

# Определение параметров межзвездного поглощения света по данным каталога Hipparcos

Ф.Амосов

23 апреля 2015 г.

## Аннотация

Основная задача исследования — построение метода автоматического поиска пылевых облаков в окрестности Солнца на основе массовых каталогов звезд. На первом этапе метод был применен к данным каталога Hipparcos, а именно для построения двумерной панорамы распределения пылевых облаков на небесной сфере. Метод исследования основан на сравнении эталонного показателя цвета звезды данного спектрального класса с наблюдаемым показателем цвета. Так как полная двумерная спектральная классификация известна не для всех звезд каталога Hipparcos, была решена вспомогательная задача: используя параллаксы звезд каталога Hipparcos дополнить информацию о звездах классом светимости на основе данных о видимой звездной величине и параллаксе звезды. В результате работы была построена карта распределения пылевой материи, вызывающей покраснения света звезд, на небесной сфере. Недостаточная плотность звезд и низкие относительные точности параллаксов для далеких звезд не позволяют полностью использовать возможности метода, которые будут полностью раскрыты по завершении миссии GAIA

## 1 Введение

Общая задача поиска облаков межзвездной пыли, ответственной за поглощение света звезд, является трехмерной. Для ее надежного выполнения необходим массовый каталог параллаксов звезд. На сегодняшний день таким каталогом является Hipparcos, но, как мы увидим, его точности недостаточно для решения этой задачи. Видимо её полное решение будет возможно только после получения результатов миссии GAIA, чьей основной задачей является определение структуры Млечного Пути в окрестности Солнца. Поэтому мы представим решение частной задачи — задачи построения двумерного распределения пыли по небесной сфере. Ее решение позволит по крайней мере определить направления наибольшего поглощения и изменения таких характеристик звезд как звездная величина и показатель цвета. Это позволит в будущем учитывать межзвездное поглощение при определении характеристик звезды таких как, к примеру, показатель цвета.

## 2 Исходные данные

### 2.1 Общие сведения о каталоге Hipparcos

В 1989 году Европейское Космическое Агентство (ESA) осуществило запуск космического аппарата HIPPARCOS (HIgh Precision PARallax COllecting Satellite — «спутник для сбора высокоточных параллаксов») с целью получения положений, собственных движений и параллаксов звезд на миллисекундном уровне точности. Космический аппарат проработал на орбите 37 месяцев, в течение которых он выполнял астрометрические и фотометрические измерения звезд по заданной программе. Обработка этих наблюдений привела к созданию двух каталогов: Hipparcos[1], содержащего информацию о 118218 звездах с точностью определения положений, годичных собственных движений и параллаксов на уровне 1 mas (milli arc second), и каталога Tycho[5], содержащего уже свыше 1 млн. звезд, с точностью измерения тех же параметров до 25 mas.

Положения и собственные движения звезд в Hipparcos приводятся в фундаментальной системе ICRS (International Celestial Reference System), реализованной в настоящее время с помощью каталога внегалактических радиоисточников, получившего название ICRF (International Celestial Reference Frame). Следует отметить, что достигнутая точность привязки осей координат системы отсчета каталога HIPPARCOS к осям ICRF оценивается величиной 0.6 mas по всем трем углам поворота и величиной 0.25 mas/год по всем трем компонентам вектора остаточного взаимного вращения двух систем отсчета.

В 2007 году вышла новая редакция астрометрических данных каталога Hipparcos [4] — каталог HIPNEWCAT (HIPparcos NEW astrometric CATalog). Утверждается, что точность положений, параллаксов и собственных движений всех звезд, ярче  $H_P = 8$ , улучшена в 4 раза, а для всех остальных звезд более, чем в 2 раза. Уменьшена взаимная корреляция параметров иногда в 10 раз. Именно эта версия использовалась в работе в качестве источника астрометрических данных.

## 2.2 Фотометрические системы каталогов Tycho и Hipparcos

Фотометрические измерения на основном инструменте спутника HIPPARCOS выполнялись в широкой полосе (обозначаемую как  $H_P$ ). В дополнение, почти для всех звезд каталога была выполнена двухцветная фотометрия (фотометрия Tycho величины  $V_T$  и  $B_T$ ). Точность определения  $H_P$  составляет 0.0004<sup>m</sup> – 0.007<sup>m</sup> (для звезд 2 – 12<sup>m</sup>), а точность одного измерения – 0.003<sup>m</sup> – 0.05<sup>m</sup>.

Фотометрические системы  $H_P$ ,  $V_T$  и  $B_T$  — это инструментальные системы, и они не совпадают с общепринятой системой Джонсона. Используя значения звездной величины  $V_J$  по шкале Джонсона и показателя цвета для 8000 стандартных звезд с хорошими фотометрическими данными в системе  $B_T$  и  $V_T$ , были получены следующие эмпирические линейные соотношения, применимые к диапазону  $-0.2 < (B - V)_T < 1.8$ :

$$V_J = V_T - 0.090(B - V)_T$$

$$(B - V)_T = 0.850(B - V)_T$$

Точность этих преобразований в среднем лучше, чем 0.015<sup>m</sup> для  $V_J$  и 0.05<sup>m</sup> для  $(B - V)_T$ . Эти преобразования применимы к звездам, чей цвет не искажен межзвездным поглощением, и игнорируют зависимость от класса светимости. Формулы вообще не применимы к звездам класса M, даже если их показатель цвета  $(B - V)_T < 1.8^m$ .

## 2.3 Спектральные характеристики звезд

Каталог Hipparcos для большинства звезд содержит информацию о спектральном типе, полученную из наземных наблюдений. Основной источник — Мичиганский каталог (Michigan catalogue for the HD stars, vol. 1-4, (Houk+, 1975-1988)) и несколько других источников. Однако, информация о спектральных классах приведена только для звезд южного экваториального полушария (рис.2).

## 2.4 Используемые данные

В каталоге Hipnewcat отсутствуют спектральные данные, поэтому в данной работе они были взяты из каталога Hipparcos. Также из каталога Hipparcos были взяты значения видимой звездной величины V. Итак, мы рассматривали только те звезды, которые имеют данные о положении, параллаксе в Hipnewcat и данные о спектральном типе и видимой звездной величине в Hipparcos. Кроме того, мы не рассматривали звезды, у которых в каталоге Hipnewcat указано число компонент более одной. Число звезд, используемых в данной работе получилось равным 98827.

Таким образом, в нашей работе использовалась следующая информация:

- положения звезд (каталог Hipnewcat)
- параллаксы звезд (каталог Hipnewcat)
- фотометрия ( $V_{mag}$  - каталог Hipparcos)
- класс светимости (каталог Hipparcos)

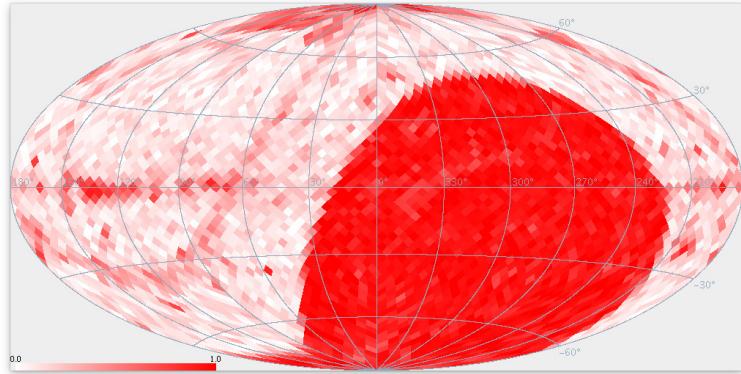


Рис. 1: Распределение звезд на небесной сфере, для которых в каталоге HIPPARCOS имеются сведения о классе светимости. Чем больше в пикселе доля звезд, имеющих класс светимости, тем он краснее. Видим отсутствие класса светимости практически у всех звезд северного экваториального полушария.

### 3 Покраснение

#### 3.1 Определение

Межзвездное поглощение может быть описано избытком цвета. Избыток цвета мы будем называть «покраснением». *Покраснение* звезды есть

$$E = E_{B-V} = (B - V)_{obs} - (B - V)_{int}, \quad (1)$$

где  $(B - V)_{obs}$  — ее видимый показатель цвета звезды (с учетом межзвездного поглощения), а  $(B - V)_{int}$  — теоретический показатель цвета звезды (без учета межзвездного поглощения). Значение  $(B - V)_{obs}$  мы можем получить на основе данных фотометрии звезды из каталога. В таблице ?? приведены значения абсолютной звездной величины  $M_V$  и  $(B - V)_{int}$  и спектральный класс, взятые из работы [?]. То есть, для получения  $(B - V)_{int}$  звезды нам нужно знать ее спектральный класс и класс светимости.

Приведем пример расчета покраснения на звезде HIP 44800,

- У нее в каталоге  $(B - V)_{obs} = 0.535^m$
- Класс F7V, поэтому (по [7])  $(B - V)_{int} = 0.493^m$
- Покраснение  $0.535^m - 0.493^m = 0.042^m$

Покраснение — это количественное измерение межзвездного поглощения, поэтому мы можем сказать, что «между нами и звездой HIP 44800 пыли на  $0.042^m$ ».

### 4 Способ получения классов светимости

Как мы увидели в обзоре данных каталога Hipparcos, практически у всех звезд северного экваториального полушария отсутствует класс светимости. Для нас его наличие чрезвычайно важно, ввиду того, что мы на основе класса светимости и спектрального класса рассчитываем истинное значение  $B - V$  для звезд  $(B - V)_{int}$ . Тем самым, отсутствие класса светимости у половины звезд делает невозможным проведение наших расчетов для всего северного экваториального полушария.

Для исправления этого недостатка используем метод машинного обучения. Натренируем классификатор, который будет определять класс светимости для звезды по двум факторам — ее показателю

Таблица 1: Структура обучающего множества

класс светимости	$B - V < 0.6$	$B - V \geq 0.6$	всего
III	1947	16681	18628
V	15549	5630	21179
III & V	17496	22311	39807

цвета  $B - V$  и ее абсолютной звездной величине  $M_V$ . Этих факторов должно быть достаточно, т.к. классы светимости теоретически разделимы на диаграмме Герцшпрунга-Рессела.

В каталоге HIPPARCOS доля звезд, которые не относятся ни к III, ни к V классам светимости мала (16,3%, 8058 из 49285). Поэтому, мы упростим задачу — обучим линейный бинарный классификатор, который будет предсказывать III или V класс. Сделаем это с помощью метода опорных векторов [8], основные положения которого приведены в Приложении.

В качестве обучающего множества возьмем все звезды, у которых присутствует класс светимости III и V. Таких звезд 39807. Распределение этих звезд по классам светимости показано в таблице ??:

Диаграмма Герцшпрунга-Рессела, соответствующая нашему обучающему множеству, показана на рис. 3, а карта звезд этого множества — на рис. ??.

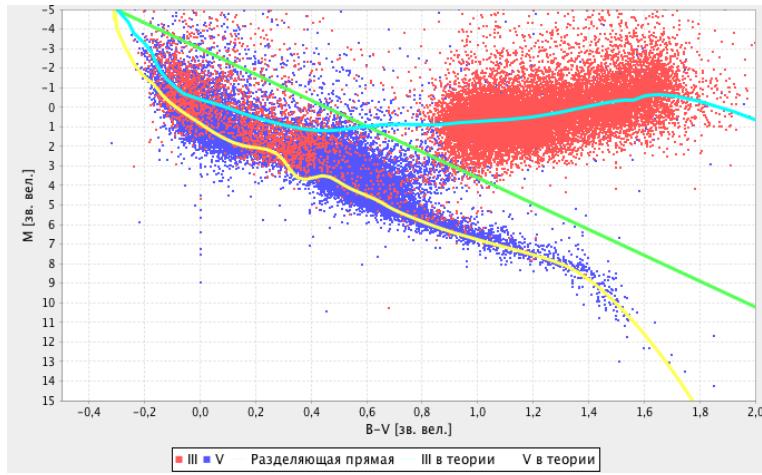


Рис. 2: Диаграмма Герцшпрунга-Рессела обучающего множества звезд (табл. ??). Разделяющая линия — зеленый цвет, теоретические кривые для III и V классов светимости показаны голубым и желтыми цветами соответственно.

Результатом обучения классификатора является разделяющая классы прямая, уравнение которой имеет следующий вид:

$$F(B - V, M) = -2.9876 \cdot (B - V) + 0.4526 \cdot M + 1.3547 = 0, \quad (2)$$

где  $B - V$  — показатель цвета,  $M$  — абсолютная звездная величина.

Мы видим, что при показателе цвета  $B - V \geq 0.6$  классификатор работает практически идеально, но при  $B - V < 0.6$  и  $F(B - V, M) > 0$  он всем звездам предсказывает V класс при большой доле звезд III класса в обучающем множестве (11.1%, 1947 из 17496). При  $B - V < 0.6$  звезды III и V класса неразделимы, поэтому для этой половины можно принять другое решение — результат работы классификатора будет взвешенным средним III и V классов, то есть некоторым средним классом,  $(B - V)_{int}$ , у которого будет равен

$$(B - V)_{int} = w_1 \times (B - V)_{int}(III) + w_2 \times (B - V)_{int}(V). \quad (3)$$

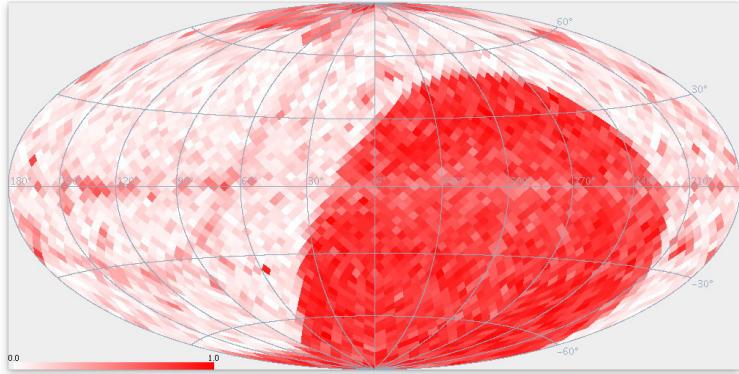


Рис. 3: Распределение звезд обучающего множества по небесной сфере. Отличие от рис.(2) в том, что здесь показаны звезды только III и V классов светимости.

В этой формуле веса  $w_1$  и  $w_2$ , ( $w_1 + w_2 = 1$ ) логично взять в соответствии с априорной вероятностью классов в этой области (0.11 и 0.89).

В таблице двумерной спектральной классификации [7] у спектрального типа G2 показатель цвета при III классе 0.733, при V — 0.630. Это максимальная разница между III и V классом в интересующем нас диапазоне. Класс, соответствующий средневзвешенному решению будет иметь показатель цвета

$$(B - V)_{int} = 0.11 \cdot 0.733 + 0.89 \cdot 0.630 = 0.64. \quad (4)$$

Как мы видим, даже в наихудшем случае отличие от V класса минимальное —  $0.01^m$  — гораздо ниже уровня ошибок показателя цвета в каталоге HIPPARCOS. Поэтому в дальнейшем мы всегда будем использовать решение классификатора в этой области по V классу.

С учетом этого соглашения проведем оценку качества работы классификатора с помощью процедуры 10-fold кросс-валидации (см. Приложение). Полученные оценки параметров классификатора (полнота, точность и F1-мера) приведены в следующей таблице:

Решение классификатора →	III	V	Класс	Точность	Полнота	F1-мера
III	16636	1992	III	95%	89%	92%
V	783	20396	V	91%	96%	93%

Как видим, классификатор имеет приемлемое качество. Результатом его работы является назначение класса светимости для испытуемой звезды с известными значениями  $B - V$  и  $M_V$ .

## 5 Градиент покраснения по расстоянию

### 5.1 Идеальная кривая покраснения

Предположим, что на некотором луче зрения бесконечно много звезд, и они расположены на нем всюду плотно. Пусть для каждой звезды мы можем идеально измерить ее покраснение. Тогда, ход покраснения на этом луче зрения должен иметь вид, показанный на рис. 5 красной линией. Покраснение должно всегда монотонно расти, т.к. пыль присутствует всюду. Очевидно, что там, где покраснение растет быстрее - пыли больше, там где медленнее - меньше. Поэтому, можно сказать, что облака на этом луче зрения находятся там, где покраснение растет «очень быстро» (синие области на рис.5).

Тем самым, построение кривых покраснения в разных направлениях на небе может позволить находить области повышенного мезвездного поглощения, то есть находить пылевые облака.

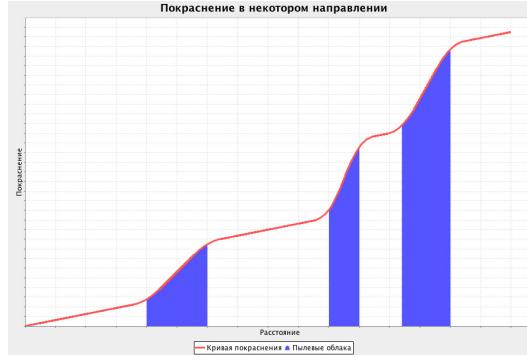


Рис. 4: Идеальный ход покраснения на луче зрения. По оси  $x$  отложено расстояние по лучу зрения, а по  $y$  — покраснение у соответствующих звезд.

Как было сказано выше, для каждой звезды, у которой известны видимый показатель цвета ( $B-V$ ) и класс светимости, можно вычислить покраснение. Если для всех таких звезд известны еще и их пространственные координаты с хорошей точностью, то мы можем говорить о пространственном распределении покраснения звезд, то есть о функции  $E = E(r)$ . Ввиду того, что практически все значения  $E(r)$  очень сильно запущлены разного рода ошибками, не удается проследить истинный ход этих функций. Поэтому приходится ограничиться линейной функцией  $E(r) = kr + b$ , параметры которой можно оценить методом наименьших квадратов. Таким образом, в данной работе ход покраснения по лучу зрения моделируется линейной функцией  $kr + b$ , а основной результат работы заключается в вычислении трендов покраснения звезд  $k$  в различных направлениях.

Построение кривых покраснения в разных направлениях позволит нам понять пространственное распределение межзвездного поглощения (пыли). Звезд в каталоге не бесконечное число, поэтому реальные кривые покраснения будут не непрерывными кривыми, а будут наборами точек, описывающими ход покраснения. Аналогично, вместо звезд на луче зрения мы должны использовать звезды в малых конусах. Поэтому, ход покраснения у нас будет выглядеть, к примеру, так, как показано на рис. ??.

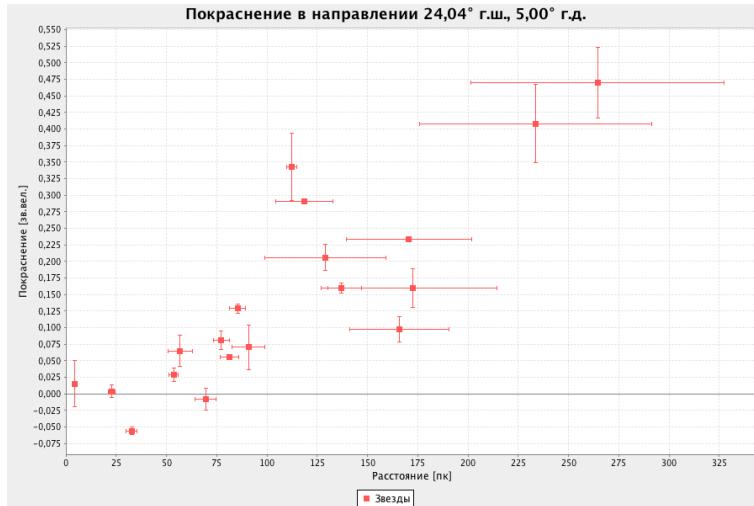


Рис. 5: Ход покраснения звезд по расстояниям в направлении  $l = 5^\circ.00, b = 24.04^\circ$ . По горизонтали — расстояния в пк, по вертикали — покраснение (зв. вел.). Данные из каталога Hipparcos

## 6 Карты градиентов покраснения звезд

В данной работе все изображения неба даются в проекции Хаммера-Айтофа. для галактической системы координат. Центр изображения — центр галактики, сверху — северный галактический полюс, снизу — южный галактический полюс, увеличение галактических долгот справа налево. Небесная сфера разбита на «пиксели» стандартным алгоритмом HEALPix [10]. Ключевым параметром (resolution parameter) алгоритма, определяющим разбиение сферы на равные площадки, является число  $N_{side}$ . Общее число пикселей  $N_{pix} = 12N_{side}^2$ . Двумя параллелями со склонениями  $\pm \arcsin(2/3)$  вся сфера разбивается на три части — экваториальную и две полярные. В полярных зонах выбирается по  $N_{side} - 1$  параллелей, в экваториальной зоне число параллелей равно  $(2N_{side} + 1)$ . На каждой параллели экваториальной области находятся центры  $4N_{side}$  площадок. Ближайшие к полюсам параллели всегда содержат по четыре площадки, а при движении от полюсов к экватору в полярных зонах число площадок на каждой параллели увеличивается на единицу. Нумерация площадок  $i = 0, \dots, N_{pix} - 1$  идет по параллелям с севера на юг. В дальнейшем мы будем обозначать площадки, соответствующие «пикселям» разбиения как  $\{P_i\}_{i=1}^{N_{side}}$ , а конусы, высекаемые соответствующими пикселями, — через  $\{C_i\}_{i=1}^{N_{side}}$ . (У нас  $N_{side} = 18$ ).

Такое разбиение позволит нам

1. рассмотреть ход покраснения в каждом конусе как одномерную функцию  $E(r) = kr + b$ . Это корректно, ввиду того, что конусы достаточно узкие;
2. сделать наши результаты «независимыми», т.к. конусы не пересекаются;
3. поместить в каждый конус примерно одинаковое число звезд, чтобы избежать недостатка звезд в некоторых конусах.

### 6.1 Вычисление тренда

Как было сказано выше, мы моделируем покраснение звезд вдоль луча зрения линейной функцией  $E(r) = kr + b$  и ищем параметры этой модели, соответствующие каждому конусу  $C_i$  с помощью метода наименьших квадратов. При этом мы используем следующие критерии отбора звезд.

#### 6.1.1 Отбор звезд по параллаксу

В каталоге Hipparcos, далекие звезды имеют очень большие ошибки в параллаксе. На расстояниях, скажем, в 400 пк они могут достигать 100%. Такие ошибки могут очень сильно испортить наши результаты. Тем самым, мы не будем рассматривать звезды, у которых относительная ошибка параллакса не превосходит 25%. Меньшее значение порога оставит нам очень малое число звезд, которых не хватит для того, чтобы вычисленные тренды были достоверными.

#### 6.1.2 Устранение выбросов

При небольшом количестве звезд на луче зрения (в нашем случае 20) метод наименьших квадратов очень чувствителен к выбросам, т.е. большие ошибки значений покраснения и/или параллакса могут очень сильно испортить тренд. Для устранения этого недостатка после построения тренда по всем звездам в конусе, мы выбрасывали те, звезды, которые давали самые большие отклонение от найденного тренда. Затем, новое значение тренда определялось уже по только по оставшимся звездам. После выброса 10% самых плохих звезд тренд, к примеру, может быть таким, как это показано на рис. ???. В данном случае, мы выбросили одну звезду (синюю). Разумеется, нельзя выбрасывать слишком много звезд из-за опасности «подгонки» данных под модель. В нашем случае порог 10% звезд — это обычно 0, 1 или 2 звезды в каждом конусе.

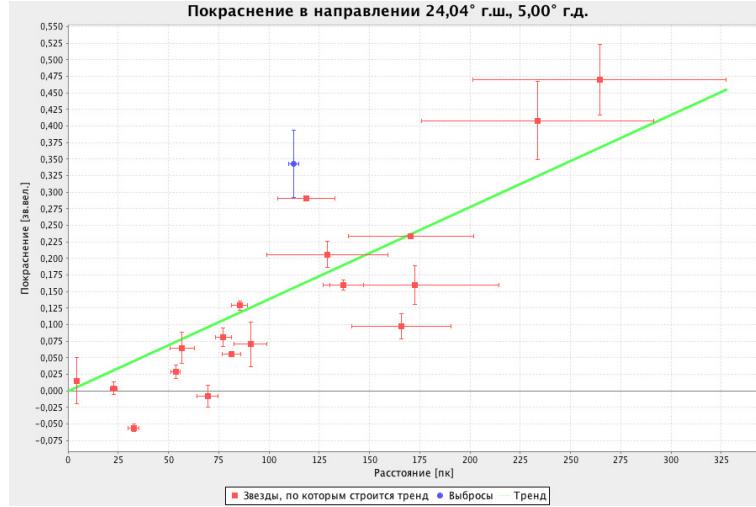


Рис. 6: Ход покраснения звезд по расстояниям в направлении  $l = 5^{\circ}.00, b = 24.04^{\circ}$ . По горизонтали — расстояния в пк, по вертикали — покраснение (зв. вел.). При вычислении градиента  $k$  отброшена одна точка (синий цвет). Данные из каталога Hipparcos.

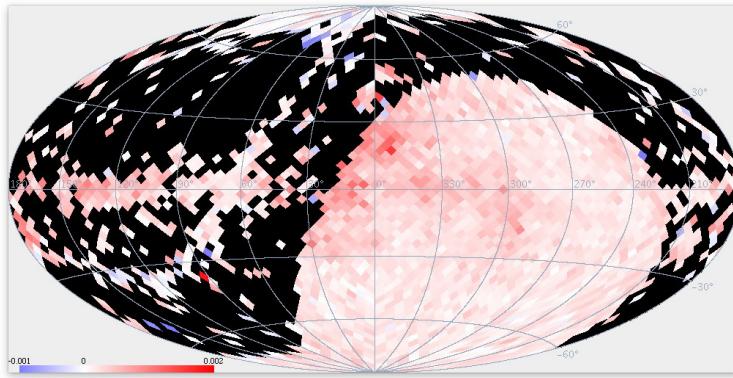


Рис. 7: Распределение градиента покраснения, построенного по звездам обучающего множества

## 6.2 Распределение коэффициента $k$ по небесной сфере

Тем самым, ход покраснения в конусе  $C_i$  мы описываем одним числом  $k_i$  — скоростью роста покраснения в этом конусе. Она, как мы ранее выясняли, должна коррелировать с наличием пыли. Поэтому, составив карту распределения коэффициента  $k$ , мы составим панораму пыли в окрестности Солнца.

На рисунке ?? показана карта распределения градиентов покраснения звезд для звезд обучающего множества. Черным цветом закрашены те площадки, для которых звезды обучающего множества не позволяют вычислить покраснение, так как для них не известны классы светимости. Число таких площадок равно ???.

Как было сказано выше, определение спектрального класса для остальных звезд, не входящих в обучающее множество, нами производилось с помощью двумерного линейного классификатора. Карты градиентов покраснения, построенные с помощью классификатора для звезд с  $B-V < 0.6$  и  $B-V > 0.6$  показаны на рисунках 10 и 11. Аналогичная карта, построенная по всем звездам, показана на рис.12.

Сравнение этих карт с рис.?? показывает, что примененный нами метод машинного обучения для сортировки звезд по III и V классам светимости позволил получить информацию о межзвездном покраснении практически для всей небесной сферы.

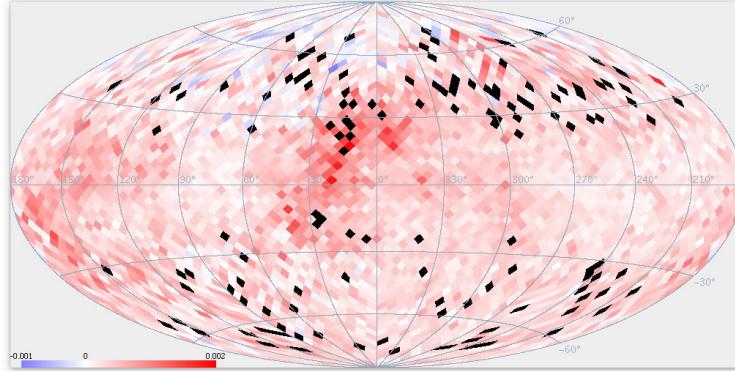


Рис. 8: Распределение градиента покраснения, построенного по звездам с  $B - V < 0.6$  (44890 звезд)

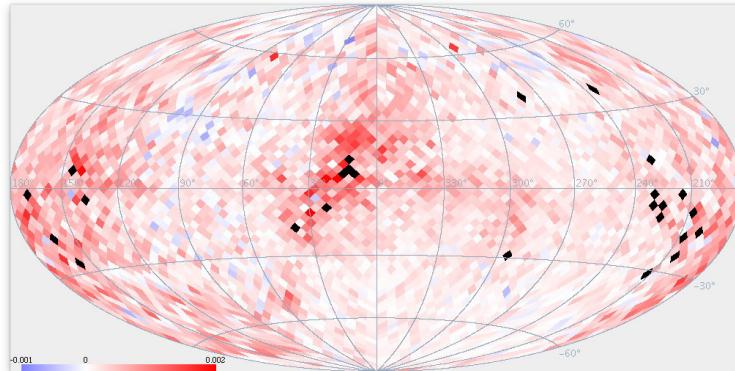


Рис. 9: Распределение градиента покраснения, построенного по звездам с  $B - V \geq 0.6$  (49309 звезд)

### 6.3 Статистическая надежность градиентов покраснения

На рисунке 12 показаны в условной шкале цветов значения градиентов покраснения для всех использованных нами 3888 площадок неба. Наибольшее значение найденных градиентов составляет 1.71 м/кпк, среднее значение - 0.15 м/кпк, среднее значение ошибки градиента - 0.04 м/кпк. В связи с этим интересно поставить вопрос о статистической достоверности значений градиентов, найденных в каждой площадке. Традиционно при сделанном предположении о нормальном распределении градиентов ответ на этот вопрос можно получить, вычислив отношение значения градиента к его среднеквадратичной ошибке. Известно, что при  $k/\sigma_k=3$  надежность результата составляет 99.87 процента (правило трех сигм), при  $k/\sigma_k=2$  надежность результата составляет 97.72 процентов. На рисунке 13 показаны черным цветом площадки, для которых  $k/\sigma_k < 2$ , то есть те площадки, на которых надежность определения градиентов покраснения не превышает 97.72 процентов.

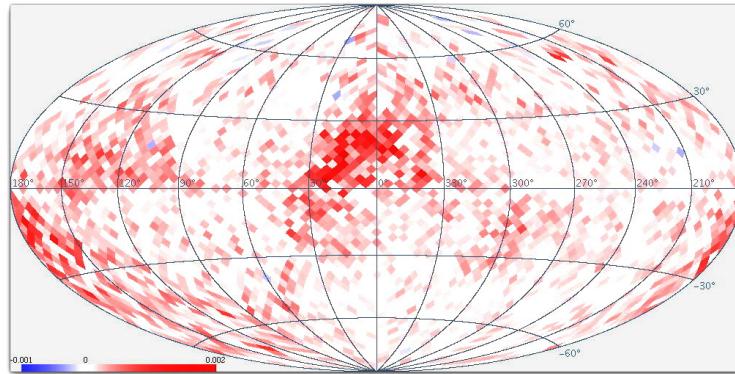


Рис. 10: Распределение градиента покраснения, построенного по всем звездам (94199 звезд)

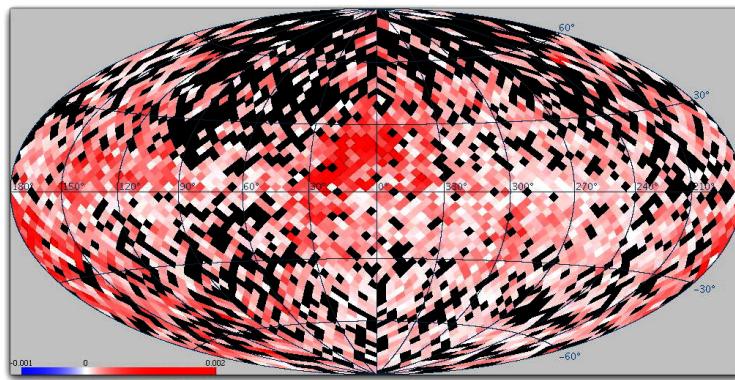


Рис. 11: Карта, на которой черным цветом отмечены пиксели, в которых  $k < 2\sigma_k$

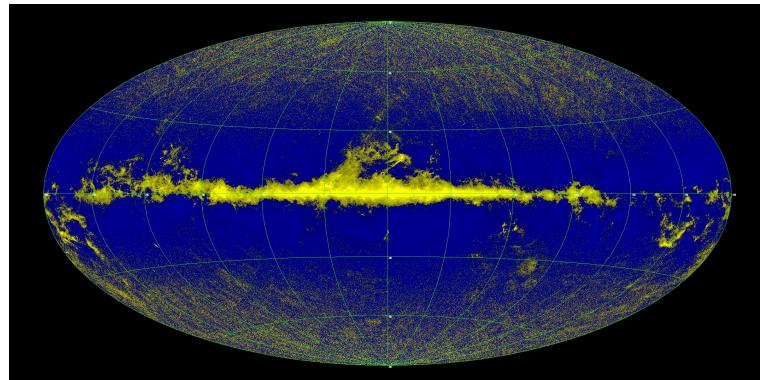


Рис. 12: Распределение показателя цвета J-H (2MASS)

#### 6.4 Сравнение с результатами, полученными с помощью 2MASS

На рисунках ??-?? показаны распределения по небесной сфере показателей цвета J-H и  $H-K_s$  звезд по данным каталога 2MASS [?]. На этих рисунках, полученных по показателям цвета в ближней ин-

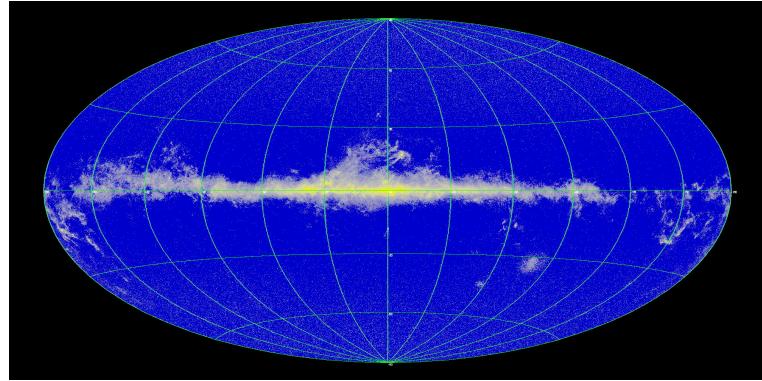


Рис. 13: Распределение показателя цвета H-K (2MASS)

фракрасной области ( $J - 1.2 \mu m$ ,  $H - 1.6 \mu m$  и  $K_s - 2.2 \mu m$ ), прослеживаются те же области, которые мы получили по покраснению  $E(B - V)$  в оптическом диапазоне. Строго говоря, наши результаты можно сравнивать только с показателями цветов, исправленными за стандартные значения, то есть с величинами

$$E = E_{J-H} = (J - H)_{obs} - (J - H)_{int}, \quad E = E_{H-K_s} = (H - K_s)_{obs} - (H - K_s)_{int}, \quad (5)$$

На рис. ?? показаны зависимости стандартных (intrinsic) показателей цвета  $B-V$ ,  $J-H$ , и  $H - K_s$  от спектрального класса [?], [?]. Мы видим, что в формулах (??) поправка за стандартное значение показателя цвета намного меньше, чем для  $B-V$ . Кроме того, эти поправки слабо зависят от как от спектра звезд, так и от спектрального класса, то есть практически одинаковы для всех звезд. Это обстоятельство (особенно ярко выражено для показателя цвета  $H-K$ ) позволяет считать, что карты цветов в ближней инфракрасной области рис.?? и рис.?? показывают детали, которые напрямую обуславливаются именно межзвездным поглощением.

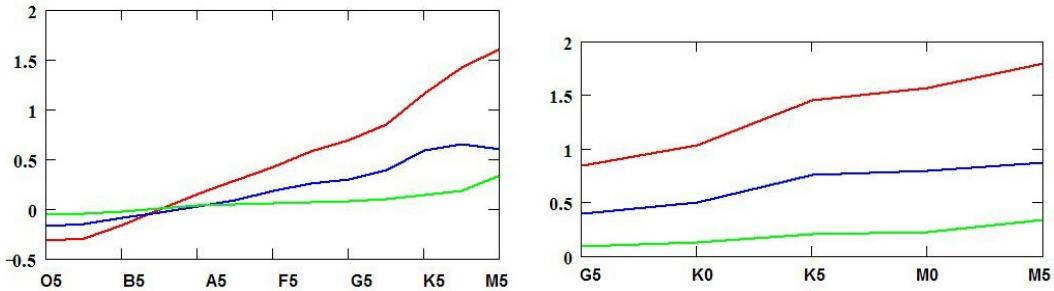


Рис. 14: Стандартные показатели цвета в зависимости от спектрального класса. Слева спектральный класс V, справа — III.  $B-V$  (красная линия),  $J-H$  (синяя линия);  $H - K_s$  (зеленая линия)

## 7 Отрицательное покраснение

На некоторых площадках коэффициент  $k$  отрицательный. В теории, такого не должно быть, т.к. межзвездное поглощение не может делать звезды более голубыми. Пример хода покраснения на одной из площадок. Мы видим, что это вызвано тем, что некоторые звезды имеют сильно отрицательное

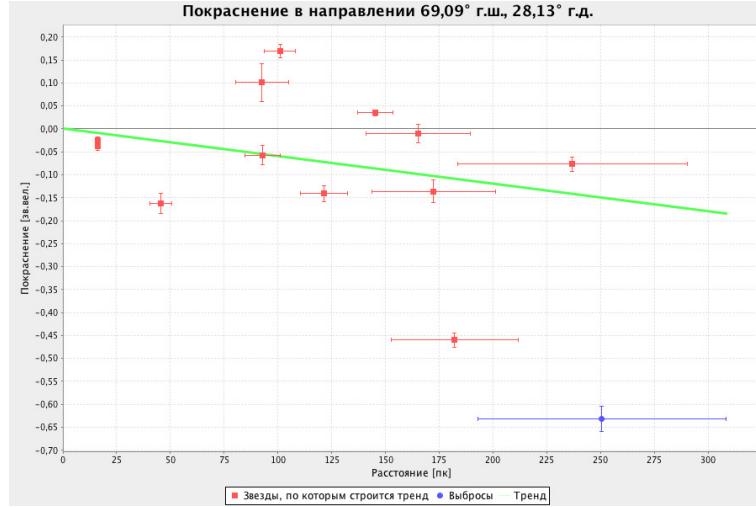


Рис. 15: Отрицательный ход покраснения

покраснение. То есть  $E_{B-V} + 3\sigma_{E_{B-V}} < 0$ , где  $\sigma$  — это ошибка. Такого не должно быть, т.к. звезда не может сильно голубеть на расстоянии. Давайте рассмотрим какую-нибудь звезду, имеющую такое аномальное отрицательное покраснение. К примеру, HIP 66713. Ее параметры:

- Параллакс  $7.99 \pm 0.77$
- Показатель цвета  $0.386 \pm 0.014$
- Спектральный тип G0V
- Видимая звездная величина 8.37

Мы видим, что она не имеет аномальных параметров. Так же, ее параметры имеют небольшие ошибки. Согласно таблице в [7], спектральному типу G0V соответствует показатель цвета 0.580. Тем самым, покраснение  $E = (B - V)_{obs} - (B - V)_{int} = (0.386 \pm 0.014) - 0.580 = -0.194 \pm 0.014$ . Мы видим аномальное покраснение у не аномальной звезды. И таких звезд (имеющих  $E_{B-V} + 3\sigma_{E_{B-V}} < 0$ ) 11993 из 94670 (12.6%), то есть достаточно много. Основные возможные причины таких отклонений — неверные таблицы «спектральный тип — покраснение», ошибки в спектральной классификации (см. обучающее множество в «Способах получения классов светимости»). Этой проблеме будет посвящена следующая статья.

## 8 Заключение

Перечислим основные результаты, полученные в работе.

- Создан бинарный классификатор, позволяющий звезде с известной абсолютной звездной величиной назначить класс светимости III или V. Для III класса светимости точность классификатора равна 95%, полнота — 89%. Для V класса светимости соответствующие характеристики равны 91% и 96%.

- Для 98827-39807 =59020 звезд каталога HIPPARCOS с помощью бинарного классификатора определены классы светимости.
- Для 98827 звезд каталога HIPPARCOS определены покраснения звезд по показателю цвета B-V.
- Получена карта значений градиента покраснения в направлениях, определяемых центрами 3888 равновеликих площадок, построенных методом HealPix.
- Определена статистическая надежность результатов для каждой площадки.
- Произведено сравнение карт покраснения, полученных нами в оптическом диапазоне, с аналогичными результатами, полученными в ближней инфракрасной области по данным каталога 2MASS. Практическое совпадение этих карт свидетельствует о надежной работе использованного нами бинарного классификатора.

## 9 Приложение

### 9.1 Классификация данных методом опорных векторов

Задача классификации состоит в том, чтобы определить, к какому классу относится данный объект на основе обучающей выборки — других объектов, про которые заранее известно, к каким классам они принадлежат. Каждый объект описывается числовыми атрибутами, поэтому задача классификации объектов сводится к задаче классификации точек в  $\mathbb{R}^n$ . Если классов всего два, то задача называется бинарной классификацией, если несколько — мультиклассификацией.

Задачу классификации на  $m$  классов можно сформулировать следующим образом, пусть есть обучающая выборка  $(x_i, y_i)$ ,  $x_i \in \mathbb{R}^n$ ,  $y_i \in \{1, \dots, m\}$ . Требуется на основе обучающей выборки построить решающую функцию  $F: \mathbb{R}^n \rightarrow \{1, \dots, m\}$ , сопоставляющую класс любой точке из  $\mathbb{R}^n$ . В случае бинарной классификации, классы будет удобно обозначить за  $\{-1, 1\}$ .

Бинарный классификатор назовем *линейным*, если его решающая функция выглядит следующим образом,

$$F(x) = \text{sign}(w \cdot x + b) = \text{sign} \left( \sum_{i=1}^n w_i x_i + b \right),$$

где  $x_i$  — компоненты вектора  $x$ ,  $b$  — параметр,  $w$  — вектор, соответствующий нормали к *разделяющей гиперплоскости* (ее уравнение  $w \cdot x + b = 0$ ). Классы  $\{-1, 1\}$  назовем линейно разделимыми, если существует гиперплоскость, разделяющая точки разных классов по разным полупространствам относительно этой гиперплоскости.

В простейшем случае, метод опорных векторов — это алгоритм обучения линейного бинарного классификатора методом *максимального зазора*. Если классы  $\{-1, 1\}$  линейно разделимы, это есть алгоритм нахождения гиперплоскости, которая наилучшим образом разделяет классы. То есть гиперплоскости  $(w, b)$ , которая находится на максимальном расстоянии до ближайшей точки «-1» класса и ближайшей точки «1» класса. Можно показать, что такая гиперплоскость будет иметь эти расстояния равными  $\frac{1}{\|w\|}$  — величине *зазора* (рис 13). Зазор должен быть максимальным, поэтому, для нахождения такой гиперплоскости требуется максимизировать  $\frac{1}{\|w\|}$ , то есть минимизировать  $\|w\|^2$ . Тем самым, метод опорных векторов сводится к решению следующей задачи,

$$\begin{cases} \|w\|^2 \rightarrow \min \\ \text{sign}(w \cdot x_i + b) = y_i \end{cases}$$

где второе условие соответствует линейной разделимости обучающей выборки. Перепишем,

$$\begin{cases} \|w\|^2 \rightarrow \min \\ y_i(w \cdot x_i + b) \geq 1 \end{cases}$$

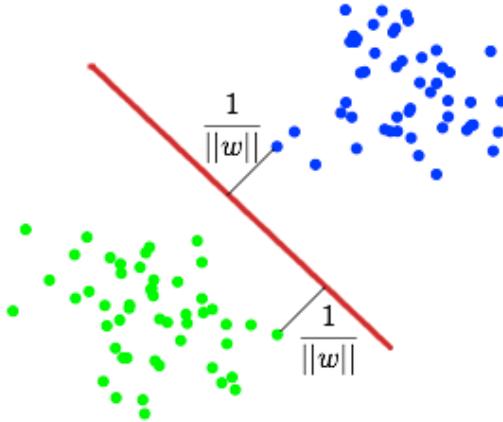


Рис. 16: Решение задачи бинарной классификации в  $\mathbb{R}^2$  методом опорных векторов

что есть задача квадратичного программирования, которая решается с помощью множителей Лагранжа.

Для того, чтобы алгоритм мог работать в случае, если классы линейно неразделимы, позволим ему допускать ошибки на обучающей выборке. Введем набор дополнительных переменных  $\xi_i \geq 0$ , характеризующих величину ошибки на точках  $x_i$ . Смягчим ограничения в неравенствах и введем в минимизируемый функционал штраф за суммарную ошибку,

$$\begin{cases} \frac{1}{2} \|w\|^2 + C \sum_i \xi_i \rightarrow \min \\ y_i(w \cdot x_i + b) \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases}$$

Где  $C$  — параметр настройки метода. Опять же, получилась задача квадратичного программирования. Эффективно такая задача может решаться такая с помощью метода «Sequential Minimal Optimization» (SMO) [9].

## 9.2 Кросс-валидация, оценка качества работы классификатора

Для получения несмешенной оценки качества работы аналитической модели проводится *кросс-валидация*. Она заключается в том, что обучающая выборка дизьюнктно делится на две части — тренировочное множество и тестовое множество, затем модель обучается на тренировочном множестве, а оценка качества работы получается на тестовом. Одним из способов проведения кросс-валидации является  $k$ -fold кросс-валидация. В этом случае обучающая выборка разбивается на  $k$  частей, затем на  $k-1$  части выборки производится обучение модели, а оставшаяся часть используется для тестирования. Процедура повторяется  $k$  раз, в итоге каждая из  $k$  частей используется в качестве тестовой. В результате получается оценка эффективности выбранной модели с наиболее равномерным использованием имеющихся данных.

Для задачи классификации на  $m$  классов, результатом кросс-валидации является таблица  $m \times m$ , в ячейке  $i, j$  которой записано число объектов класса  $i$  из тестового множества, про которые классификатор считает, что они принадлежат классу  $j$ . В случае  $k$ -fold кросс-валидации итоговая таблица

$m \times m$  есть сумма всех  $k$  таблиц, полученных из обучения + тестирования классификатора на каждом из  $k$  разбиений на тренировочное и тестовое множество.

По полученной таблице  $C[m, m]$  можно считать различные метрики — числа, характеризующие качество классификации,

- *Точность* определения класса  $i$  —  $\frac{C_{ii}}{\sum\limits_{j=1}^m C_{ji}}$
- *Полнота* определения класса  $i$  —  $\frac{C_{ii}}{\sum\limits_{j=1}^m C_{ij}}$
- *F<sub>1</sub>-мера* определения класса  $i$  - среднее гармоническое точности и полноты определения класса  $i$

## Список литературы

- [1] Perryman M.A.C., Lindegren L., Kovalevsky J., Hog E., Bastian U., Bernacca P.L., Creze M., Donati F., Grenon M., Grewing M., van Leeuwen F., van der Marel H., Mignard F., Murray C.A., Le Poole R.S., Schrijver H., Turon C., Arenou F., Froeschle M., Petersen C.S., "The Hipparcos Catalogue"(1997A&A...323L..49P)
  - [2] Hog E., Baessgen G., Bastian U., Egret D., Fabricius C., Grossmann V., Halbwachs J.L., Makarov V.V., Perryman M.A.C., Schwerkendiek P., Wagner K., Wicenec A., "The Tycho Catalogue"(1997A&A...323L..57H)
  - [3] van Leeuwen F., Evans D.W., Grenon M., Grossmann V., Mignard F., Perryman M.A.C., "The Hipparcos mission: photometric data."(1997A&A...323L..61V)
  - [4] Hipparcos, the New Reduction of the Raw Data van Leeuwen F., Astron. Astrophys. 474, 653 (2007),
  - [5] Wright et al. *Tycho-2 Spectral Type Catalog* 2003: The Astronomical Journal
  - [6] Бинни, Меррифилд (J. Binney and M. Merrifield), Galactic Astronomy (Princeton: Princeton Univ. Press, 1998).
  - [7] Скрутски и др., (Skrutskie, M. F., and 30 colleagues), The Two Micron All Sky Survey (2MASS). The Astronomical Journal 131, 1163-1183, (2006).
  - [8] Страйджест – книжка
  - [9] Straizys and Romualda Lazauskaite. INTRINSIC COLOR INDICES AND LUMINOSITY SEQUENCES OF STARS IN THE 2MASS TWO-COLOR DIAGRAM V., Baltic Astronomy, vol. 18, 19–31, 2009.
  - [10] А.А.Сминов, А.С.Цветков, А.В.Попов *Неточности в спектральной классификации звезд каталога Tycho-2 Spectral Type* 2006.
  - [11] V.N.Vapnik *The Nature of Statistical Learning Theory* 1995.
  - [12] Hastie, T.; Tibshirani, R. & Friedman, J. (2001), *The Elements of Statistical Learning* , Springer New York Inc. , New York, NY, USA .
  - [13] Gorski et al. *HEALPix: A Framework for High-Resolution Discretization and Fast Analysis of Data Distributed on the Sphere* 2005.
- intrinsic 1) внутренний  
 2) действительный 3) природный 4) естественный 5) присущий 6) собственный 7) характеристический