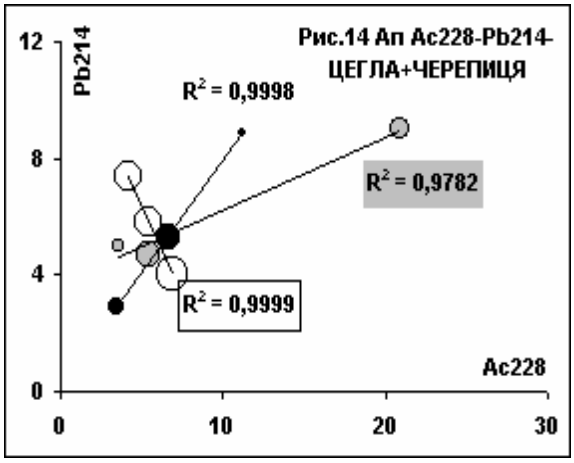

Рис.14. Діаграма “Апі – Апj - k” Ac228, Pb214 для зразків “цегла+черепиця” після регресійного аналізу.

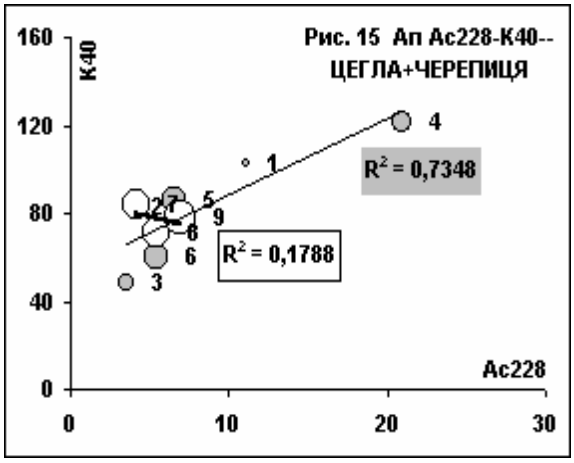

Рис.15. Діаграма “Апі – Апj - k” Ac228,K40 для зразків ”цегла+черепиця” після регресійного аналізу.

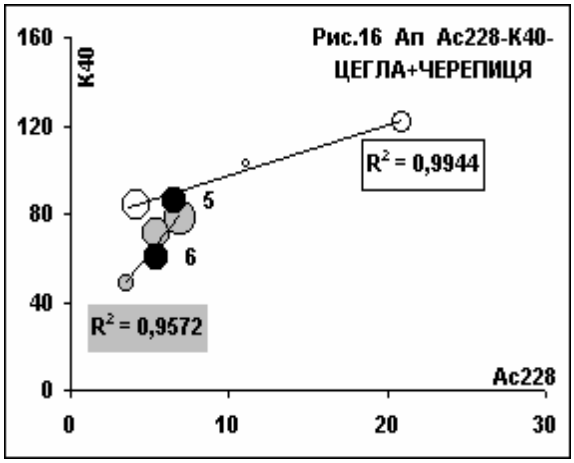

Рис.16. Діаграма “Апі – Апj - k” Ac228, K40 для зразків” цегла+черепиця” після регресійного аналізу.

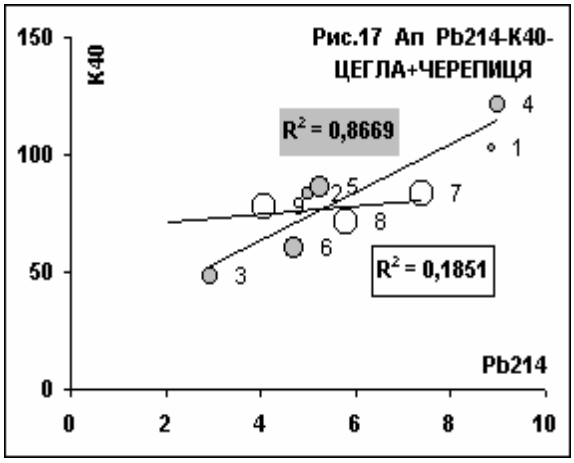

Рис.17. Діаграма “Апі – Апj - k” Pb214, K40 для зразків “цегла+черепиця” до регресійного аналізу.

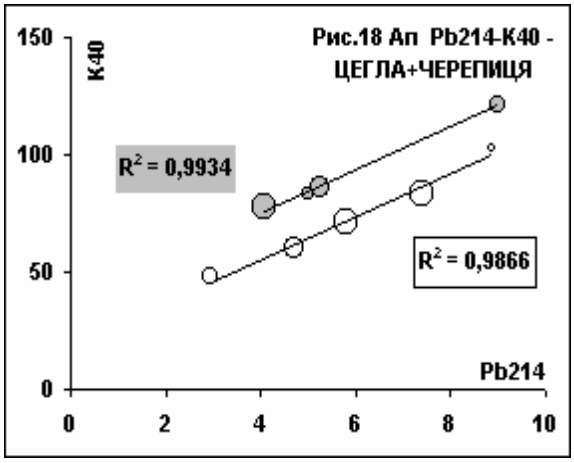

Рис.18. Діаграма “Апі – Апj - k” Pb214, K40 для зразків “цегла+черепиця“ після регресійного аналізу.

3. Висновки.

3.1. На основі експериментальних даних прикладної ядерної гамма-спектрометрії природньої гамма-активності зразків довкілля (грунти, глини, керамічні вироби з них) розглянуто можливості регресійного аналізу з метою пошуку емпіричних залежностей між значеннями характерних ознак, в якості яких використані значення питомих активностей K40, Ac228 (ряд Th232) , Pb214 (ряд.U238).

3.2. Показано, що в багатьох випадках є можливість виявлення таких залежностей, як лінійних регресій з високим коефіцієнтом детермінації (R^2>0,7).

3.3. По інформативності діаграми “Апі-Апj-k” не поступаються інформативності дендрограмам кластерного аналізу, позаяк дозволяють оцінити дистанцію (відстань) між точками (зразками).

3.4. Алгоритм пошуку залежностей, реалізований “вручну“ на невеликих множинах точок, може бути автоматизований .

3.5. Регресійні та кореляційні залежності не дають пояснення природи цих залежностей. Причинна природа залежностей для кожного із досліджених типів зразків, та в цілому - це окрема тема.

3.6. Перспективність розглянутих методів пошуку залежностей, на основі отриманих результатів, як методики, на наш погляд, очевидна, однак потребує доопрацювання, та підтвердження, зокрема, на масивах даних природньої гамма-активності калібрувальних (стандартних) зразків довкілля.

Література

1. Стець М. В. Емпіричний вибір моделей прикладної ядерної гамма-спектрометрії матеріальних систем. .// Вісник Ужну. Серія Хімія.- 2007.- Вип.18. – С. 116- 125.
2. Дюк В., Самойленко. А .Data Mining. Учебный курс. СПб, Питер, 2001, 365 стр.
3.Пфанцагль И. Теория измерений. М., Мир, 1976, 248стр..
4. Уокенбах Дж. Диаграммы в Excel. М., Диалектика, 2003, 448 стр.
5.. Васильев А.Н. Научные вычисления в Microsoft Excel. (Диалектика, Москва, 2004)
6. Стец М.В., Маслюк В.Т., Бузаш В.М., Матьовка О.М. Прикладная ядерная гамма-спектрометрия керамических артефактов. // Тезисы докладов 7 конференции по физике высоких энергий, ядерной физике и ускорителям.. 23-27 .02. 2009. Харьков. Стр.32.

SEARCH FOR DEPENDENCES IN THE DATA ARRAY FOR THE NATURAL GAMMA-ACTIVITY OF ENVIRONMENTAL SAMPLES

Stets M.V., Matyovka O.M.

Based on the experimental data on the applied nuclear gamma-spectrometry of the natural gamma-activity of environmental samples (soils, clays and related products) the possibilities of regressive analysis are considered aimed at the search of empiric relations between the specific sign values, i.e. the specific activities of K40, Ac228 (Th232 series), Pb214 (U238 series). It has been shown that in many cases it is possible to find such relations as linear regressions with high determination ratio (R^2>0.7).