2Кластерный анализ

Целью методов кластерного анализа является разбиение выборок многомерных данных на группы объектов близких в смысле некоторой заданной меры сходства. Такие компактные группы называются кластерами, классами или таксонами.

Методы кластерного анализа называют также методами обучения без учителя, автоматической группировки или таксономии.

Методы кластерного анализа могут использоваться в качества вспомогательных инструментов при решении задач прогнозирования или распознавания. Однако нередко кластеризация может иметь самостоятельное значение.

2 Кластерный анализ

Большинство известных алгоритмов кластеризации предполагает задание расстояния $\rho(\mathbf{x},\mathbf{y})$ между произвольными векторами-описаниями

объектов. В качестве расстояния могут выступать, например, евклидова метрика:

$$\rho(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Используются и другие функции расстояния.

2Кластерный анализ

Одним из наиболее известных методов кластеризации является **алгоритм k внутригрупповых средних**. Предположим, что у нас задана выборка векторов- описаний $\tilde{S} = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$. Алгоритм находит такие кластеры, для объектов которых центр «своего кластера» будет ближе центра любого «чужого кластера».

Метод предполагает, что число кластеров изначально задано.

2Кластерный анализ

Поиск оптимальной кластеризации методом *к внутригрупповых* средних

Предположим, что предполагаемое число кластеров равно ***r.***

Зададим произвольным образом исходное разбиение выборки $\tilde{S} = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ на группы $G_1^0, \ldots, G_r^0$

Вычисляем геометрические центры исходных группПусть группа G_i^0 состоит из объектов $\{\mathbf{x}_1^0, \ldots, \mathbf{x}_{m(i)}^0\}$. Тогда центр G_i^0 вычисляется по формуле $\overline{\mathbf{x}}_i^0 = \frac{1}{m(i)} \sum_{j=1}^{m(i)} \mathbf{x}_j^0$

Вычисляются расстояния между объектами из $\tilde{S}$ и центрами $\overline{\mathbf{x}}_1^0, \ldots, \overline{\mathbf{x}}_r^0$

2Кластерный анализ

Поиск оптимальной кластеризации методом *k* внутригрупповых средних

Объекты из $\tilde{S}$ затем переносятся в группу с наименее удалённым центром. В результате мы получаем новый набор групп $G_1^1,\ldots,G_r^1$. Повторяем для набора групп $G_1^1,\ldots,G_r^1$ те же самые операции, которые ранее выполнялись для групп $G_1^0,\ldots,G_r^0$

……

Процесс завершается на некотором шаге ***k+1***, когда переносы объектов из $\tilde{S}$ в другие группы не требуются.

То есть каждый объект наименее удалён от центра той же самой группы, которой он и принадлежит. В результате мы получаем набор компактных групп - кластеров

Иерархическая кластеризация

Для того, чтобы осуществить иерархическую кластеризацию необходимо сначала задать расстояние $\mathrm{P}(G',G'')$ между произвольными кластерами G',G''.

Возможные способы задания расстояния:

1) $\mathrm{P}(G',G'') = \min_{\mathbf{x}'\in G',\mathbf{x}''\in G''} \rho(\mathbf{x}',\mathbf{x}'')$ - то есть расстоянием между двумя кластерами является минимальное расстояние между двумя объектами, один из которых принадлежит G', а второй G''.

2) $\mathrm{P}(G',G'') = \max_{\mathbf{x}'\in G',\mathbf{x}''\in G''} \rho(\mathbf{x}',\mathbf{x}'')$ - то есть расстоянием между двумя кластерами является максимальное расстояние между двумя объектами, один из которых принадлежит G', а второй G''.

Иерархическая кластеризация

3) $\mathrm{P}(G', G'') = \rho(\bar{\mathbf{x}}', \bar{\mathbf{x}}'')$ - расстояние между центрами кластеров G', G''

4) $\mathrm{P}(G', G'') = \frac{1}{m'm''} \sum_{i=1}^{m'} \sum_{j=1}^{m''} \rho(\mathbf{x}'_i, \mathbf{x}''_j)$ - среднее расстояние между объектами из двух кластеров

Отметим, что в случае, когда все кластеры состоят только из одного объекта, расстояния между ними всегда равны расстояниям между этими единственными объектами.

Иерархическая кластеризация

На начальном этапе кластерами являются объекты $\tilde{S}$

На каждом последующем шаге происходит объединение двух ближайших кластеров из набора, образованного на предыдущем шаге.

Процесс завершается при достижении одного из следующих условий.

1) Кластеры, образованные на новом шаге теряют компактность. Тогда мы оставляем в силе кластеризацию, полученную на предыдущем шаге.
2) Образуется требуемое число кластеров
3) Процесс завершается, если достигнутая кластеризация удовлетворяет требованиям эксперта исследователя.

ИССЛЕДОВАНИЯ ФОЛЬКЛОРНО-МИФОЛОГИЧЕСКИХ ТРАДИЦИЙ С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ

Целью настоящей работы является разработка и обоснование методов интеллектуального анализа данных, эффективных при исследовании фольклорно-мифологических традиций по наборам представленных в них мотивов. База данных, содержащая информацию о встречаемости мотивов, создана и Ю.Е. Березкиным [Березкин 2007; 2009] и размещена на сайте http://starling.rinet.ru/kozmin/tales/index.php?index=berezkin

В 2007 г. база включала сведения о встречаемости 1355 мифологических мотивов в 337 традициях (на ноябрь 2009 в ней 1483 мотива и 470 традиций). Для этого на протяжении почти двадцати лет были проанализированы более 5500 публикаций на германских, романских, славянских и прибалтийско-финских языках, использованы также некоторые неопубликованные материалы. Под мотивом понимаются повторяющиеся образы, эпизоды или их сочетания максимальной протяженности, встречающиеся в двух и более (практически - во многих) текстах. В базу данных включались только такие мотивы, которые обнаружены не менее, чем в четырех традициях. Под традицией понимается совокупность текстов, записанных у одной этно-языковой группе.

В базе данных для всех традиций в бинарной форме фиксируется наличие или отсутствие каждого мотива в проанализированных источниках.

Следует подчеркнуть, что наличие 0 в некоторой позиции традиции не обязательно достоверно свидетельствует о реальном отсутствии мотива ввиду недостаточной изученности некоторых традиций. Последнее обстоятельство не позволяет использовать в качестве функции близости стандартные метрики Евклида или Хэмминга, которые предполагают суммирование совпадений по всем сюжетам, что приведёт к установлению высокой близости между двумя слабо исследованными традициями. В связи с этим были выдвинуты альтернативные функции расстояния между традициями *T[i]* и *T[j]* .

Функция *Sc*(*T[i], T[j])=1- 0.5*{ k*C*(*t[i],t[j] /N +1}*, где *C*(*t[i],t[j])* представляет собой величину статистики критерия Хи-квадрат, при оценивании достоверности связи между двумя дихотомическими разбиениями. *N* - общее количество мотивов в исследуемой базе,

k=1, если мотивы в среднем чаще встречаются в *T[j]* при условии наличия их в *T[i].*

k=-1, если мотивы в среднем реже появляются в *T[j]* при условии наличия их в *T[i].*

Выявление однородных групп традиций. Для выявления групп традиций с близким характером встречаемости мифологических мотивов использовался широко распространённый метод иерархической группировки.

На начальном этапе каждая традиция считалась отдельным кластером.

На каждом шаге происходит объединение кластеров с миниимальным значением усреднённой (по всем парам объектов из разных кластеров) функции расстояния.

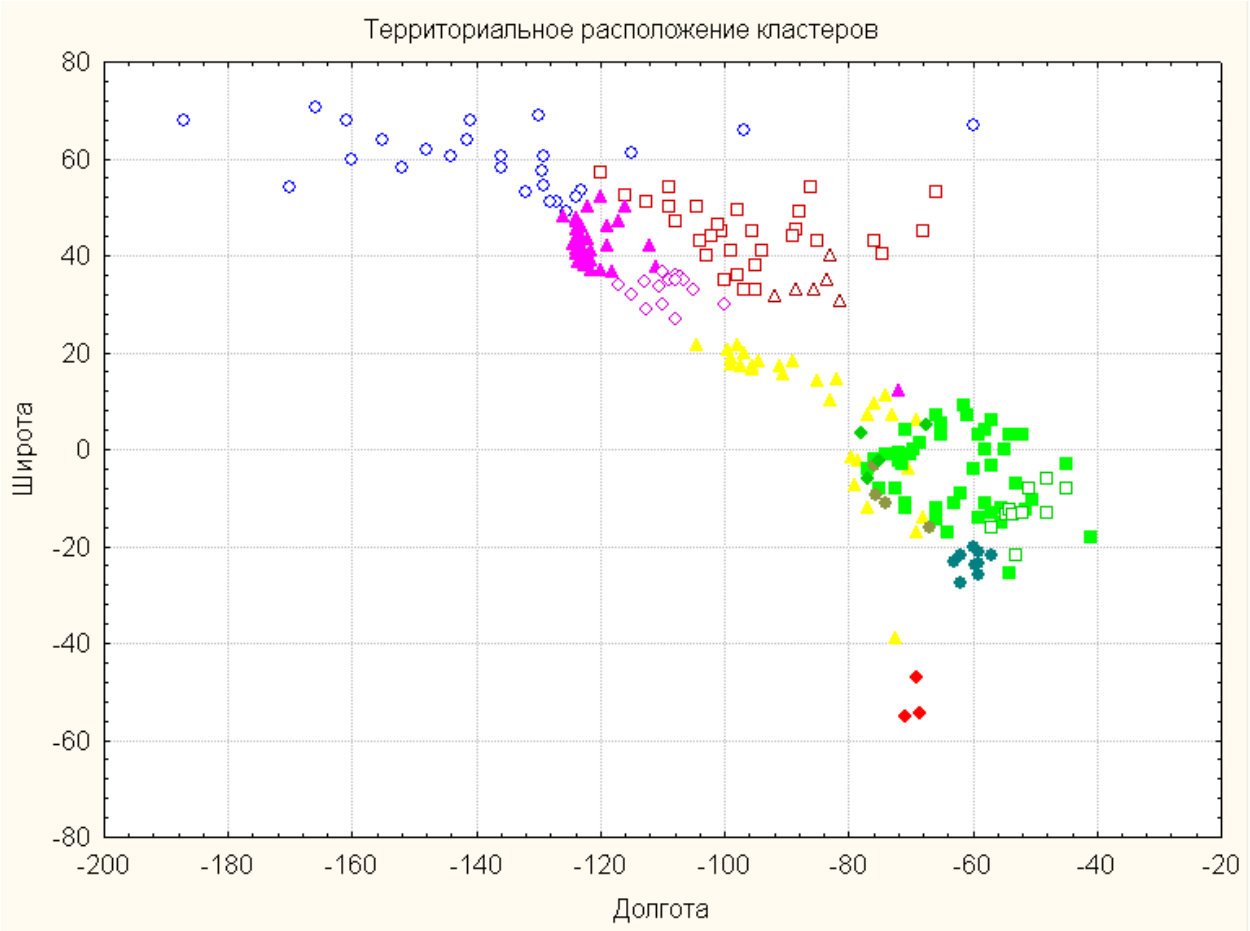
Процесс продолжался до тех пор пока традиции не оказывались объединёнными в заданное исследователем число кластеров.

На первом этапе исследования проводились для индейских традиций Американского континента

Проведённые исследования показали, что традиции в кластерах, полученных согласно сходству мифологических мотивов, оказываются, как правило, также близкими географически.

На приводимых далее рисунках результаты кластеризации показаны в системе географических координат при заданном числе кластеров равным 8 и 11.


Территориальное распределение кластеров
80
60
40
20
0
-20
-40
-60
-80
ШИРОТА
-200
-180
-160
-140
-120
-100
-80
-60
-40
-20
ДОЛГОТА


Территориальное расположение кластеров
80
60
40
20
0
-20
-40
-60
-80
Широта
-200
-180
-160
-140
-120
-100
-80
-60
-40
-20
Долгота

Дополнительным способом оценки сходства между традициями (или группами традиций) *T1* и *T2* является вычисление коэффициента корреляции расстояния до *T1* и *T2* набора других традиций.

Использовался набор всех индейских традиций американского континента.

Таблица 1. *Коэффициенты корреляции между средними расстояниями американских фольклорных традиций до соответствующих пар кластеров.*

	○	□	△	▲	◇	▲	■	○	●	◆	●	◆
○	**1.00**	0.29	0.00	0.33	0.01	-0.42	-0.32	-0.28	-0.17	-0.27	-0.26	0.00
□	0.29	**1.00**	0.48	0.46	0.46	-0.37	-0.58	-0.43	-0.21	-0.46	-0.42	0.01
△	0.00	0.48	**1.00**	0.26	0.47	0.17	-0.30	-0.30	-0.10	-0.13	-0.12	-0.09
▲	0.33	0.46	0.26	**1.00**	0.55	-0.29	-0.37	-0.31	0.08	-0.30	-0.32	0.06
◇	0.02	0.46	0.47	0.55	**1.00**	0.03	-0.41	-0.34	-0.04	-0.27	-0.24	0.10
▲	-0.42	-0.37	0.17	-0.29	0.03	**1.00**	0.38	0.15	0.10	0.45	0.43	0.04
■	-0.32	-0.58	-0.30	-0.37	-0.41	0.38	**1.00**	0.75	0.37	0.64	0.68	0.05
○	-0.28	-0.43	-0.30	-0.31	-0.34	0.15	0.75	**1.00**	0.34	0.38	0.39	-0.01
●	-0.17	-0.21	-0.10	0.08	-0.04	0.10	0.37	0.34	**1.00**	0.17	0.24	0.14
◆	-0.27	-0.46	-0.13	-0.30	-0.27	0.45	0.64	0.38	0.17	**1.00**	0.48	0.05

	-0.26	-0.42	-0.12	-0.32	-0.24	0.43	0.68	0.39	0.24	0.48	1.00	0.08
	0.00	0.01	-0.09	0.06	0.10	0.04	0.05	-0.01	0.14	0.05	0.08	**1.00**

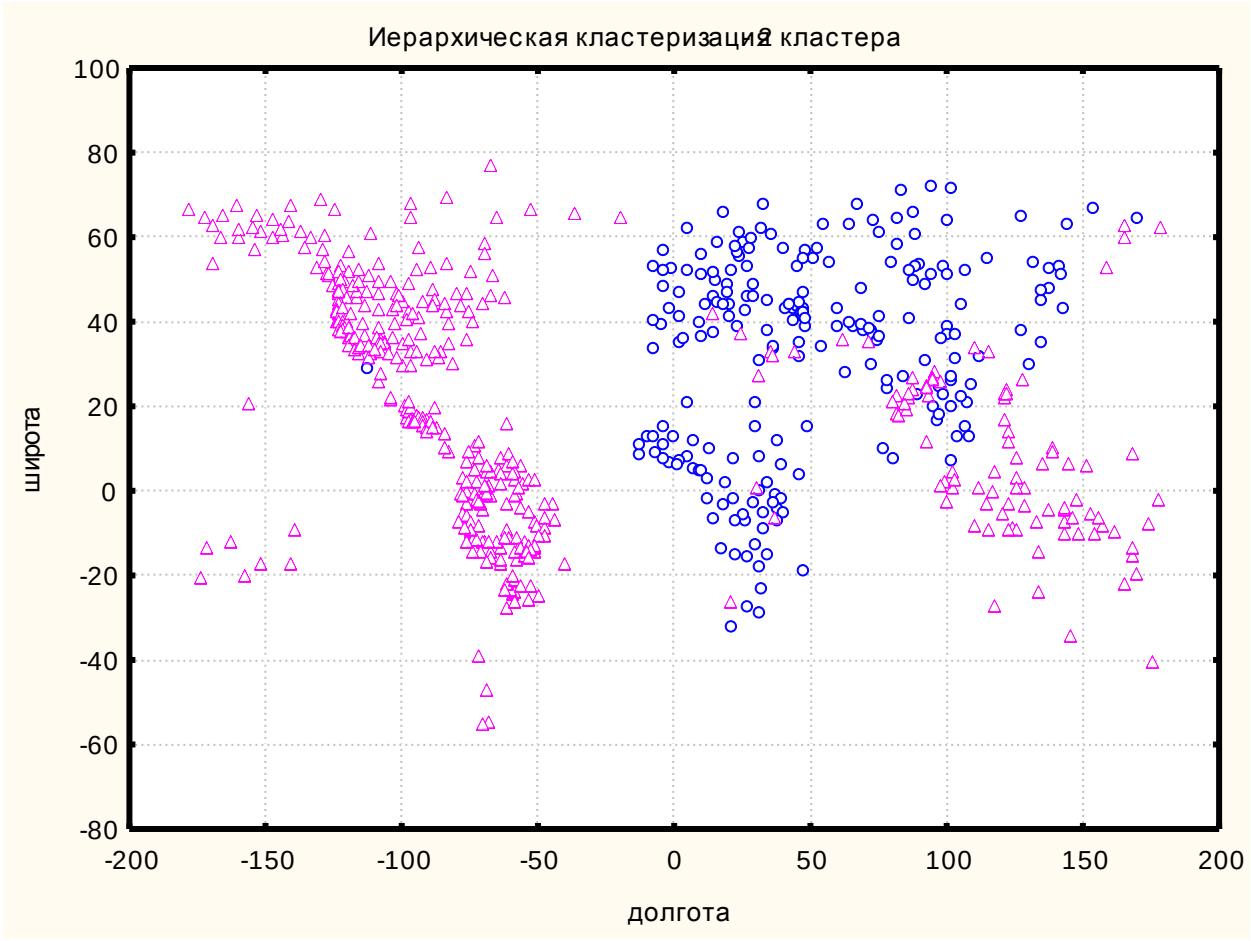
Таблица 2. *Коэффициенты корреляции между средними расстояниями американских Фольклорных традиций для пар (кластер –внеамерикнская традиция).*

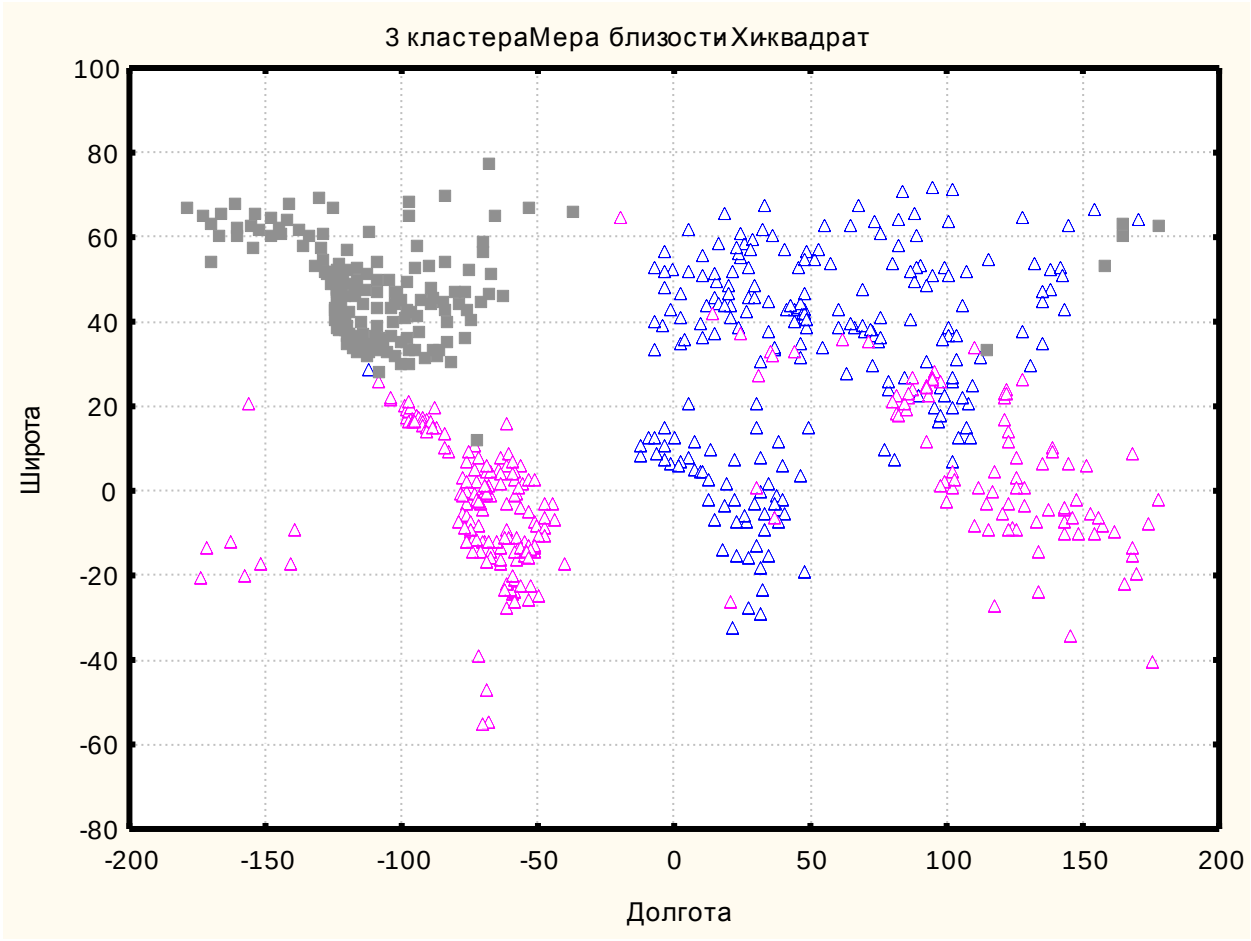
	○	□	△	▲	◇	▲	■	○	●	◆	●	◆
Chinese_	0.13	-0.13	0.22	-0.09	0.16	0.58	0	-0.14	0.01	0.09	0.16	0.04
Garo_Chin_Mizo_Kachari_	0.0025	-0.066	0.33	-0.03	0.15	0.62	0.21	-0.039	0.01	0.26	0.46	0.018
Hadza_Sandawe	-0.42	-0.34	-0.06	-0.16	0.057	0.37	0.47	0.55	0.32	0.46	0.35	0.20
Chukchi	**0.78**	0.38	0.24	0.40	0.31	-0.22	-0.47	-0.49	-0.19	-0.28	-0.32	0.028
Evenk:_Baikal_Amur	0.35	0.54	0.57	0.48	0.61	-0.04	-0.50	-0.46	-0.19	-0.26	-0.27	0.037
Ainu	0.61	0.15	0.11	0.33	0.15	-0.03	-0.14	-0.20	0.16	-0.19	-0.08	0.11
New_Guinea_Papuans	-0.26	-0.49	-0.17	-0.21	-0.24	0.47	0.81	0.60	0.37	0.71	0.52	0.12

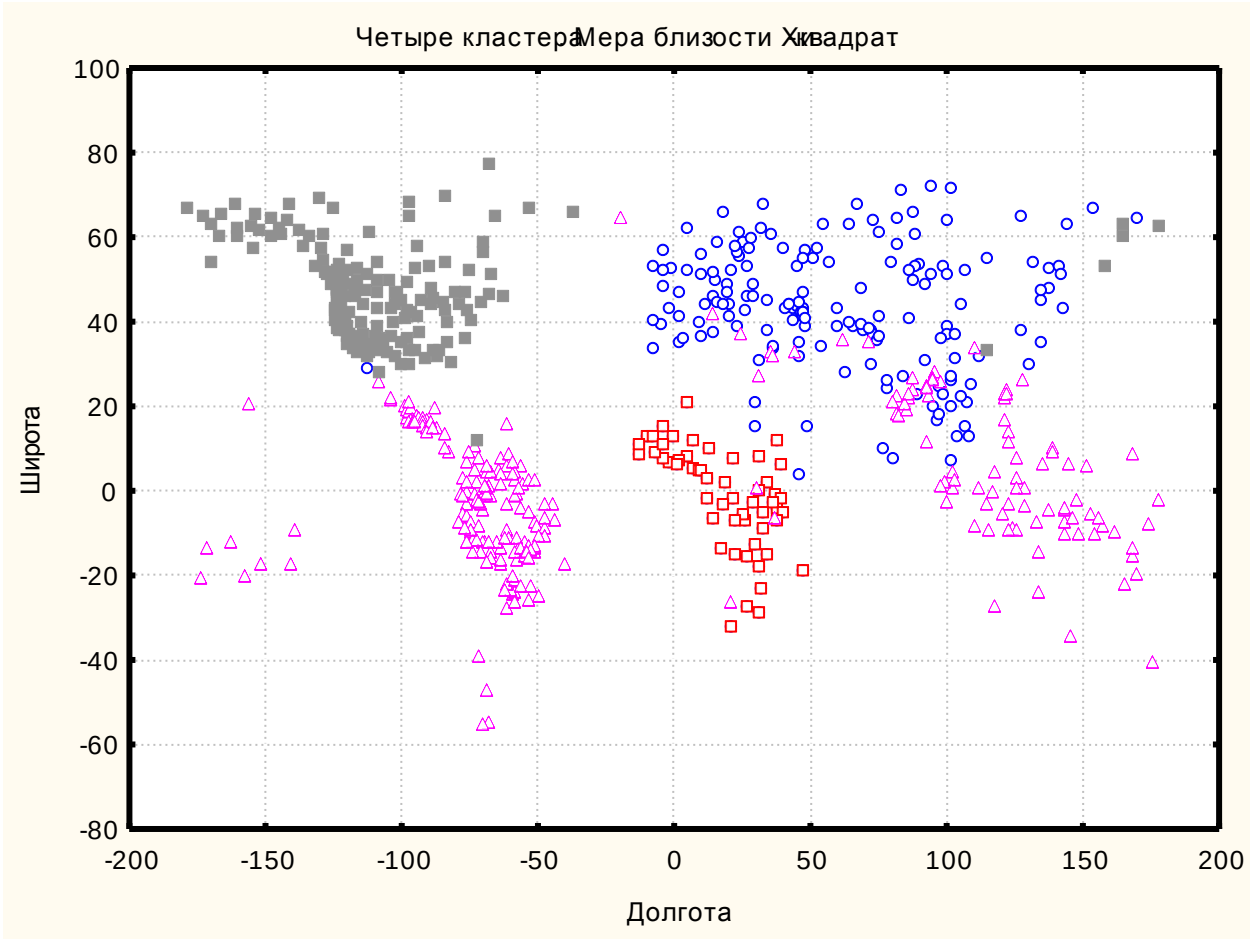
На втором этапе исследования проводились для традиций, распространённых по всему миру

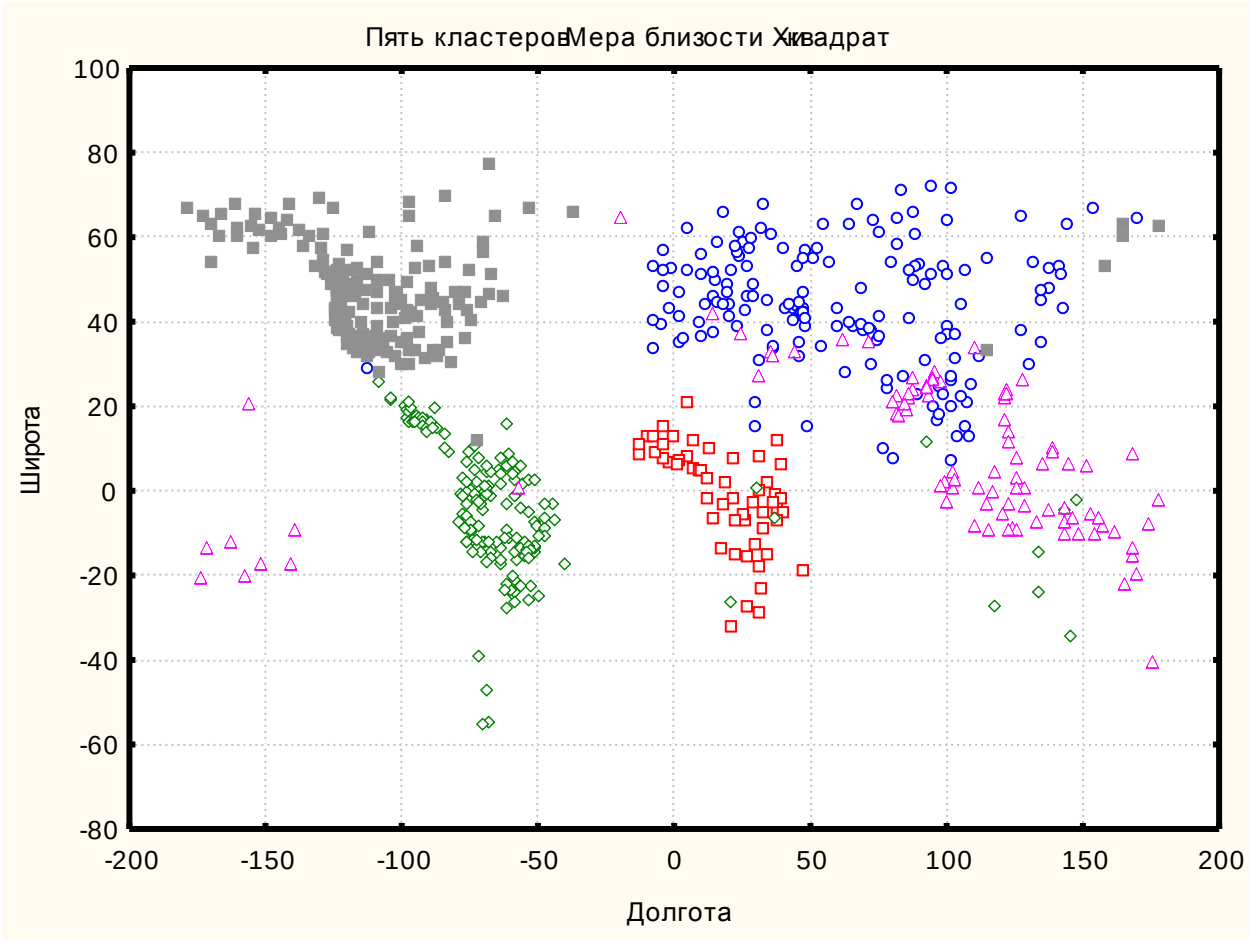
На приводимых далее рисунках результаты кластеризации показаны в системе географических координат при заданном числе кластеров равным от 2 до 8.

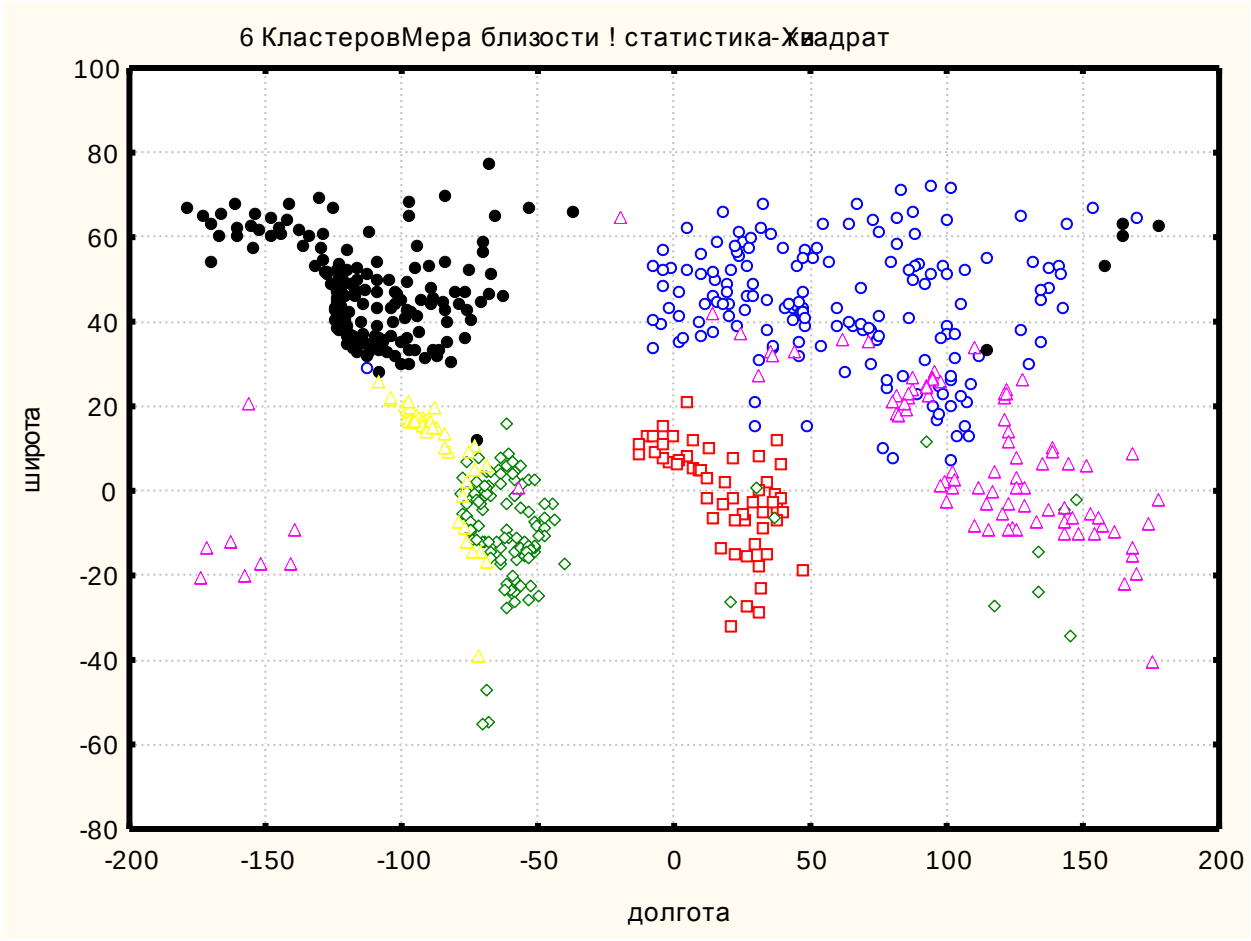
Исследования подтвердили выраженную тенденцию, что традиции в кластерах, полученных согласно сходству мифологических мотивов, оказываются, как правило, также близкими географически.

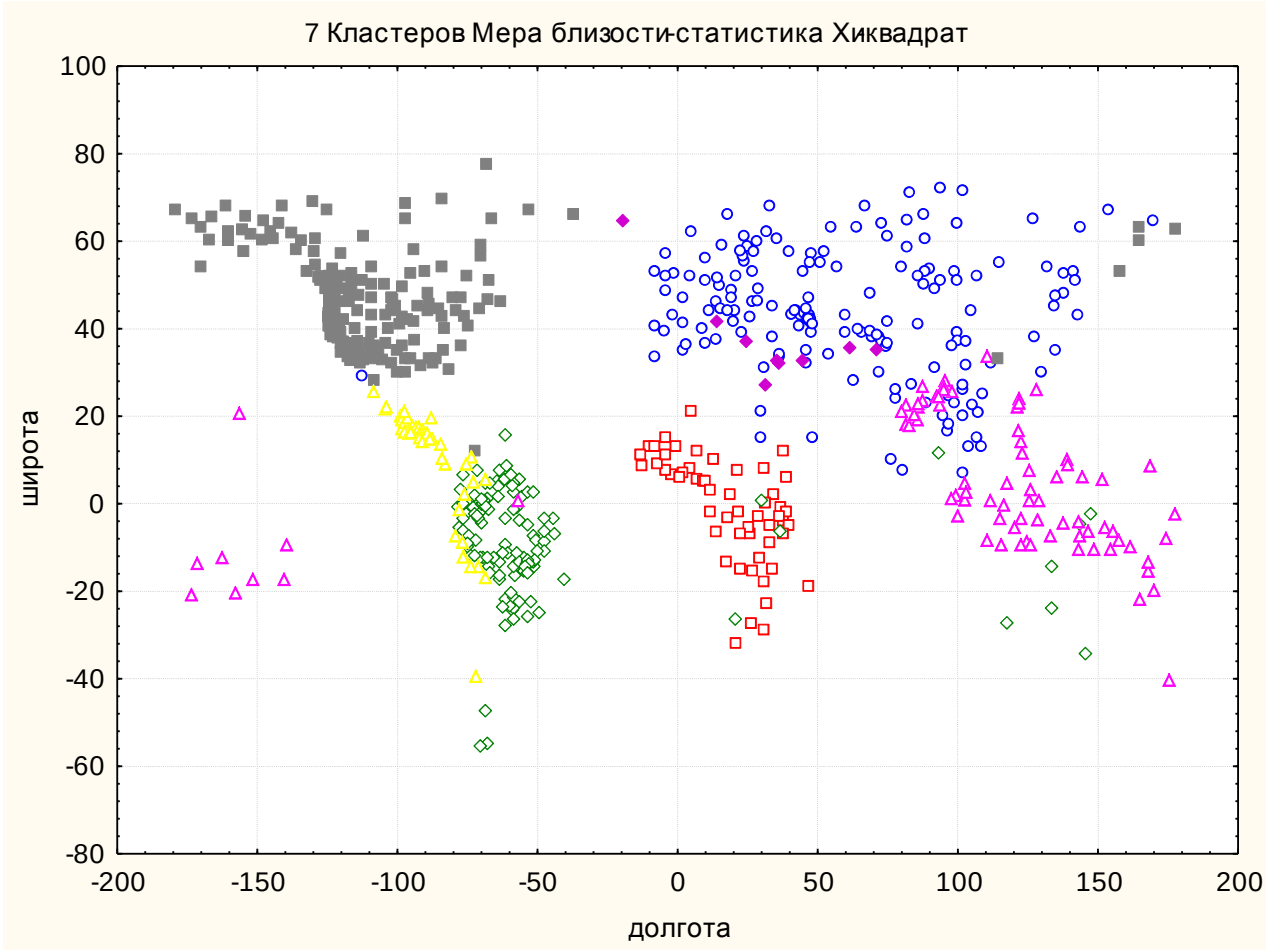

Иерархическая кластеризация 2 кластера
100
80
60
40
20
0
-20
-40
-60
-80
широта
-200
-150
-100
-50
0
50
100
150
200
долгота


3 кластераМера близостиХи-квадрат
100
80
60
40
20
0
-20
-40
-60
-80
Широта
-200
-150
-100
-50
0
50
100
150
200
Долгота


Четыре кластера Мера близости Хи-квадрат
100
80
60
40
20
0
-20
-40
-60
-80
Широта
-200
-150
-100
-50
0
50
100
150
200
Долгота


Пять кластеров Мера близости Хи-квадрат
100
80
60
40
20
0
-20
-40
-60
-80
Широта
-200
-150
-100
-50
0
50
100
150
200
Долгота


6 КластеровМера близости ! статистика-Хи
100
80
60
40
20
0
-20
-40
-60
-80
широта
-200
-150
-100
-50
0
50
100
150
200
долгота



85 0.159000 Ancient_Egypt_____ Afrasian_____ 27.500000 30.700001

99 0.170000 Accad,Assiria,Babylon_____ Afrasian_____ 33.000000 44.000000

101 0.198125 Ugarit,_Phoenicia_____ Afrasian_____ 33.000000 35.000000

140 0.190500 Ancient_Italy_____ Indoeuropean+_____ 42.000000 13.500000

158 0.211875 Ancient_Greece_____ Indoeuropean_____ 37.500000 24.000000

100	0.162375	Old_Testament	Afrasian	32.500000 35.500000
111	0.166125	Zoroastrism,Shah_Name	Indoeuropean	36.000000 61.000000
174	0.124500	Edda	Indoeuropean	65.000000 -20.000000
270	0.092750	Kafirs	Indoeuropean	35.500000