

## ТЕОРЕТИЧЕСКАЯ И МАТЕМАТИЧЕСКАЯ ФИЗИКА

**Использование кластерного анализа для обработки данных GPS-интерферометрии**В. И. Захаров<sup>1,a</sup>, П. А. Будников<sup>2,b</sup><sup>1</sup> *Московский государственный университет имени М. В. Ломоносова, физический факультет, кафедра физики атмосферы. Россия, 119991, Москва, Ленинские горы, д. 1, стр. 2.*<sup>2</sup> *Институт прикладной геофизики имени академика Е. К. Федорова (ФГБУ «ИПГ»). Россия, 129128, Москва, ул. Ростокинская, д. 9.**E-mail: <sup>a</sup> zvi\_555@list.ru, <sup>b</sup> atm5571@phys.msu.ru*

Статья поступила 20.05.2011, подписана в печать 31.10.2011

Исследованы возможности применения методов кластерного анализа для определения параметров неоднородных структур верхней атмосферы — ионосферы, выделенных методом GPS-интерферометрии. Проведено тестирование различных алгоритмов кластеризации на квазиреальных данных. Выявлено значительное уменьшение дисперсии параметров зарегистрированных ионосферных структур по сравнению с обычной статистической обработкой, особенно в случае малой выборки данных и высокого уровня шума. Приведены примеры обработки экспериментальных данных методом кластерного анализа.

*Ключевые слова:* ионосфера, неоднородности среды, GPS-интерферометрия, алгоритмы кластеризации.

УДК: 551.510.5; 550.388.2; 550.3+551.5:629.78. PACS: 94.05.-a, 94.20.Vv, 91.10.Fc, 94.05.Sd.

**Введение**

Состояние земной ионосферы характеризуется высокой степенью изменчивости и наличием неоднородных структур, что связано со многими геофизическими процессами и явлениями. Важнейшими факторами, влияющими на ионосферу, является солнечная активность, изменение магнитного поля Земли и разные процессы, протекающие в системе геосфер [1–4]. Иначе говоря, ионосфера как часть атмосферы является индикатором динамических процессов в системе «планета Земля — атмосфера».

Явления, происходящие в ионосфере, изменяют параметры прохождения радиоволн, что используется в геофизическом мониторинге на базе систем радионавигации и является источником различных погрешностей, например в работе упомянутых систем или при зондировании поверхностных слоев из космоса [3]. Именно указанные явления делают возможным исследование ионосферы прецизионными наземными радиофизическими методами — томографическими [4, 5] и интерферометрическими [6–9]. Различные воздействия на ионосферу гео- и гелиофизических феноменов можно трактовать как активные эксперименты и использовать для решения ряда задач физики ионосферной плазмы, распространения радиоволн, физики ударных волн и пр. В последние годы стало ясно, что надежный прогноз погоды и климатической изменчивости без учета ионосферных процессов и атмосферно-термосферных связей маловероятен.

Так на повестку дня выдвигается вопрос не только обнаружения возмущений на фоне смеси «ионосферных сигналов», но и классификации получаемых сведений о параметрах ионосферного возмущения, таких как амплитуда, период, групповая скорость перемещения, угловые характеристики волнового вектора и т. п. в раз-

личных гео- и гелиофизических условиях с целью увеличения достоверности оценки параметров изучаемых ионосферных структур.

В настоящей статье рассмотрены некоторые аспекты применения кластерного анализа для идентификации и определения параметров структур со сходными признаками, выделенные методом GPS-интерферометрии.

**1. Особенности исследования ионосферы методом GPS-интерферометрии**

Метод основан на использовании комбинации регистрируемых фаз на рабочих частотах  $L1$  и  $L2$ , последующей фильтрации высокочастотного шума (с периодами менее 3–5 мин) и построения оценки производной по времени изучаемого процесса (подробнее см. [5–7]).

Стандартная процедура обработки массива данных в заданном регионе [7–9] заключается в следующем. Из большого числа станций GPS, географически расположенных в одном регионе, выбираются группы из трех станций, удовлетворяющие следующим критериям: расстояния между станциями должны быть менее 100–200 км; на каждой из станций данные не должны иметь длительных пропусков и сбоев. Первое условие дает возможность использовать приближение плоского фронта для волнового возмущения. Выбранные три станции будем называть «измерительной ячейкой региональной подсети» или просто «ячейкой подсети». В пакете программного обеспечения для комплексного регионального анализа спутниковых сигналов системы GPS (Complex Regional Analysis Satellite Signals, CRASS GPS), созданного на кафедре физики атмосферы физического факультета МГУ, реализована именно указанная идеология.

Представленная конфигурация приемников GPS в ячейке подсети представляет собой решетку с мини-

мально необходимым количеством элементов. Первичными данными являются ряды «наклонного» значения полного электронного содержания (интеграла пространственного ионосферного распределения электронов от приемной станции до навигационного искусственного спутника Земли (ИСЗ)), определенные для каждой станции в ячейке подсети, а также соответствующие им ряды значений угла места и азимута направления на ИСЗ.

После выделения волновой структуры, для ее идентификации обычно используются статистические процедуры [6–9], но в настоящей работе мы наряду с ними также используем и кластерный анализ, применяемый в современной геофизике для структуризации данных по определенным параметрам.

Необходимо отметить, что метод GPS-интерферометрии имеет ракурсную зависимость амплитуды исследуемого сигнала от направления визирования на навигационный ИСЗ, что в ряде случаев может сильно затруднить его применение для исследования волновых возмущений [7].

## 2. Исследование методов кластерного анализа применительно к данным GPS-интерферометрии

Кластер-анализ [10–13] — это способ группировки наборов многомерных объектов, основанный на представлении результатов отдельных наблюдений точками некоторого геометрического пространства с целью выделения в исходных данных однородных подмножеств, объекты внутри которых похожи, в известном смысле, друг на друга. Поэтому сами кластеры можно трактовать как непрерывные области некоторого пространства с относительно высокой плотностью точек, отделенные от других таких же областей областями с относительно низкой плотностью точек.

Единого, общепризнанного критерия для применения кластеризации в прикладных задачах в настоящее время нет. Например, в кластерной мере сходства объектов по классификации Снита и Сокела [12] может быть выделено четыре группы критериев: 1) коэффициенты корреляции; 2) меры расстояния; 3) коэффициенты ассоциативности и 4) вероятностные коэффициенты сходства. Каждая группа определяет физически ясный признак, в целом связанный с конкретными свойствами изучаемых объектов, именно поэтому в последнее время кластерный анализ все шире используется в геофизике [10].

### 2.1. Основные методы кластерного анализа

Все разнообразие кластерных методов может быть условно разделено на шесть основных семейств [13]: 1) иерархические агломеративные методы; 2) иерархические дивизимные методы; 3) итеративные методы группировки; 4) методы поиска модальных значений плотности; 5) методы сгущений; 6) методы, использующие теорию графов.

Эти семейства имеют различные подходы к созданию групп, и применение различных методов к одним и тем же данным может привести к различающимся результатам. Итак, *большинство методов кластерного анализа являются эвристическими*, основанными на опыте использования. Они являются не более чем

*правдоподобными алгоритмами, используемыми для создания объектов с заданными свойствами.*

Из всех видов кластерных методов в приложениях наиболее часто употребляются иерархические агломеративные методы, в которых присоединение объектов происходит к уже существующим кластерам на основе некоторого начального разбиения. Указанные методы различаются и по правилам построения кластеров. Существует много различных правил группировки, каждое из которых порождает специфический иерархический метод, четыре из которых наиболее распространены. К ним относятся методы одиночной связи, полной связи, средней связи и метод минимальной дисперсии.

### 2.2. Тестирование методов кластерного анализа при обработке данных GPS-интерферометрии

В ходе анализа указанные четыре метода кластеризации были реализованы в виде прикладных программ для изучения эффективности каждой методики при ее работе с наборами тестовых квазиреальных данных.

Тестовые файлы данных имитируют реальные данные, получаемые при обработке методом GPS-интерферометрии спутниковой информации, и содержат в себе модели различных структур, а также аддитивный некоррелированный шум. Тестовые данные включают структуры различных типов — как компактные квазисферические, так и размытые объекты, обладающие разными дисперсиями по каждому из признаков.

Каждый тестовый файл содержит несколько сотен единичных объектов обладающих четырьмя параметрами: двумя координатами, скоростью и направлением движения. Эти объекты принадлежат множествам с задаваемыми при моделировании параметрами (средние координаты, скорости, направления, а также дисперсии соответствующих величин). Расположение точек по координатам моделируется гауссовым процессом с максимумом распределения в центре кластера. Скорости и направление движения объектов, принадлежащих кластеру, также определяют как случайные гауссовы отклонения от заданных средних величин. Шум характеризуется случайным выбором параметров и составляет от 5 до 30% выборки по объему данных. Значения дисперсий координат, амплитуды случайных отклонений и уровень шума известны и задаются при создании файла.

На рис. 1 представлен типичный файл тестовых данных, содержащий 150 объектов. Всего для исследований методик создано около 100 таких файлов с количеством объектов 150 и 300 соответственно. Скорости каждой структуры задаются в диапазоне от 200 до 2000 м/с, со случайными нормальными отклонениями. Направления движения объектов отличаются от среднего в пределах до 20 градусов, что соответствует точности методики GPS-интерферометрии [6–9]. Точками с заполнением обозначены объекты, принадлежащие заданным четырем различным кластерам. Стрелками указаны направления и амплитуды скоростей. Объекты, представленные в виде квадратов с белым центром, являются шумом. Координаты  $X, Y$  на рис. 1 и далее представляют относительное географическое местоположение исследуемых точек в регионе анализа, самая

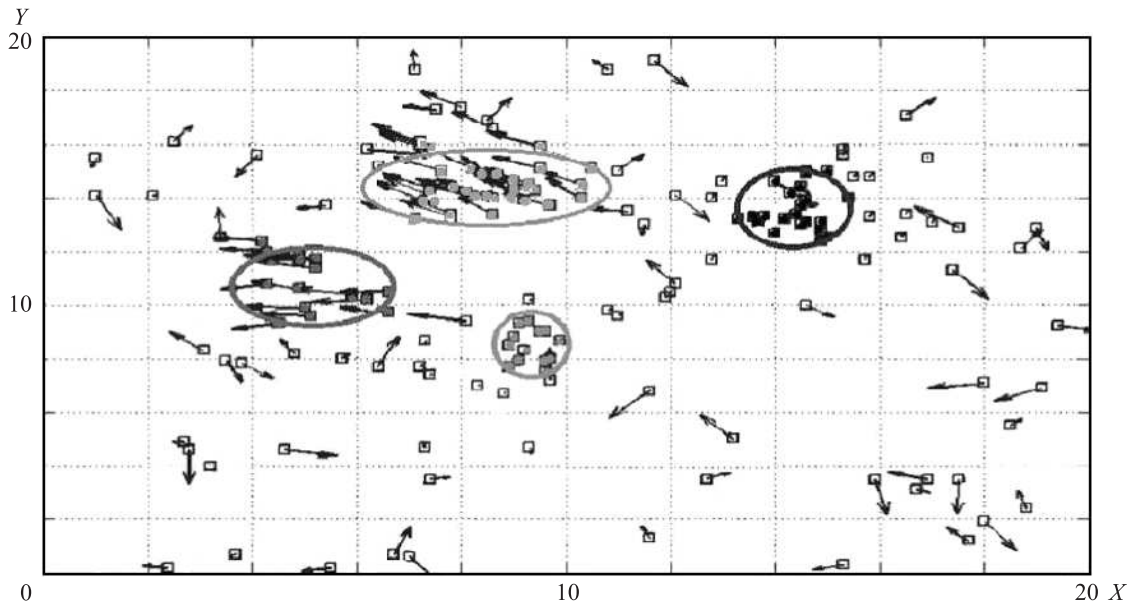


Рис. 1. Пример тестовых данных, содержащих четыре кластера различной формы и дисперсии (отмечены эллипсами)

южная и западная точка которого имеет координаты соответственно  $(0, 0)$ .

Таким образом, созданные файлы данных по своей структуре, составу и численным значениям максимально приближены к данным, получаемым экспериментально, и используются нами для проверки реализованных алгоритмов кластеризации. Задаваемые в численном эксперименте структуры обладают известными параметрами, что позволяет оценить качество тестируемых методов кластеризации.

### 2.3. Исследованные алгоритмы кластеризации

В работе нами исследованы следующие алгоритмы кластеризации, применяемые к созданным тестовым данным: 1) метод средней связи по всем четырем параметрам (координатам, скоростям и направлениям); 2) метод одиночной связи по координатам, совмещенный с методом средней связи по скоростям и направлениям; 3) модифицированный метод средних плотностей; 4) модифицированный метод « $k$ -means» [14].

**Метод средней связи** [10–13] представляет собой схему кластерного анализа, при которой точка добавляется к уже существующему кластеру или два кластера  $i$  и  $j$  объединяются, если расстояние между их средними координатами не превосходит заданной величины (на первом этапе разбиения все объекты считаются кластерами). Далее строится матрица расстояний между объектами, в которой находится наименьшее значение, и происходит объединение  $i$ -го и  $j$ -го кластеров. При этом  $j$ -му кластеру присваиваются усредненные с весом этих кластеров значения координат, направлений и скоростей, а  $i$ -й элемент далее не рассматривается. На следующем шаге строится матрица расстояний на единицу меньшей размерности, опять находится минимальное значение и происходит объединение. Этот цикл заканчивается в тот момент, когда минимальное значение расстояния превысит заданную пороговую величину.

**Метод одиночной связи** отличается тем, что после вычисления матрицы расстояний и нахождения минимального значения объединение точек в кластеры происходит без их изменения по координатам.

При использовании **метода средних плотностей** [12] вычисляется поле средних плотностей с различным шагом разбиения, определяемым адаптивными алгоритмами. Из полученного поля плотностей происходит выделение кластеров. Для выделенных структур вычисляются средние координаты, скорость и направление. Объекты со значениями скоростей и направлений, отличающихся от среднего больше, чем на определенную величину, считаются шумом. Кластеры, расположенные друг от друга на расстоянии менее определенного, считаются одним объектом.

**Алгоритм « $k$ -means»** (иначе — алгоритм  $k$ -средних) — один из наиболее популярных в данное время методов кластеризации. Он представляет собой модификацию алгоритма, используемого в математической статистике для нахождения оценок максимального правдоподобия параметров вероятностных моделей, примененную для разделения смеси гауссиан. В нем множество элементов векторного пространства разбивают на заранее известное число кластеров  $k$ , стремясь минимизировать дисперсию для точек каждого кластера:

$$V = \sum_{i=1}^k \sum_{x_j \in S_i} (x_j - \mu_i)^2,$$

где  $k$  — число кластеров,  $S_i$  — полученные кластеры,  $i = 1, 2, \dots, k$ , и  $\mu_i$  — центры масс векторов  $x_j \in S_i$ .

Алгоритм состоит в том, что на каждой итерации заново вычисляется центр масс для каждого кластера, полученного на предыдущем шаге, затем векторы разбиваются на кластеры вновь в соответствии с тем, какой из новых центров оказался ближе по выбранной метрике. Алгоритм завершается, когда на какой-то итерации не происходит изменения кластеров. Хотя получение точного решения для произвольного

начального разбиения является трудной задачей, получение приблизительного результата происходит весьма быстро [14].

К основным недостаткам алгоритма относятся следующие обстоятельства: во-первых, время вычисления зависит степенным образом от количества объектов [14]; во-вторых, требуется указать точное количество кластеров до начала их поиска, причем существует неопределенность начального выбора центров кластера.

Для устранения известных недостатков метод *k*-means был модифицирован за счет специального выбора первичного разбиения, начальное число кластеров в котором нами задается как половина объема исследуемой выборки. Хотя первоначальное разбиение занимает дополнительное время, последующие разбиения происходят гораздо быстрее и в целом этот способ приводит к значительному уменьшению ошибки. Для проверки устойчивости получаемого решения этим алгоритмом мы проводим его инициализацию и последующую обработку статистически значимое число раз (от 30 до 100), при этом определяется наиболее вероятное число кластеров, соответствующих оптимальному решению.

**3. Проверка алгоритмов кластеризации на тестовых данных**

В ходе проверки различные методы распознавания, рассмотренные в п. 2.3, использовались для решения тестовых задач определения параметров структур, имитирующих реальный эксперимент.

Например, на рис. 2 представлены результаты работы алгоритма *k*-means для обработки данных, содержащих три кластерных центра в двух случаях — при наличии слабого (рис. 2, а) и сильного (рис. 2, б) шума, т. е. число шумовых отсчетов — 30% от полезного объема информации. Видно, что метод *k*-means хорошо выделяет плотные кластеры при малом и среднем уровне шума, причем вероятность правильного выделения близка к 1. При увеличении числа случайных точек (обозначены белым на рис. 2) свыше 25% от

общего числа вероятность нахождения истинного числа объектов заметно падает, причем возможно появление выбросов (ошибочных включений) на границах кластеров (см., например, область ложного 4-го кластера, отмеченного цифрой 4 на рис. 2, б). Впрочем, на созданных имитациях экспериментальных данных с большим уровнем шума все исследованные методики кластеризации делают заметные ошибки в распознавании структур.

В случае, когда выделяемая структура имеет заранее неизвестную форму, отличную в многомерном случае от сфероида, использование метода средней связи и средней плотности неизбежно ведет к потере данных, а следовательно, искаженному выделению и интерпретации структур. Метод одиночной связи по координатам склонен к распознаванию двух различных близко расположенных структур как единого целого при большом уровне шума, когда соединение кластеров происходит через своеобразные «мостики», образованные объектами, не принадлежащими ни одному кластеру.

Немаловажной особенностью данных алгоритмов является устойчивость разбиения в зависимости от параметров кластеризации, что позволяет сделать вывод о правдоподобии картины разбиений без дополнительного, зачастую довольно трудоемкого и в ряде случаев неформализуемого анализа.

Все результаты применения кластеризации, полученные при тестировании различных алгоритмов, нами были статистически обработаны и для удобства сравнения сведены в табл. 1. Тестирование производилось на 100 различных наборах данных (файлов), содержащих от 150 до 300 точек (из них до 25% — аддитивный шум). Для получения статистически значимых результатов для метода *k*-means проводилось от 30 до 100 запусков, для остальных по 1, поскольку для них результаты разбиения являются повторяющимися при неизменных входных данных. Значок  $\delta$  означает ошибку оценки приводимого параметра, *X, Y* — пространственные

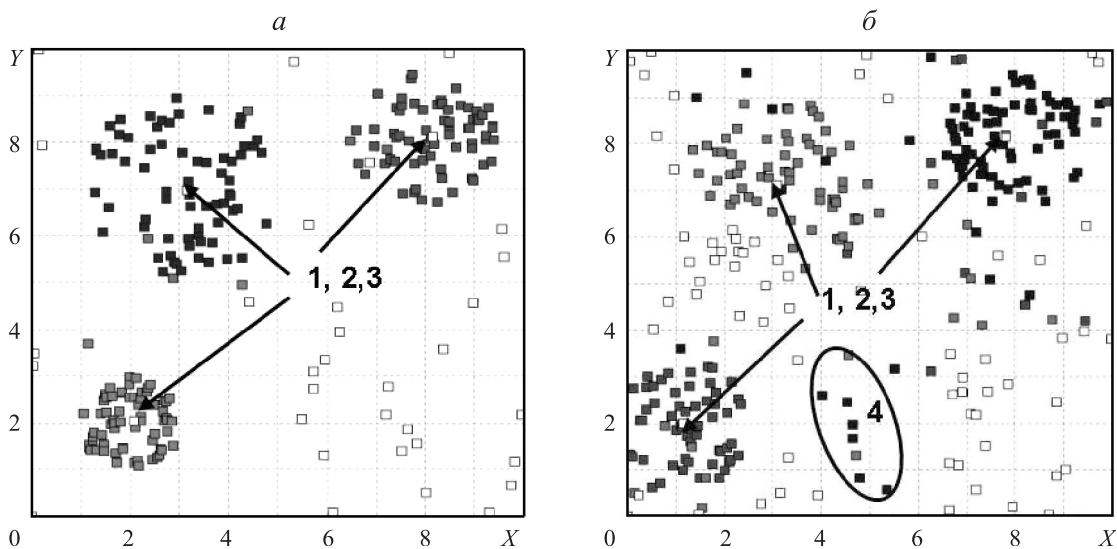


Рис. 2. Распознавание структур с разным уровнем шума методом *k*-means. Центры заданных и выделенных кластеров обозначены цифрами 1–3; на рис. б цифрой 4 обозначен пример кластера-артефакта, появившегося при большом уровне шума

Таблица 1  
Средние значения исследуемых параметров  
для различных методов кластеризации

Реализованный метод	Количество сгруппированных объектов, %	$\delta X$ , %	$\delta Y$ , %	$\delta D(x, y)$ , %
<i>k</i> -means	90	3	2.5	12
Средней связи	80	5	5	8
Одиночной связи	95	6	8	15
Средних плотностей	60	20	20	30

координаты,  $D$  — дисперсия. В табл. 1 приведены средние величины оценок исследуемых параметров.

Проведенное тестирование показало нецелесообразность использования метода средних плотностей для целей дальнейших исследований, поскольку он дает наименее точные результаты при определении средних величин и дисперсий кластеров, а также склонен игнорировать кластеры малой плотности.

Результаты нашего анализа показывают, что использование метода средней и одиночной связи, а также метода *k*-means для тестовых задач в целом эквивалентно и приводит к неустойчивому распознаванию лишь сильно размытых, зашумленных и вытянутых структур. Метод средней связи, имеющий минимальную оценку дисперсии, почти вдвое менее точен при оценке средних параметров структуры по сравнению с методом *k*-means, а метод одиночной связи уступает по точности оценок им обоим.

Поэтому наиболее перспективными для целей настоящей работы признаны два метода — *k*-means и метод средней связи. Оба дают оценки местоположения центров кластеров в пределах статистической

погрешности, причем у метода *k*-means собственно оценки средних лучше. Несколько хуже он определяет кластерную дисперсию, особенно в случае «размытых» структур. Наша реализация метода весьма быстрая, что позволяет провести целую серию испытаний для различных начальных вариантов разбиений и убедиться в устойчивости получаемого решения. Именно этот метод реализован в пакете прикладных программ CRASS GPS.

#### 4. Применение кластеризации для обработки эксперимента

В экспериментальной обработке использованы данные за 12 дней, относящиеся к периоду с 1 по 11 сентября 2008 г., от 100 станций сети GEONET (Япония) [15], на которых наблюдались фазы сигналов частот  $L1$  и  $L2$  с 30-секундной дискретизацией. Всего в обработке, таким образом, участвовало более 5 млн индивидуальных измерений фазы, которые делились практически поровну между геофизически спокойным и умеренно возмущенным периодами, при этом гелиообстановка была полностью спокойной [16].

Все используемые станции для проведения дополнительных методических исследований были разделены на пять подсетей из 20 приемников каждая по территориальному признаку компактного расположения и в местах наибольшей плотности приемников, что позволяет получить размеры измерительной ячейки (т. е. группы из трех станций) менее чем  $30 \times 30$  км. Отметим, что все ячейки полностью соответствуют требованиям методики GPS-наблюдений [6, 7].

В ходе обработки для каждой подсети в определенные периоды времени выделялись волновые неоднородности ионосферы по общепринятой статистической методике [7]. На рис. 3 представлены результаты такого анализа для всего дня 6 сентября 2008 г., причем, чтобы

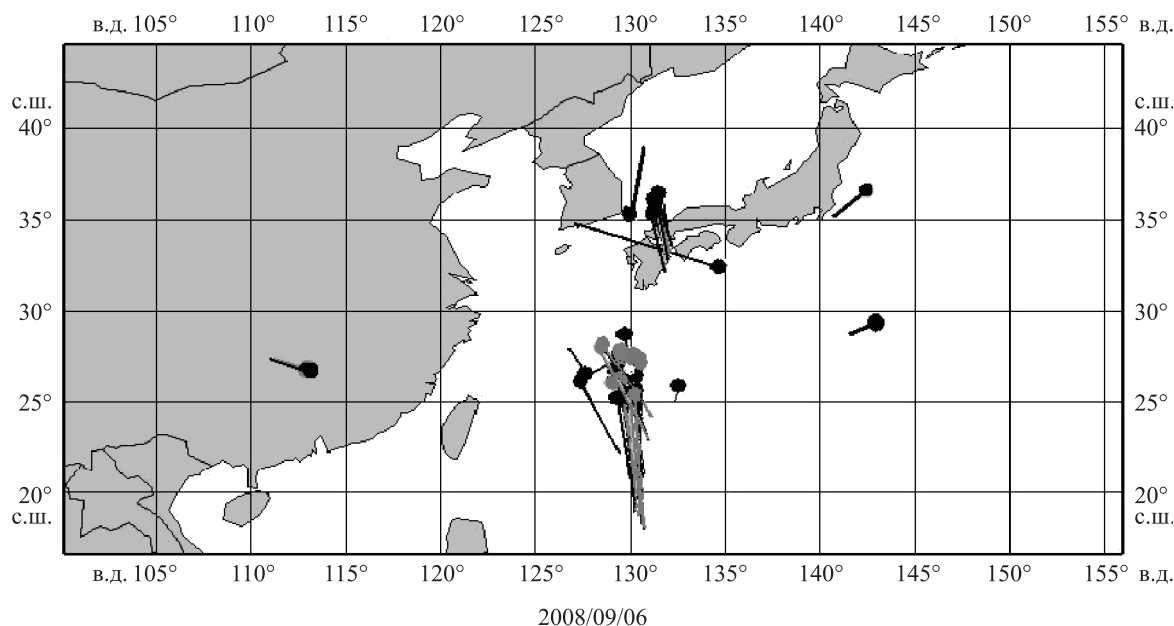


Рис. 3. Ионосферные неоднородности, выделенные 06.09.2008 одной наблюдательной подсетью. На карте обозначены координаты подыоносферной точки, направление задает азимут движения волновой структуры, длина вектора пропорциональна скорости

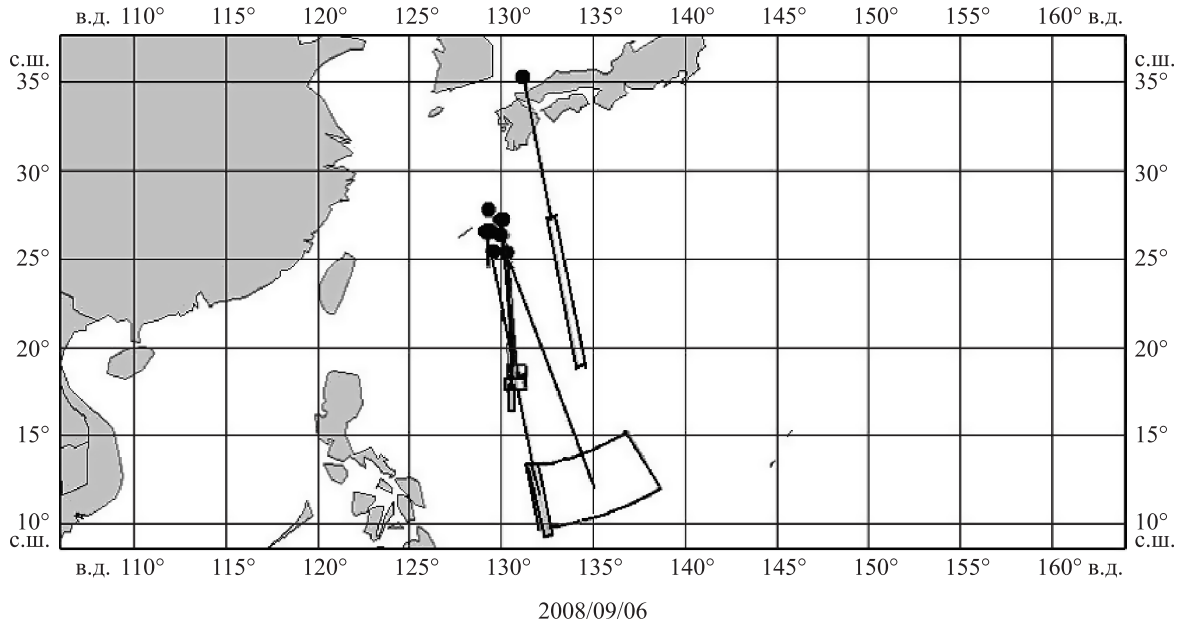


Рис. 4. Выделенные кластеры для входных данных 06.09.2008. Размеры криволинейного сегмента содержат оценку дисперсии скорости и направления движения кластера

не загромождать рисунок, данные приведены только для одной (самой южной) подсети из 20 станций.

На приведенной карте видно как скопление неоднородностей, потенциально обладающих близкими свойствами, так и ряд «шумовых» срабатываний наблюдательной подсети. Далее для каждого 4-часового интервала наблюдений периода 01.09–12.09 был проведен кластерный анализ, в ходе которого выделены устойчивые волновые структуры и исключены те неоднородности, которые детектировались неустойчиво. Результат кластерного анализа для данных рис. 3 приведен на рис. 4.

В дальнейшем анализе учитывались все данные от всех 100 станций, причем использовались как результаты простого статистического детектирования структур, аналогичного рис. 3, так и результаты применения процедур кластеризации.

Результаты проведенного обоими методами статистического анализа содержатся в табл. 2, где представлены оценки параметров волновых структур — скорости и направления движения, выделенных методом GPS-интерферометрии, при стандартной статистической обработке и кластерным анализом. Второй столбец содержит результаты оценок точностей методики, выполненные другими авторами и приведенные в обзоре [7]. Отметим, что табл. 2 содержит вполне предсказуемый результат — данные после кластерного

анализа имеют меньшую дисперсию, поскольку процедура кластеризации по своему построению обладает фильтрующими свойствами в пространстве выбранных параметров. Кроме того, применение алгоритмов кластеризации, в отличие от методов статистики, позволяет изучать пространственно-временное распределение неоднородностей, причем именно по тем параметрам, которые являются наиболее значимыми для физической интерпретации.

Отметим, что при наличии геомагнитных возмущений методика GPS-интерферометрии несколько теряет свою эффективность из-за увеличения шумовой компоненты в данных. Применение кластеризации в этом случае дает более уверенные результаты именно за счет фильтрующих свойств данной методики.

Итак, основное преимущество кластерного анализа заключается в возможности полноценно использовать малые выборки данных, а также данных с высоким уровнем аддитивного шума, которые при проведении статистической обработки дают значительные погрешности. Более того, проблема неполноты GPS-данных при решении задач высокоорбитальной томографии [17] для малого числа приемников в регионе исследований в настоящее время представляет значительную трудность и вообще характерна для подобных задач.

В заключение отметим, что в ходе обработки данных сравнивались результаты, полученные на близко рас-

Таблица 2

**Ошибки определения ионосферных неоднородностей различными методами**

Параметр	Обзорные данные [7], min-max значения %	Статистическая обработка		Кластерный анализ	
		Невозмущенные условия, %	Возмущенные условия, %	Невозмущенные условия, %	Возмущенные условия, %
Скорость структуры	6–20	15–20	18–25	8–12	10–15
Азимут	5–16	10–15	12–18	7–10	8–14

положенных измерительных GPS-ячейках, содержащих общие станции и полностью независимые друг от друга. При этом количественных различий не выявлено, т. е. полученные нами и приведенные выше оценки погрешностей методики устойчивы и не зависят от способа формирования измерительных ячеек.

### Основные выводы

Целью настоящей работы являлась разработка методик анализа для проведения экспериментальных определений параметров движения неоднородных структур верхней атмосферы на основе применения кластеризации для классификации и группирования структур по наиболее существенным в экспериментальной практике параметрам.

В работе проведены методические исследования возможностей распознавания ионосферных структур методами кластерного анализа.

Было исследовано четыре алгоритма кластеризации: метод средней связи, метод одиночной связи, модифицированный метод средних плотностей и модифицированный метод  $k$ -means. Последний метод реализован в пакете исследовательских прикладных программ CRASS GPS. Проведено тестирование указанных алгоритмов на квазиреальных данных, моделирующих результаты, полученные при обработке экспериментальных данных.

Лучшие результаты при выделении кластеров по координатам в случае компактных множеств показали методы  $k$ -means и средней связи, а при вытянутых и рассеянных структурах метод одиночной связи. Для обработки реальных данных был выбран метод  $k$ -means как наиболее быстрый и автоматизированный.

В ходе статистической обработки реальных данных получено, что результаты определения параметров ионосферных неоднородностей имеют меньшие погрешности после проведения кластеризации, нежели оценки, полученные только статистическими методами. Этот результат вполне ожидаем, поскольку процедура кластеризации по своему построению обладает фильтрующими свойствами в пространстве выбранных параметров.

В целом полученные нами оценки согласуются с результатами других авторов, но в случае кластерного анализа имеют меньшие дисперсии, т. е. ошибки определения параметров. Количественных различий между ячейками одной подсети, содержащими общие станции и полностью независимыми друг от друга, не выявлено, т. е. приведенные нами оценки погрешностей устойчивы и не зависят от способа формирования измерительных ячеек.

Основное преимущество кластерного анализа заключается в возможности полноценно использовать малые выборки данных, а также данные с повышенным уровнем аддитивного шума, соответствующие геомагнитно-возмущенным дням. При проведении статистической обработки такие данные могут давать значительные погрешности. Вместе с тем, использование методов кластеризации в случае сильного шума позволяет осуществить эффективную фильтрацию входных данных и тем самым уменьшить дисперсии получаемых оценок параметров неоднородных структур.

Работа выполнена при частичной поддержке ФЦП «Научные и научно-педагогические кадры инновационной России» (ГК № П1072 и П167) и Министерства образования и науки РФ (проект 14.740.11.0203).

### Список литературы

1. Акасофу С., Чепмен С. Солнечно-земная физика. М., 1975.
2. Гершман Б.Н., Ерухимов А.М., Яшин Ю.Я. Волновые явления в ионосфере и космической плазме. М., 1984.
3. Илюшин Я.А., Куницын В.Е. // Радиотех. и электрон. 2004. **49**, вып. 2. С. 171.
4. Куницын В.Е., Андреева Е.С., Кожарин М.А., Нестеров И.А. // Вестн. Моск. ун-та. Физ. Астрон. 2005. № 1. С. 74.
5. Куницын В.Е., Терещенко Е.Д., Андреева Е.С., Нестеров И.А. // Успехи физ. наук. 2010. **180**. № 5. С. 40.
6. Hoffman-Wellenhoft V. GPS Theory and Practice. Vienna; N.Y., 1998.
7. Афраймович Э.Л., Перевалова Н.П. GPS-мониторинг верхней атмосферы Земли. Иркутск, 2006.
8. Захаров В.И., Зиенко А.С. // Вестн. Моск. ун-та. Физ. Астрон. 2007. № 2. С. 47.
9. Захаров В.И., Зиенко А.С., Куницын В.Е. // Электромагн. волны и электронные системы. 2008. № 8. С. 51.
10. Gvishiani A., Dubois J. Artificial intelligence and dynamic systems for geophysical applications. Berlin, 2002.
11. Ким Дж.-О., Мюллер Ч.У., Клекка У.Р. и др. Факторный, дискриминантный и кластерный анализ. М., 1989.
12. Sneath Ph.A., Sokal R.R. The principles and practice of numerical classification. San Francisco, 1973.
13. Гумис Л.Х. Статистическая классификация и кластерный анализ. М., 2003.
14. Arthur D., Vassilovitskii S. // Proceedings of the eighteenth annual ACM-SIAM symposium on discrete algorithms. New Orleans (Louisiana), 2007. P. 1027.
15. <http://www.gsi.go.jp/ENGLISH/index.html>.
16. <http://spidr.ngdc.noaa.gov/spidr>.
17. Nesterov I.A., Kunitsyn V.E. // Adv. in Space Res. 2011. doi: 10.1016/j.asr.2010.11.034.

**The using of the cluster analysis for GPS-radio interferometry data processing****V. I. Zakharov**<sup>1,a</sup>, **P. A. Budnikov**<sup>2,b</sup><sup>1</sup>*Department of Physics of Atmosphere, Faculty of Physics, M. V. Lomonosov Moscow State University, Moscow 119991, Russia.*<sup>2</sup>*Institute of Applied Geophysics by Acad. E. K. Fedorov (FSBO «IAG»), Moscow 129128, Russia.**E-mail: <sup>a</sup> zvi\_555@list.ru, <sup>b</sup> atm5571@phys.msu.su.*

In the article there were studied the possibilities and the applications of the cluster analysis methods for estimation of the upper atmosphere — ionosphere — inhomogeneous parameters, that were obtained by GPS-radio interferometry method. It is spent the testing of various clusterization algorithms for the quasireal data. Considerable reduction of the dispersion of the registered ionospheric structures parameters in comparison with usual statistical processing is revealed, especially in case of small data sample and high noise level. Examples of the experimental data processing by the cluster analysis are resulted.

*Keywords:* ionosphere, inhomogeneous structures (irregularities), GPS-interferometry, clusterization or clustering algorithms.

PACS: 94.05.-a, 94.20.Vv, 91.10.Fc, 94.05.Sd.

*Received 20 May 2011.*

English version: *Moscow University Physics Bulletin* 1(2012).

**Сведения об авторах**

1. Захаров Виктор Иванович — канд. физ.-мат. наук, доцент, доцент; e-mail: zvi\_555@list.ru.
2. Будников Павел Александрович — научный сотрудник; e-mail: pavel9860@gmail.com.