МИНОБРНАУКИ РОССИИ
САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
«ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)
Кафедра МОЭВМ

ОТЧЕТ

по лабораторной работе №7

по дисциплине «Методы статистической обработки данных»

Тема: Кластерный анализ. Метод поиска сгущений.

Студент гр. 5381 _____ Лянгузов А. А.

Преподаватель _____ Середа В. И.

Санкт-Петербург

2019

Цель работы.

Освоение основных понятий и некоторых методов кластерного анализа.

Основные теоретические положения.

Кластерный анализ (или кластеризация) – задача распределения однородных объектов из данного набора по группам (кластерам) таким образом, что объекты, принадлежащие одной группе, больше похожи друг на друга, чем на объекты из других групп. Процесс кластеризации подразумевает формирование этих групп. Это общая нечеткая формулировка задачи, на практике существует множество критериев «похожести» объектов – требований, предъявляемых кластерам, поэтому существует большое количество различных алгоритмов кластеризации, а также их модификаций.

Существующие методы можно разделить следующим образом:

I. По степени принадлежности объектов кластерам:
   - Строгая кластеризация – каждый объект может принадлежать только одному кластеру.
   - Нестрогая кластеризация – каждый объект может принадлежать нескольким кластерам в разной степени.

II. По способу формирования кластеров (по критериям кластеров):
   - Иерархические алгоритмы (по связанности объектов)
     1) Агломеративные (восходящие, объекты объединяются в кластеры)
     2) Дивизивные (нисходящие, кластеры дробятся на более мелкие)
   - Алгоритмы, основанные на поиске/использовании центров кластеров (k-means)
   - Кластеризация по распределениям
   - Кластеризация по плотностям

В данной лабораторной работе использовалось два *критерия качества* кластерных разбиений:

3) Средний кластерный радиус $r_{кл}$:

$$r_{кл} = \frac{1}{K}\sum_{i=1}^{K}\left(\frac{1}{m_i}\sum_{j=1}^{m_i} dist\left(c_i,\ p_j^i\right)\right),\ где\ K - количество\ кластеров,\ m_i - количество\ объектов\ в\ кластере\ i,\ c_i - центр\ кластера\ i,\ p_j^i - j\text{ый}\ объект,\ принадлежащий\ кластеру\ i,\ dist(x,\ y) - евклидово\ расстояние\ между\ x\ и\ y.$$

Для данного показателя вычисляются для каждого кластера расстояния от центров до точек, им принадлежащих, находятся средние радиусы для каждого кластера, а затем вычисляется среднее значение из средних.

4) Среднее внутрикластерное расстояние $d_{кл}$:

$$d_{кл} = \frac{1}{K}\sum_{i=1}^{K}\frac{2}{(m_i-1)m_i}\sum_{p,q\in\mathcal{K}_i} dist(p,\ q),\ где\ K - количество\ кластеров,\ m_i - количество\ объектов\ в\ кластере\ i,\ \mathcal{K}_i - кластер\ i,\ dist(p,\ q) - евклидово\ расстояние\ между\ p\ и\ q.$$

Положим, в кластере $\mathcal{K}_i$ содержится m_i элементов, тогда для каждого кластера вычисляются расстояния между его элементами и высчитывается среднее, а затем для всех кластеров находится среднее из средних.

Метод поиска сгущений (ForEl).

В данном методе каждый кластер определяется своим центром (как и в к-средних) и радиусом (одинаковым для всех кластеров), которые вместе образуют круг (шар). Эти области покрывают все множество исходных точек. Данному кластеру принадлежат точки, лежащие внутри его круга.

Очевидно, что ситуация, когда области нескольких кластеров пересекаются, может возникнуть довольно часто. Необходимо обуславливаться, каким образом решать подобные конфликты. В данной работе исследовалась строгая кластеризация и предпочтение отдавалось кластеру с наибольшим количеством элементов, при равенстве элементов выбирался первый встретившийся кластер из соперничающих.

Входные данные:

- исходное множество, подлежащее кластеризации;
- R – радиус.

Выбор кластерного радиуса является нетривиальной задачей, поскольку зависит от вида кластеризуемого множества.

Также алгоритму входе поиска кластеров (итеративно) необходимо выбирать стартовые точки для инициализации центров. Конечный результат может сильно варьироваться в зависимости от «удачности» начальных приближений.

В данной работе было реализовано две модификации метода поиска сгущений: «стандартный» и «с полным просмотром».

I. «Стандартный» вариант:
    1. Из множества рассматриваемых точек по какому-то принципу выбирается первое начальное приближение – центр кластера.
    2. Все рассматриваемые точки, лежащие «внутри» кластерного круга (т.е. расстояние до которых от центра кластера меньше радиуса), включаются в текущий кластер, все точки, лежащие вне круга, – кластеру не принадлежат.
    3. Пересчитывается центр кластера.
    4. Пункты 2 и 3 повторяются до тех пор, пока кластер не стабилизируется, т.е. нельзя ни добавить, ни исключить точки.
    5. Все точки, добавленные в кластер, исключаются из рассмотрения. Начинается новая итерация поиска кластера с пункта 1.

6. Поиск продолжается, пока рассматриваемое множество не станет пусто.

Принципы выбора начальных приближений:

- Random – случайным образом.
- Minmedian – необходимо посчитать расстояния между всеми рассматриваемыми точками, затем для всех точки найти медианные расстояния и выбрать минимальное из них, точку с минимальным медианным расстоянием принять за начальное приближение.
- Maxmedian – то же самое, что minmedian, только выбирать максимальную медиану.

II. «С полным просмотром»:

1. Просматриваются все точки множества: каждая точка по очереди принимается за начальное приближение и находятся N (объем выборки) кандидатов в кластеры (уникальных кластеров может оказаться меньше), как описано выше.
2. В качестве найденного кластера выбирается содержащий наибольшее количество точек. Данные точки исключаются из дальнейшего поиска.
3. Повторяются первый и второй пункты для оставшихся точек, пока все точки исходного множества не будут разделены на кластеры.

Для измерения расстояния используется евклидово расстояние между точками в двумерном пространстве:

$$\mathrm{dist}(\mathrm{p},\ \mathrm{q}) = \sqrt[2]{\left(p_x - q_x\right)^2 + \left(p_y - q_y\right)^2}$$

За центр кластера принимается центр масс (барицентр, центроид) точек, ему принадлежащих:

$$\left(c_x,\ c_y\right)_i = \left(\frac{1}{m_i}\sum_{j=1}^{m_i} x_j^i,\ \ \frac{1}{m_i}\sum_{j=1}^{m_i} y_j^i\right),\ \text{где } m_i - \text{количество элементов в кластере } i,\ \left(x_j^i,\ y_j^i\right) - \text{координаты } j^{\text{ой}} \text{ точки кластера}$$

Задание.

Дано конечное множество из объектов, представленных двумя признаками (в качестве этого множества принимаем исходную двумерную выборку, сформированную ранее в лабораторной работе №4). Выполнить разбиение исходного множества объектов на конечное число подмножеств (кластеров) с использованием метода поиска сгущений. Полученные результаты содержательно проинтерпретировать.

Ход выполнения.

1.Двумерная выборка объема *N = 107*.

v	501.00	369.00	344.00	473.00	426.00	528.00	497.00	467.00	506.00	431.00	454.00
E	130.40	84.30	86.80	137.90	121.10	163.40	147.30	140.50	158.40	125.00	131.10
v	371.00	482.00	393.00	441.00	463.00	440.00	481.00	340.00	468.00	397.00	496.00
E	89.20	139.90	103.20	122.80	129.10	128.50	135.20	85.10	142.00	108.60	143.10
v	434.00	541.00	352.00	438.00	453.00	423.00	351.00	525.00	409.00	469.00	386.00
E	122.30	146.80	87.70	134.90	119.50	131.10	89.00	165.90	121.00	131.50	95.50
v	505.00	436.00	488.00	449.00	493.00	512.00	472.00	423.00	465.00	351.00	359.00
E	137.50	114.30	134.10	124.50	129.70	169.90	134.20	130.80	140.70	102.90	71.90
v	457.00	467.00	400.00	418.00	492.00	434.00	510.00	392.00	463.00	459.00	397.00
E	126.40	135.10	114.60	118.60	137.50	110.50	140.60	82.70	125.00	145.40	106.80
v	424.00	436.00	429.00	398.00	493.00	522.00	518.00	463.00	437.00	386.00	493.00
E	119.00	116.70	112.90	109.00	154.50	154.50	144.40	121.20	121.80	105.80	151.20
v	414.00	480.00	585.00	562.00	508.00	421.00	463.00	422.00	406.00	544.00	345.00
E	113.50	153.90	177.70	175.90	159.00	117.80	136.70	122.90	110.10	166.70	95.90
v	478.00	393.00	437.00	448.00	458.00	422.00	468.00	430.00	371.00	543.00	471.00
E	126.60	122.80	115.10	121.90	121.70	115.70	144.90	104.30	91.90	155.40	143.90
v	475.00	521.00	353.00	437.00	362.00	490.00	484.00	459.00	480.00	482.00	522.00
E	132.00	139.60	98.00	118.40	111.70	139.90	140.40	136.70	153.30	148.20	143.80
v	576.00	390.00	514.00	442.00	421.00	443.00	438.00	429.00			
E	166.40	91.40	153.60	115.40	107.90	121.90	126.70	120.90			

Было произведено масштабирование значений выборки так, чтобы они все попадали в интервал [0; 1].

v	0.6571 4286	0.1183 6735	0.0163 2653	0.5428 5714	0.3510 2041	0.7673 4694	0.6408 1633	0.5183 6735	0.6775 5102	0.3714 2857	0.4653 0612
E	0.5529 301	0.1172 023	0.1408 318	0.6238 185	0.4650 284	0.8648 393	0.7126 654	0.6483 932	0.8175 803	0.5018 904	0.5595 463
v	0.1265 3061	0.5795 9184	0.2163 2653	0.4122 4490	0.5020 4082	0.4081 6327	0.5755 1020	0.0000 0000	0.5224 4898	0.2326 5306	0.6367 3469
E	0.1635 161	0.6427 221	0.2958 412	0.4810 964	0.5406 427	0.5349 716	0.5982 987	0.1247 637	0.6625 709	0.3468 809	0.6729 679
v	0.3836 7347	0.8204 0816	0.0489 7959	0.4000 0000	0.4612 2449	0.3387 7551	0.0448 9796	0.7551 0204	0.2816 3265	0.5265 3061	0.1877 5510
E	0.4763 705	0.7079 395	0.1493 384	0.5954 631	0.4499 055	0.5595 463	0.1616 257	0.8884 688	0.4640 832	0.5633 270	0.2230 624
v	0.6734 6939	0.3918 3673	0.6040 8163	0.4448 9796	0.6244 8980	0.7020 4082	0.5387 7551	0.3387 7551	0.5102 0408	0.0448 9796	0.0775 5102
E	0.6200 378	0.4007 561	0.5879 017	0.4971 645	0.5463 138	0.9262 760	0.5888 469	0.5567 108	0.6502 836	0.2930 057	0.0000 000
v	0.4775 5102	0.5183 6735	0.2448 9796	0.3183 6735	0.6204 0816	0.3836 7347	0.6938 7755	0.2122 4490	0.5020 4082	0.4857 1429	0.2326 5306
E	0.5151 229	0.5973 535	0.4035 917	0.4413 989	0.6200 378	0.3648 393	0.6493 384	0.1020 794	0.5018 904	0.6947 070	0.3298 677
v	0.3428 5714	0.3918 3673	0.3632 6531	0.2367 3469	0.6244 8980	0.7428 5714	0.7265 3061	0.5020 4082	0.3959 1837	0.1877 5510	0.6244 8980
E	0.4451 796	0.4234 405	0.3875 236	0.3506 616	0.7807 183	0.7807 183	0.6852 552	0.4659 735	0.4716 446	0.3204 159	0.7495 274
v	0.3020 4082	0.5714 2857	1.0000 0000	0.9061 2245	0.6857 1429	0.3306 1224	0.5020 4082	0.3346 9388	0.2693 8776	0.8326 5306	0.0204 0816
E	0.3931 947	0.7750 473	1.0000 000	0.9829 868	0.8232 514	0.4338 374	0.6124 764	0.4820 416	0.3610 586	0.8960 302	0.2268 431
v	0.5632 6531	0.2163 2653	0.3959 1837	0.4408 1633	0.4816 3265	0.3346 9388	0.5224 4898	0.3673 4694	0.1265 3061	0.8285 7143	0.5346 9388
E	0.5170 132	0.4810 964	0.4083 176	0.4725 898	0.4706 994	0.4139 887	0.6899 811	0.3062 382	0.1890 359	0.7892 250	0.6805 293
v	0.5510 2041	0.7387 7551	0.0530 6122	0.3959 1837	0.0897 9592	0.6122 4490	0.5877 5510	0.4857 1429	0.5714 2857	0.5795 9184	0.7428 5714
E	0.5680 529	0.6398 866	0.2466 919	0.4395 085	0.3761 815	0.6427 221	0.6474 480	0.6124 764	0.7693 762	0.7211 720	0.6795 841
v	0.9632 6531	0.2040 8163	0.7102 0408	0.4163 2653	0.3306 1224	0.4204 0816	0.4000 0000	0.3632 6531			
E	0.8931 947	0.1843 100	0.7722 117	0.4111 531	0.3402 647	0.4725 898	0.5179 584	0.4631 380			

Теперь обе величины, составляющие двумерную выборку, вносят одинаковый вклад при расчете расстояния.

2. В работе было проведено два эксперимента:

1. в первом при построении кластеров варьировался радиус и использовались различные модификации алгоритма (выбор начального приближения),

2. во втором эксперименте фиксировался радиус, а начальные приближения выбирались случайно.

Результаты первого эксперимента.

Радиус варьировался в промежутке [0.1; 0.325] с шагом в $\delta = 0.025$.

K – количество найденных кластеров.

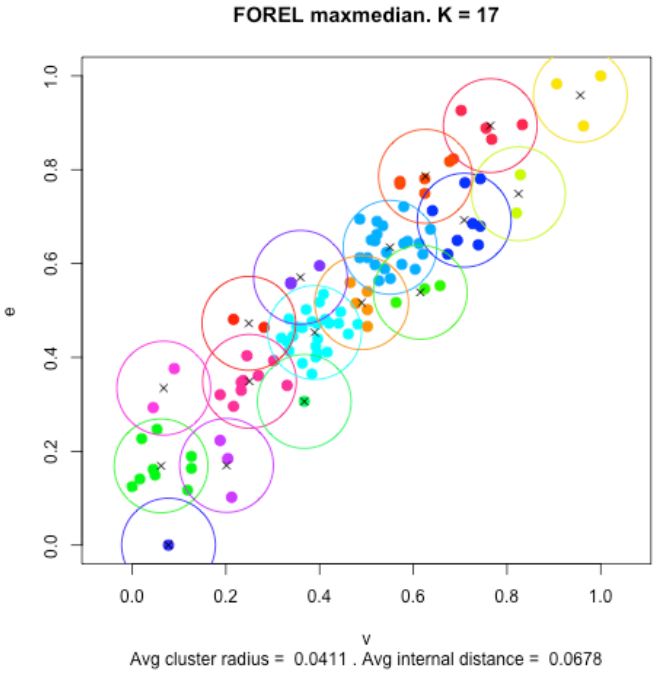
FOREL maxmedian - вариант алгоритма, при котором в качестве начального приближения выбирается точка, имеющую максимальное значение медианы (среднее расстояние до остальных точек),

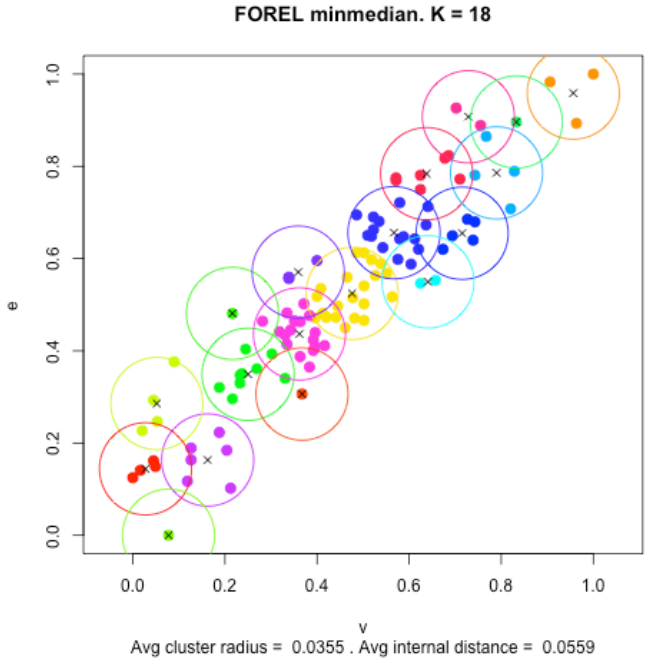
FOREL minmedian - вариант алгоритма, при котором в качестве начального приближения выбирается точка, имеющую минимальное значение медианы (среднее расстояние до остальных точек),

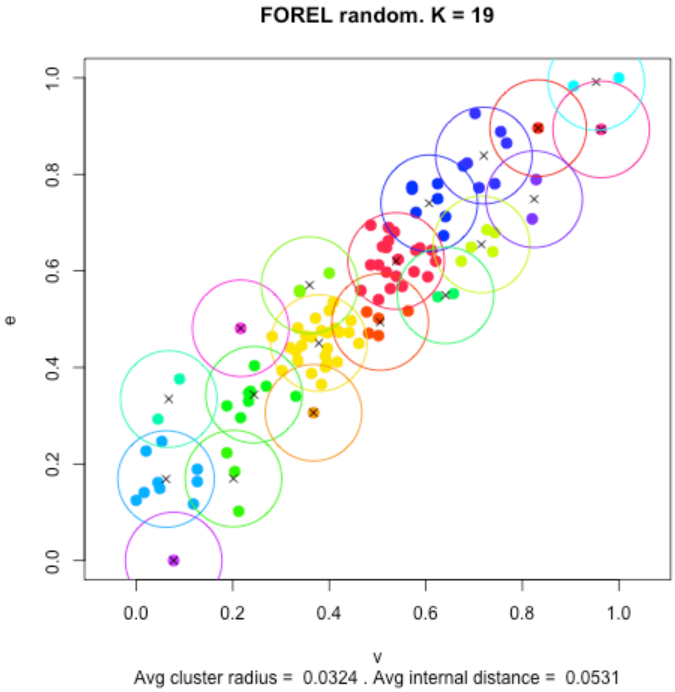
FOREL random – вариант алгоритма с соответствующим способом выбора начальных приближений.

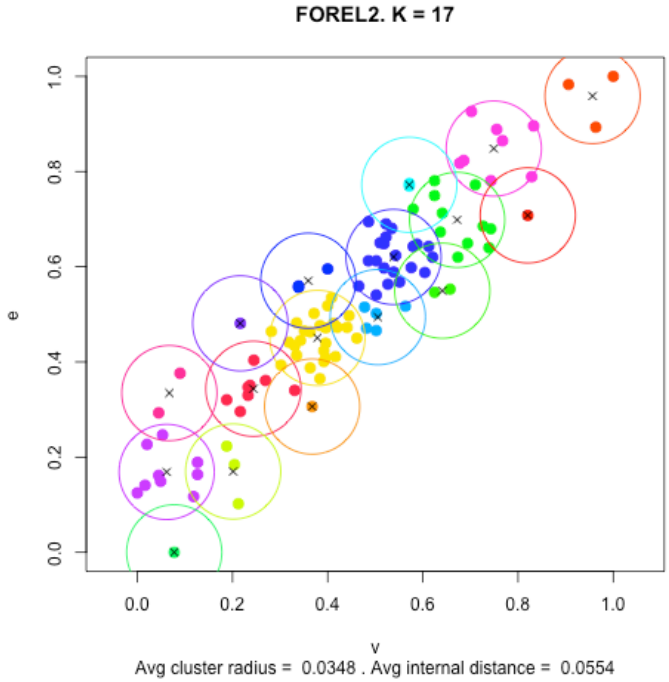
FOREL2 – модификация алгоритма «с полным просмотром».

R = 0.1


FOREL maxmedian. K = 17
e
v
0.0
0.2
0.4
0.6
0.8
1.0
Avg cluster radius = 0.0411 . Avg internal distance = 0.0678

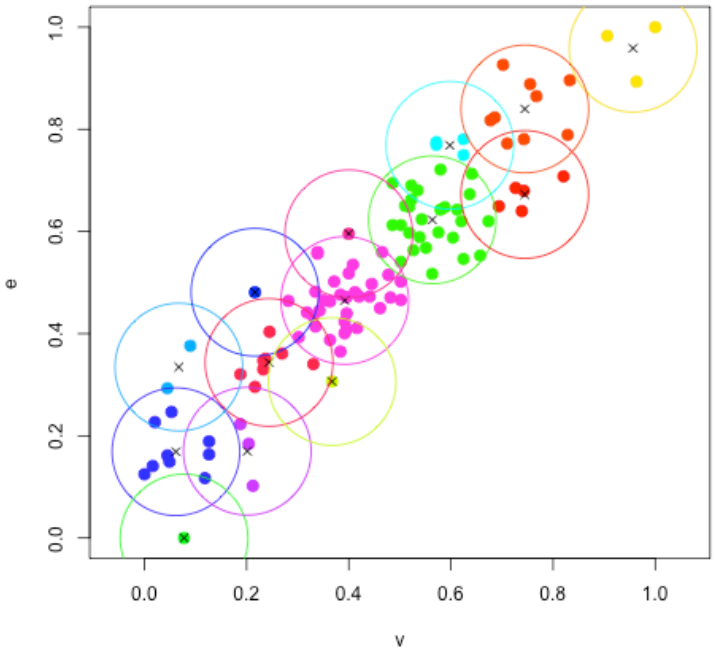

FOREL minmedian. K = 18
e
v
0.0
0.2
0.4
0.6
0.8
1.0
Avg cluster radius = 0.0355 . Avg internal distance = 0.0559


FOREL random. K = 19
e
v
0.0
0.2
0.4
0.6
0.8
1.0
Avg cluster radius = 0.0324 . Avg internal distance = 0.0531


FOREL2. K = 17
e
v
0.0
0.2
0.4
0.6
0.8
1.0
Avg cluster radius = 0.0348 . Avg internal distance = 0.0554

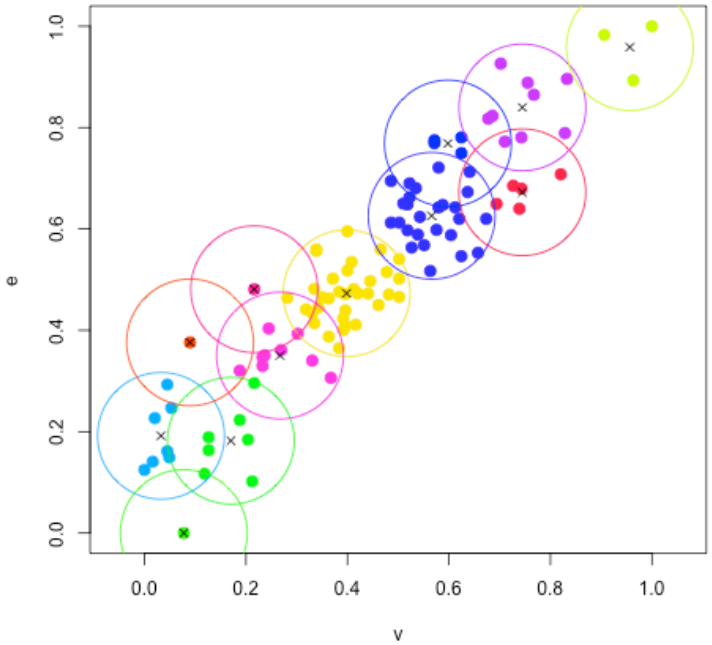
R = 0.125

FOREL maxmedian. K = 14



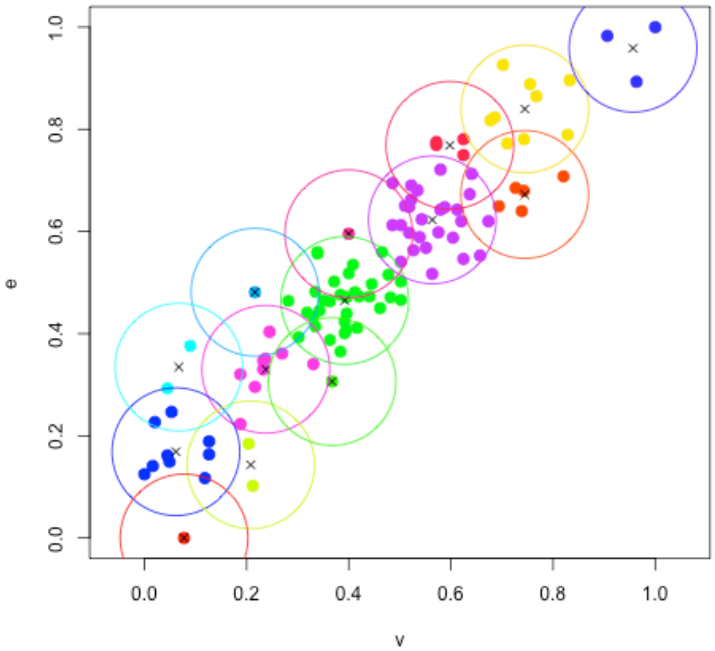
Avg cluster radius = 0.0381 . Avg internal distance = 0.0606

FOREL minmedian. K = 12



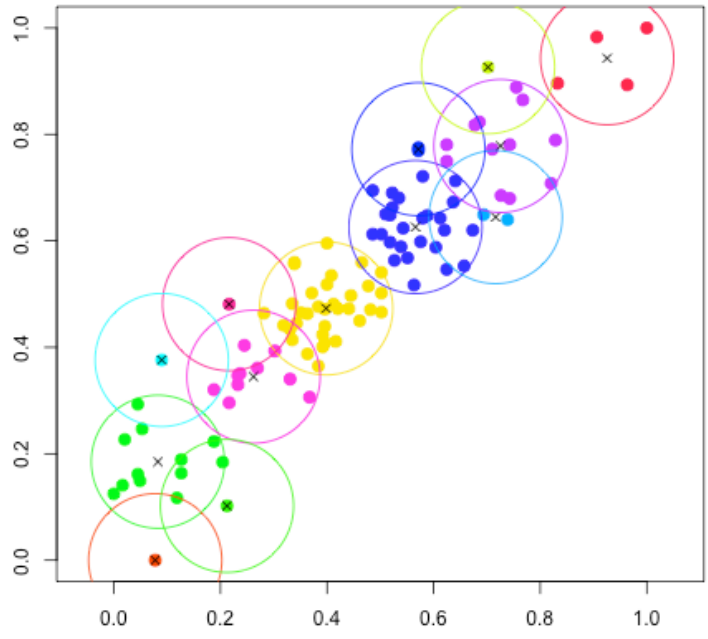
Avg cluster radius = 0.0433 . Avg internal distance = 0.0655

FOREL random. K = 14



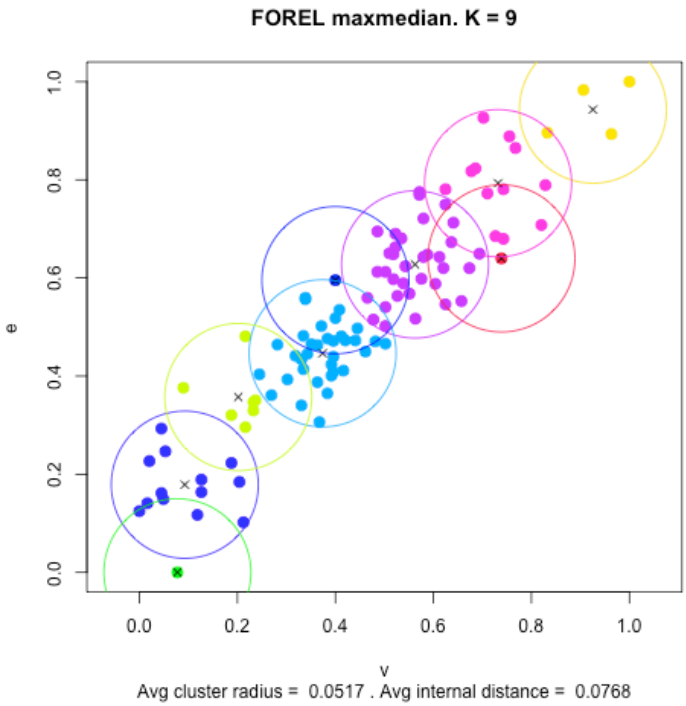
Avg cluster radius = 0.0385 . Avg internal distance = 0.0617

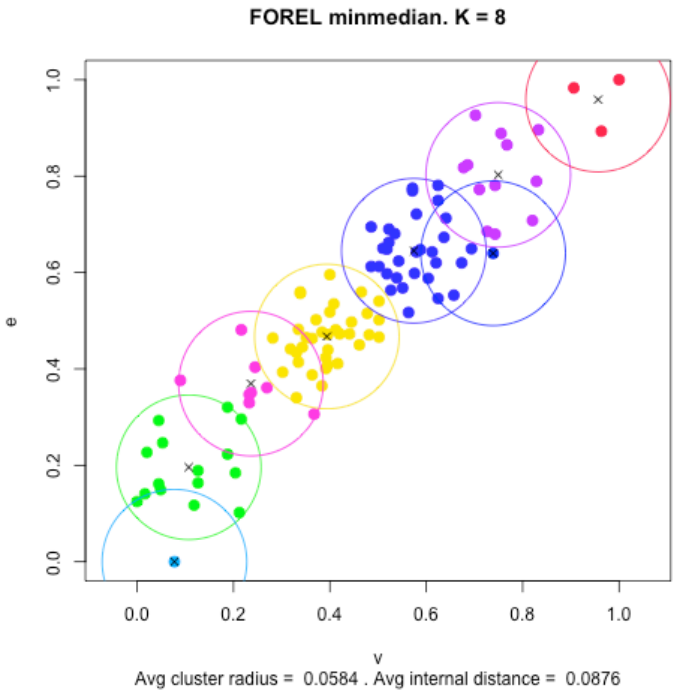
FOREL2. K = 13

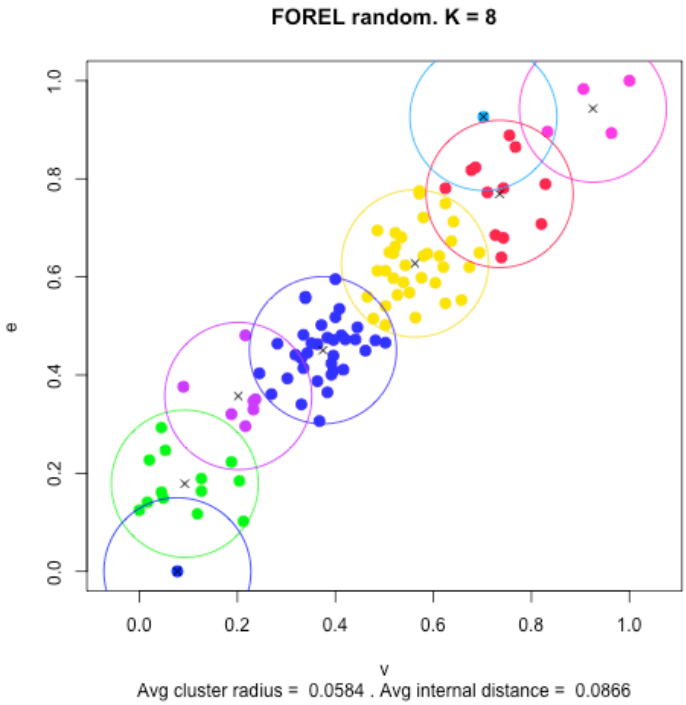


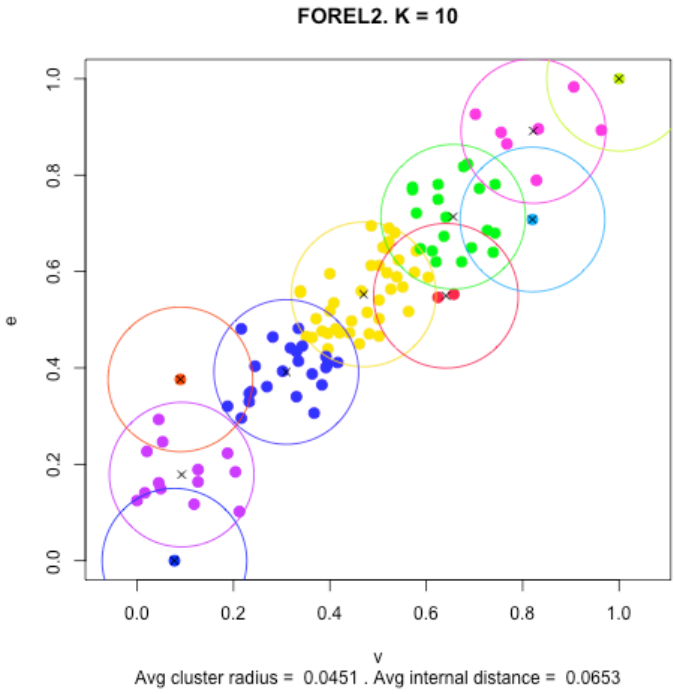
Avg cluster radius = 0.0351 . Avg internal distance = 0.0531

R = 0.15

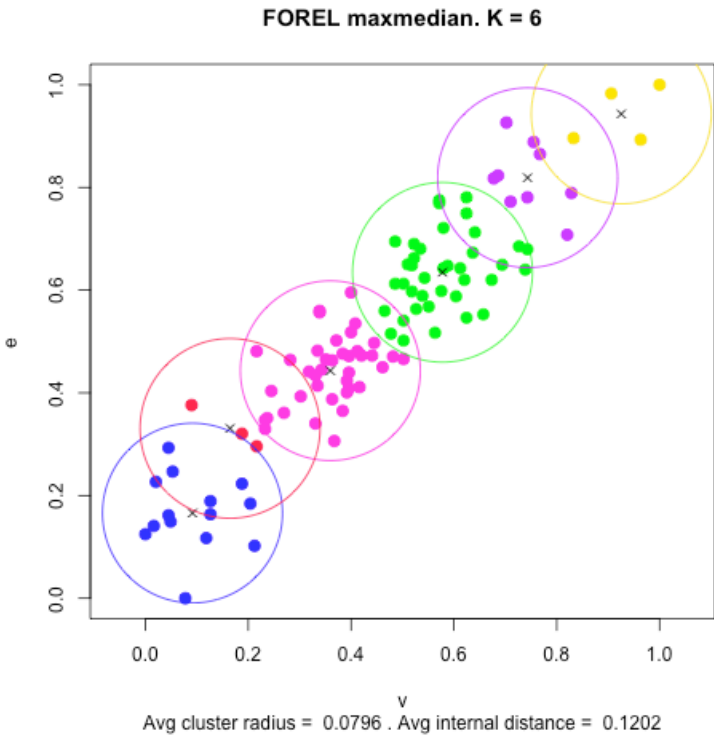

FOREL maxmedian. K = 9
e
v
0.0
0.2
0.4
0.6
0.8
1.0
Avg cluster radius = 0.0517 . Avg internal distance = 0.0768


FOREL minmedian. K = 8
e
v
0.0
0.2
0.4
0.6
0.8
1.0
Avg cluster radius = 0.0584 . Avg internal distance = 0.0876

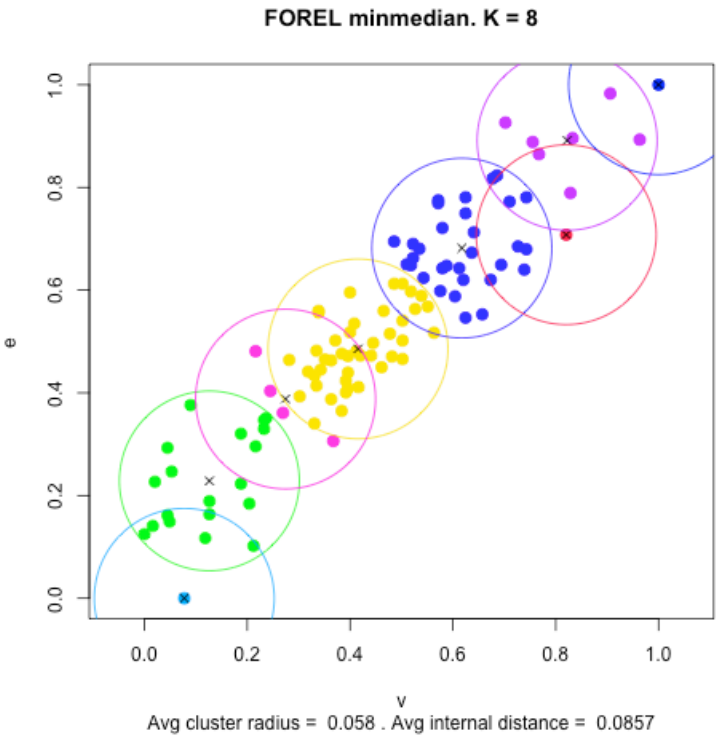

FOREL random. K = 8
e
v
0.0
0.2
0.4
0.6
0.8
1.0
Avg cluster radius = 0.0584 . Avg internal distance = 0.0866


FOREL2. K = 10
e
v
0.0
0.2
0.4
0.6
0.8
1.0
Avg cluster radius = 0.0451 . Avg internal distance = 0.0653

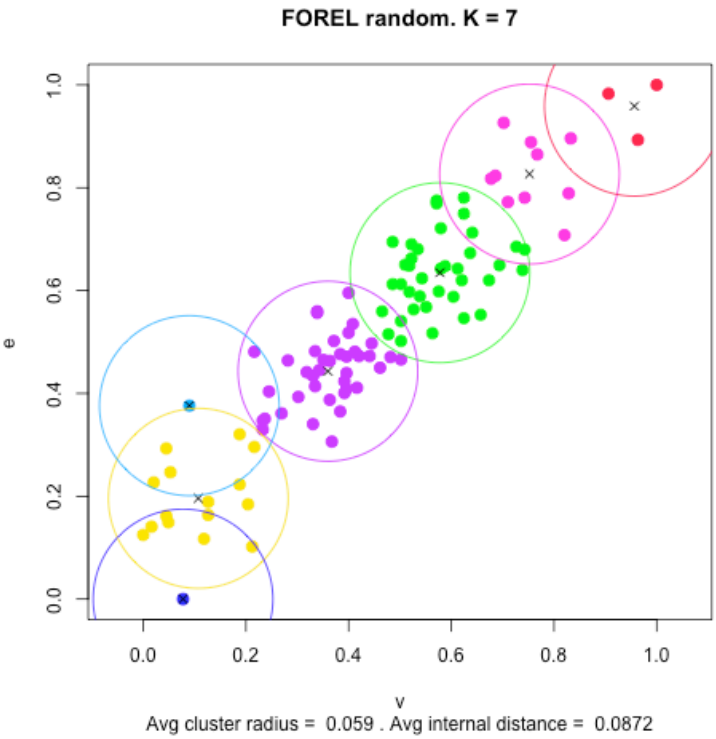
R = 0.175



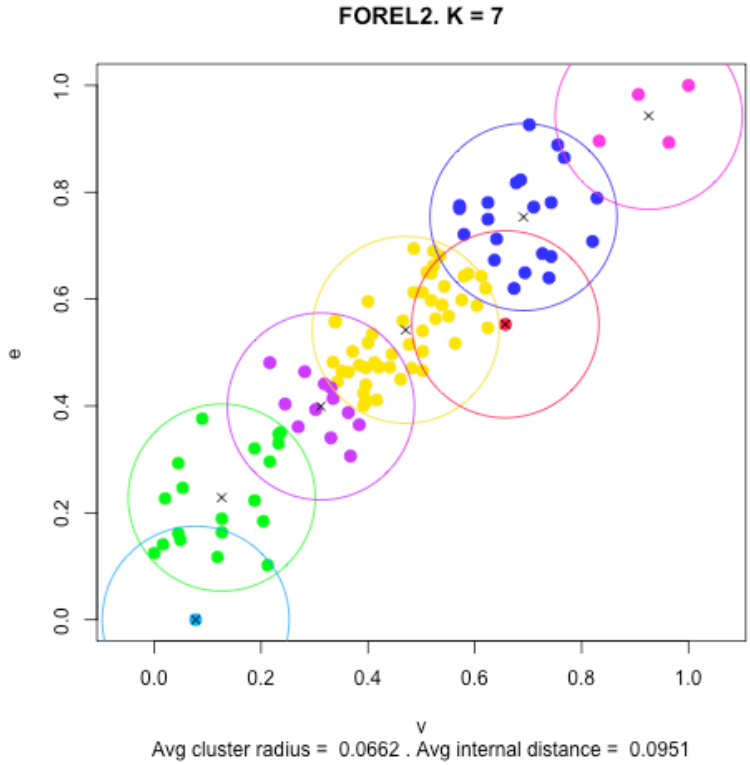
Avg cluster radius = 0.0796 . Avg internal distance = 0.1202



Avg cluster radius = 0.058 . Avg internal distance = 0.0857

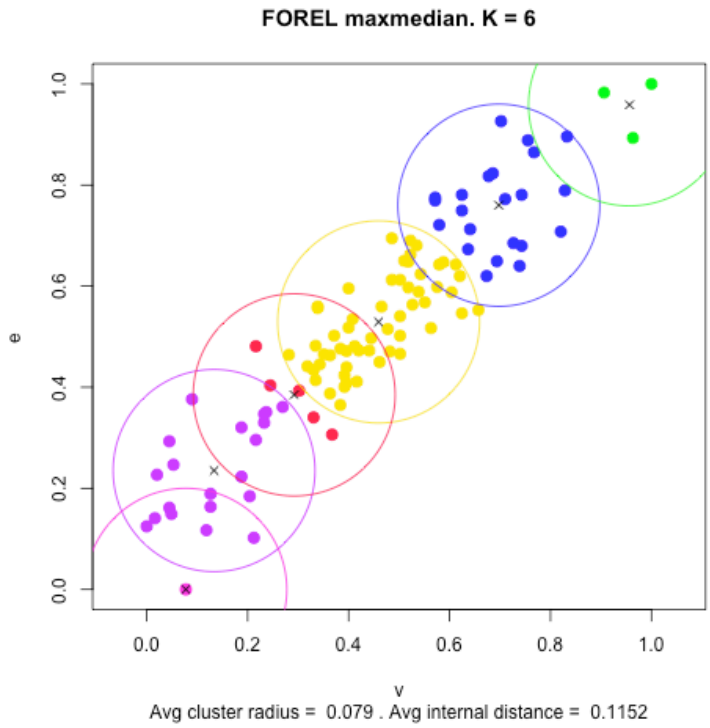


Avg cluster radius = 0.059 . Avg internal distance = 0.0872

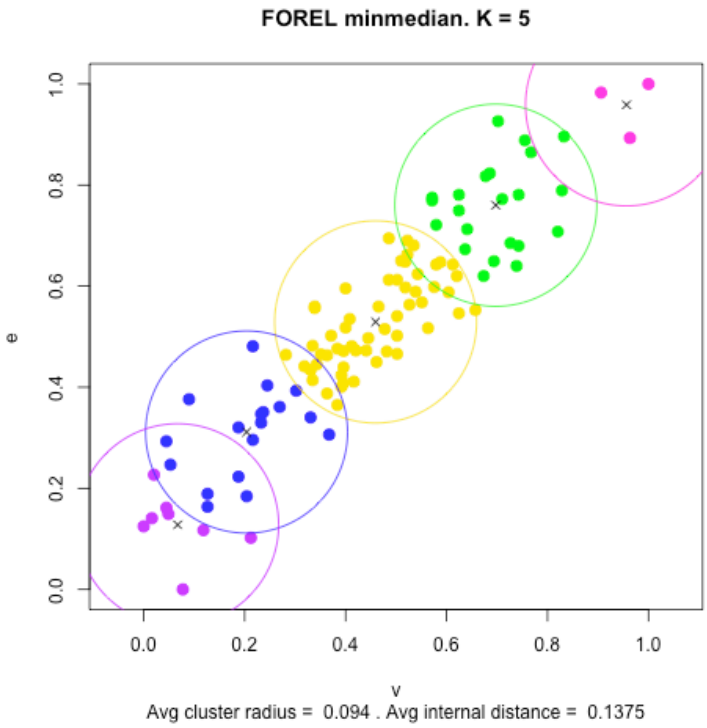


Avg cluster radius = 0.0662 . Avg internal distance = 0.0951

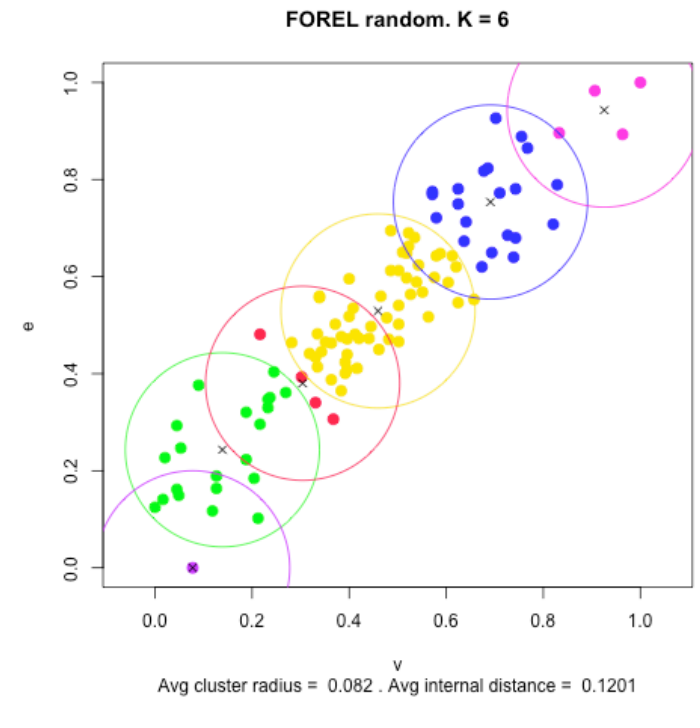
R = 0.2



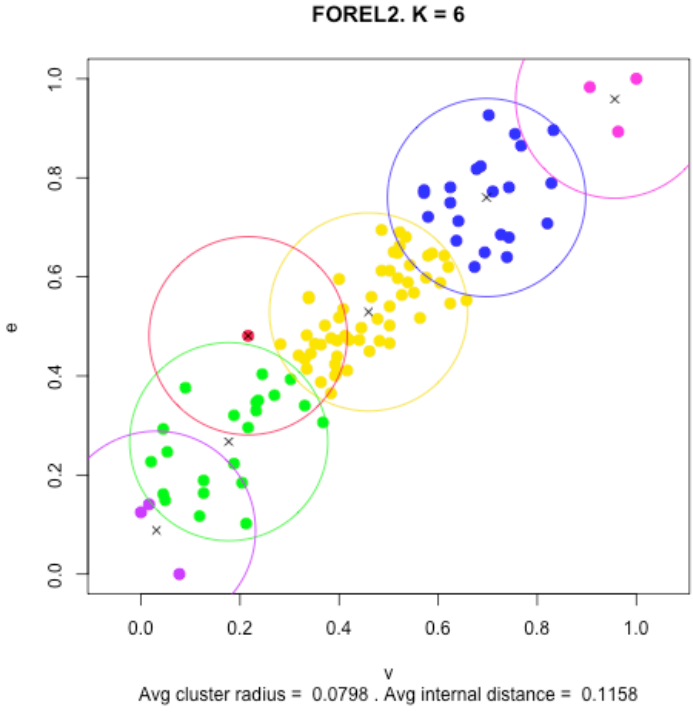
Avg cluster radius = 0.079 . Avg internal distance = 0.1152



Avg cluster radius = 0.094 . Avg internal distance = 0.1375

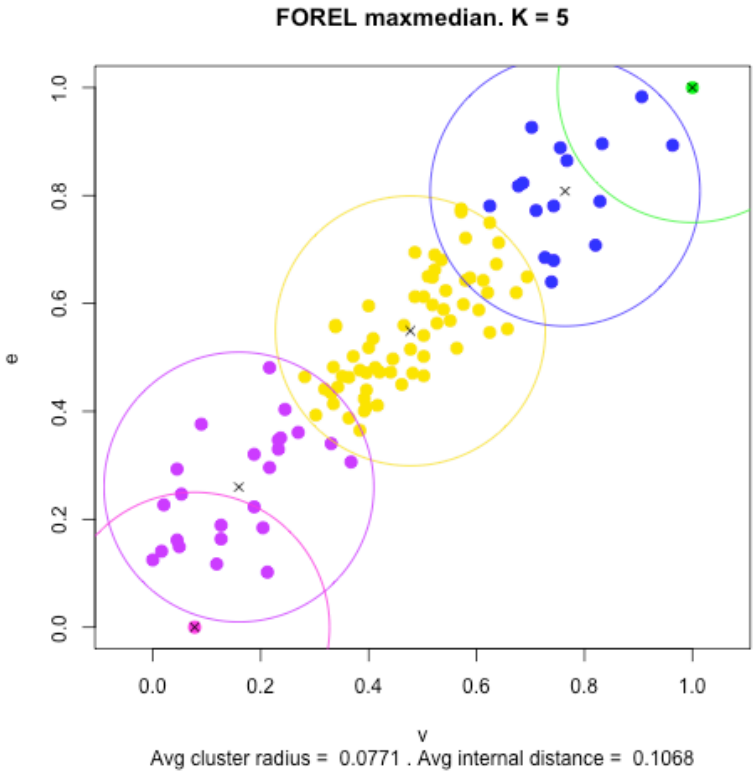


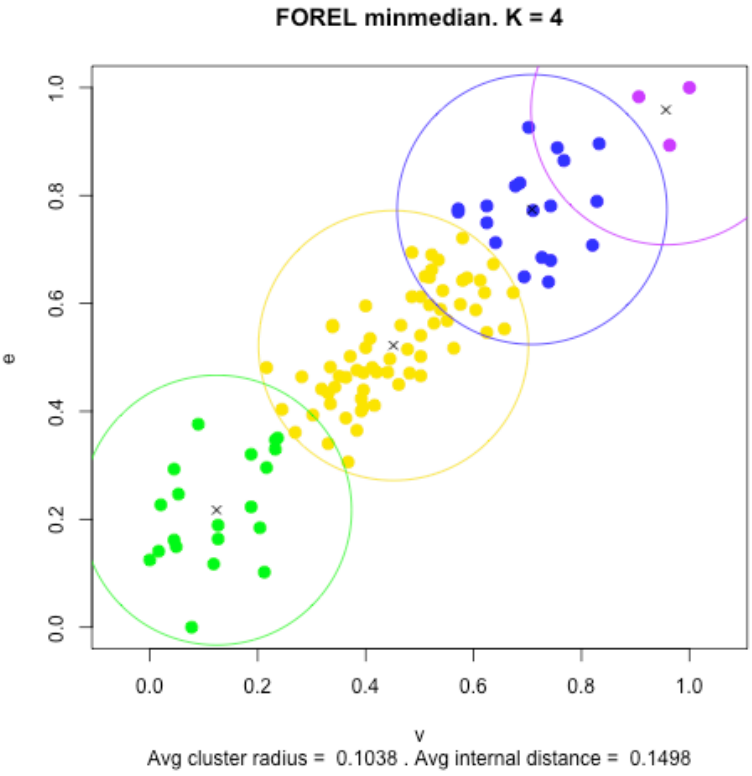
Avg cluster radius = 0.082 . Avg internal distance = 0.1201

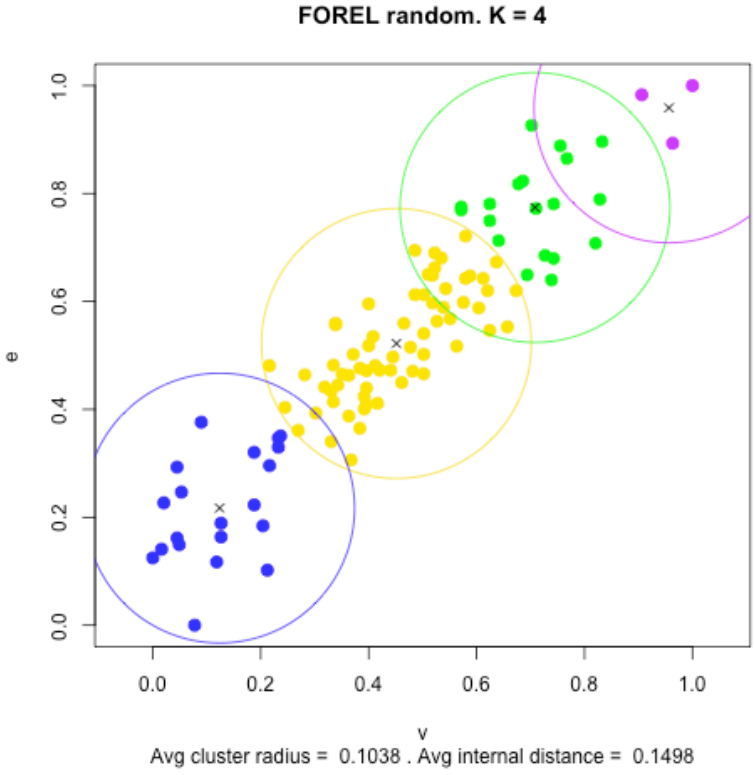


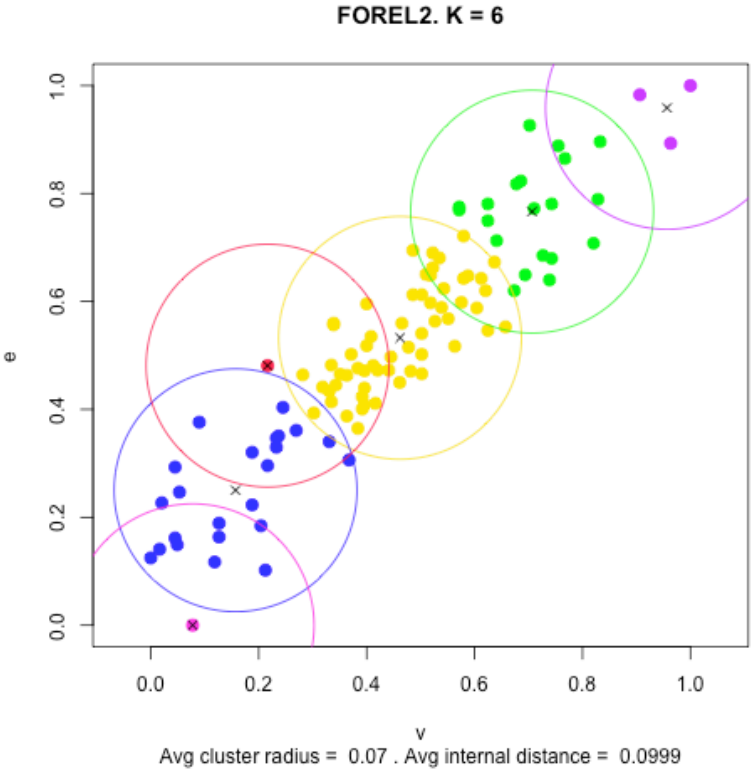
Avg cluster radius = 0.0798 . Avg internal distance = 0.1158

R = 0.225

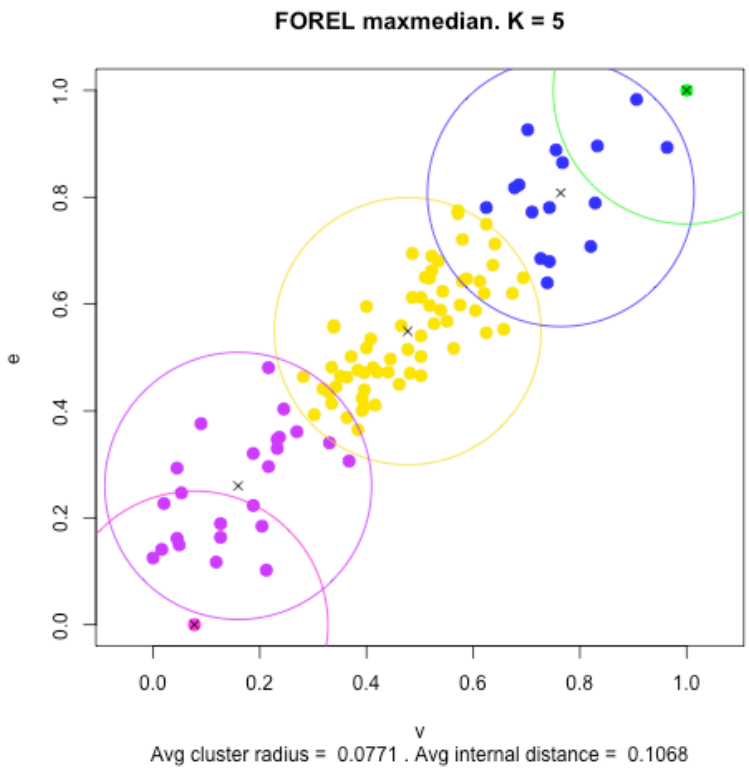

FOREL maxmedian. K = 5
e
v
0.0
0.2
0.4
0.6
0.8
1.0
Avg cluster radius = 0.0771 . Avg internal distance = 0.1068

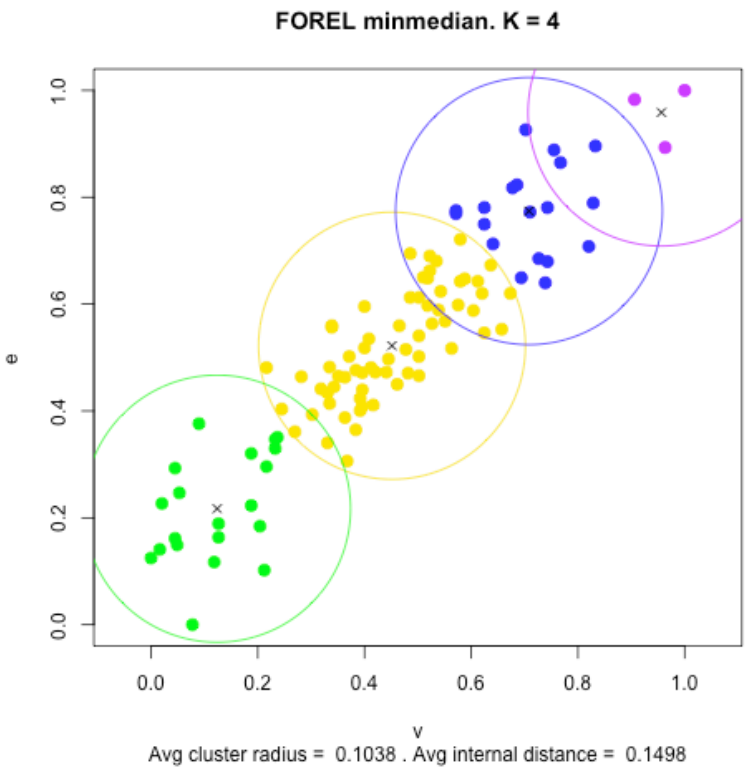

FOREL minmedian. K = 4
e
v
0.0
0.2
0.4
0.6
0.8
1.0
Avg cluster radius = 0.1038 . Avg internal distance = 0.1498

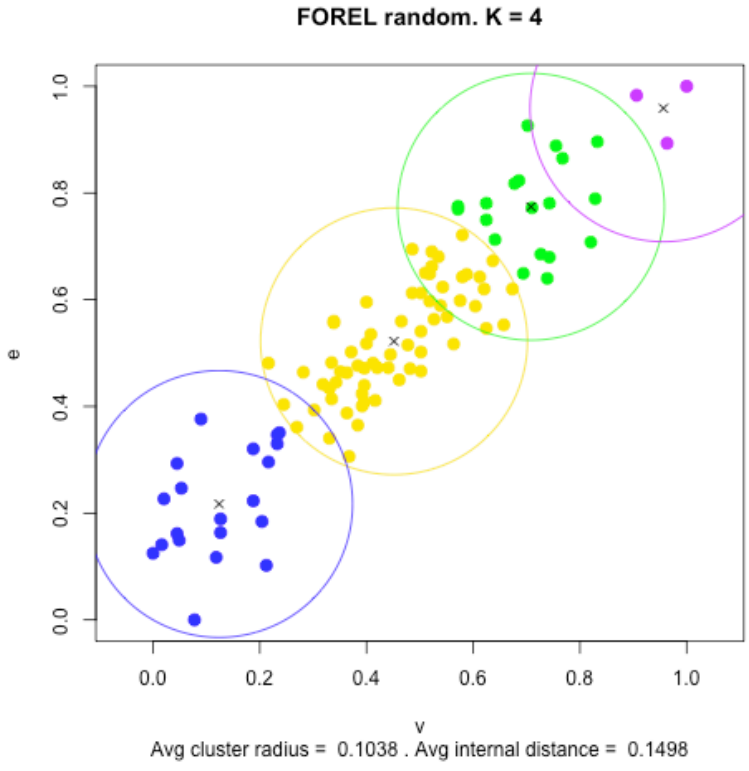

FOREL random. K = 4
e
v
0.0
0.2
0.4
0.6
0.8
1.0
Avg cluster radius = 0.1038 . Avg internal distance = 0.1498

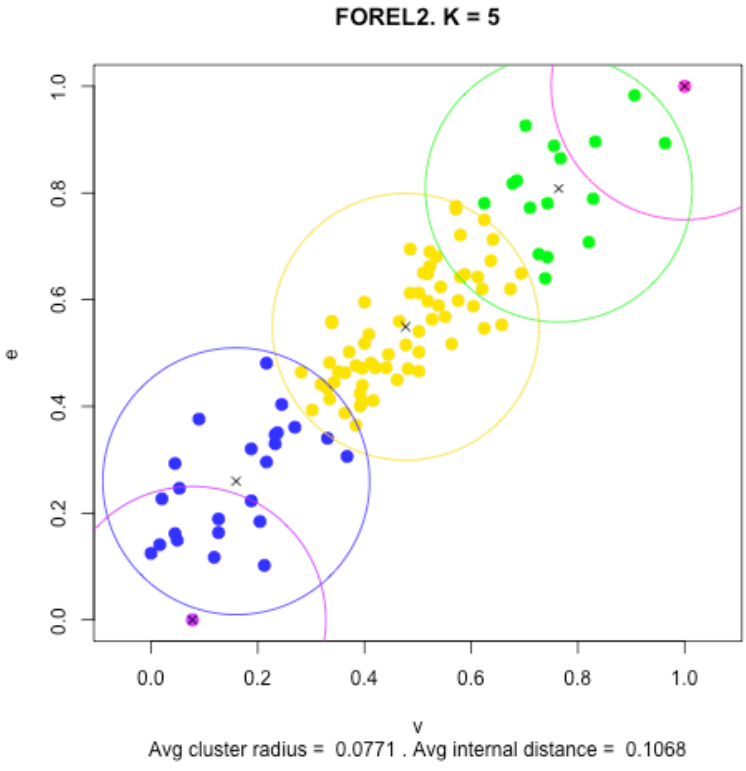

FOREL2. K = 6
e
v
0.0
0.2
0.4
0.6
0.8
1.0
Avg cluster radius = 0.07 . Avg internal distance = 0.0999

R = 0.25

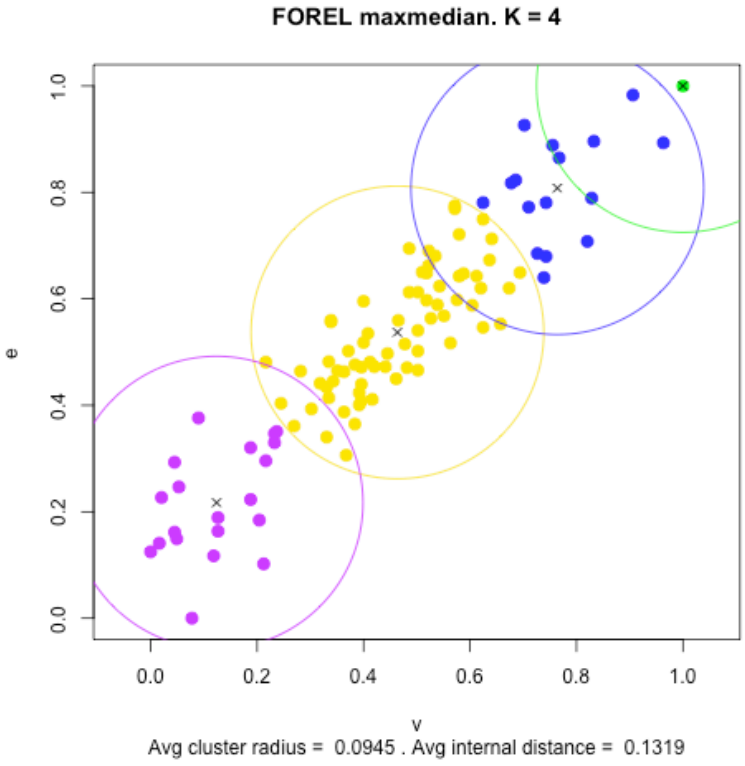

FOREL maxmedian. K = 5
e
v
0.0
0.2
0.4
0.6
0.8
1.0
Avg cluster radius = 0.0771 . Avg internal distance = 0.1068

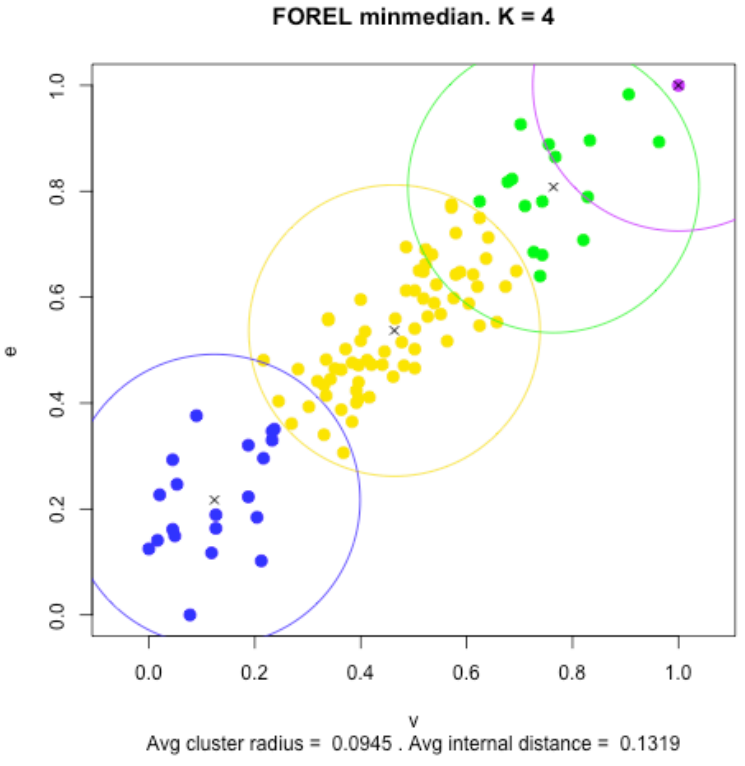

FOREL minmedian. K = 4
e
v
0.0
0.2
0.4
0.6
0.8
1.0
Avg cluster radius = 0.1038 . Avg internal distance = 0.1498

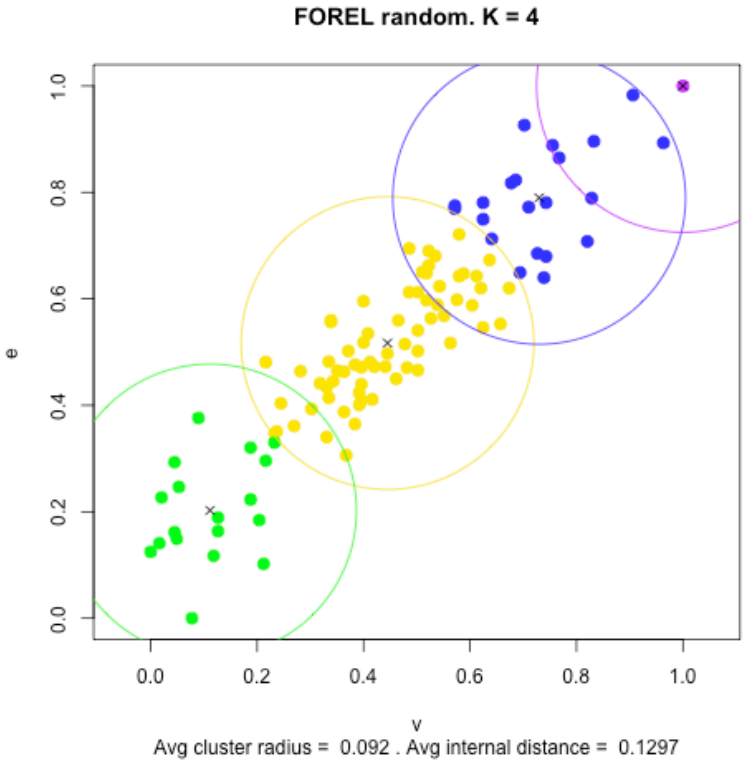

FOREL random. K = 4
e
v
0.0
0.2
0.4
0.6
0.8
1.0
Avg cluster radius = 0.1038 . Avg internal distance = 0.1498

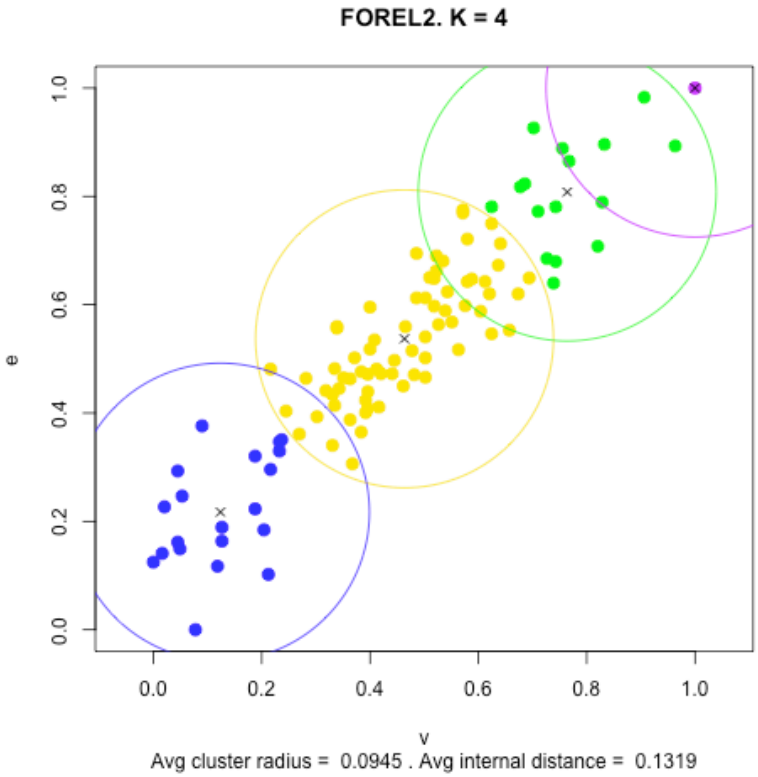

FOREL2. K = 5
e
v
0.0
0.2
0.4
0.6
0.8
1.0
Avg cluster radius = 0.0771 . Avg internal distance = 0.1068

R = 0.275

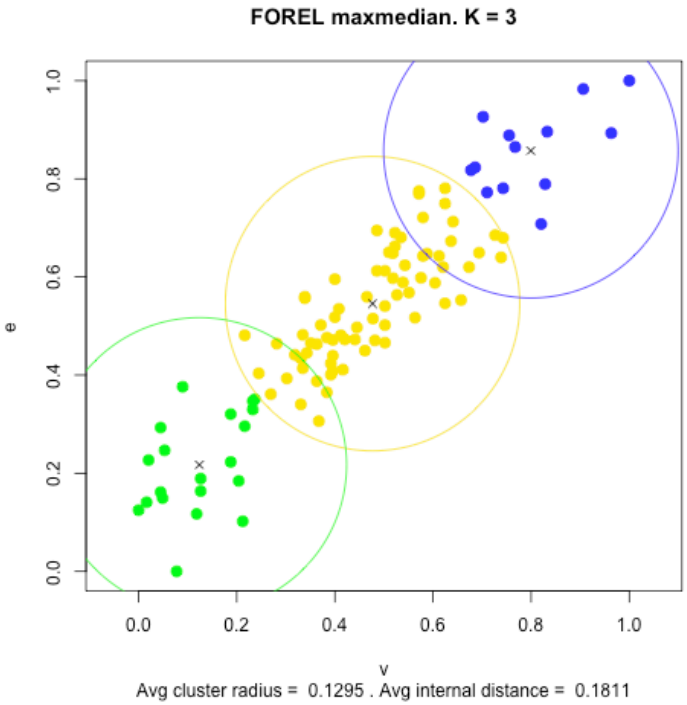

FOREL maxmedian. K = 4
e
v
0.0
0.2
0.4
0.6
0.8
1.0
Avg cluster radius = 0.0945 . Avg internal distance = 0.1319


FOREL minmedian. K = 4
e
v
0.0
0.2
0.4
0.6
0.8
1.0
Avg cluster radius = 0.0945 . Avg internal distance = 0.1319

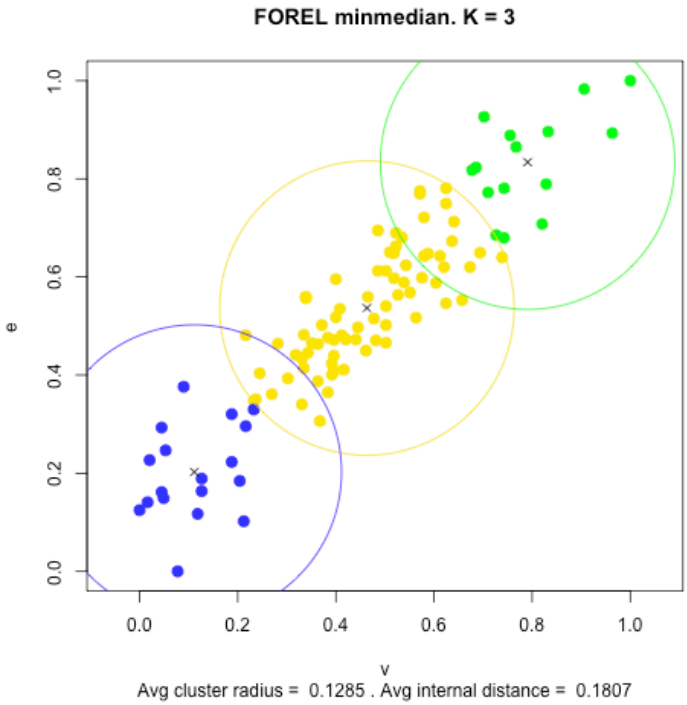

FOREL random. K = 4
e
v
0.0
0.2
0.4
0.6
0.8
1.0
Avg cluster radius = 0.092 . Avg internal distance = 0.1297


FOREL2. K = 4
e
v
0.0
0.2
0.4
0.6
0.8
1.0
Avg cluster radius = 0.0945 . Avg internal distance = 0.1319

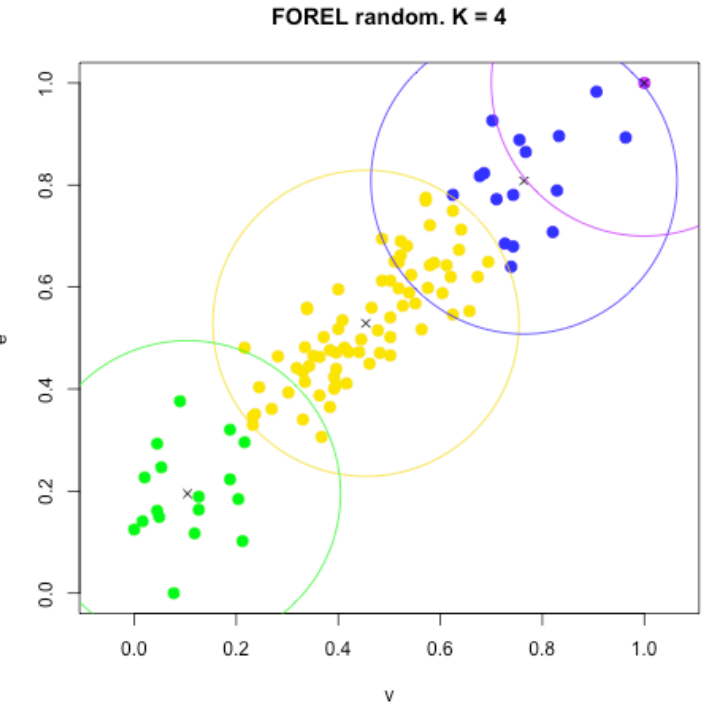
R = 0.3



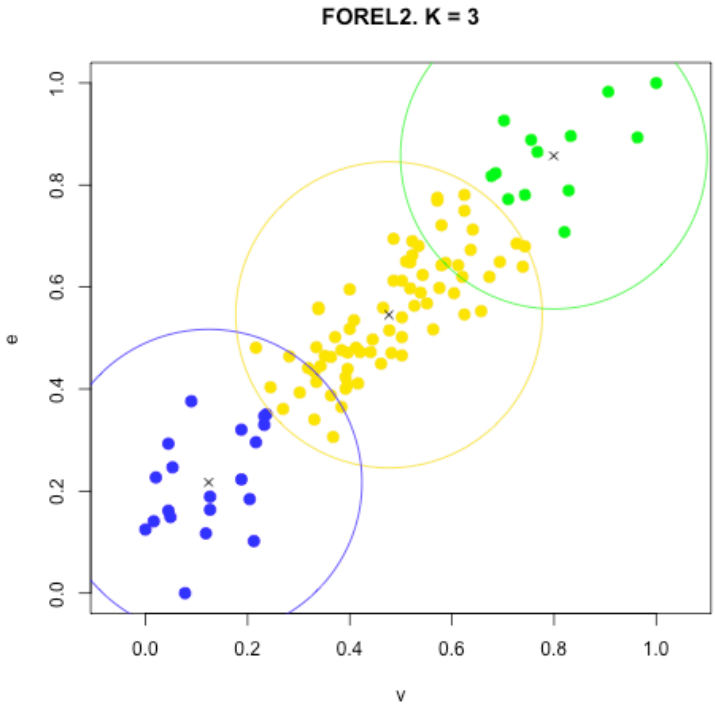
Avg cluster radius = 0.1295 . Avg internal distance = 0.1811



Avg cluster radius = 0.1285 . Avg internal distance = 0.1807

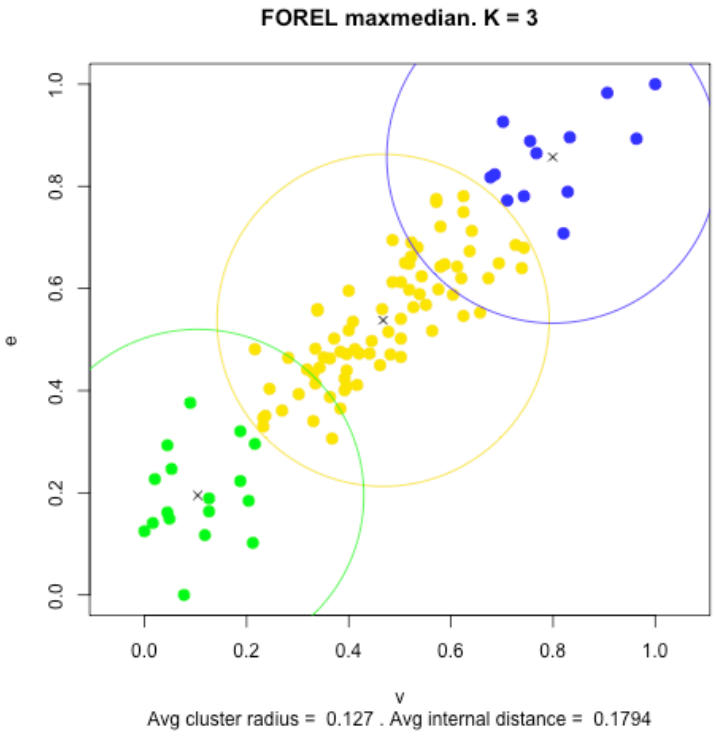


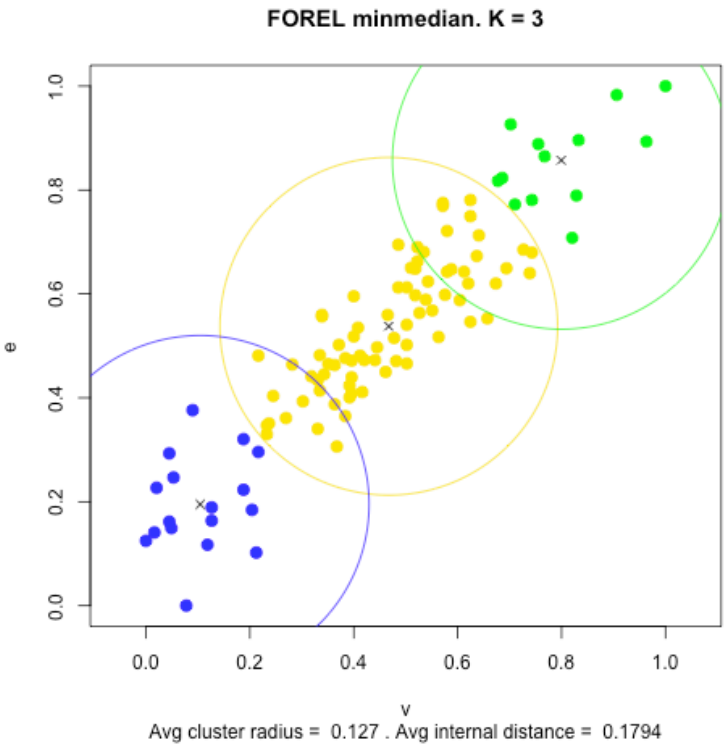
Avg cluster radius = 0.0926 . Avg internal distance = 0.1307

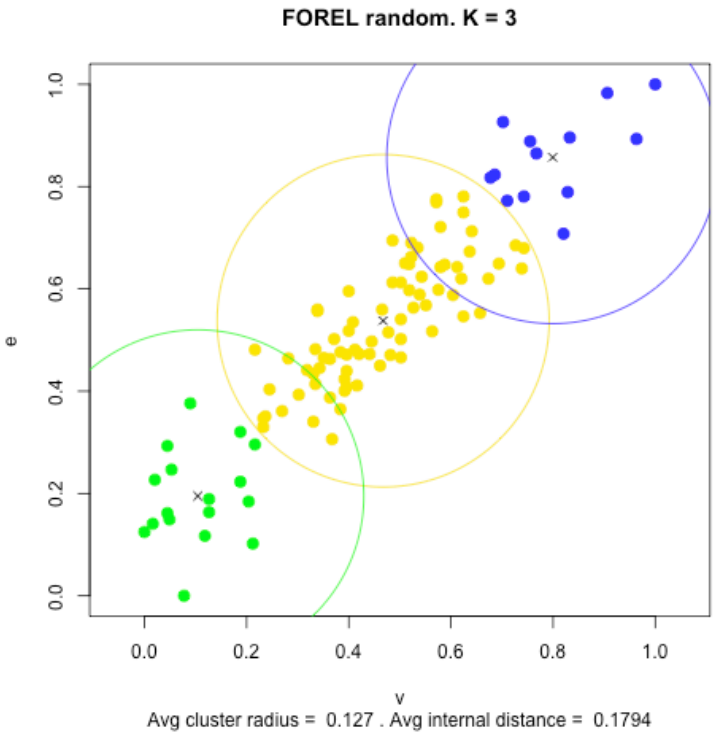


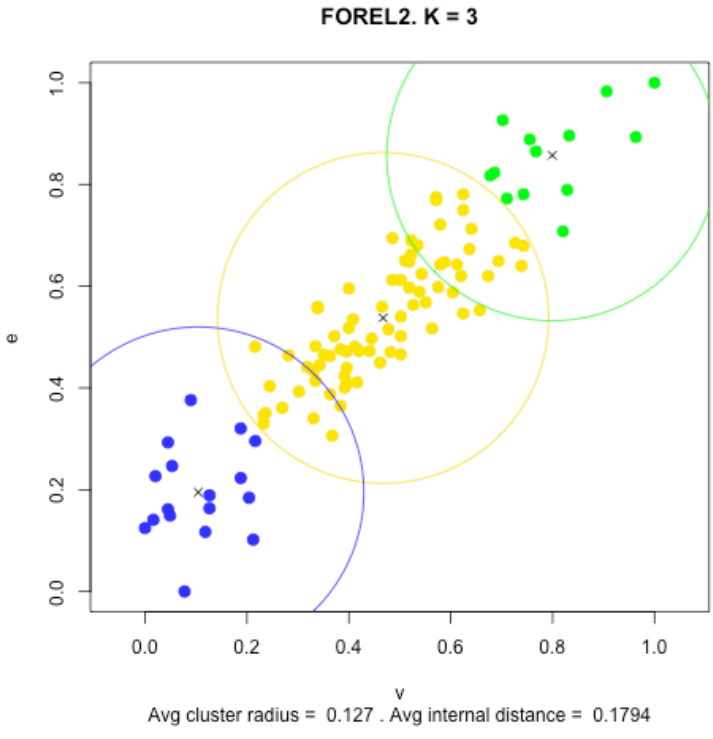
Avg cluster radius = 0.1295 . Avg internal distance = 0.1811

R = 0.325


FOREL maxmedian. K = 3
e
v
0.0
0.2
0.4
0.6
0.8
1.0
Avg cluster radius = 0.127 . Avg internal distance = 0.1794


FOREL minmedian. K = 3
e
v
0.0
0.2
0.4
0.6
0.8
1.0
Avg cluster radius = 0.127 . Avg internal distance = 0.1794


FOREL random. K = 3
e
v
0.0
0.2
0.4
0.6
0.8
1.0
Avg cluster radius = 0.127 . Avg internal distance = 0.1794


FOREL2. K = 3
e
v
0.0
0.2
0.4
0.6
0.8
1.0
Avg cluster radius = 0.127 . Avg internal distance = 0.1794

Сравнительная таблица

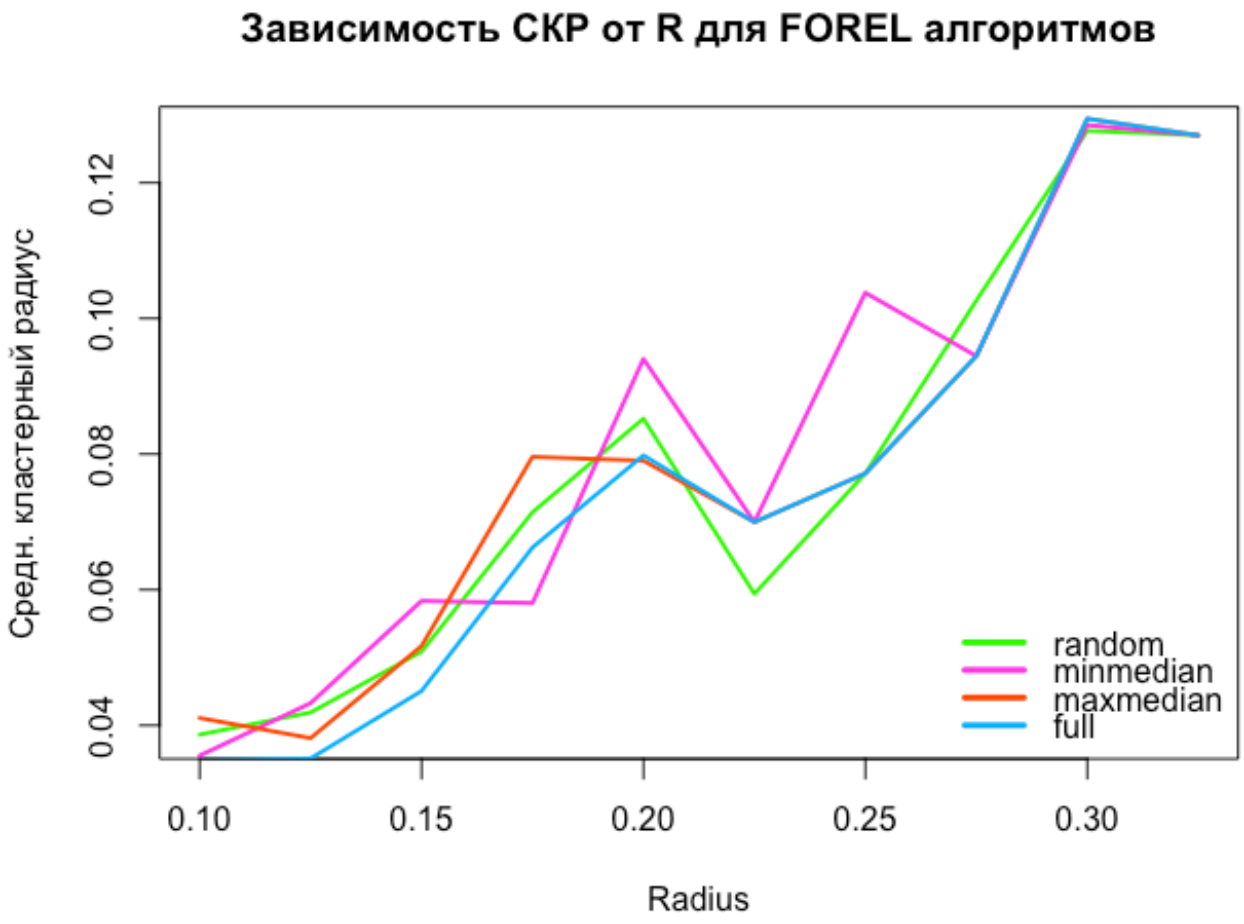
R	Средний кластерный радиус				Среднее внутрикластерное расстояние			
	Стандартный			С полным просмотром	Стандартный			С полным просмотром
	Maxmedian	Minmedian	Random		Maxmedian	Minmedian	Random	
0.1	0.04109562	0.03553614	0.03236344	0.03479994	0.06781942	0.05594195	0.05310429	0.05536840
0.125	0.03812300	0.04327640	0.03848928	0.03512750	0.06058621	0.06551348	0.06169720	0.05311393
0.15	0.05170254	0.05835503	0.05841075	0.04506971	0.07675329	0.08758054	0.08659528	0.06531378
0.175	0.07960202	0.05803598	0.05900465	0.06623173	0.12023622	0.08565217	0.08718532	0.09511398
0.2	0.07902859	0.09402532	0.08197344	0.07981205	0.11520496	0.13749269	0.12009480	0.11583243
0.225	0.06997255	0.06997255	0.05938874	0.06997255	0.09987377	0.09987377	0.08471683	0.09987377
0.25	0.07714720	0.10379381	0.10379381	0.07714720	0.10676493	0.14981927	0.14981927	0.10676493
0.275	0.09445128	0.09445128	0.09201983	0.09445128	0.13185564	0.13185564	0.12969719	0.13185564
0.3	0.12945532	0.12850745	0.09258588	0.12945532	0.18105053	0.18074903	0.13066532	0.18105053
0.325	0.12699336	0.12699336	0.12699336	0.12699336	0.17940961	0.17940961	0.17940961	0.17940961

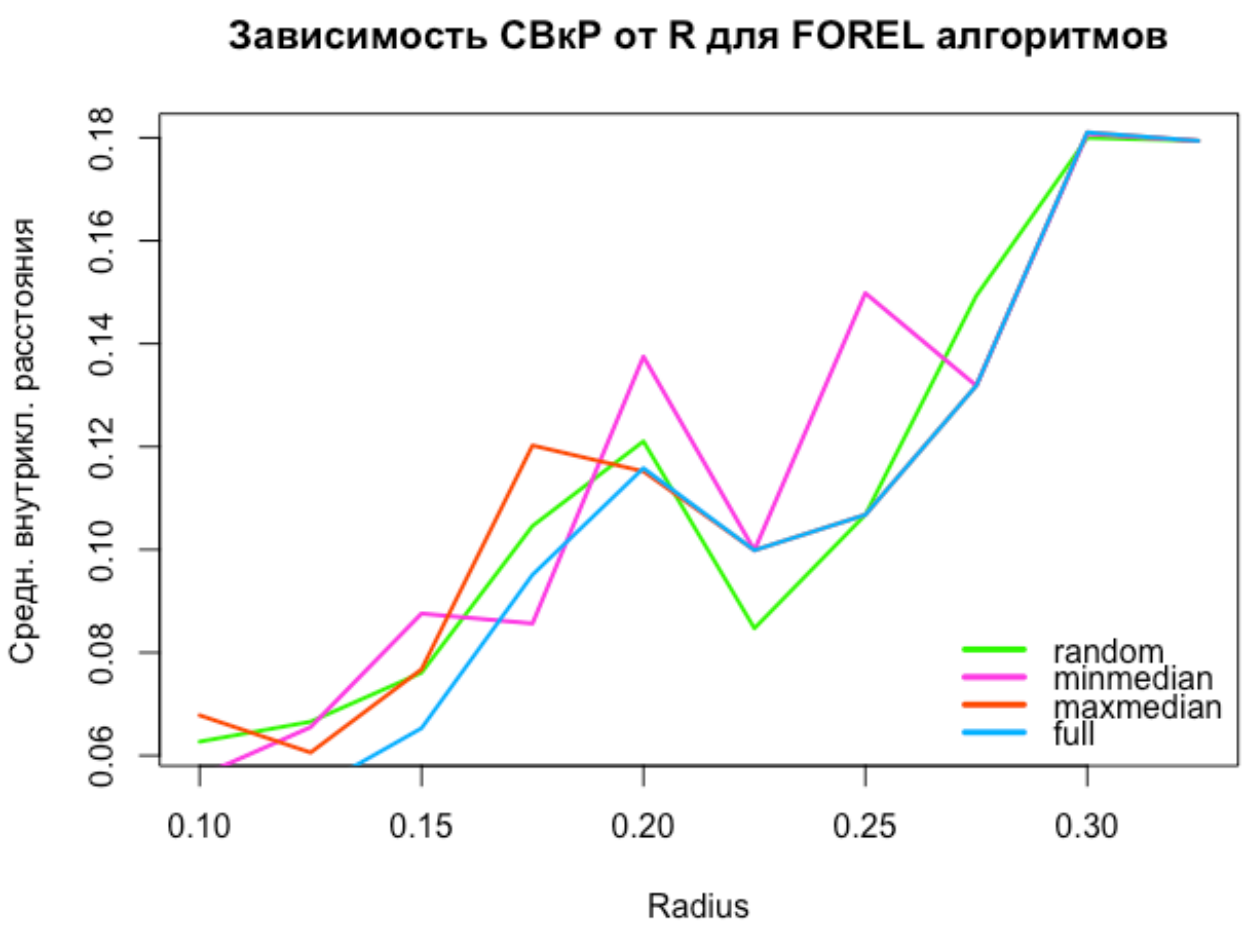
R	Количество кластеров			
	Стандартный			С полным просмотром
	Maxmedian	Minmedian	Random	
0.1	17	18	19	17
0.125	14	12	14	13
0.15	9	8	8	10

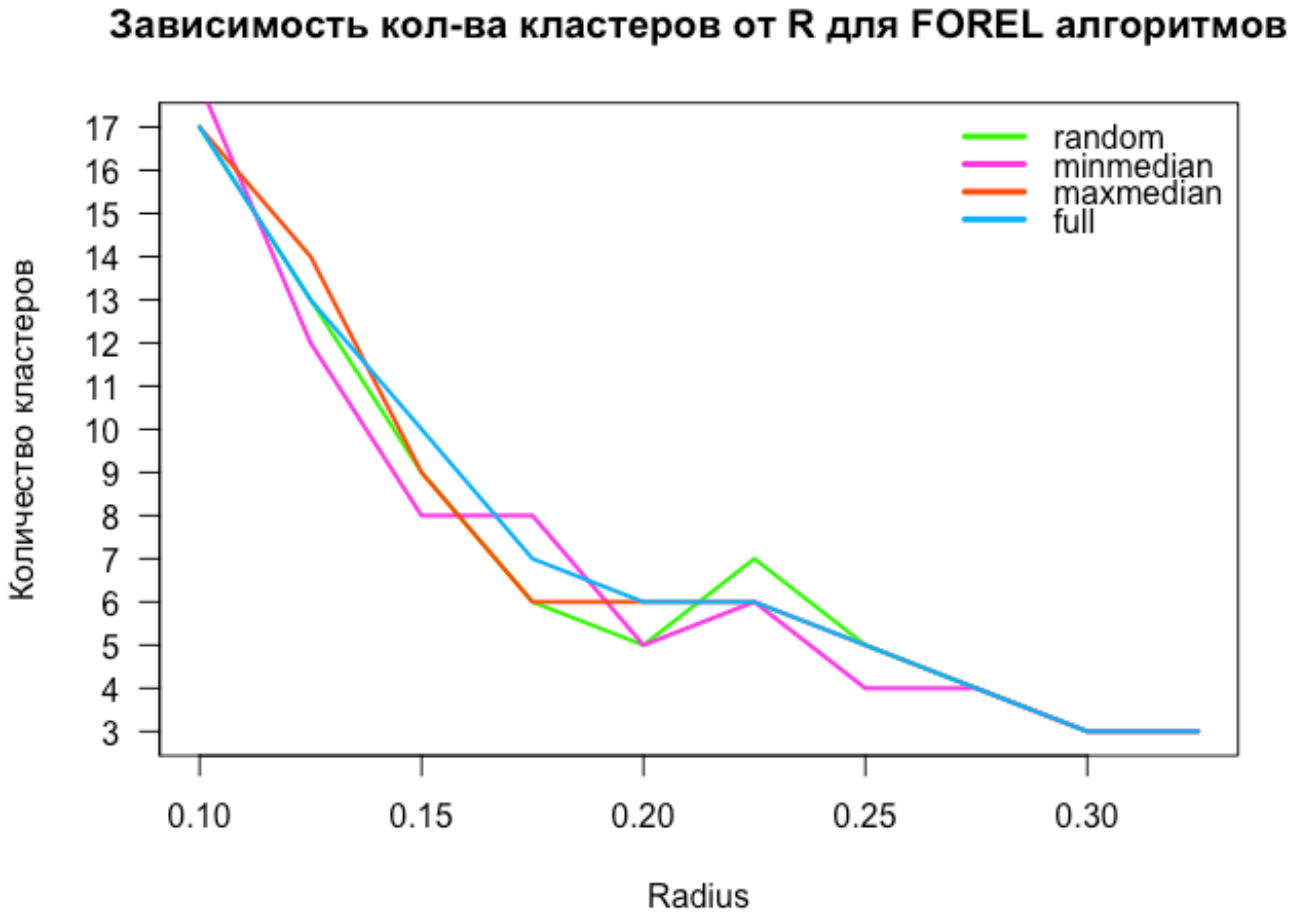
0.175	6	8	7	7
0.2	6	5	6	6
0.225	6	6	7	6
0.25	5	4	4	5
0.275	4	4	4	4
0.3	3	3	4	3
0.325	3	3	3	3

Для заданного радиуса в зависимости от способа выбора начальных приближений получается различное количество кластеров, в силу этого сложно оценивать качество полученных кластеров с помощью рассматриваемых характеристик.

Например, если рассмотреть три нижние строки в таблице, представленной выше, можно отметить, что, во-первых, количество полученных кластеров для всех четырех способов одинаково и равно трем, во-вторых, наилучший результат показала модификация «с полным просмотром» при заданном радиусе R = 0.325.




Зависимость СВкР от R для FOREL алгоритмов
Средн. внутрикл. расстояния
0.06
0.08
0.10
0.12
0.14
0.16
0.18
0.10
0.15
0.20
0.25
0.30
Radius
random
minmedian
maxmedian
full


Зависимость кол-ва кластеров от R для FOREL алгоритмов
Количество кластеров
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
0.10
0.15
0.20
0.25
0.30
Radius
random
minmedian
maxmedian
full

с. Результаты второго эксперимента.

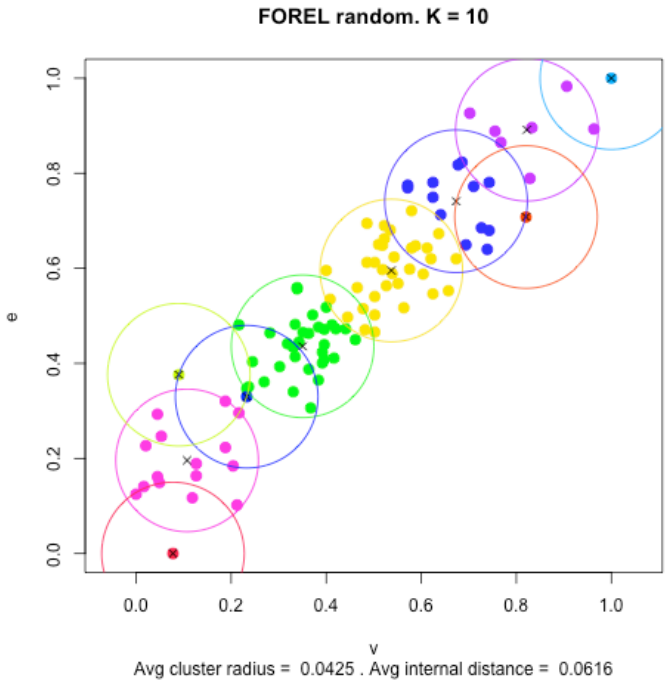
Значение радиуса было зафиксировано на уровне $R = 0.15$.

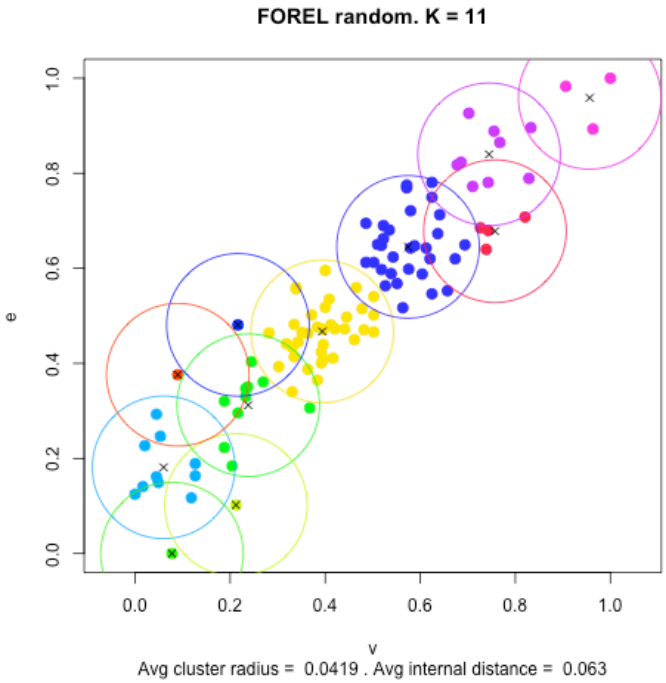
В этом эксперименте стандартный вариант алгоритма был запущен 10 раз. В результате получилось 10 различных разбиений на кластеры. Начальные приближения выбирались случайно. Далее представлена таблица с характеристиками разбиения.

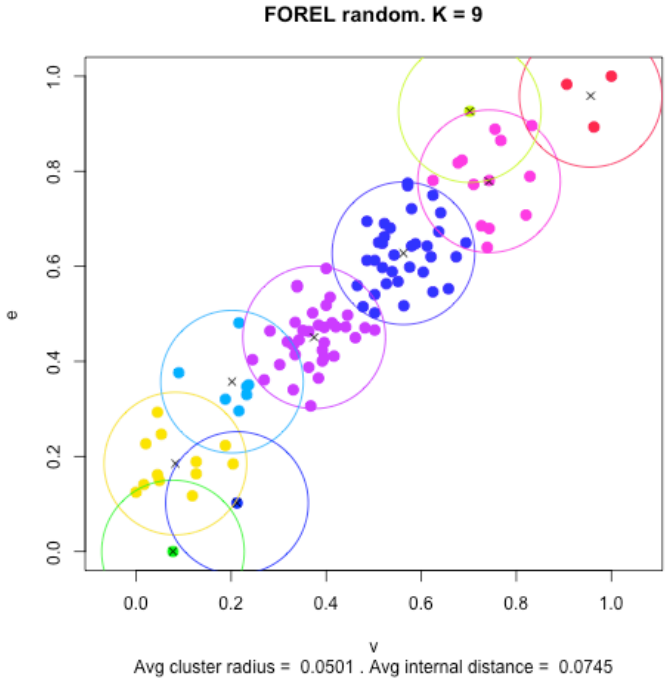
Получившиеся разбиения можно поделить на группы по количеству выделенных кластеров (K). На графиках представлены разбиения с лучшими характеристиками в своей группе.

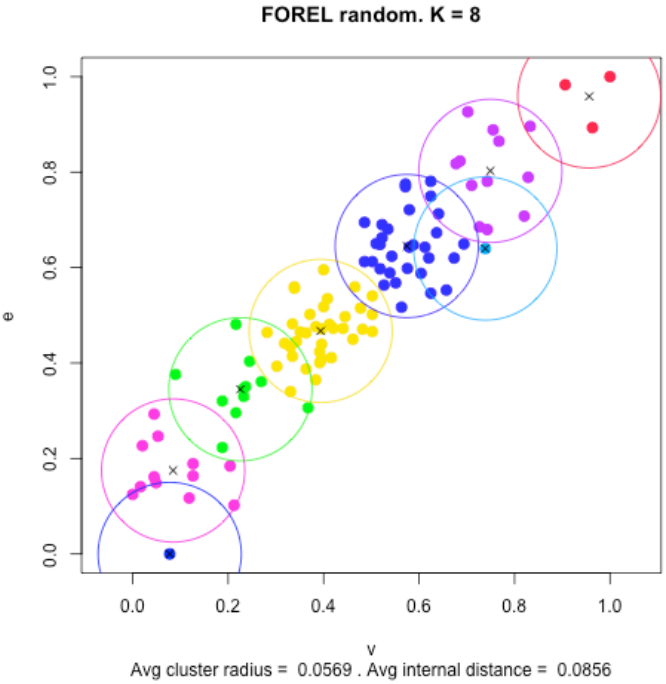
Сравнительная таблица.

№	Средний кластерный радиус	Среднее внутрикластерное расстояние	Количество кластеров
1	0.05252209	0.07770190	9
2	0.05382640	0.07967489	9
3	0.04245416	0.06156258	10
4	0.05841075	0.08659528	8
5	0.05054584	0.07553307	9
6	0.05072532	0.07607701	9
7	0.05087707	0.07609360	9
8	0.04185156	0.06298835	11
9	0.05013994	0.07454487	9
10	0.05688853	0.08563213	8


FOREL random. K = 10
e
v
0.0
0.2
0.4
0.6
0.8
1.0
Avg cluster radius = 0.0425 . Avg internal distance = 0.0616


FOREL random. K = 11
e
v
0.0
0.2
0.4
0.6
0.8
1.0
Avg cluster radius = 0.0419 . Avg internal distance = 0.063


FOREL random. K = 9
e
v
0.0
0.2
0.4
0.6
0.8
1.0
Avg cluster radius = 0.0501 . Avg internal distance = 0.0745


FOREL random. K = 8
e
v
0.0
0.2
0.4
0.6
0.8
1.0
Avg cluster radius = 0.0569 . Avg internal distance = 0.0856

Выводы.

В ходе лабораторной работы был реализован алгоритм поиска сгущений (ForEl) для кластеризации точек на двумерной плоскости в двух модификациях: «стандартной» и «с полным просмотром».

Для «стандартного» варианта было реализовано три способа выбора начальных приближений.

Было проведено два набора экспериментов: в первом варьировался радиус и способы поиска кластеров сравнивались между собой, во втором для фиксированного радиуса и случайных начальных приближений несколько раз запускался «стандартный» алгоритм для определения надежности характеристик полученного разбиения.

По результатам проведенного исследования можно сделать вывод, что при использовании метода поиска сгущений, значения среднего кластерного расстояния и среднего кластерного радиуса достигаются в случае успешного нахождения плотных зон (сгущений). Однако, при таком разбиении остаются точки, образующие кластеры нулевого радиуса и с нулевым кластерным расстоянием.

Также полученные значения среднего внутрикластерного расстояния и среднего кластерного радиуса довольно устойчивы, при фиксации радиуса кластера.

Основной задачей работы является визуализация результатов работы алгоритма. Результаты представлены в тексте работы в виде графиков и таблиц.