

# РОЗВИТОК РЕГІОНАЛЬНОЇ ЕКОНОМІКИ

---

М.Д. БІЛИК,  
д.е.н., професор, Київський національний економічний університет ім. В. Гетьмана,  
Н.В. СЕНЬКО,  
бакалавр, Київський національний університет ім. Т. Шевченка

## Аналіз енергоспоживання регіонів України за допомогою статичних методів кластеризації

*У статті розглянуто основні засоби інтелектуального аналізу даних. Застосовано кластерний підхід на основі методу  $k$ -середніх для виявлення найбільш енергоємних регіонів України. Відповідно до розглянутого методу з'являється можливість вдосконалення розподілу енергоресурсів.*

**Ключові слова:** кластерний підхід, метод  $k$ -середніх, регіони України, енергоресурси.

*В статье рассмотрены основные средства интеллектуального анализа данных. Использован кластерный подход на основе метода  $k$ -средних для выявления наиболее энергоемких регионов Украины. Согласно рассмотренному методу появляется возможность совершенствования распределения энергоресурсов.*

**Ключевые слова:** кластерный подход, метод  $k$ -средних, регионы Украины, энергоресурсы.

*The article reviews the basic tools of intellectual analysis. Applied clustering approach based on  $k$ -means method to identify the most energy-intensive regions of Ukraine. According to the reporting method it is possible to improve the distribution of energy.*

**Актуальність теми.** Одним з основних чинників функціонування економіки будь-якої країни є енергоресурси. Наявність види, доступ до них значно впливають на стан еконо-

міки окремих регіонів. Розвиток промисловості в Україні зумовлює зростання обсягу споживання енергоресурсів, запаси яких не є нескінченими. Тому виникає потреба в застосуванні кластерного аналізу, який дасть змогу об'єднати регіони відповідно до їх енергетичних потреб з метою забезпечення ефективного управління ресурсами.

**Постановка проблеми.** З погляду енергоефективності Україна відстає від інших європейських країн у зв'язку із застосуванням застарілих технологій, що використовуються в багатьох галузях економіки, неекономним використанням енергії, неоптимальною структурою виробництва, тому здійснення кластерного аналізу дозволяє провести найефективнішу структуризацію регіонів, що забезпечить вирішення поставлених проблем.

**Аналіз досліджень та публікацій з проблеми.** Кластеризації присвячені праці таких видатних вчених: С.С. Бакая, В.І. Бойка, А.Д. Войотовича, Ю.П. Воскобойніка, Т.В. Кравец, В.М. Кутяїна, М.Г. Лобаса, І.І. Лукінова, Б.В. Погріщука, М.Я. Полоцького, П.Т. Саблука, Р.П. Саблука, О.К. Слюсаренка, О.О. Сторожука, І.І. Червено, однак не було проведено кластеризацію України за показниками енергоспоживання.

**Метою статті** є застосування інтелектуального аналізу даних, а також кластеризація даних методом  $k$ -середніх для формування кластерів регіонів України відповідно до результатів.

**Виклад основного матеріалу.** Інтелектуальний аналіз даних (datamining) – це процес отримання та подальше застосування знань або раніш невідомої інформації із уже наявних доступних даних. Під цим поняттям ховається широке розмаїття технологій і процесів, за допомогою яких вхідні «сирі» дані оброблюються, чистяться та аналізуються.

Засоби DataMining на основі наявних даних самі можуть будувати моделі, які дають змогу кількісно та якісно оцінювати ступінь впливу різних досліджуваних факторів на задану властивість об'єкта. Крім того, вони дають змогу формулювати нові гіпотези про характер досі невідомих, але таких, що реально існують, залежностей між даними.

Методи багатовимірної аналізу – найбільш дієвий кількісний інструмент дослідження соціально-економічних процесів, які описуються великою кількістю характеристик. До них відноситься кластерний аналіз, таксономія, розпізнавання образів, факторний аналіз [7].

Кластерний аналіз найбільш яскраво відображає риси багатовимірної аналізу в класифікації. Його можна застосовувати в різноманітних ситуаціях, які зустрічаються як у наукових, так і часто у дослідженнях прикладного характеру.

Нехай множина  $I = \{I_1, I_2, \dots, I_n\}$  визначає  $n$  об'єктів (індивідів), які належать деякій популяції  $\pi_1$ . Припустимо також, що існує деяка множина спостережуваних показників або характеристик  $C = [C_1, C_2, \dots, C_p]^T$ , якими наділений кожен індивід з  $I$ . Спостережні характеристики можуть бути як кількісними, так і якісними. Результат виміру  $i$ -ої характеристики  $I_j$  об'єкта позначимо символом  $x_{ji}$ , а вектор  $X_j = [x_{ji}]$ , розмірності  $p \times 1$  буде відповідати кожному ряду вимірів (для  $j$ -го індивіда). Таким чином, для множини індивідів  $I$  дослідник має множину векторів вимірів  $X = \{X_1, X_2, \dots, X_n\}$ , які описують множину  $I$ . Відзначимо, що множина  $X$  може бути представлена як  $n$  точок у  $p$ -вимірному евклідовому просторі  $E_p$ .

Нехай  $m$  – ціле число, яке менше за  $n$ . Тоді можна сформулювати задачу кластерного аналізу, яка полягає в тому, що на основі даних, які містить множина  $X$ , треба розбити множину об'єктів  $I$  на  $m$  кластерів (підмножин)  $\pi_1, \pi_2, \dots, \pi_m$  так, щоб кожен об'єкт  $I_i$ , який належав одній і тільки одній підмножині розбиття і щоб об'єкти, які належать одному і тому ж кластеру, були подібними, в той час як об'єкти, які належать різним кластерам, були різнорідними (неподібними) [6].

Розв'язком задачі кластерного аналізу є розбиття, яке задовольняє певному критерію оптимальності [4]. Цей критерій може бути представлений як деякий функціонал, який виражає рівні бажаності розбиття та групування. Даний функціонал часто називають цільовою функцією. Наприклад, як цільова функція може бути взята внутрішньогрупова сума квадратів відхилень:

$$W = \sum_{j=1}^n (x_j - \bar{x})^2 = \sum_{j=1}^n x_j^2 - \frac{1}{n} \left( \sum_{j=1}^n x_j \right)^2 \quad (1)$$

де  $x_j$  – являє собою виміри  $j$ -го об'єкта.

Для розв'язання задачі кластерного аналізу необхідно визначити поняття подібності та неоднорідності.

Зрозуміло, що  $i$ -й та  $j$ -й об'єкти попадали б в один кластер, коли відстань (віддаленість) між точками  $X_i$  та  $X_j$  була б достатньо маленькою та попадали б до різних кластерів, коли ця відстань була б достатньо великою. Таким чином, попадання в один або різні кластери об'єктів визначається поняттям відстані між  $X_i$  та  $X_j$  з  $E_p$ , де  $E_p$  –  $p$ -вимірний евклідов простір. Невід'ємна функція  $d(X_i, X_j)$  називається функцією відстані (метрикою), якщо:

$$d(X_i, X_j) \geq 0, \text{ для всіх } X_i \text{ та } X_j \in E_p$$

$$d(X_i, X_j) = 0, \text{ тоді і тільки тоді, коли } X_i = X_j$$

$$d(X_i, X_j) = d(X_j, X_i)$$

$$d(X_i, X_j) \leq d(X_i, X_k) + d(X_k, X_j),$$

$$\text{де } X_i, X_j \text{ та } X_k - \text{будь-які три вектори з } E_p \text{ [4].}$$

Значення  $d(X_i, X_j)$  для  $X_i$  та  $X_j$  називається відстанню між  $X_i$  та  $X_j$  і еквівалентна відстані між  $I_i$  та  $I_j$  відповідно вибраними характеристиками  $C = [C_1, C_2, \dots, C_p]^T$ .

Найбільш вживаними є такі функції відстані:

$$\text{Евклідова відстань: } d_2(X_i, X_j) = \left[ \sum_{k=1}^p (x_{ki} - x_{kj})^2 \right]^{\frac{1}{2}}.$$

$$l_1\text{-норма: } d_1(X_i, X_j) = \sum_{k=1}^p |x_{ki} - x_{kj}|.$$

$$\text{Сюпремум-норма: } d_\infty(X_i, X_j) = \sup_{k=1, p} \{ |x_{ki} - x_{kj}| \}.$$

$$l_p\text{-норма: } d_p(X_i, X_j) = \left[ \sum_{k=1}^p |x_{ki} - x_{kj}|^p \right]^{\frac{1}{p}}.$$

Евклідова метрика являється найбільш популярною. Метрика  $l_1$  найбільш легка для розрахунків. Сюпремум-норма легко розраховується і включає в себе процедуру впорядкування, а  $l_p$ -норма охоплює функції відстані 1, 2 та 3.

Нехай  $n$ -вимірів  $X_1, X_2, \dots, X_n$  представлені у вигляді матриці даних розміром  $p \times n$ :

$$x = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{p1} & x_{p2} & \dots & x_{pn} \end{pmatrix} = (X_1, X_2, \dots, X_n).$$

Тоді відстань між парами векторів  $X_i$  та  $X_j$  може бути представлена у вигляді симетричної матриці відстаней:

$$D = \begin{pmatrix} 0 & d_{12} & \dots & d_{1n} \\ d_{21} & 0 & \dots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \dots & 0 \end{pmatrix}$$

Протилежним поняттям до відстані є поняття подібності між об'єктами  $I_i$  та  $I_j$ . Невід'ємна дійсна функція  $S(X_i, X_j) = S_{ij}$  називається мірою подібності, якщо:

$$0 \leq S(X_i, X_j) < 1, \text{ для } X_i \neq X_j$$

$$S(X_i, X_i) = 1$$

$$S(X_i, X_p) = S(X_p, X_i)$$

Пари значень мір подібності можна об'єднати в матрицю подібності:

$$S = \begin{pmatrix} 1 & s_{12} & \dots & s_{1n} \\ s_{21} & 1 & \dots & s_{2n} \\ \dots & \dots & \dots & \dots \\ s_{n1} & s_{n2} & \dots & 1 \end{pmatrix}$$

Величину називають коефіцієнтом подібності [1].

За способом розбиття на кластери алгоритми бувають двох типів: ієрархічні та неієрархічні. Класичні ієрархічні алгоритми працюють лише з категоріальними атрибутами, коли будується повне дерево вкладених кластерів. Тут поширені агломеративні методи побудови ієрархій кластерів – в них відбувається послідовне об'єднання вихідних об'єктів та відповідне зменшення числа кластерів. Ієрархічні алгоритми забезпечують порівняно високу якість кластеризації та не потребують попереднього встановлення кількості кластерів [2]. Більшість з них мають складність  $O(n^2)$ .

Неієрархічні алгоритми базуються на оптимізації деякої цільової функції, яка визначає оптимальне в деякому сенсі розбиття множини об'єктів на кластери. До цієї групи відносяться алгоритми сімейства  $k$ -середніх ( $k$ -means, fuzzy-means, Густафсон – Кесселя), які як цільову функцію використовують суму квадратів зважених відхилень координат об'єктів від центрів шуканих кластерів. Кластери шукаються сферичної або еліпсоїдної форми. Зупинимось детальніше на методі  $k$ -means [2].

Метод  $k$ -середніх базується на мінімізації суми квадратів відстаней між кожним елементом початкових даних та цен-

тром його кластера, тобто функції  $J = \sum_{k=1}^M \sum_{i=1}^N d^2(x_i, c_k)$ ,

$x_i \in X$  – об'єкт кластеризації (точка),  $c_j \in C$  – центр кластера (центроїд).

На момент старту алгоритму має бути відомим число  $M$  (кількість кластерів). Вибір числа  $M$  може базуватися на результатах попередніх досліджень, теоретичних міркувань або інтуїції [7].

Опис алгоритму. Першочергове розподілення об'єктів по кластерах. Обираються  $M$  точок. На першому кроці ці точки

вважаються центрами кластерів. Вибір початкових центроїдів може проводитися шляхом підбору спостережень для максимізації початкової відстані, випадковим вибором спостережень або вибором перших спостережень [4].

Ітеративний розподіл об'єктів по кластерах. Об'єкти розподіляються по кластерах шляхом розрахунку відстані від об'єкту до центрів кластерів та вибору найменшого.

Коли всі об'єкти розподілені по кластерах, заново розраховуються їх центри.

$$c_j = \frac{\sum_{i=1}^L x_i}{L}, \quad x_i \in C_j, L - \text{обсяг кластеру.}$$

Якщо  $c_j = c_{j-1}$ , то це означає, що кластерні центри стабілізувались і відповідно розподіл закінчено. Інакше повертаємось до кроку 1.

Важким є вибір числа кластерів. У випадку, коли припущень немає, зазвичай роблять декілька спроб, порівнюючи результати. Після отримання результатів кластерного аналізу методом  $k$ -середніх слід перевірити правильність кластеризації (тобто оцінити, наскільки кластери відрізняються один від одного). Для цього розраховуються середні значення для кожного кластеру. За умов гарної кластеризації мають бути отримані досить різні середні значення для всіх вимірів або хоча б для більшої частини [5].

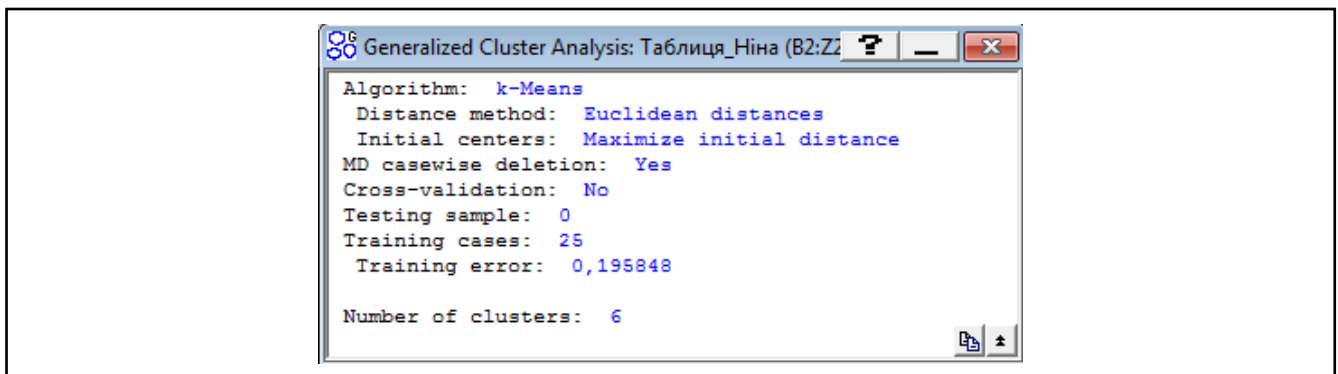
Основні переваги алгоритму  $k$ -середніх:

- простота використання;
- швидкість використання;
- зрозумілість та прозорість алгоритму.

Недоліки методу:

- значна чутливість до викидів, які можуть спотворювати середнє;
- може повільно працювати на великих базах даних [6].

Для кластеризації областей України в роботі використовується база даних, розроблена автором на основі даних Державного комітету статистики. Всі розрахунки проводилися за допомогою комп'ютерної програми обробки статистичних даних Statistica, програми для роботи з самоорганізованими картами ViscoverySOMine 5.2 та програми для створення й обробки електронних таблиць Microsoft Excel.



#### Результати кластеризації

\* Джерело: власні розрахунки автора.

## РОЗВИТОК РЕГІОНАЛЬНОЇ ЕКОНОМІКИ

Для проведення результатів кластеризації за допомогою самоорганізованих карт Кохонена проведемо додаткову кластеризацію методом  $k$ -середніх. Для цього скористаємося командами меню Data Mining – Generalized EM &  $k$ -Means Cluster Analysis.

За основу обиралися кількість кластерів (6), яку «запропонувала» програма ViscoverySOMine 5.2. Алгоритм був запрограмований на 50 ітерацій з максимізацією початкових відстаней, які розраховувались як Евклідові. Вікно з результатами розташоване на рисунку.

Щоб охарактеризувати отримані кластери, наведемо статистику середніх величин по кластерах (табл. 1).

**Кластер 1.** До його складу входять дві області – Одеська та Хмельницька. Характерною особливістю даного кластеру є найнижчий середній обсяг використання вугілля. Проте регіони цього кластеру знаходяться на другому місці за споживанням електроенергії. Стосовно результатів використання природного газу та нафти можна сказати, що вони є низькими.

**Кластер 2** представлений трьома областями: Львівською, Полтавською та Харківською. Середнє споживання нафти та нафтопродуктів становить 0,193 млн. т умовного палива (т.у.п.), що є другим результатом. Обсяги використаної електроенергії одні з найнижчих серед кластерів.

**Кластер 3.** Його обсяг є одним з найбільших і включає вісім елементів. Це такі області, як Вінницька, Волинська, Запорізька, Сумська, Черкаська, Чернівецька, Чернігівська та АР Крим. Середні обсяги споживання вугілля є четвертими (0,701 млн. т.у.п.) серед кластерів. Знаходиться на передостанньому місці в рейтингу використання природного га-

зу і нафти та нафтопродуктів. Характерним є третій за розмірами показник витраченої електроенергії.

**Кластер 4** має однаковий розмір з другим кластером, який включає Дніпропетровську, Київську та Луганську області. Характеризується найвищим середнім обсягом використання нафти та нафтопродуктів (3,632 млн. т.у.п.). Займає другу позицію у споживанні вугілля та природного газу та третю – в електроенергії.

**Кластер 5.** Його представником є Донецька область. Кластер є лідером за показниками спожитого вугілля, природного газу та електроенергії. Займає четверту сходинку у використанні нафти та нафтопродуктів.

**Кластер 6.** Містить такі області: Житомирська, Закарпатська, Івано-Франківська, Кіровоградська, Миколаївська, Рівненська, Тернопільська і Херсонська. Вони в середньому споживають найменше електроенергії, природного газу, сирої нафти та нафтопродуктів. Показник використаного вугілля становить 0,344 млн. т.у.п., що є п'ятим результатом.

Щоб перевірити якість кластеризації (однією з її характеристик є висока різниця середніх значень величин по кожному кластеру), необхідно провести дисперсійний аналіз отриманих кластерів.

Як видно з табл. 3 усі значення  $p < 0,05$ , тобто з ймовірністю 95% ми можемо стверджувати про значну різницю між кластерами.

Застосуємо алгоритм кластеризації для 23 регіонів, включивши ті, що мають яскраво виражені особливості, та припустивши, що число кластерів дорівнює 7. Результати кластеризації продемонструємо за допомогою таблиці середніх значень (табл. 3).

Таблиця 1. Середні значення змінних

Кластер	1	2	3	4	5	6
Вугілля	0,198	1,072	0,701	5,354	22,462	0,345
Нафта та нафтопродукти	1,709	2,649	0,446	3,632	1,273	0,405
Газ природний	2,126	3,353	1,569	6,360	8,355	1,219
Електроенергія	0,579	0,193	0,355	0,446	0,612	0,146
Число елементів	2	3	8	3	1	8
Частота, %	8	12	32	12	4	32

\* Джерело: власні розрахунки авторів.

Таблиця 2. Результати дисперсійного аналізу кластерів

Case names	Between	df	Within	df	F	p value
Вугілля	596,9393	5	38,44750	19	58,99914	0,000000
Нафта та нафтопродукти	69,7768	5	37,08297	19	7,15023	0,000648
Газ природний	262,9582	5	18,26669	19	54,70291	0,000000
Електроенергія	2,6228	5	0,14346	19	69,47327	0,000000

\* Джерело: власні розрахунки авторів.

Таблиця 3. Середні величини по кластерах

Кластер	1	2	3	4	5	6	7
Вугілля	0,218	1,659	0,052	0,030	0,056	3,810	1,849
Нафта та нафтопродукти	1,12	0,542	0,293	5,811	0,364	4,776	0,940
Газ природний	1,80	2,251	0,941	2,980	1,098	6,044	3,045
Електроенергія	0,502	0,353	0,309	0,213	0,135	0,541	0,193
Кількість елементів	4	3	3	1	7	2	3
Частота, %	17,391	13,043	13,043	4,347	30,434	8,695	13,04

\* Джерело: власні розрахунки авторів.

Таблиця 4. Значення характеристик дисперсійного аналізу кластерів

Casenames	Between	df	Within	df	F	p value
Вугілля	46,9309	6	13,82675	16	9,05122	0,000206
Нафта та нафтопродукти	87,9483	6	15,49484	16	15,13594	0,000009
Газ природний	147,0427	6	15,93671	16	24,60444	0,000000
Електроенергія	2,2462	6	0,09101	16	65,81740	0,000000

\* Джерело: власні розрахунки авторів.

**Кластер 1.** Посідає друге місце за обсягом та включає такі чотири області: АР Крим, Одеська, Хмельницька і Чернігівська. Середні обсяги споживання регіонів даної групи знаходяться на другій позиції з-поміж кластерів. Витрати вугілля та природного газу порівняно низькі.

**Кластер 2** містить три регіони: Вінницьку, Запорізьку та Черкаську області. Споживання вугілля та електроенергії становить 1,659 та 0,354 млн. т.у.п. відповідно, що є третім результатом. Використання нафти та нафтопродуктів знаходиться на низькому рівні.

**Кластер 3.** Його представники: Волинська, Сумська та Чернівецька області. Його особливістю є самий низький обсяг витрат природного газу (0,941 млн. т.у.п.). Також середнє споживання вугілля регіонами даного кластеру демонструє шостий результат.

**Кластер 4** представляє лише з одна Полтавська область. Витрати вугілля даного кластеру є найнижчими поряд із найбільшим споживанням нафти та нафтопродуктів (5,811 млн. т.у.п.).

**Кластер 5.** Є найбільшим за розмірами і містить такі сім областей: Житомирська, Закарпатська, Кіровоградська, Миколаївська, Рівненська, Тернопільська, Херсонська. Середнє споживання електроенергії регіонами даного кластеру найнижче. Показники витрат вугілля, природного газу і нафти є досить низькими.

**Кластер 6** представлений Київською та Луганською областями. Особливість цього кластеру полягає у найвищому серед кластерів середніх обсягів споживання вугілля, електроенергії і природного газу. Показник використання нафти та нафтопродуктів знаходиться на другому місці.

**Кластер 7** представлений трьома регіонами: Івано-Франківська, Львівська та Харківська області. Він є другим за обсягами споживання вугілля та на шостому місці через використання електроенергії.

Щоб бути впевненим, як кластеризацію проведемо дисперсійний аналіз, результати якого подамо у вигляді таблиці (табл. 4).

Оскільки всі значення  $p < 0,05$ , то з ймовірністю 95% ми можемо стверджувати про значну різницю між кластерами, а отже кластеризація є якісною.

### Висновки

Кластерний аналіз застосовується для розбиття заданої вибірки об'єктів на підмножини, які називаються кластерами, так, щоб кожний кластер складався з подібних об'єктів, а об'єкти різних кластерів істотно відрізнялися.

Метод к-середніх базується на мінімізації суми квадратів відстаней між кожним елементом початкових даних та центром його кластера. На момент старту алгоритму має бути відомим число кластерів, вибір якого може спиратися на результати попередніх досліджень, теоретичних міркувань або інтуїції.

Кластеризація методом к-середніх із використанням спочатку шостого, а потім відповідно сьомого кластерів, дала різний кількісний, а отже якісний склад кластерів, отриманих за допомогою карт Кохонена. Дисперсійний аналіз кластерів вказав на істотну різницю між ними.

### Список використаних джерел

1. Дюран Б.Д. Кластерный анализ [Текст] / Б.Д. Дюран. Пер. с англ. Е.З. Демиденко. Под ред. А.Я. Боярского. Предисловие А.Я. Боярского. М., «Статистика», 1977. – 128 с.
2. Ежов А.А. Нейрокомпьютинг и его применение в экономике и бизнесе [Текст] / А.А. Ежов, С.А. Шумский. – М.: МИФИ.
3. Кравець Т.В., Кузнецов Г.М. Рейтингове оцінювання діяльності підприємств за допомогою модифікованого методу кластеризації // Держава та регіони. – 2010. – №6. – С. 173–180.
4. Погрішук Б.В. Кластеризація регіонів за показниками інноваційно-інвестиційної діяльності зерно продуктового підкомплексу України [Текст] / Б.В. Погрішук. – 2010. – С. 34–45.
5. Сайт. [Електрон. ресурс] – Режим доступу: <http://www.base-group.ru>
6. Сайт. [Електрон. ресурс]: <http://itnews.com.ua>
7. Сайт [Електрон. ресурс] – Режим доступу: <http://wiki.auditory.ru>