Моделирование смесей распределений в задачах геологии и геофизики

Igor Magdeev

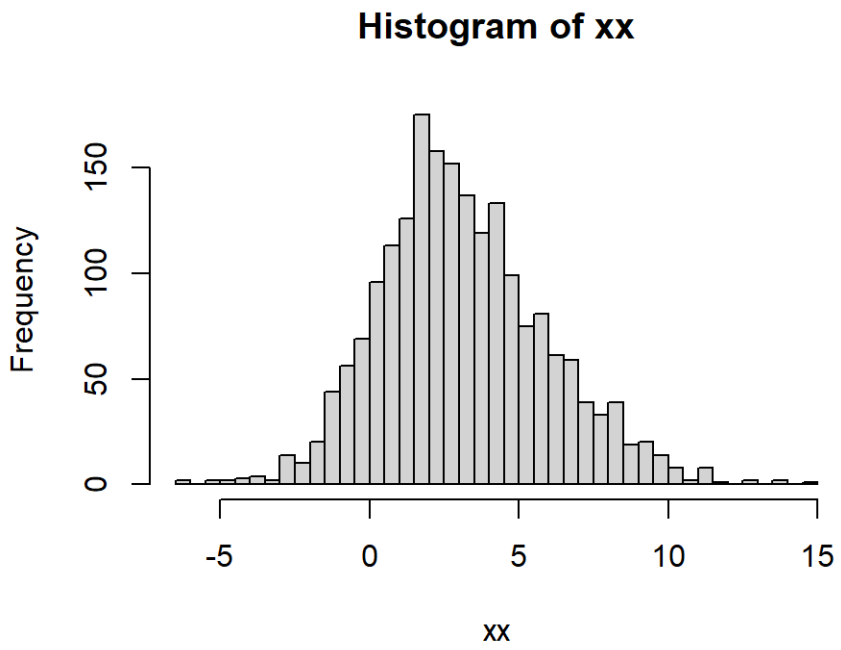
Февраль 03, 2022

Смеси распределений

- В каких задачах встречаются?
- Реальный пример задачи: Gaussian component analysis
- Инструмент моделирования - пакет `mclust`
- Синтетический пример (одномерный вариант)
- Пример 1: геофизические данные
- Пример 2: гранулометрический состав

Синтетический пример : Генерируем данные

```
set.seed(1248)
x1 <- rnorm(1000, 2, 2)
x2 <- rnorm(1000, 4, 3)
xx <- c(x1, x2)
hist(xx, breaks = 60)
```



Синтетический пример : Строим модель “по умолчанию”

```
mcl <- Mclust(data = xx, G = 1:6)
summary(mcl)
```

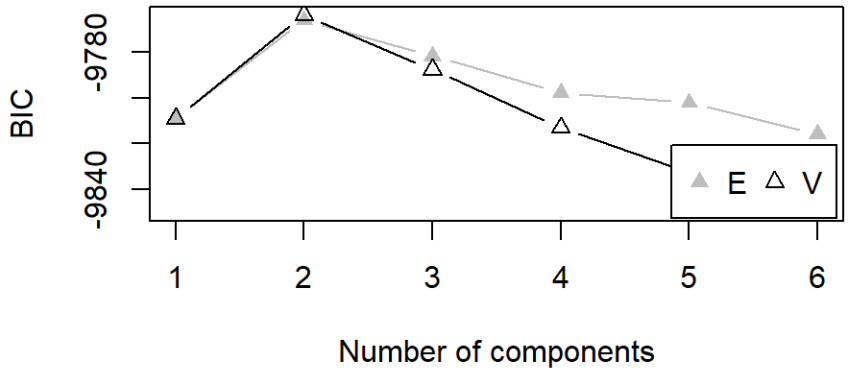
```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust V (univariate, unequal variance) model with 2 components:
##
##  log-likelihood    n df       BIC      ICL
##       -4862.754 2000  5 -9763.513 -11312.7
##
## Clustering table:
##    1    2
## 1253  747
```

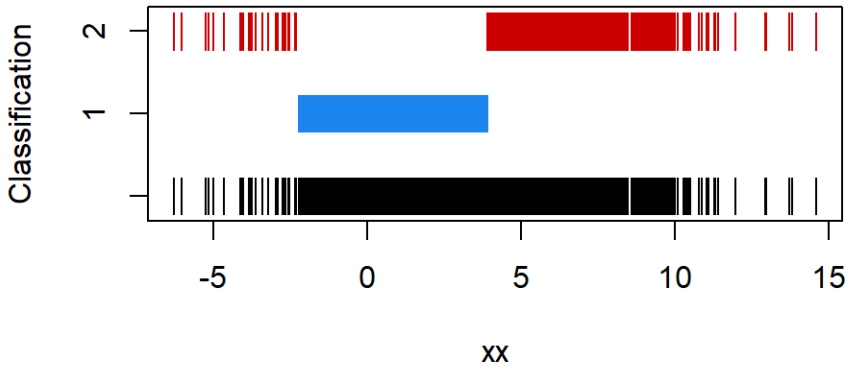
Синтетический пример : Параметры модели

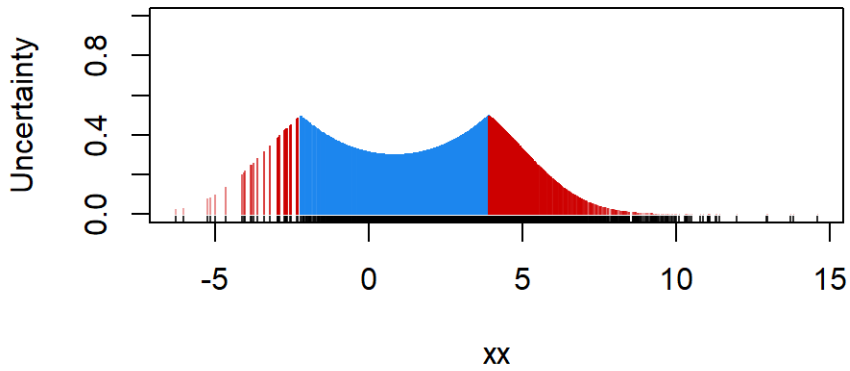
```
mcl$parameters
```

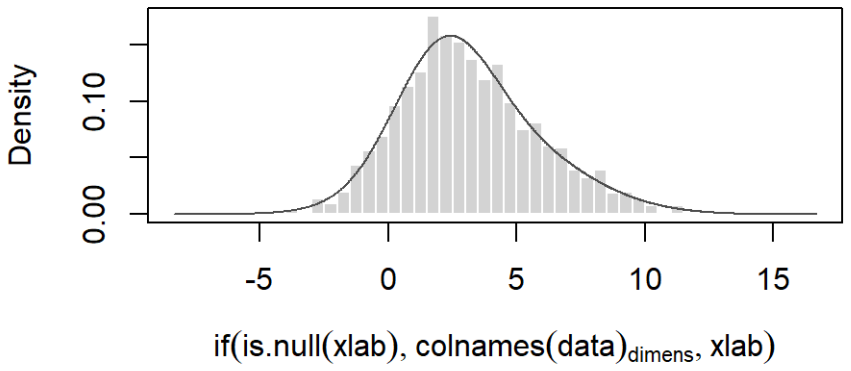
```
## $pro
## [1] 0.4911182 0.5088818
##
## $mean
##        1        2
## 2.061640 4.159457
##
## $variance
## $variance$modelName
## [1] "V"
##
## $variance$d
## [1] 1
##
## $variance$G
## [1] 2
##
## $variance$sigmasq
## [1] 3.554736 9.810583
##
## $variance$scale
## [1] 3.554736 9.810583
```

Синтетический пример : Как “читать графики”?





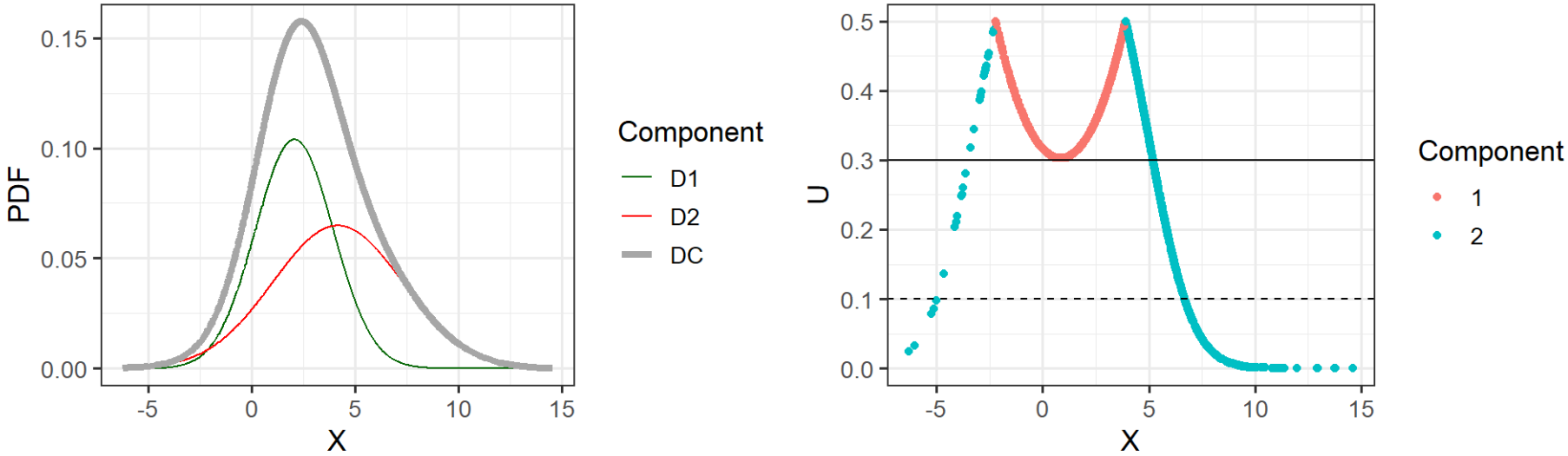




Синтетический пример : Теперь с ggplot2

Полезным может оказаться набор функций, которых нет в исходном пакете:

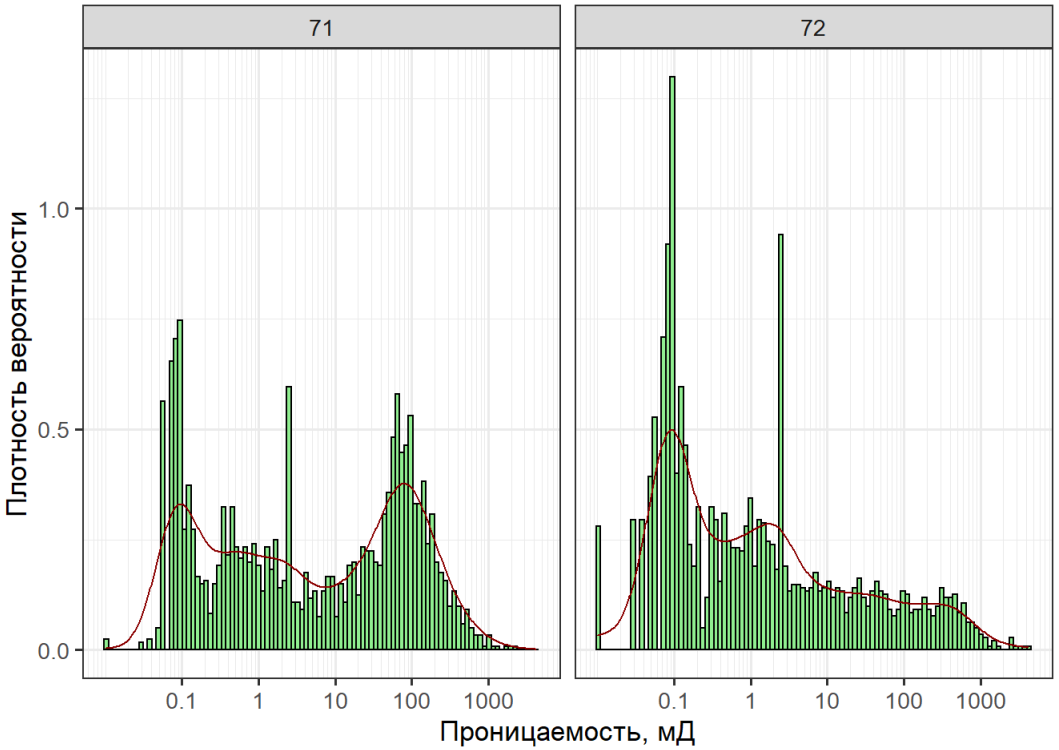
- воспроизведение плотности вероятности с отрисовкой отдельных компонент
- нахождение точек вдоль оси X, в которых неопределенность превышает заданный предел (-3.49, 5.2)



Реальный пример #1 : Геофизические данные

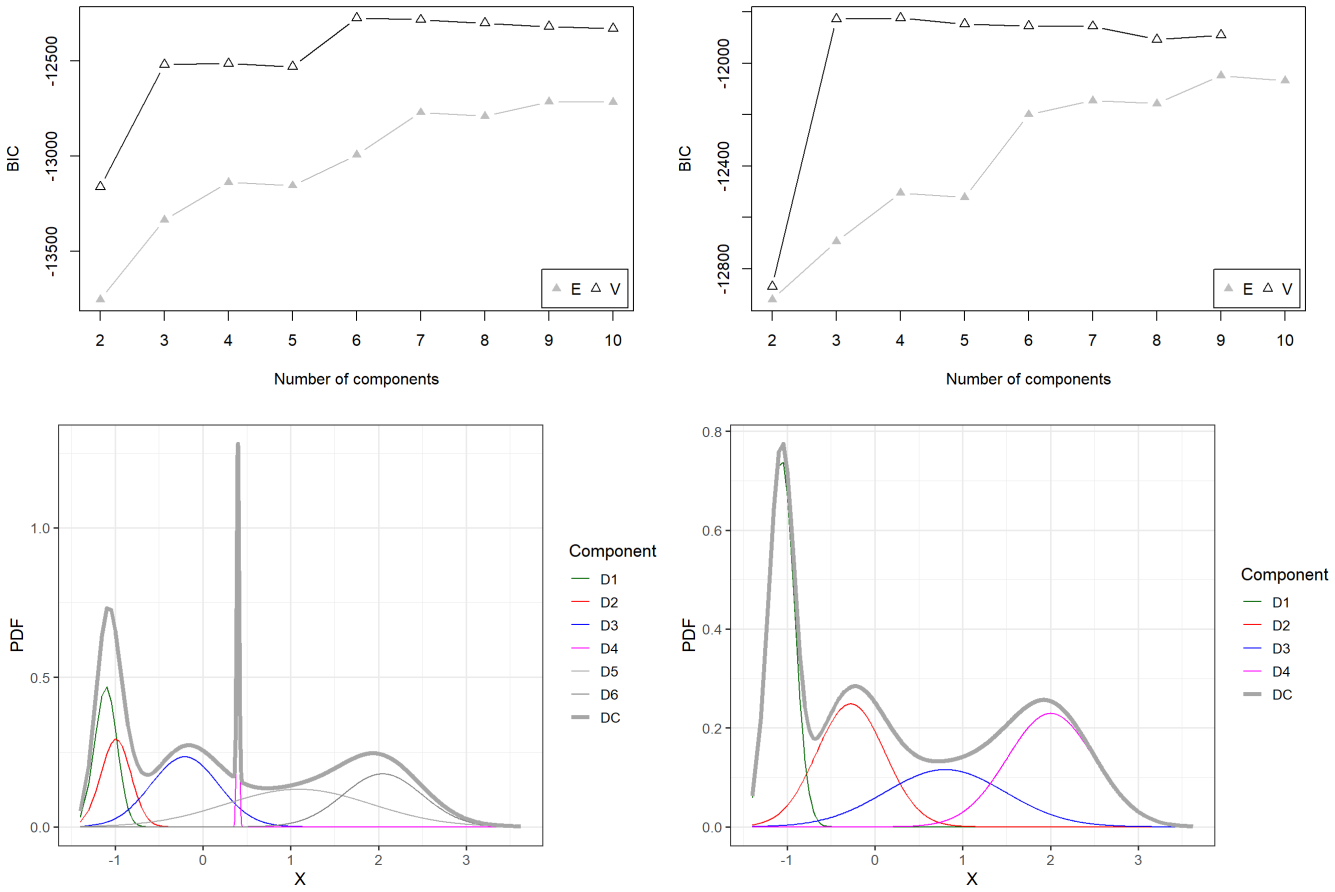
Предварительный анализ измерения профильной проницаемости на керне

(массовые “конвейерные” измерения прибором на выбуренной горной породе)



Проницаемость

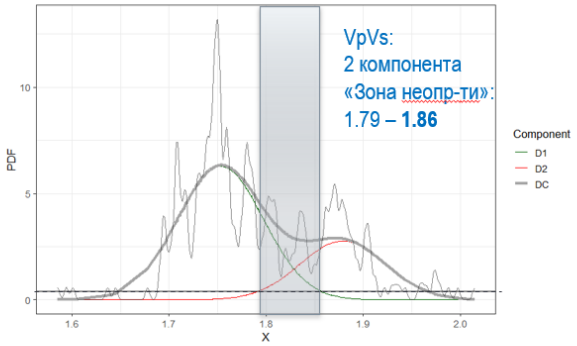
Реальный пример #1 : Модель `Mclust`

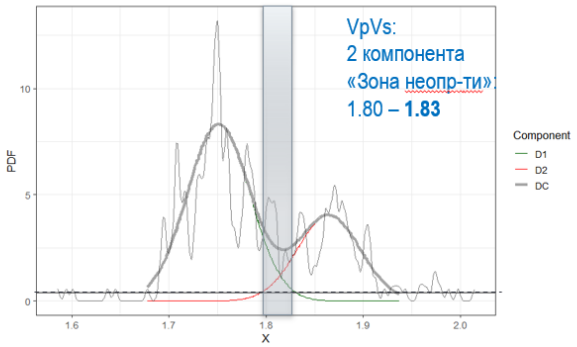


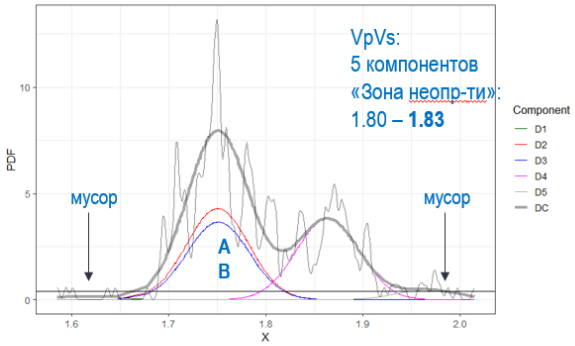
Слева - исходный вариант, справа - после удаления данных (4%) вокруг пика

Реальный пример #1 : Снова “работа с данными”

Зачем нужно фильтровать данные?







5 компонент «формально описывают» данные хорошо, но …
В действительности:
- Крайние компоненты это «мусор» (VpVs < 1.64 или VpVs > 1.94)
- Два сходных (красным A и синим B) должны объединяться

ИТОГО:
- Фильтр в интервале 1.64 < VpVs < 1.94)
- Два компонента, как и вначале
- Заметное сужение области неопределенности

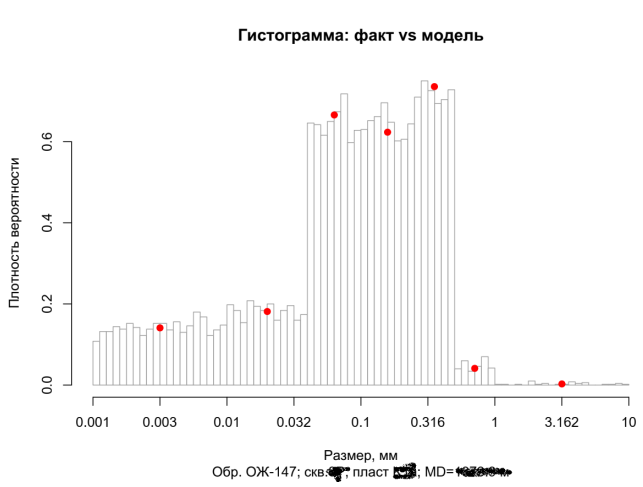
Реальный пример #2 : Гранулометрический состав

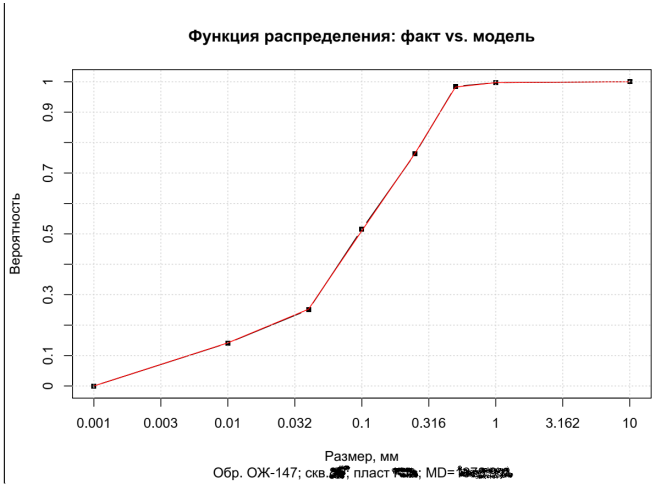
- **Гранулометрический состав горных пород - это:**
- Результат физических измерений (масса порошкообразной субстанции в определенном интервале размера сит; или коэф-ты поглошения света в определенном диапазоне при лазерно-спектрометрическом методе)
- На входе есть только “гистограмма” размеров зерен образца
- Выборку чисел для решения задачи потребуется генерировать
- **Что делать если данных много?** (работа с набором гистограмм-образцов - десятки, сотни)
- задача автоматизации выбора модели для каждого образца (“стратегия”)
- задача быстрой обработки всех образцов (параллельные вычисления `futures`)
- задача обработки ансамбля полученных моделей (расширение классификации)

Реальный пример #2 : Модельная выборка

По мотивам https://stats.stackexchange.com/questions/191725/sample-from-distribution-given-by-histogram

```
simhist <- function(x, y, size = 10000)
{
  npts <- length(x)
  mids <- (x[-npts] + x[-1])/2
  bins <- sample(length(mids), size, prob = y, replace = TRUE)
  res <- runif(length(bins), x[bins], x[bins + 1])
  na.omit(res)
}
```





Реальный пример #2 : Автоматическая обработка

Для каждого образца:

- подбирается модель `Mclust`
- считываются параметры модели: среднее (диаметр зерна), стандартное отклонение (коэф-т сортировки зерен), процентное содержание (зерен данного типа)
- логгируется информация о процессе (для многопоточной обработки)

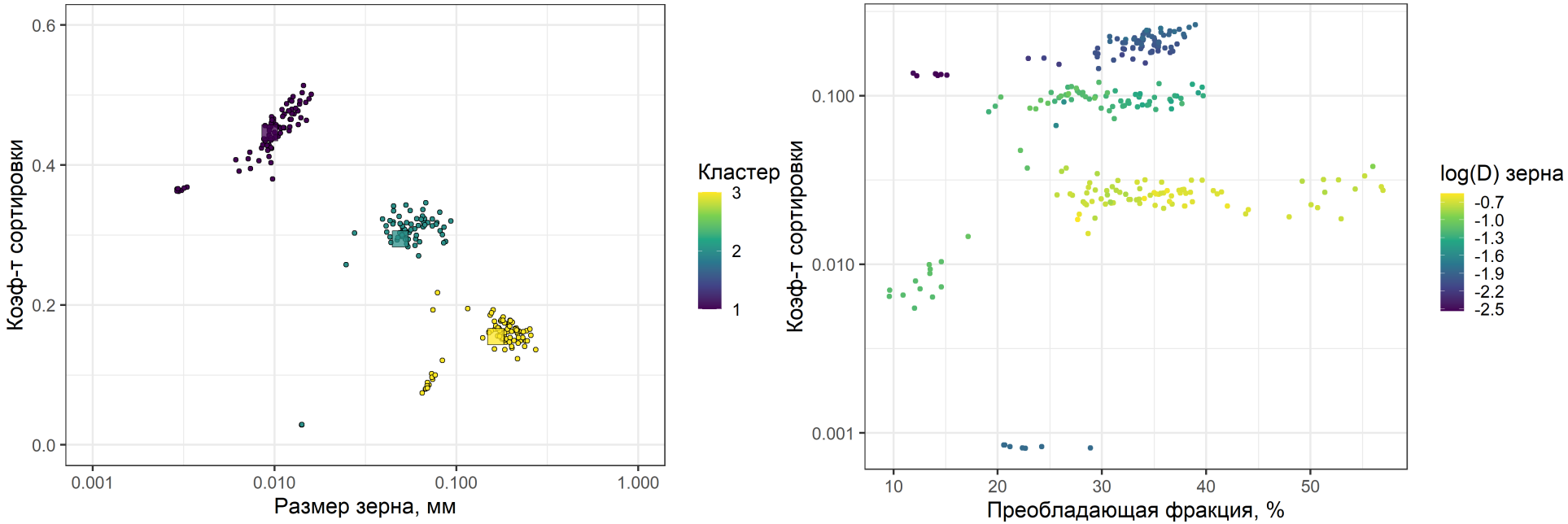
Итого: по всем образцам записывается новая таблица

893 образца, 8 потоков => 15 минут (DELL Precision, Intel(R) Xeon(R) CPU E3-1505M v5 @ 2.80GHz)

Фрагмент лога:

```
Job started at 2021-10-07 12:31:05 for 893 cases using 8 workers ...
Worker: 11900, Case 82; 37601-14; 1874.5; 1888.7; NA; НП5 ...
...
Worker: 18240, Case 9001; 29017-12; 2256.37; 2271.02; NA; NA ...
Job ended at 2021-10-07 12:47:35 for 893 cases using 8 workers
```

Реальный пример #2 : Обработка N образцов



Кластеризация результатов автоматического расчета моделей `Mclust`

Реальный пример #2 : Варианты отбора моделей

1. Стратегия 1 (MAXiMIN) :
   - Выбрать 2 модели по максимальному BIC для двух типов - E и V
   - Из двух моделей выбрать одну наименьшей сложности (с минимумом компонентов)
2. Стратегия 2 (MINiMIN):
   - Выбрать 2 модели с указанным минимальным относительным приростом информации (BIC gain), например, 0.03 (за 1.0 принимается модель с 1 компонентом)
   - Из двух моделей выбрать одну наименьшей сложности
3. Стратегия 3 (MCLUST):
   - Выбрать модели, предложенные алгоритмом Mclust (TOP-3)
   - Выбрать одну модель наименьшей сложности
4. Альтернативы - критерий ICL (https://arxiv.org/abs/1411.4257v2)

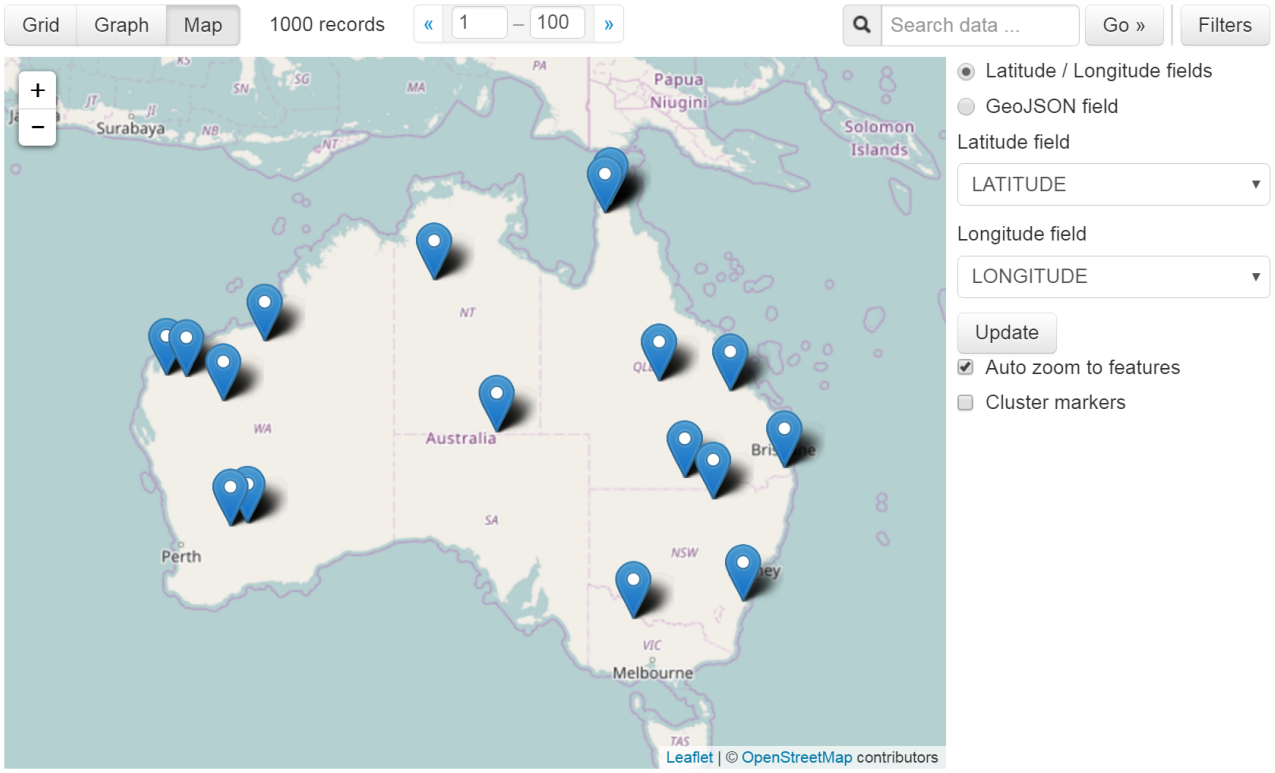
BACKUP

BACKUP SLIDES

Реальный пример #3 : Данные по размерам частиц

https://data.gov.au/dataset/ds-ga-c2b57800-84a2-2e21-e044-00144fdd4fa6/details?q=particle%20size%20data

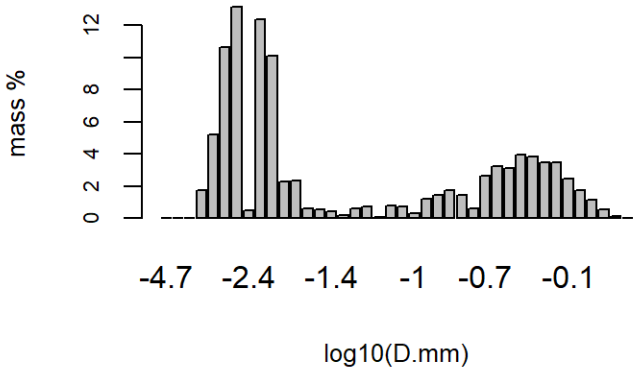
Реальный пример #3 : Места измерений

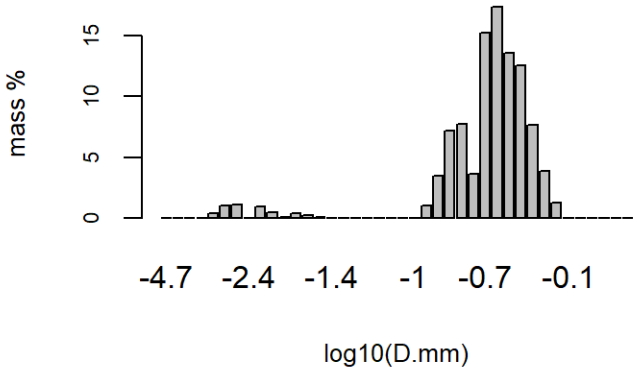


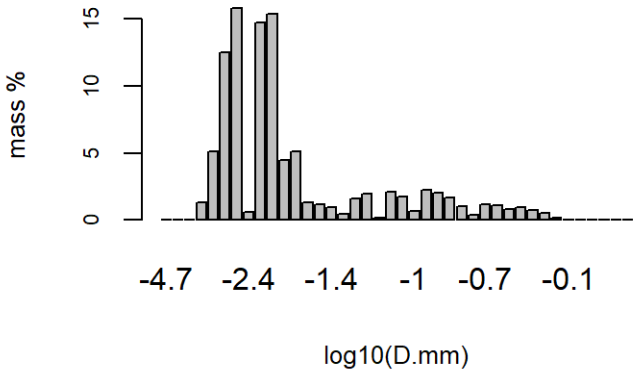
Реальный пример #3 : Набор данных

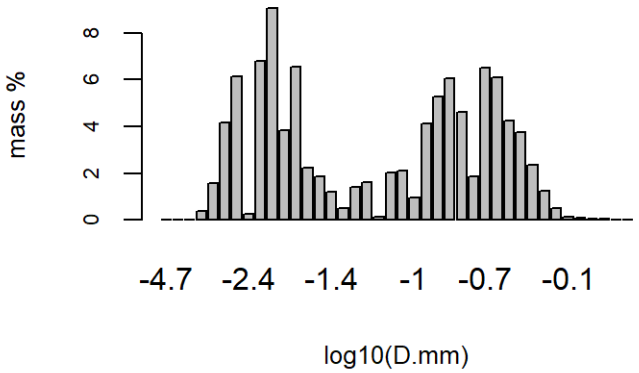
```
##      SITEID          DATE.SAMPLED          LATITUDE        LONGITUDE
##  Min.   :2.007e+09   Length:2627        Min.   :-43.33   Min.   :113.8
##  1st Qu.:2.007e+09   Class :character   1st Qu.:-31.39   1st Qu.:129.9
##  Median :2.007e+09   Mode  :character   Median :-26.75   Median :138.0
##  Mean   :2.007e+09                      Mean   :-26.24   Mean   :135.9
##  3rd Qu.:2.007e+09                      3rd Qu.:-20.88   3rd Qu.:144.3
##  Max.   :2.007e+09                      Max.   :-11.03   Max.   :153.3
##
##     STATE           DUPLICATE.CODE     DUPLICATE.SITEID      SAMPLEID
##  Length:2627        Length:2627        Min.   :2.007e+09   Length:2627
##  Class :character   Class :character   1st Qu.:2.007e+09   Class :character
##  Mode  :character   Mode  :character   Median :2.007e+09   Mode  :character
##                                        Mean   :2.007e+09
##                                        3rd Qu.:2.007e+09
##                                        Max.   :2.007e+09
##                                        NA's   :2136
##   GRAIN.SIZE           DEPTH
##  Length:2627        Length:2627
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
##
```

Реальный пример #3 : Случайные элементы









Заключение : Области применения простых моделей

- Простые одномерные модели - капля в море того, что может функционал `mclust`
- Где нужен гранулометрический состав?
- Как понимать вероятность принадлежности классу? (априорная, апостериорная)
- Что может и не может такая модель?
    - сценарий 1:
        - есть одно новое число
        - найти, в какой класс и с какой вероятностью попадает
    - сценарий 2:
        - есть новая выборка
        - найти, в какие классы и с какой вероятностью попадает