УДК 551.58

Э.Н. Серга, *к.геогр.н*, **Е.П. Школьный ,** *д.т.н.*, **П.П. Попович**
Одесский государственный экологический университет

КОМПЛЕКСНЫЙ ПОДХОД К РЕШЕНИЮ ВОПРОСА О КЛАСТЕРИЗАЦИИ ДАННЫХ

Исследовано применение комплексного подхода к кластеризации данных с помощью факторного и кластерного анализов. Выделены и проанализированы факторы и факторные нагрузки по четырем климатическим характеристикам по территории Украины для февраля. Построены обобщённые кластеры для указанных климатических характеристик.
Ключевые слова: *общий фактор, кластер, дисперсия.*

Вступление. Исследования региональных климатов Украины проводятся в течение длительного времени. Данные о распределении климатических элементов и климатических особенностей физико-географических зон Украины впервые были обобщены И.Е. Бучинским [1]. В которой были использованы результаты наблюдений, проводившихся до Второй мировой войны. Более обширные и глубокие исследования климатических особенностей регионов страны содержатся в более поздней монографии по климату Украины [2]. Вопросы закономерностей формирования микроклимата, методы микроклиматических исследований и определение основных микроклиматических показателей, особенности основных типов современных ландшафтов и их элементов детально рассматриваются в работе [3].

Большое количество работ зарубежных и отечественных авторов посвящено изучению изменений температуры воздуха и осадков, выявлению пространственной и временной неоднородности, цикличности в их колебаниях [4, 5], а также исследованию статистической структуры изменения климатических элементов и их изменения во времени и пространстве. Однако при этом уделяется незначительное внимание атмосферной циркуляции, её роли в формировании регионального климата, в том числе температуры и осадков. Независимо от того, что наиболее вероятные характеристики могут относится к аномальным, а значит и циркуляция воздуха, формирующая их, тоже является аномальной, такие величины являются основными параметрами регионального климата и полезными при долгосрочном прогнозировании погоды.

Характерные особенности глобального потепления, реакция на него климата Украины, анализ трансформации климатических полей среднемесячных приземных температур и интенсивности осадков на территории Украины за последние 100 лет, а также динамики уровней Черного и Азовского морей и основные принципы построения сценариев изменения климатических условий Украины в ближайшем будущем при сохранении современных темпов глобального потепления, влияния его на повторяемость природных катастрофических явлений – аспекты, которые требуют дополнительного изучения с помощью существующих методов, в том числе и физико-статистического анализа.

Материалы и методы исследования. Несмотря на сравнительно небольшую территорию, на которой располагается Украина, климатические характеристики в

разных её регионах существенно различаются. Причиной этого являются различия атмосферных процессов, наблюдающихся в разных регионах Украины, с одной стороны, и большое разнообразие подстилающей поверхности, с другой. Иногда, при решении некоторых видов задач (например, разработка и проверка оправдываемости методов прогнозов метеовеличин), необходимо, прежде всего, выделить районы на территории Украины, которые обладают сходством рассматриваемых характеристик климата.

Известно, что существующий подход к определению границ региональных климатов некоторой территории, связан непосредственно с методами кластерного и корреляционного анализа, а также с использованием параметрических критериев Фишера-Снедекора и Стьюдента (определение однородности) [6]. В этом направлении необходимо учесть потребность определения однородно-климатических регионов по нескольким характеристикам, например температуре, осадкам и влажности. К сожалению, методы кластерного анализа, в подавляющем большинстве случаев, позволяют производить регионализацию только по одной характеристике, т.е. работают с двумерным массивом.

Мы предлагаем использовать факторный анализ как предварительный этап для подготовки исходной выборки. Именно он позволяет сжать информацию; определить веса, с помощью которых можно в дальнейшем проанализировать нагрузки каждой характеристики на полученный общий фактор; а также - временной ход общих факторов.

Согласно методике физические переменные x_i в факторном анализе представляются в виде:

$$x_i = \sum_{j=1}^{k} P_{ij} f_j + \upsilon_i, \ \left(i = \overline{1,n}\right), \tag{1}$$

где

f_j – общие факторы,

P_{ij} – веса общих факторов в переменных (факторные нагрузки),

υ_i – остатки.

При этом $k << n$. Общие факторы независимы, с единичной дисперсией и нулевым математическим ожиданием, остатки также независимы и имеют дисперсии d_i.

Указанные особенности указанных элементов факторного анализа приводит к тому, что матрицу ковариаций физических переменных можно записать в виде:

$$K = PP' + Д, \tag{2}$$

где

P – матрица весов мерности $n \times k$,

$Д$ – n - мерная диагональная матрица дисперсий остатков,

$(')$ – означает операцию транспонирования.

Перечисленные особенности указанных факторных характеристик дают возможность с помощью матричного равенства

$$F = \left(E + N\right)^{-1} P^{'} Д^{-1} X\,, \tag{3}$$

где

E – единичная матрица,

X – n - мерный вектор исходных физических величин,

найти значения $k \times n$ мерной матрицы общих факторов F.

Квадратной матрица N мерности k определяется с помощью равенства

$$N^2 = P^{'} Д^{-1} \left(\hat{K} - Д\right) Д^{-1} P\,, \tag{4}$$

где

$\hat{K}$ - выборочная матрица ковариаций.

Как известно, объективная классификация векторов состояния достигается путем применения методов кластерного анализа. Отличительной особенностью кластерного анализа является то, что в нем не существует однозначного критерия, подобного ошибке классификации или среднего риска принятия решения, как, например, в дискриминантном анализе. В литературе описан целый ряд методов кластеризации [7,8]. Большинство из них можно отнести к трем разновидностям. Первая из них объединяет методы, основанные на отыскании моды распределения.

Идея этих методов состоит в том, что кластеры соответствуют максимумам плотности распределения источника, порождающего данные. Применение этих методов сопряжено с большими трудностями, если в качестве данных, подлежащих кластеризации, выступают многомерные векторы.

Другой разновидностью кластерного анализа являются методы, в которых в качестве критерия кластеризации используется отношение внутрикластерной дисперсии к межкластерной дисперсии. Этот критерий может дополняться другими критериями.

К третьей группе относятся иерархические алгоритмы кластеризации. При их использовании полагают, что некоторым образом определен способ измерения расстояния между кластерами.

Для нашего исследования мы использовали метод кластерного анализа под названием «Универсальный адаптивный итерационный метод кластерного анализа (УАИМКА)» [9], который основан как на метрических (эвклидовое расстояние)

$$D_{ij} = \sqrt{\sum_{s=1}^{n} \left(x_{js} - x_{is} \right)^2} \tag{5}$$

так и на неметрических мерах сходства и критериях кластеризации и не требующий на входе гипотетических данных.

В этом методе в качестве исходной информации выступает матрица $X = (x_{ij})_{m \times n}$, содержащая m векторов-строк мерности n, характеризующая статистические ряды объёмом n в m пунктах, которые и должны быть кластеризованы. В качестве априорной информации, в отличие от других методов, задается только минимальное количество векторов τ (по умолчанию 2), которые могут составить кластер.

Результаты исследования и их анализ. Исследование проводилось для полей среднего количества общей облачности, месячного количества осадков, среднемесячной температуры воздуха и среднемесячной массовой доли водяного пара. Поля задавались значениями указанных климатических характеристик, полученных для 33 метеорологических станций территории Украины. Всего было взято по 32 поля указанных характеристик климата за период с 1960 по 1992 год для января и февраля. Иначе говоря, исследованию подвергались 33 тридцатидвухмерных вектора для каждой характеристики климата Украины февраля месяца.

Согласно предлагаемой методике были рассчитаны факторные веса и дисперсии остатков. Как показывают расчеты, более 93% дисперсии температуры воздуха (T) и массовой доли водяного пара (q) с большими нагрузками одного знака обуславливает первый фактор, что позволяет судить, в первую очередь, о совпадении направленности хода процессов этих характеристик с временным ходом вышеупомянутого фактора. Второй же фактор (где он присутствует) обуславливает в основном дисперсию общего количества облачности (N) и месячного количества осадков (R).

Веса полученных общих факторов ($F1$ и $F2$)в переменных содержатся в таблице 1.

Таблица 1 – Веса общих факторов в физических переменных. Февраль

Физическая переменная	Фактор	Номер станции								
		1	2	3	4	5	6	7	8	9
N	*F1*	0,83	0,84	0,71	0,00	0,85	0,87	0,86	0,05	0,89
	F2	-	-	-	0,89	-	-	-	0,91	-
R	*F1*	0,63	0,41	0,35	-0,06	0,61	0,57	0,57	-0,13	0,51
	F2	-	-	-	0,88	-	-	-	0,90	-
T	*F1*	0,87	0,94	0,89	0,98	0,89	0,92	0,92	0,99	0,91
	F2	-	-	-	-0,1	-	-	-	-0,08	-
q	*F1*	0,93	0,94	0,92	0,99	0,94	0,95	0,95	0,99	0,94
	F2	-	-	-	0,04	-	-	-	0,01	-

Продолжение таблицы 1

Физическая переменная	Фактор	Номер станции								
		10	11	12	13	14	15	16	17	18
N	*F1*	0,45	0,79	0,51	0,68	0,19	0,88	0,08	0,77	0,62
	F2	0,77	-	0,68	0,62	0,83	-	0,85	0,33	0,49
R	*F1*	-0,11	0,52	-0,10	0,80	0,02	0,37	-0,21	0,07	-0,07
	F2	0,93	-	0,93	0,06	0,88	-	0,80	0,98	0,94
T	*F1*	0,99	0,91	0,98	-0,10	0,99	0,93	0,99	0,96	0,97
	F2	0,09	-	0,05	0,95	0,12	-	-0,08	-0,05	-0,09
q	*F1*	0,99	0,94	0,98	-0,69	0,99	0,96	0,99	0,97	0,96
	F2	0,08	-	0,08	0,10	0,11	-	-0,05	0,06	-0,01

Продолжение таблицы 1

Физическая переменная	Фактор	Номер станции								
		19	20	21	22	23	24	25	26	27
N	*F1*	0,45	0,81	0,21	0,30	0,86	0,88	0,21	0,86	0,17
	F2	0,73	-	0,82	0,85	-	-	0,83	-	0,83
R	*F1*	-0,29	0,58	-0,01	-0,17	0,62	0,64	-0,12	0,55	-0,18
	F2	0,84	-	0,87	0,90	-	-	0,87	-	0,83
T	*F1*	0,97	0,94	0,99	0,98	0,92	0,92	0,99	0,92	0,99
	F2	-0,04	-	0,09	0,09	-	-	0,03	-	-0,01
q	*F1*	0,98	0,94	0,98	0,98	0,95	0,94	0,99	0,95	0,99
	F2	0,02	-	0,12	0,08	-	-	0,06	-	0,00

Продолжение таблицы 1

Физическая переменная	Фактор	Номер станции					
		28	29	30	31	32	33
N	*F1*	0,39	0,73	0,43	0,66	0,80	0,70
	F2	0,73	0,25	0,82	0,58	0,17	0,49
R	*F1*	0,03	0,03	-0,29	-0,13	0,02	0,05
	F2	0,91	0,98	0,90	0,94	0,99	0,97
T	*F1*	0,98	0,95	0,94	0,97	0,96	0,97
	F2	0,12	-0,09	0,07	-0,08	-0,09	0,02
q	*F1*	0,96	0,96	0,96	0,97	0,96	0,98
	F2	0,23	-0,03	-0,05	-0,05	-0,02	0,08

Анализ таблицы весов предоставляет нам возможность провести дальнейшее исследование с выбором преобладающего влияния характеристик, значимость которых существенней для климатического районирования территории Украины.

На первом этапе поставленной задачи кластеризация для последующего анализа была проведена для каждой из рассмотренных выше характеристик в отдельности.

Как следует из анализа, в феврале в полях общего количества облачности выделяются 3 кластера, один из них располагается в западной части, другой в восточной части территории Украины (Харьковская и Донецкая области), третий захватывает центральную части территории. Границы между кластерами имеют меридиональную ориентацию

Границы кластеров в полях среднемесячных значений температуры и влажности практически совпадают. В том и другом случаях в один из кластеров вошла Правобережная Украина, исключая ее степную часть, в другой кластер – Левобережная Украина (Полтавская, Харьковская и Донецкая области). Третий кластер занимает юг Украины (Одесская, Николаевская, Херсонская области и Крымская АР). В поле

температуры выделяется наиболее холодная северо-восточная часть Украины и наиболее теплая – южные приморские области. Следует заметить, что границы температурных и влажностных кластеров имеют меридионально-зональную ориентацию.

Что касается кластеров в поле месячных количеств осадков, то их выделялось также три. Первой из них охватывает западные области Украины, исключая Украинское Полесье, второй – южные области страны (большую часть Одесской, Николаевскую, Харьковскую области и Крымскую АР). Наибольший по площади кластер охватывает остальную территорию страны.

Следует подчеркнуть, что образовавшиеся кластеры полей общей облачности, месячного количества осадков, среднемесячной температуры и массовой доли водяного пара в феврале не являются случайными. Их положение находит очевидное физическое обоснование, а именно их структура обусловлена влиянием преобладающих над Украиной крупномасштабных атмосферных процессов, и влиянием особенностей подстилающей поверхности. Этот факт является подтверждением объективности районирования.

Как было показано выше, кластеры, относящиеся к полям различных климатических характеристик, имеют различную структуру. Однако перед нами стояла цель получить обобщённые кластера для всех четырёх рассматриваемых характеристик.

Для этого мы провели отдельную кластеризацию используя данные факторного анализа, по первому фактору и по второму, причём в последнем случае в исходную выборку для пунктов имеющих один фактор включался именно этот фактор. Таких кластеров в обоих случаях оказалось три. Результаты данного исследования представлены соответственно на рис. 1,2.

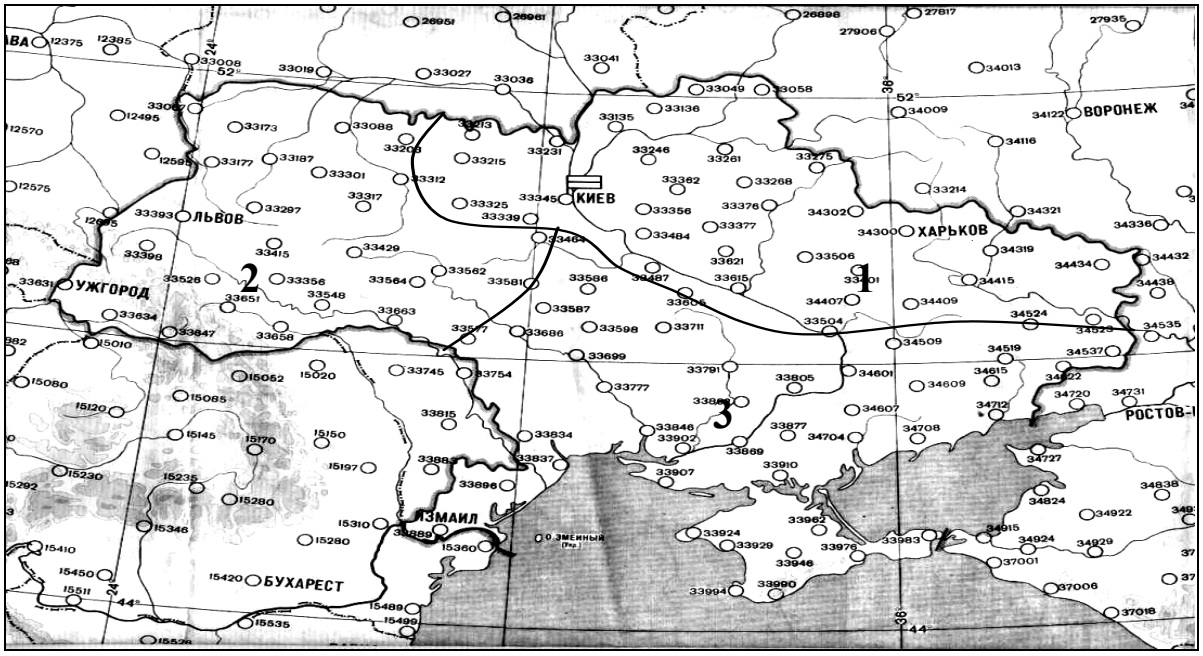

Рис. 1. Обобщённые кластеры территории Украины по 1-му фактору

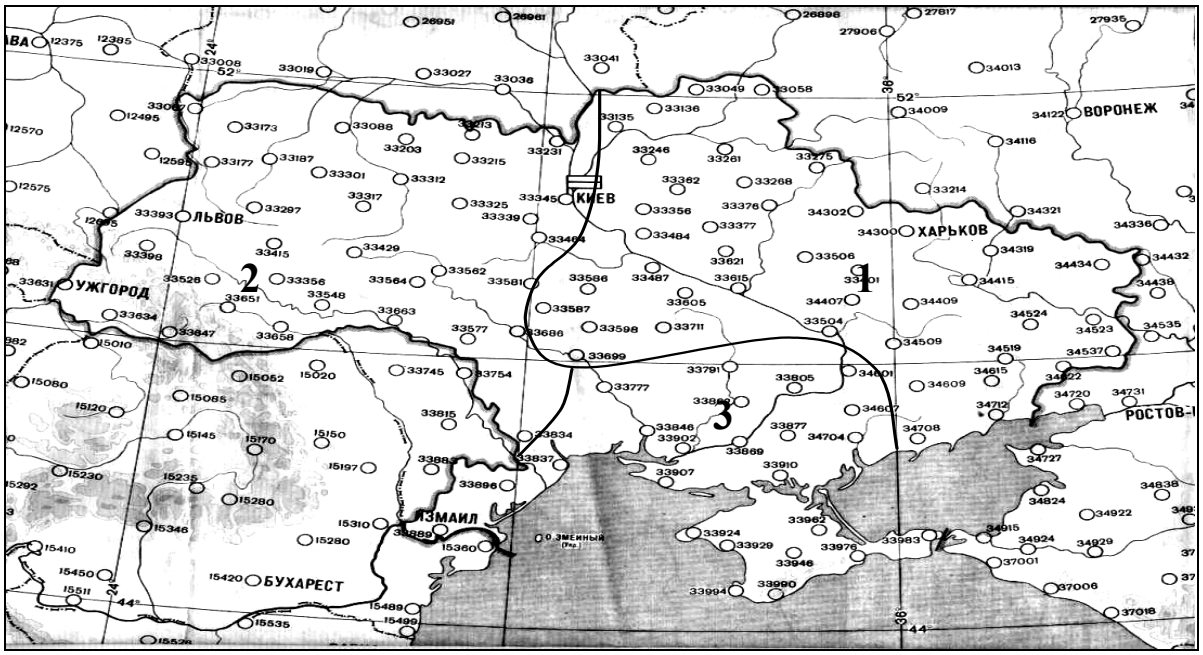

Рис.2. Обобщённые кластеры территории Украины по 2-му и 1-му фактору

Как следует из рис. 1, в первом случае, первый из обощённых кластеров (1) занимает – северо-восток, восток Украины, второй (2) – западную Украину, третий кластер (3) – южную приморскую ее часть и юго-восток; центральная часть территории большей частью поделена между (1) и (3) кластерами. Во втором случае, структура обобщенных кластеров несколько перестраивается. Центральная часть страны поделена уже между (1) и (2) кластерами. Первый кластер (1) в том числе охватывает северо-восточную и восточную части, а также восточное Приазовье. Второй кластер (2) включает приморские области страны (Одесскую, Николаевскую, Херсонскую и Автономную Республику Крым). Третий кластер (3) – объединяет области западной части территории.

С целью анализа результатов кластеризации векторов-факторов, на основе данных исходных характеристик, были рассчитаны внутрикластерные средние значения компонент средних векторов по кластерам, а также внутрикластерные и межкластерные дисперсии. Они приводятся в табл. 2-5 в той же размерности, что и соответствующие им компоненты средних векторов.

Таблица 2 - Средние значения компонент средних векторов при районировании по первому фактору

	Номер кластера		
	1	2	3
	для среднемесячного значения общей облачности		
Средние значения	7,39	7,56	7,30
	для месячного количества осадков		
Средние значения	34,88	37,11	36,35
	для температуры воздуха		
Средние значения	-5,63	-3,71	-3,17
	для массовой доли водяного пара		
Средние значения	3,78	4,21	4,46

Таблица 3 - Внутрикластерные и межкластерные дисперсии при районировании по первому фактору

Для среднемесячного значения общей облачности			
№ кластеров	1	2	3
1	**0,98**	1,01	0,99
2	0,64	**0,61**	0,68
3	0,64	0,70	**0,63**
Для месячного количества осадков			
1	**182,94**	188,06	185,16
2	302,78	**297,65**	298,25
3	276,47	274,85	**274,25**
Для температуры воздуха			
1	**16,71**	20,53	22,99
2	17,75	**13,93**	14,24
3	20,28	14,31	**14,01**
Для массовой доли водяного пара			
1	**1,22**	1,41	1,71
2	1,23	**1,03**	1,10
3	1,57	1,15	**1,08**

Таблица 4 - Средние значения компонент средних векторов при районировании по второму и первому фактору

	Номер кластера		
	1	2	3
	для среднемесячного значения общей облачности		
Средние значения	7,35	7,50	7,32
	для месячного количества осадков		
Средние значения	31,13	37,11	35,95
	для температуры воздуха		
Средние значения	-5,36	-3,80	-2,62
	для массовой доли водяного пара		
Средние значения	3,88	4,20	4,58

Таблица 5 - Внутрикластерные и межкластерные дисперсии при районировании по второму и первому фактору

Для среднемесячного значения общей облачности			
№ кластеров	1	2	3
1	**0,98**	1,00	0,98
2	0,62	**0,60**	0,63
3	0,58	0,65	**0,58**

Продолжение таблицы 5

Для среднемесячного значения общей облачности			
№ кластеров	1	2	3
Для месячного количества осадков			
1	**182,12**	193,14	189,80
2	287,07	**283,05**	284,44
3	339,06	339,77	**338,38**
Для температуры воздуха			
1	**16,11**	18,62	23,86
2	16,37	**13,86**	15,29
3	21,86	15,54	**14,11**
Для массовой доли водяного пара			
1	**1,09**	1,20	1,61
2	1,12	**1,02**	1,17
3	1,85	1,50	**1,34**

Как следует из табл. 2–5, средние значения общего балла облачности практически одинаковы во всех кластерах, однако различными являются его внутрикластерные дисперсии. Наибольшая дисперсия имеет место в кластере 1, к которому относится северо-восточная и восточная часть Украины. Это объясняется тем, что на эту часть Украины зимой периодически распространяется отрог сибирского максимума, в области которого наблюдается небольшое количество облачности. Он отступает к востоку при усилении циклонической деятельности. Над остальной частью территории поле давления оказывается более стабильным.

Наблюдаются отличия в кластерах по средним значениям месячных количеств осадков и по их дисперсиям. Как следует из табл. 2-5, наибольшими средними значениями и внутрикластерными дисперсиями характеризуются кластеры 2 и 3 (следует в большей степени ориентироваться на случай совместной кластеризации по второму и первому фактору, т.к. наибольшие веса по осадкам приходятся именно на второй фактор). Именно через эти регионы Украины проходят траектории средиземноморских и черноморских циклонов. Как известно, в их области наблюдаются обильные осадки и метели. По сравнению с другими кластерами, кластеру 1 (западные области страны) присуще наименьшие, а кластеру 3 (южные области) – наибольшее значение дисперсии месячного количества осадков.

Значительные различия между кластерами имеют место и по температуре. В этом отношении наиболее “холодным” являются кластер 1 (северо-восточная, восточная части Украины), а наиболее “теплым” – кластер 3 (юг страны). Наиболее “холодному” второму кластеру соответствует и наибольшая дисперсия температуры. Как и следовало ожидать, наиболее “холодный” кластер является и наименее “влажным”, и наоборот.

Анализ внутрикластерных средних значений откликов и их дисперсий показывает, что выделенные кластеры в феврале, хорошо отражают особенности крупномасштабных атмосферных процессов, развивающихся на территории Украины, а также характер различий условий подстилающей поверхности.

С целью анализа временного хода и выявления периодичности в процессах рассматриваемых характеристик климата, нами были построены графики временного хода первого фактора по кластерам.

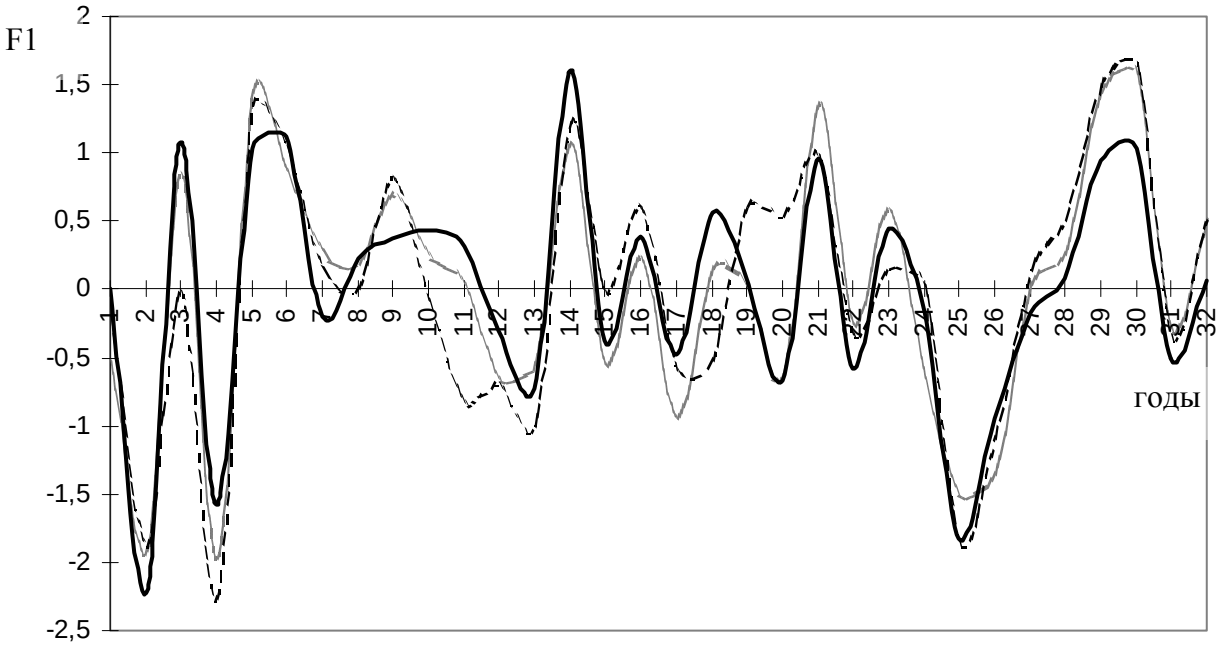

Рис. 3 Временной ход 1-го фактора 1-3 кластеров (…. – 1 кластер, _ _ _ - 2 кластер, _____ - 3 кластер) в период с 1960 по 1992 г.г.

Согласно рис. 3 временной ход первого фактора по территории Украины для февраля в основном имеет квазидвухлетнюю периодичность на которую накладываются колебания большего масштаба. Причём при совместном анализе факторных весов и временных графиков можно перенести это суждение непосредственно на рассматриваемые климатические характеристики.

Выводы. Данная методика совмещённого климатического районирования с целью получения обобщённых кластеров по нескольким характеристикам вполне применима для исследования. Применение в исследовании общих факторов даёт дополнительные возможности для анализа используемых характеристик и полученных результатов.

Список литературы

1 *Бучинский И.Е.* Климат Украины в прошлом, настоящем, будущем. – К: Госсельхозиздат, 1963. - 308 с.
2 *Климат Украины* / Под редакцией Прихотько Г.Ф., Ткаченко А.Р., Бабиченко В.Н.. – Л.: Гидрометеоиздат, 1967. – 913 с.
3 *Щербань М.И.* Микроклиматология. – К.: Вища школа, 1985. –221 с.
4. *Геодонов А.Д.* Изменения температуры воздуха на северном полушарии за 90 лет.- Л.: Гидрометеоиздат, 1973. – 146 с.
5. *Рубинштейн Е.С., Полозова Л.Г.* Современное изменение климата. – Л.: Гидрометеоиздат, 1966. – 268 с.
6. *Школьний Є.П., Лоєва І.Д., Гончарова Л.Д.* Обробка та аналіз гідрометеорологічної інформації. Підручник. – К.: Вища школа, 1999. – С.455-513.
7. *Горелик А.Л., Скрипник В.А.* Методы распознания. - М.: Высшая школа, 1984.- 273 с.

8. *Райзин Дж. Вэн.* Классификация и кластер. - М.: Мир, 1980. – 244 с.
9. *Серга Э.Н.* Универсальный адаптивный итерационный метод кластерного анализа // Міжвідомчий науковий зб. України: Метеорологія, кліматологія та гідрологія. – 2003. – Вип.47. – С.83-89.

Комплексний підхід к вирішенню питання о кластеризації даних.
Сєрга Е.М. , Школьний Є.П., Попович П.П.
Досліджено застосування комплексного підходу до кластеризації даних за допомогою факторного й кластерного аналізів. Виділені та проаналізовані фактори й факторні навантаження по чотирьох кліматичних характеристиках по території України для лютого. Побудовано узагальнені кластери для зазначених кліматичних характеристик.
***Ключові слова**: загальний фактор, кластер, дисперсія.*

The complex approach to the decision of the question about claster analyses of data.
Serga E.N., Shkolnyj E.P., Popovich P.P.
Application of the complex approach to the claster analyses of data by means of factorial and claster analyses is investigated. Factors and factorial loadings under four climatic characteristics on territory of Ukraine for February are allocated and analysed. Are constructed generalized clasters for the specified climatic characteristics.
***Keywords:** the general factor, claster, dispersion.*