# Определение параметров межзвездного поглощения света по данным каталога Hipparcos

Ф.Амосов

9 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

Для определения параметров межзвездного поглощения, нам потребуется иметь следующие данные о звездах,

- положение
- параллакс
- фотометрия
- спектральный класс и класс светимости

Мы будем использовать модифицированный каталог Hipparcos [1], ввиду того, что у подавляющего большинства звезд присутствуют все необходимые данные.

## 2.1 Сведения о каталоге

Каталог Hipparcos содержит данные о 117955 звездах. У некоторых звезд отсутствуют необходимые нам данные, поэтому в данной статье мы будем работать с меньшим числом звезд.

### 2.1.1 Параллакс

Звезды в каталоге Hipparcos имеют низкую точность параллакса. Это можно видеть из следующей гистограммы. На ней отложены средние значения относительной ошибки в расстоянии для звезд на разных расстояниях. Мы видим что, например, на расстоянии в 500 пк ошибка в параллаксе достигает

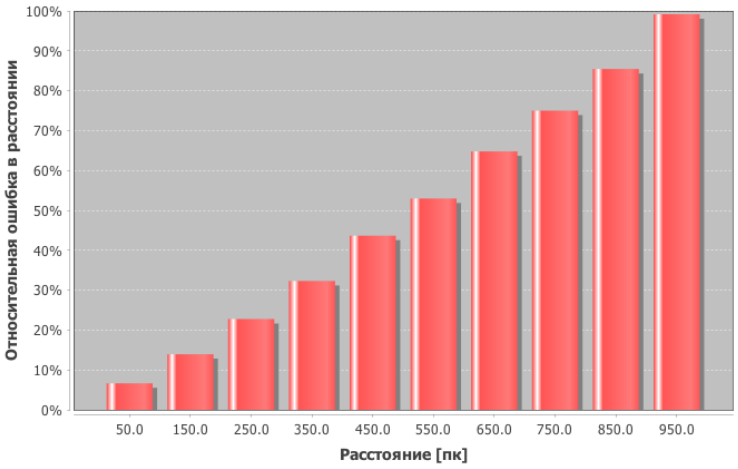


Рис. 1: Относительная ошибка в расстоянии от расстояния

50%.

### 2.1.2 Положение

Ошибки в положении звезд на небе на порядок меньше ошибок в параллаксе, поэтому их можно не учитывать, если нас интересует пространственное распределение звезд.

### 2.1.3 Фотометрия

В каталоге Hipparcos 2007 [1] приведены данные фотометрии звезд. Показатель цвета $B - V$ присутствует у всех звезд.

## 2.2 Спектральная классификация

В каталоге Hipparcos имеются данные о спектрах и спектральных данных для 110707 звезды. Эти данные заимствованы из наземных наблюдений. Дополнительная информация о спектральных классах содержится в каталоге Tycho 2 Spectral Type. К сожалению, в этом каталоге информация о спектральных данных звезд приведена только для звезд южного экваториального полушария, На этом рисунке изображена небесная сфера в галактической системе координат (в центре центр галактики), разбитая на «пиксели» алгоритмом HEALPix [7]. Далее мы подробнее его разберем. Чем больше в пикселе доля звезд, имеющих класс светимости в каталоге Tycho 2 Spectral Type, тем он краснее. Видим отсутствие класса светимости практически у всех звезд северного экваториального полушария и присутствие у южного.

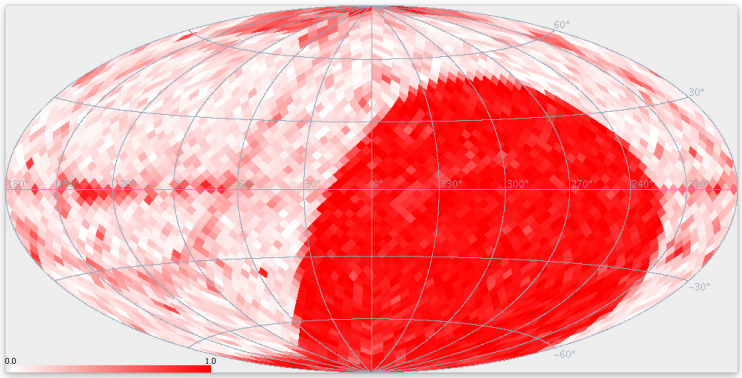

Рис. 2: Распределение наличия классов светимости

# 3 Покраснение

## 3.1 Определение

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

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

где $(B-V)_{obs}$ — ее видимый нами показатель цвета, а $(B-V)_{int}$ — теоретический показатель цвета звезды. Значение $(B-V)_{obs}$ мы можем получить на основе данных фотометрии звезды из каталога. Величины $(B-V)_{int}$ оцениваются статистически. В нашей работе мы воспользуемся приведенной в [4] двумерной таблицей «спектральный класс, класс светимости — показатель цвета». То есть, для получения $(B-V)_{int}$ звезды нам потребуются ее спектральный класс и класс светимости. Их мы можем получить из данных каталога. Приведем пример расчета покраснения на звезде HIP 44800,

- У нее в каталоге $(B-V)_{obs} = 0.535^m$
- Класс F7V, поэтому (по [4]) $(B-V)_{int} = 0.493^m$
- Покраснение $0.535^m - 0.493^m = 0.042^m$

========== ВОПРОС: ОТКУДА ВЗЯТЫ ЭТИ ДАННЫЕ? ========== *ИЗ HIPPARCOS, СПЕКТРЫ ИЗ, ВИДИМО, TYCHO 2 SPECTRAL TYPES*

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

# 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 полушария.

Исправим это. Определим неизвестные классы светимости. Сделаем это с помощью методов машинного обучения. Натренируем классификатор, который будет выдавать класс светимости для звезды по

двум факторам - ее показателю цвета и ее абсолютной звездной величине. Этих факторов будет достаточно, т.к. классы светимости теоретически разделимы на диаграмме Гершпрунга-Рассела.

Доля звезд, которые относятся ни к III, ни к V классам мала (16,3%, 8058 из 49285, имеющих класс светимости). Поэтому, мы упростим задачу — обучим линейный бинарный классификатор, который будет предсказывать III или V класс. Сделаем это с помощью метода опорных векторов [6] с обычным линейным ядром.

В качестве обучающего множества возьмем все звезды, у которых присутствует класс светимости в каталоге (КАКОМ?), и он либо III, либо V. Таких звезд 41227 (20027 III класса, 21200 V класса). ==========ПРИВЕДИТЕ ЧИСЛО ЗВЕЗД КАЖДОГО КЛАССА ДО И ПОСЛЕ B-V=0.6 *ЧИСЛО ЗВЕЗД, У КОТОРЫХ ПОКАЗАТЕЛЬ ЦВЕТА МЕНЕЕ 0.6 РАВНО 18016 — 2455 III КЛАССА, 15561 V КЛАССА*

*ЧИСЛО ЗВЕЗД, У КОТОРЫХ ПОКАЗАТЕЛЬ ЦВЕТА НЕ МЕНЕЕ 0.6 РАВНО 23211 — 17572 III КЛАССА, 5639 V КЛАССА*

Обучим на них классификатор — получим разделяющую классы прямую (13).

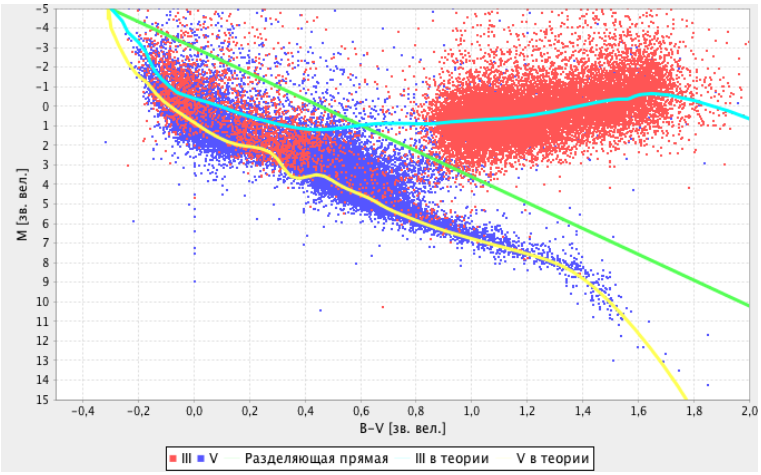


Рис. 3: Обучающее множество с полученной разделяющей прямой. Также здесь изображены теоретические кривые для классов на основе таблиц [4]

Уравнение разделяющей прямой,

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

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 $B-V < 0.6$ он всем звездам присваивает V класс при большом числе звезд III класса в обучающем множестве (12,7%, 2221 из 17499). 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

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

.

В этой формуле веса $w_1$ и $w_2$, $(w_1 + w_2 = 1)$ логично взять в соответствии с долей классов в этой области (12.7%/87.3%).

В таблице двумерной спектральной классификации [4] 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

$$(B-V)_{int} = 12.7\% \cdot 0.733 + 87.3\% \cdot 0.630 = 0.643$$

.

Как видим, даже в наихудшем случае отличие от V класса минимальное — 0.01 зв. вел. — гораздо ниже уровня каталожных ошибок. Поэтому в дальнейшем мы всегда будем использовать решение классификатора в этой области по V классу.

ЗАМЕЧАНИЕ. Эти рассуждения можно применять только в процессе кросс-валидации, когда известны пропорции звезд III класса V класса в выборках, то есть когда можно назначить веса. При реальной работе классификатора мы не знаем пропорций звезд в тестируемом множестве. Поэтому реальное покраснение $E$ испытуемой звезды лежит в пределах

$$E_V < E < E_{III},$$

где границы диапазона определяются по формуле (1). Так как у нас нет никакой инфомации о преобладающей близости величины $E$ к одной из границ диапазона, то естественно принять в качестве оценки искомого покраснения середину промежутка

$$E = \frac{1}{2}(E_V + E_{III}), (*)$$

с очевидной оценкой погрешности такого приближения

$$|\Delta E| \leq \frac{1}{2}((B-V)_{int}(III) - (B-V)_{int}(V))$$

.

Для спектрального класса G2 теперь имеем оценку погрешности покраснения, которую дает классификатор.

$$|\Delta E| \leq 0.052$$

Из рисунка 3 видно, что обе границы для покраснения можно получить только до $B - B < 0.4$. Поэтому оценку покраснения следует вести следующим образом:

1. До $B - V < 0.4$ - по формуле (*)
2. При $0.4 < B - V < 0.6$ только по V классу
3. При $B - V > 0.6$ по III или V классу по указаниям классификатора.

ЭТУ ОЦЕНКУ НАДО СРАВНИТЬ С ТОЧНОСТЬЮ ФОТОМЕТРИИ!!!

*Тем не менее так делать неправильно. Мы обучаем классификатор именно для нашего датасета, а в нем априорная вероятность появления звезды III класса в этой области равна 12.7%. На таком подходе и строится все машинное обучение. Мы не можем обучить классификатор, который будет работать в общем случае. Обученная модель работает только в той области, где наше обучающее множество является репрезентативной выборкой. Поэтому мы не имеем права ставить классам веса 50/50 — получится классификатор не для нашей задачи, не для нашего датасета, а для чего-то странного*

Для количественной оценки качества работы классификатора, проведем 10-fold кросс-валидацию (ОБЪЯСНИТЬ!) *В ПРИЛОЖЕНИИ ЭТО ДЕЛАЕТСЯ.*

Ниже приведены ее результаты,

| Классифицированы как → | III | V |
|---|---|---|
| III | 17529 | 2498 |
| V | 854 | 20346 |

| Класс | Точность | Полнота | F1-мера |
|---|---|---|---|
| III | 95% | 88% | 91% |
| V | 89% | 96% | 92% |

====== ОБЪЯСНИТЬ СМЫСЛ ЭТИХ ПОНЯТИЙ: Точность , Полнота, F1-мера======
======ПЕРЕПУТАНЫ точность и полнота!!!==========

Точность(III)=17529/(17529+2498)= 0.875 *НЕТ, Точность(III) = 17529 / (17529 + 854). ПО ССЫЛКЕ http://bazhenov.me/blog/2012/07/21/classification-performance-evaluation.html ТАБЛИЦА ПОВЕРНУТА ОТНОСИТЕЛЬНО МОЕЙ*

Полнота(III)=17529/(17529+854)= 0.954

Точность(V)=20346/(20346+854)= 0.960

Полнота(V)=20346/(20346+2498)= 0.891

Классификатор имеет приемлемое качество. Результат его работы — наличие класса светимости у всех рассматриваемых звезд.

# 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 следующий вид,

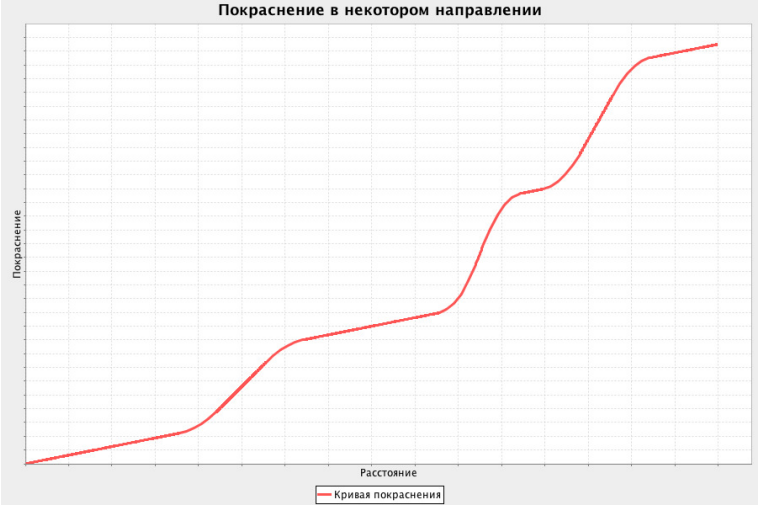


Рис. 4: Идеальный ход покраснения на луче зрения

Здесь по оси $x$ отложено расстояние по лучу зрения, а по $y$ — покраснение у соответствующих звезд. 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 было бы сказать, что облака на этом луче зрения находятся там, где покраснение растет «очень быстро» (синие области),

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

Существуют таблицы «спектральный класс, класс светимости — показатель цвета», поэтому для каждой звезды из каталога Hipparcos, у которой есть видимый показатель цвета и спектральный класс, можно вычислить покраснение. Если мы для всех таких звезд знаем еще и их пространственные координаты с хорошей точностью, то мы можем говорить о пространственном распределении покраснения. В данной работе мы рассмотрим это распределение с точки зрения трендов покраснения в различных направлениях. Результатом работы будут коэффициенты $a$ и $b$ этих трендов $ar + b$.

Построение кривых покраснения в разных направлениях позволит нам понять пространственное распределение межзвездного поглощения (пыли). Звезд в каталоге не бесконечное число, поэтому ре-

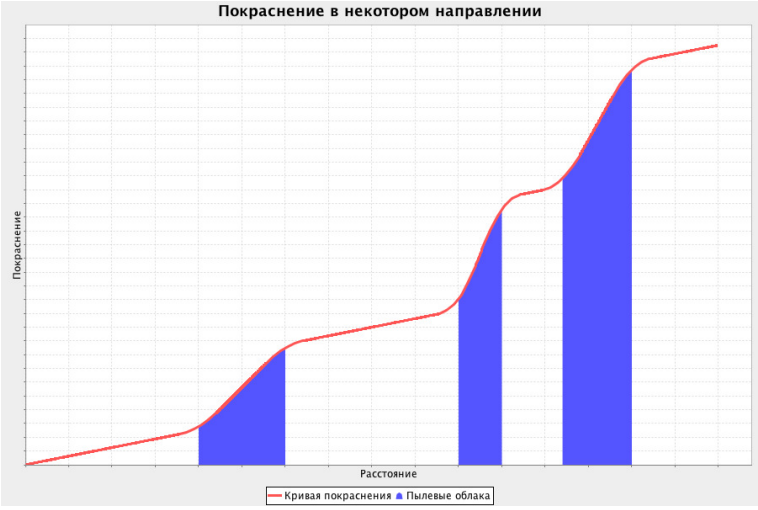


Рис. 5: Облака на идеальной кривой покраснения

альные 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,

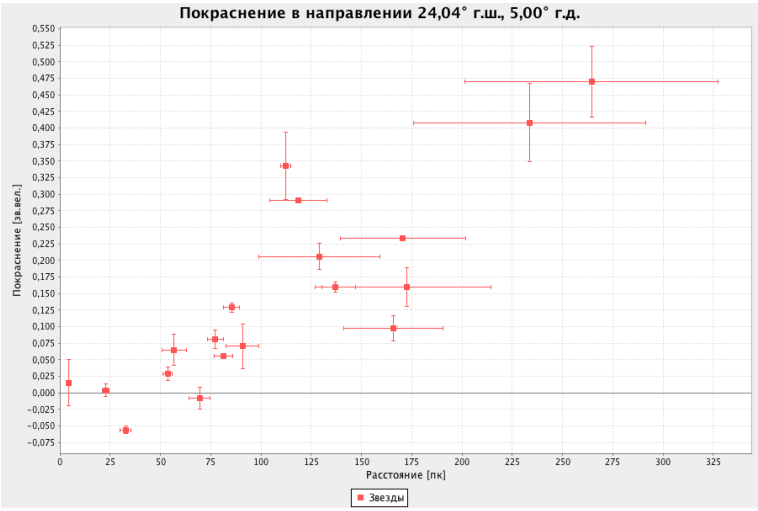


Рис. 6: Покраснение из данных каталога Hipparcos

Следующий метод позволит построить «кривые» покраснения во всех направлениях на небесной сфере. Он состоит из трех этапов,

## 5.2 Картирование небесной сферы

Для картирования небесной сферы мы воспользуемся стандартным алгоритмом Healpix [7]. В этой схеме ключевым параметром (resolution parameter), определяющим разбиение сферы на равные площадки, является число $N_{pix}$. Общее число пикселей $N = 12N_{pix}^2$. Двумя параллелями со склонениям

$\pm \arcsin(2/3)$ вся сфера разбивается на три части - экваториальную и две полярные. В полярных зонах выбирается по $N_{pix} - 1$ параллелей, в экваториальной зоне число параллелей равно $(2N_{pix} + 1)$. На каждой параллели экваториальной области находятся центры $4N_{pix}$ 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 $j = 0, ..., N - 1$ идет по параллелям с севера на юг. Будем обозначать площадки $\{P_i\}_{i=1}^{12N_{pix}^2}$ (у нас $N_{pix}^2 = 18$). Обозначим конусы, высекаемые соответствующими частями через $\{C_i\}_{i=1}^{12N_{pix}^2}$
Такое разбиение позволит нам,

1. Рассмотреть ход покраснения в каждом конусе как одномерную функцию $E(r)$. 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.

## 5.3 Тренд

Тем самым, мы ищем $E_i(r)$, соответствующую каждому $C_i$. Ввиду того, что практически все $E_i(r)$ очень сильно зашумлены разного рода ошибками, не удается проследить истинный ход этих функций. Но тренд вида $kr$ все же можно вычислить $(E_i(r) \approx k_i r)$. Он находится с помощью метода наименьших квадратов.

## 5.4 Критерии выбора звезд

### 5.4.1 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.

### 5.4.2 По тренду

Покраснения некоторых звезд вносят большие ошибки в искомые тренды. Иногда они получаются вообще отрицательными, что противоречит здравому смыслу. Но об этом мы поговорим позже.

Как известно, метод наименьших квадратов не устойчив к выбросам, т.е. совсем неверные покраснение/параллакс могут очень сильно испортить тренд. Для более устойчивого построения тренда, мы сделаем следующее. После построения тренда по всем звездам в конусе, мы выбрасываем те, у которых отклонение от тренда самое большое. Затем, мы строим тренд заново, но уже только по оставшимся звездам. После выброса 10% самых плохих звезд тренд, к примеру, может быть таким

В данном случае, мы выбросили одну звезду (синюю). 10% звезд — это обычно 0, 1 или 2 звезды в каждом конусе. Опять же, нельзя выбрасывать слишком много звезд из-за опасности «подгонки» данных под модель.

## 5.5 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.

Это небо в галактической системе координат с центром в центре галактики. Сфера разбита на $12 \cdot 18^2 = 3888$ «пикселей» алгоритмом HEALPix. В каждом кусочке построен тренд покраснения $kr$. На

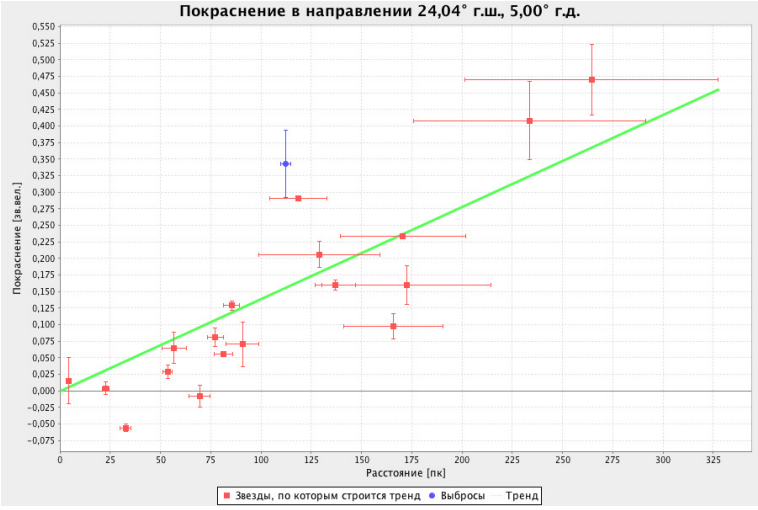


Рис. 7: Тренд покраснения, построенный без учета выбросов

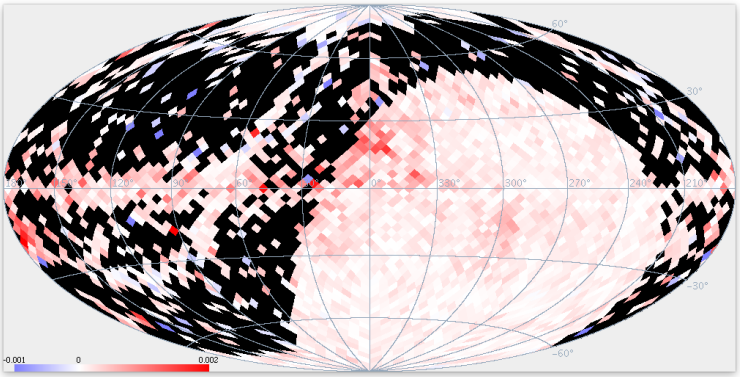

Рис. 8: Распределение градиента покраснения, построенного по звездам, имеющим класс светимости в каталоге Tycho 2 Spectral class. НУЖНО ПОСТРОИТЬ АНАЛОГИЧНУЮ КАРТУ ТОЛЬКО ПО ДАННЫМ КАТАЛОГА HIPPRCOS!!! *В HIPPARCOS НЕТ СПЕКТРАЛЬНЫХ ДАННЫХ - ЭТО НЕВОЗМОЖНО* Сравнение этих двух карт с рис.11 покажет основной результат работы ПРАВИЛЬНО ЛИ Я ПОНИМАЮ, ЧТО ЭТА КАРТА ПОСТРОЕНА ТОЛЬКО ПО ОБУЧАЮЩЕМУ МНОЖЕСТВУ?

*ДА*

этом рисунке изображено распределение значения коэффициента $k$ по пикселям. Синий цвет означает отрицательное значение $k$, красный - положительное. Чем насыщеннее цвет, тем больше значение коэффициента по модулю.

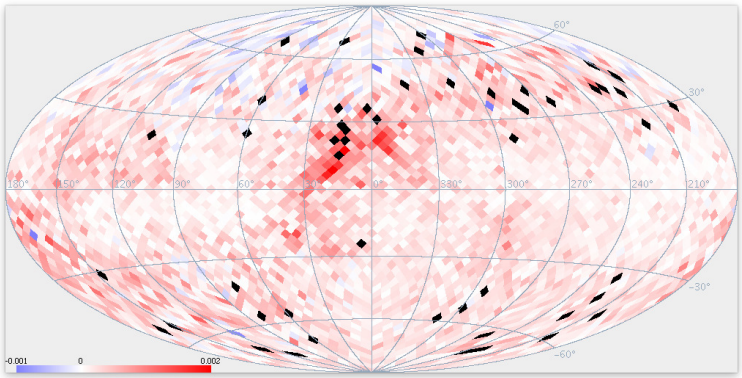

Рис. 9: Распределение градиента покраснения, построенного по звездам с $B - V < 0.6$. СКОЛЬКО ЗВЕЗД?

*44890*

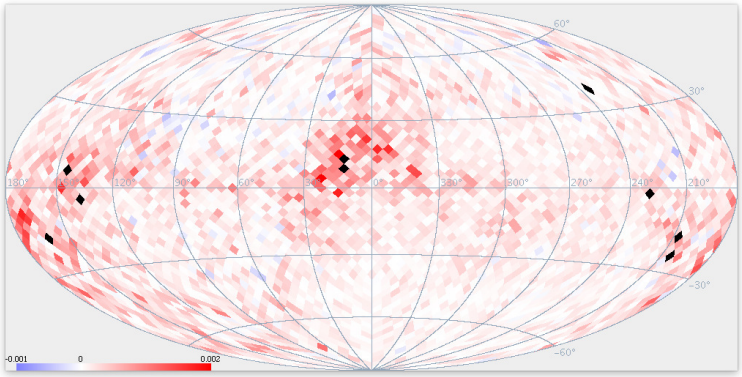

Рис. 10: Распределение градиента покраснения, построенного по звездам с $B - V \geq 0.6$. СКОЛЬКО ЗВЕЗД?

*49309*

## 6 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 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

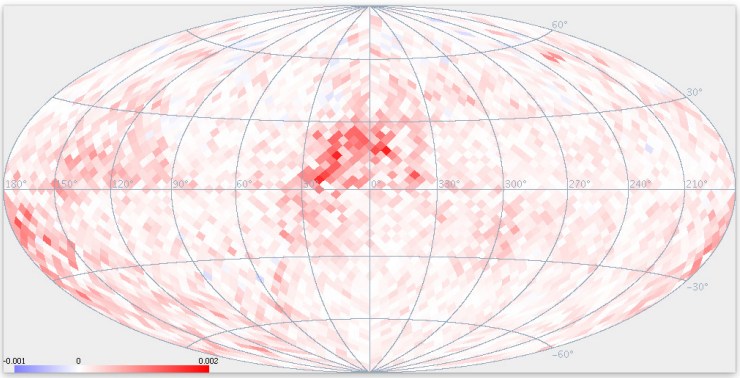


Рис. 11: Распределение градиента покраснения, построенного по всем звездам. СКОЛЬКО ЗВЕЗД? *94199*

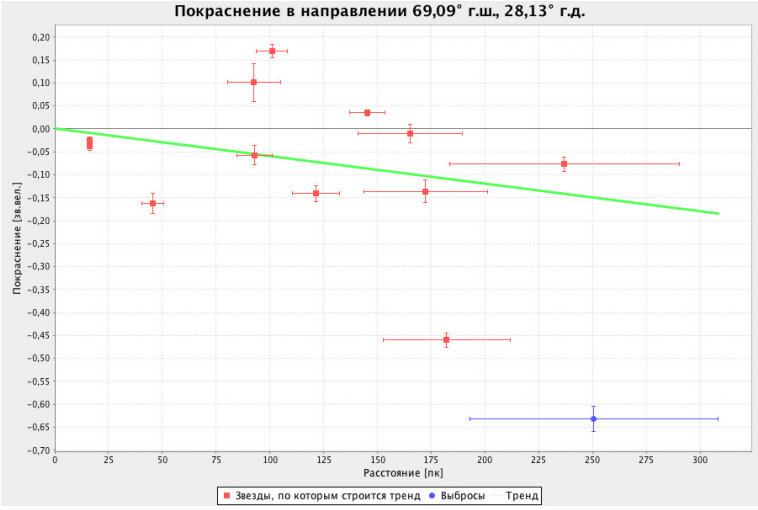


Рис. 12: Отрицательный ход покраснения

Мы видим, что она не имеет аномальных параметров. Так же, ее параметры имеют небольшие ошибки. Согласно таблице в [4], 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.

# 7 Приложение

## 7.1 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 $(x_i, y_i), x_i \in \mathbb{R}^n, y_i \in \{1, \ldots, m\}$. Требуется на основе обучающей выборки построить решающую функцию $F : \mathbb{R}^n \to \{1, \ldots, m\}$, 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 нахожде-

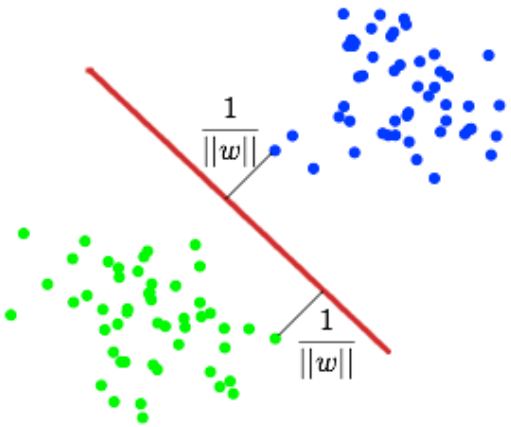


Рис. 13: Решение задачи бинарной классификации в $\mathbb{R}^2$ методом опорных векторов

ния такой гиперплоскости требуется максимизировать $\frac{1}{\|w\|}$, то есть минимизировать $\|w\|^2$. Тем самым,

метод опорных векторов сводится к решению следующей задачи,

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

где второе условие соответствует линейной разделимости обучающей выборки. Перепишем,

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

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,

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

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

## 7.2 Кросс-валидация, оценка качества работы классификатора

Для получения несмещенной оценки качества работы аналитической модели проводится *кросс-валидация*. Она заключается в том, что обучающая выборка дизъюнктно 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 разбиваются 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 $i$ — $\frac{C_{ii}}{\sum\limits_{j=1}^{m} C_{ij}}$
- $F_1$-*мера* определения класса $i$ - среднее гармоническое точности и полноты определения класса $i$

=======ПРИВЯЖИТЕ ЭТУ ТЕОРИЮ К ВАШИМ КОНКРЕТНЫМ ДАННЫМ ИЗ ТАБЛИЦЫ === *ТАК В РАЗДЕЛЕ «Способ получения классов светимости» ЭТО И ДЕЛАЕТСЯ*

# 8 Заключение

Этот раздел еще не создан

# Список литературы


[1] van Leeuwen, F. *Validation of the new Hipparcos reduction* 2007: Astronomy and Astrophysics

[2] Wright et al. *Tycho-2 Spectral Type Catalog* 2003: The Astronomical Journal

[3] Страйджест – книжка

[4] А.А.Сминов, А.С.Цветков, А.В.Попов *Неточности в спектальной классификации звезд каталога Tycho-2 Spectral Type* 2006.

[5] Гончаров

[6] V.N.Vapnik *The Nature of Statistical Learning Theory* 1995.

[7] Gorski et al. *HEALPix: A Framework for High-Resolution Discretization and Fast Analysis of Data Distributed on the Sphere* 2005.