Институт системной социологии

Изучение социальной структуры России с помощью кластерного анализа базы RLMS-HSE

Алексеев Алексей
физический факультет МГУ им. М.В. Ломоносова

Евстюшева Екатерина
факультет архивного дела РГГУ

МГУ имени М.В. Ломоносова
13 марта 2018 г.

План доклада

- Введение. Различные подходы к стратификации.
- Выбор критериев стратификации
- База данных RLMS-HSE
- Выбор вопросов и метода количественной оценки
- Кластеризация
  - Селекция данных
  - Нормализация данных
  - Метод кластеризации
  - Поиск наилучшего количества кластеров
- Результаты
- Сравнение с данными Росстата

Введение. Различные подходы к стратификации

В.И.Ленин

Определил понятия класса в марксистской теории (работа “Великий почин”)

Джон Голдторп

Новатор многомерной стратификации, фактор степеня контроля (власти)

Н.Е. Тихонова

Теоретик ресурсного подхода к стратификаци

Пьер Бурдье

Расширил понятие “капитал” в социологии

Выбор критериев стратификации

	Критерий	Ленин (теория)	Тихонова (теория)	Голдторп (теория)	Бобровский (расчёты)	Данная работа (расчёты)
1	Место в общественном производстве	есть				
2	Отношение к средствам производства	есть		есть		есть
3	Роль в организации труда	есть		есть	есть	есть
4	Способ получения общественного богатства	есть				
5	Размер получаемой части общественного богатства	есть	есть		есть	есть
6	Доступ к ресурсам общества		есть		есть	есть
7	Субъективная оценка общественного положения		есть			есть
8	Характер труда			есть		
9	Уровень и специфика образования		есть	есть	есть	есть
10	Степень социальной защищенности					

Выбор вопросов и метод количественной оценки

Выбранные критерии

- Близость к предпринимательству (k2)
- Количество подчинённых (k3)
- Зарплата за месяц (k5)
- Доступность важных общественных благ (k6)
- Субъективная оценка своего общественного положения (k7)
- Уровень образования (k9)

Выбор вопросов

ID	Критерий	Вопрос
j26	Отношение к средствам производства	А Вы лично являетесь владельцем или совладельцем предприятия, на котором Вы работаете?
j29	Отношение к средствам производства	Как Вы считаете, на этой работе Вы занимаетесь предпринимательской деятельностью?
j6.0	Роль в организации труда	Сколько у Вас подчиненных? Пожалуйста, посчитайте всех Ваших подчиненных, а не только тех, кто находится в Вашем непосредственном подчинении
j10	Размер получаемой части общественного богатства	Месячная зарплата
j721631	Доступ к ресурсам общества	Имеете ли Вы или Ваша семья возможность при желании улучшить свои жилищные условия - купить комнату, квартиру, дом?
j721632	Доступ к ресурсам общества	Имеете ли Вы или Ваша семья возможность при желании оплачивать дополнительные занятия детей - музыкальную школу, иностранные языки, спортивные секции, кружки и т.п.?
j721633	Доступ к ресурсам общества	Имеете ли Вы или Ваша семья возможность при желании откладывать деньги на крупные покупки - машину, дачу?
j721634	Доступ к ресурсам общества	Имеете ли Вы или Ваша семья возможность при желании провести всей семьей отпуск за границей?
j721636	Доступ к ресурсам общества	Имеете ли Вы или Ваша семья возможность при желании провести всей семьей отпуск на российском курорте?
j721635	Доступ к ресурсам общества	Имеете ли Вы или Ваша семья возможность при желании оплачивать учебу ребенка в ВУЗе?
I2.2	Доступ к ресурсам общества	У Вас есть договор на доп. добровольное медицинское страхование, обслуживание с какой-нибудь страховой фирмой, поликлиникой, больницей, медицинским центром? Не учитывайте полисы ОМС, полисы для выезжающих за границу, полисы страхования от клеща и т. п.
j62	Субъективная оценка общественного положения	Представьте себе лестницу из 9 ступеней, где на нижней, первой ступени, стоят нищие, а на высшей, девятой - богатые. На какой из девяти ступеней находитесь сегодня Вы лично?
j63	Субъективная оценка общественного положения	Представьте себе лестницу, из 9 ступеней, где на нижней ступени стоят совсем бесправные, а на высшей - те, у кого большая власть. На какой из девяти ступеней находитесь сегодня Вы лично?
EDUC	Уровень и специфика образования	ОБРАЗОВАНИЕ (ПОДРОБНО): старше 14 лет - 25 ВОЛНА

Селекция данных

- Отобраны данные 2015 и 2016 волны RLMS (индивидуальные, не по домохозяйствам)
- Объединены данные двух лет (в случае повторного опроса того же человека, отбираются более новые данные) — 20754 строки
- Отбираются только строки, где респондент указал величину зарплаты
- Отбираются респонденты с зарплатой >2000 р.
- Остаётся 8357 строк, данные по всем регионам России

Способ количественной оценки

- Численный показатель по каждому критерию вычисляется как сумма показателей по всем вопросам, относящимся к данному критерию
- Примеры определения численного показателя исходя из ответа опросника:

Имеете ли Вы или Ваша семья возможность при желании улучшить свои жилищные условия - купить комнату, квартиру, дом?

j721631	1		0
j721631	2	Да	1
j721631	3	ЗАТРУДНЯЮСЬ ОТВЕТИТЬ	0
j721631	4	Нет	-1
j721631	5	НЕТ ОТВЕТА	0
j721631	6	ОТКАЗ ОТ ОТВЕТА	0

Способ количественной оценки

Численная оценка показателя образования человека

	0
0 классов школы	0
ЗАТРУДНЯЮСЬ ОТВЕТИТЬ	0
НЕТ ОТВЕТА	0
1 класс школы	1
2 класса школы	2
3 класса школы	3
4 класса школы	4
5 классов школы	5
6 классов школы	6
7 классов школы	7
8 классов школы	8
9 классов школы	9
10 и более классов школы без аттестата о среднем образовании	10
среднее образование - есть аттестат о ср. образовании	10
10 и более классов школы и какое-либо професс. обр. без диплома	11
10 и более классов школы и техникум без диплома	11
7-9 классов школы (незак. среднее) и менее 2 лет в техникуме	11
7-9 классов школы (незак. средн) + ПТУ без диплома	11
10 и более классов школы и какое-либо професс. обр. с дипломом	12
7-9 классов школы (незак. средн) + ПТУ с дипломом	12
техникум с дипломом	12
1-2 года в высшем учебном заведении	13
3 и более лет в высшем учебном заведении	14
есть диплом о высшем образовании	15
аспирантура и т.п. без диплома	16
аспирантура и т.п. с дипломом	17

Исходный набор для кластеризации. Нормировка.

k2	k3	k5	k6	k7	k9
0	0	60000	1	15	13
0	0	20000	1	11	13
0	0	35000	1	15	13
0	14	35000	1	17	13
0	0	14230	1	13	13
0	0	54100	1	14	13
0	0	20000	1	12	13
0	11	34000	1	15	13
0	10	25000	1	14	13
0	0	28000	1	17	13
0	0	21600	1	14	13
0	11	42000	1	18	13
0	1	38000	1	20	13
0	14	45000	1	19	13
0	0	18000	1	10	13
0	0	52000	1	10	13
0	1	15000	1	15	13

Каждый элемент стобца преобразуется по формуле

$$X` = (x - mi) / sdi$$

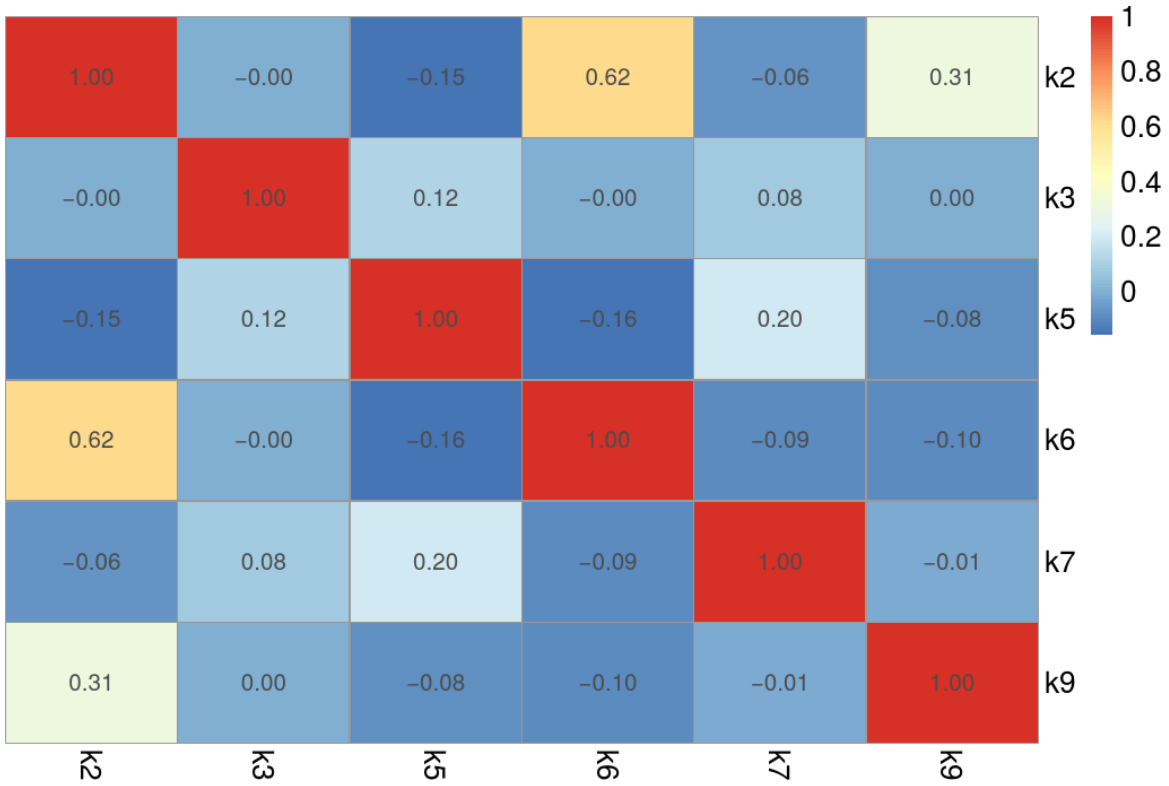
где mi – среднее арифметическое по i-му столбцу, sdi – стандартное отклонение по i-му столбцу

k2	k3	k5	k6	k7	k9
0.6818223	-0.17277582	1.772682309	1.2151	0.265019627	0.1394972
0.6818223	-0.17277582	-0.255371885	1.2151	-0.804142345	0.1394972
0.6818223	-0.17277582	0.505148438	1.2151	0.265019627	0.1394972
0.6818223	0.75936591	0.505148438	1.2151	0.799600614	0.1394972
0.6818223	-0.17277582	-0.547918702	1.2151	-0.269561359	0.1394972
0.6818223	-0.17277582	1.473544315	1.2151	-0.002270866	0.1394972
0.6818223	-0.17277582	-0.255371885	1.2151	-0.536851852	0.1394972
0.6818223	0.55962125	0.454447083	1.2151	0.265019627	0.1394972
0.6818223	0.49303970	-0.001865111	1.2151	-0.002270866	0.1394972
0.6818223	-0.17277582	0.150238954	1.2151	0.799600614	0.1394972
0.6818223	-0.17277582	-0.174249717	1.2151	-0.002270866	0.1394972
0.6818223	0.55962125	0.860057922	1.2151	1.066891107	0.1394972

Предварительный анализ данных

Выбранные критерии

- Близость к предпринимательству (k2)
- Количество подчинённых (k3)
- Зарплата за месяц (k5)
- Доступность важных общественных благ (k6)
- Субъективная оценка своего общественного положения (k7)
- Уровень образования (k9)



Положительная корреляция

- k2-k9
- k2-k6
- k5-k7

Отрицательная корреляция

- k5-k2
- k5-k6

Вывод — блага связаны с доходом, а не с зарплатой, но доход мы пока оценивать не умеем

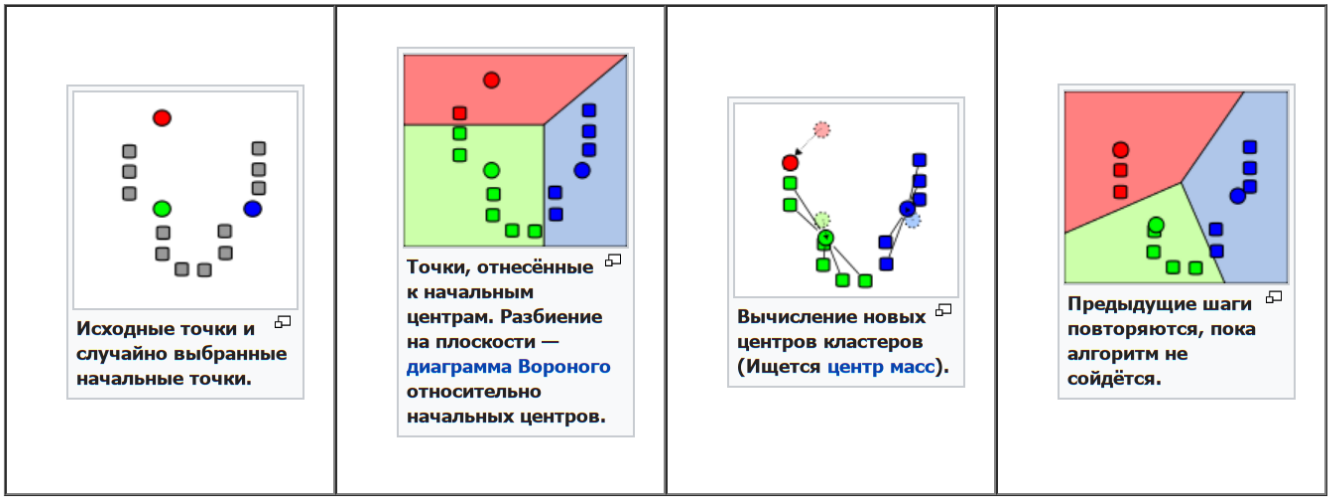
Метод кластеризации K-средних

Метод k-средних (англ. k-means) — один из самых популярных метод кластеризации. Был изобретён в 1950-х годах математиком Гуго Штейнгаузом и почти одновременно Стюартом Ллойдом. Особую популярность приобрёл после работы Маккуина.

Действие алгоритма таково, что он стремится минимизировать суммарное квадратичное отклонение точек кластеров от центров этих кластеров:

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

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

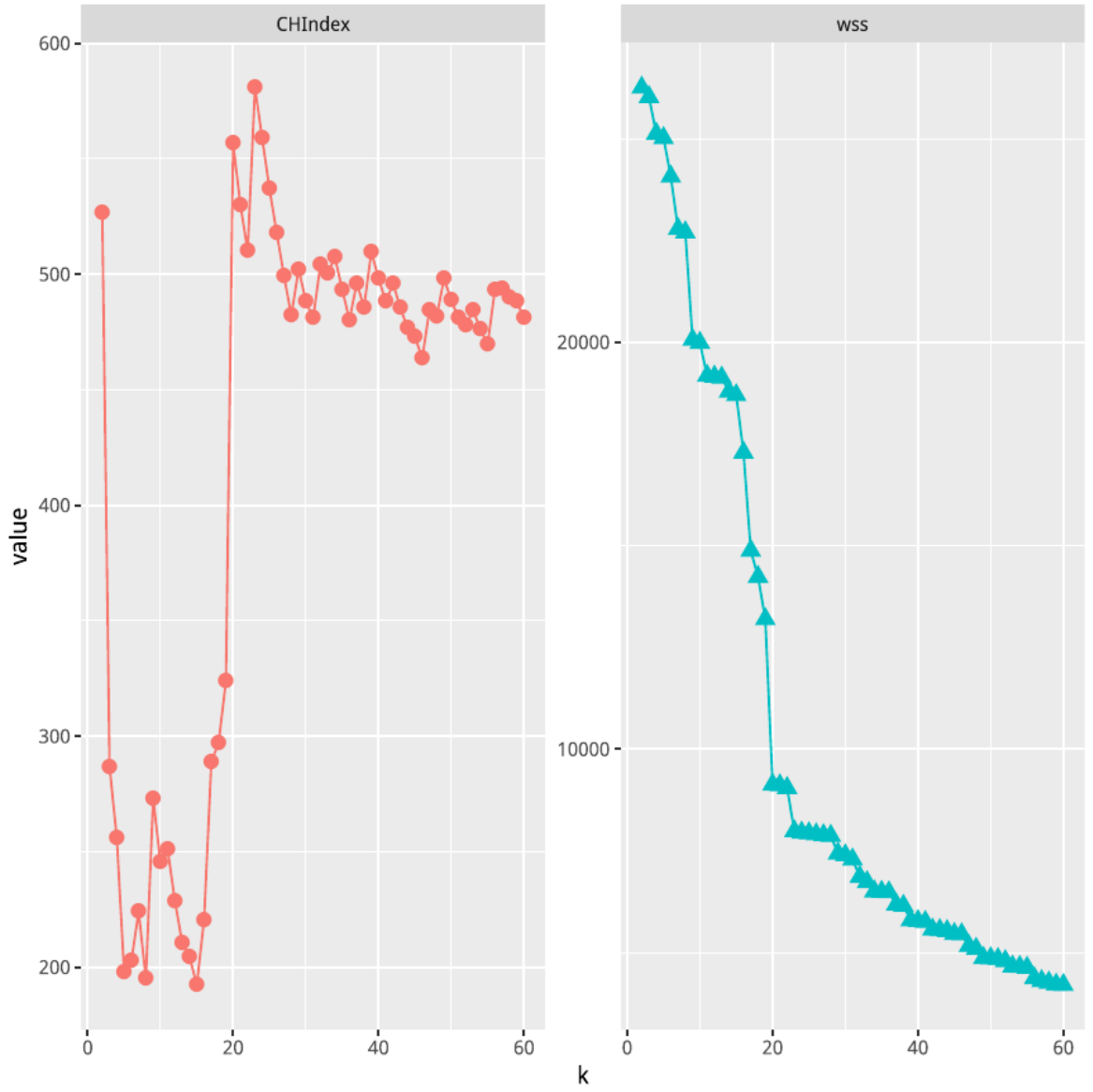


Steinhaus H. (1956). Sur la division des corps materiels en parties. Bull. Acad. Polon. Sci., C1. III vol IV: 801—804.

Lloyd S. (1957). Least square quantization in PCM's. Bell Telephone Laboratories Paper.

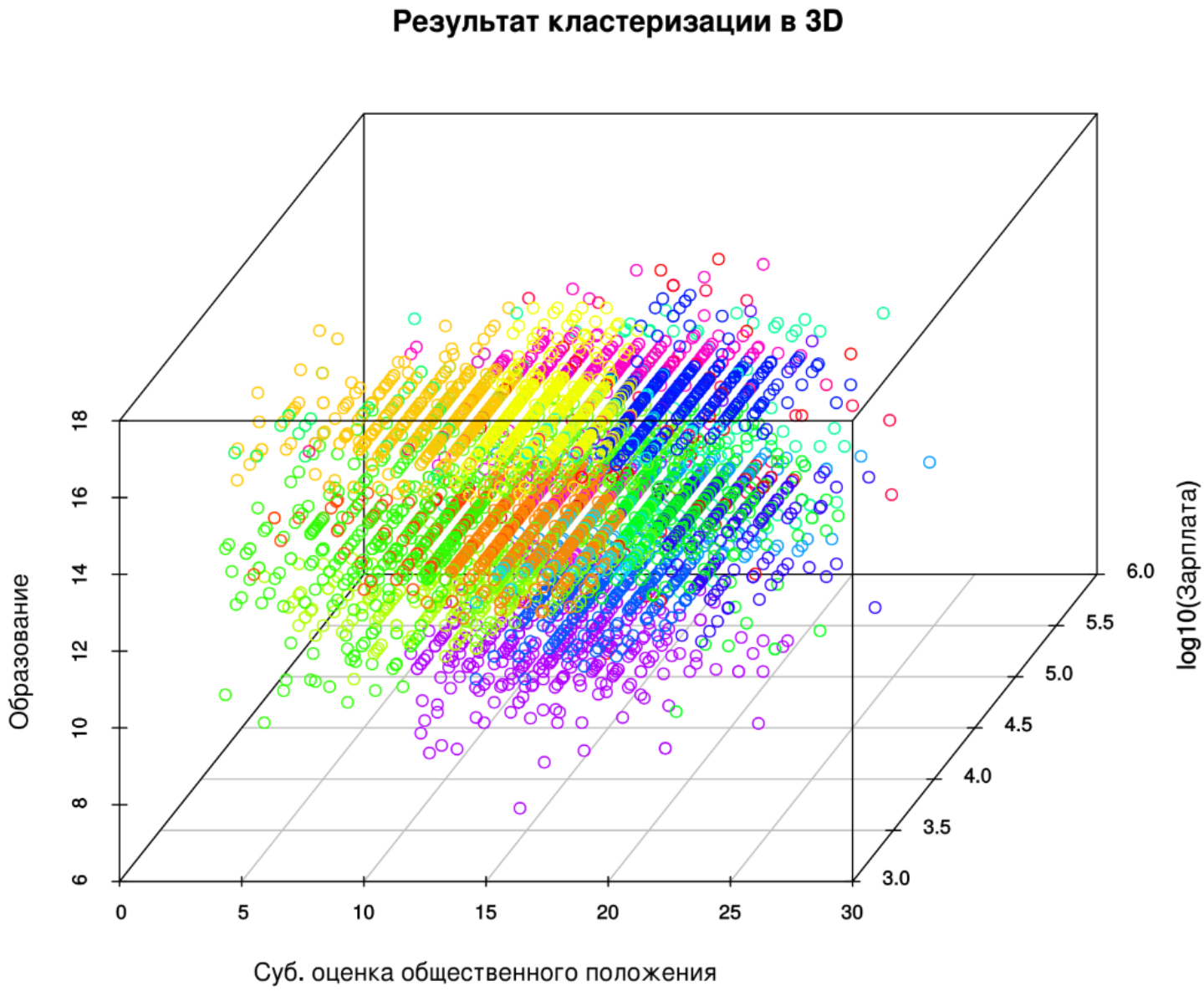
MacQueen J. (1967). Some methods for classification and analysis of multivariate observations. In Proc. 5th Berkeley Symp. on Math. Statistics and Probability, pages 281—297.

Выбор количества кластеров. Calinski-Harabasz Index .

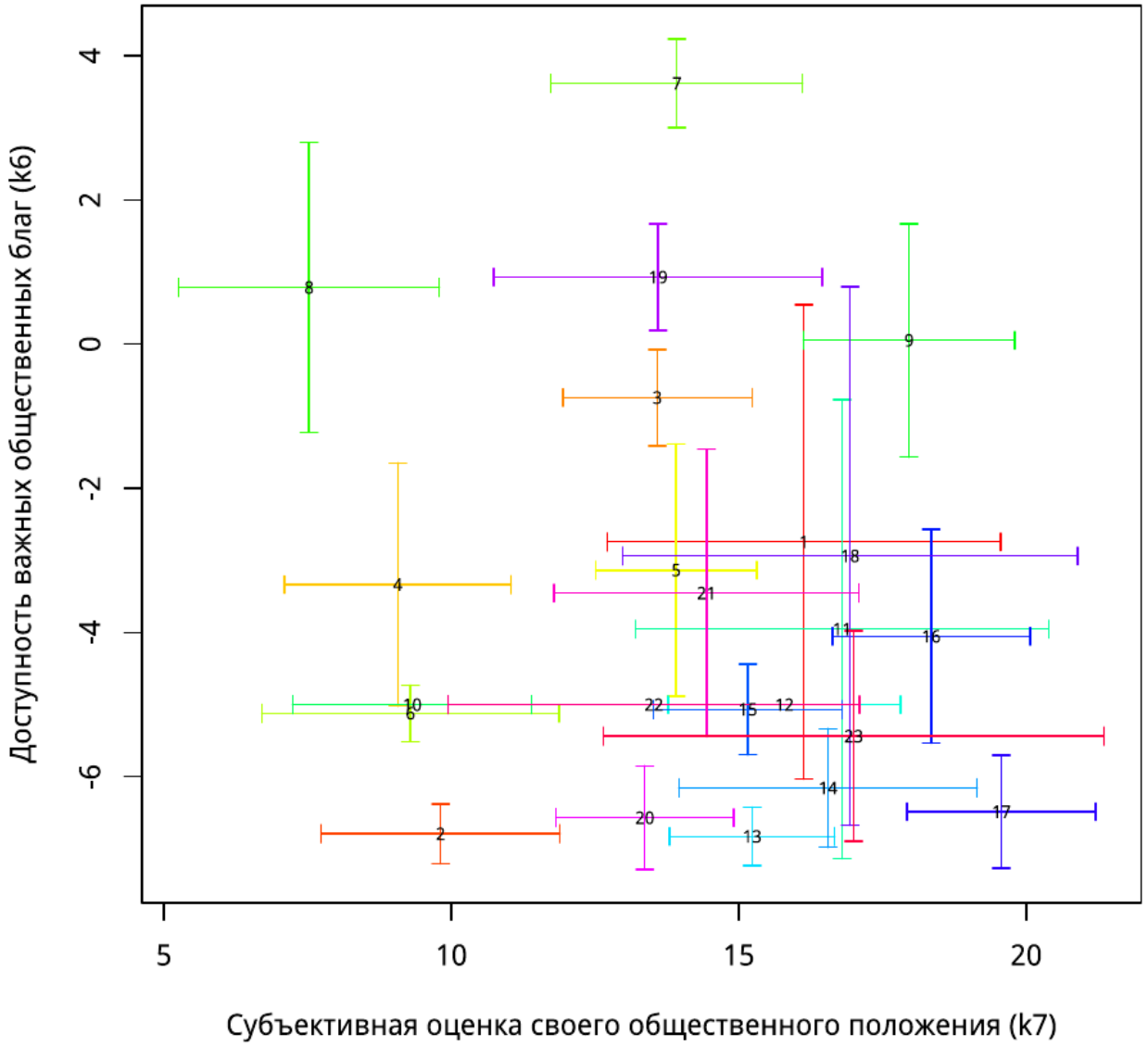


CH-index достигает максимума при k=23

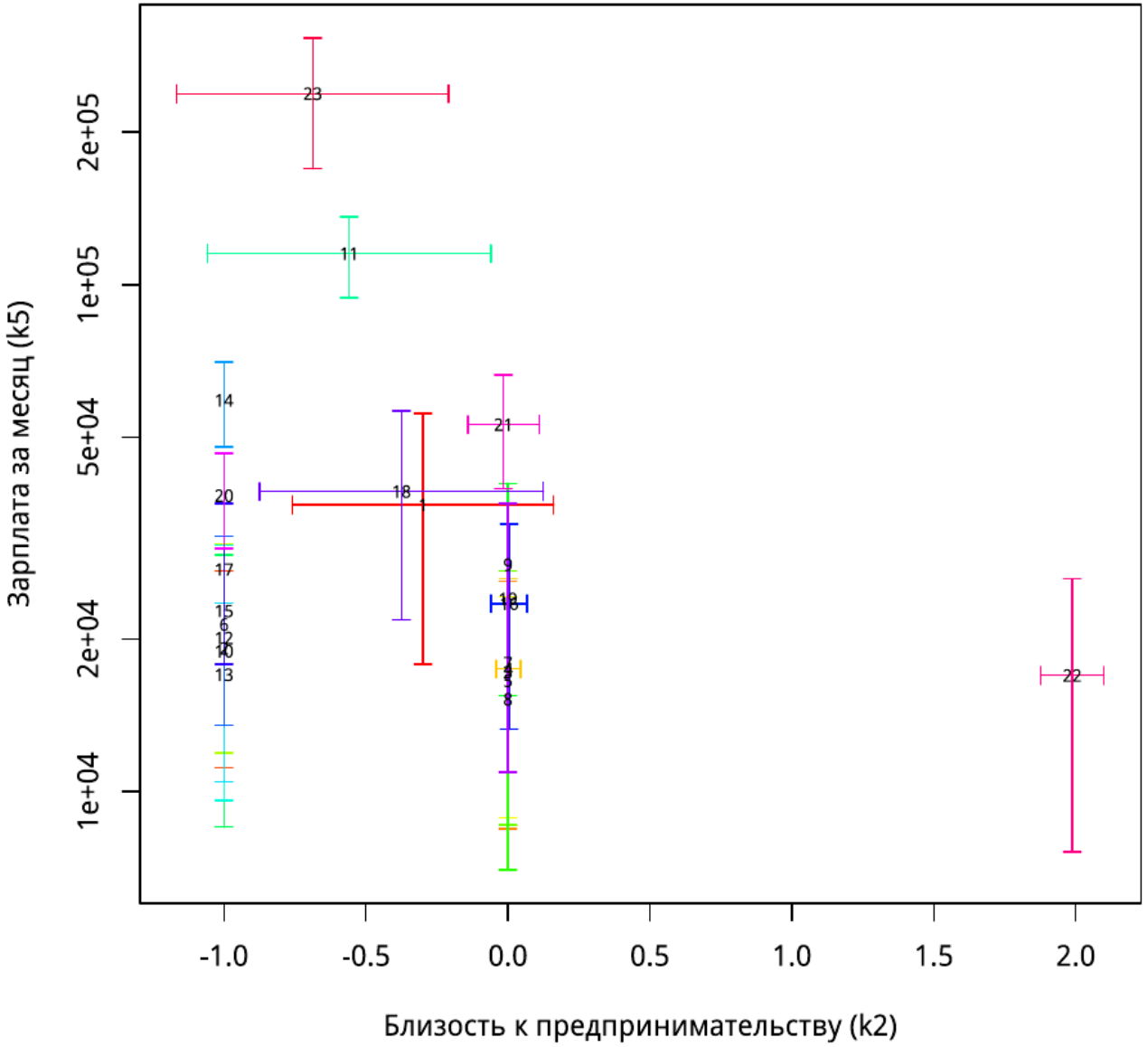
Результаты


Результат кластеризации в 3D
Образование
6
8
10
12
14
16
18
0
5
10
15
20
25
30
Суб. оценка общественного положения
3.0
3.5
4.0
4.5
5.0
5.5
6.0
log10(Зарплата)

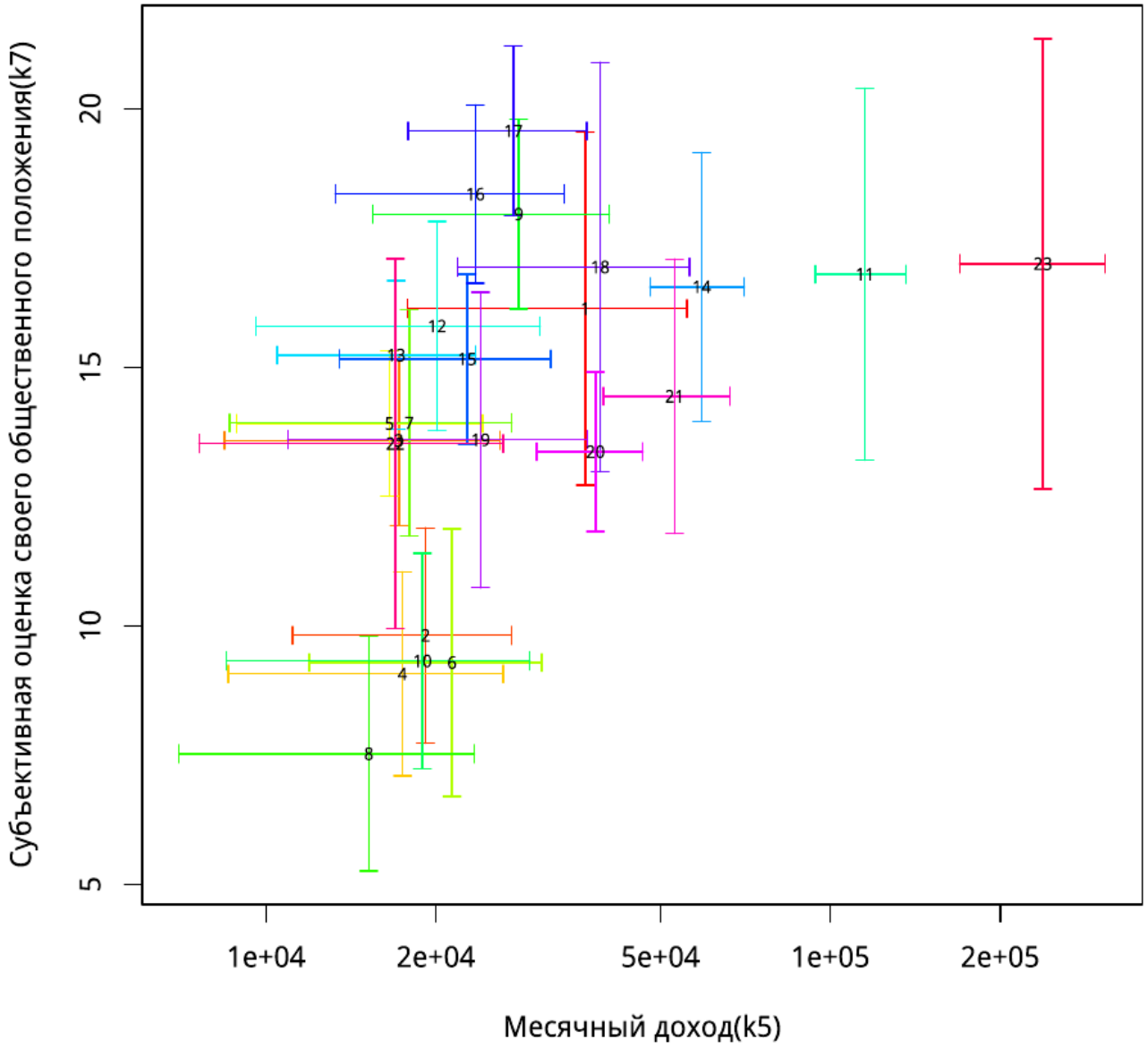
Результаты


Доступность важных общественных благ (k6)
4
2
0
-2
-4
-6
5
10
15
20
Субъективная оценка своего общественного положения (k7)
7
8
19
9
3
1
18
5
4
21
11
16
10
6
22
15
12
23
14
20
17
2
13

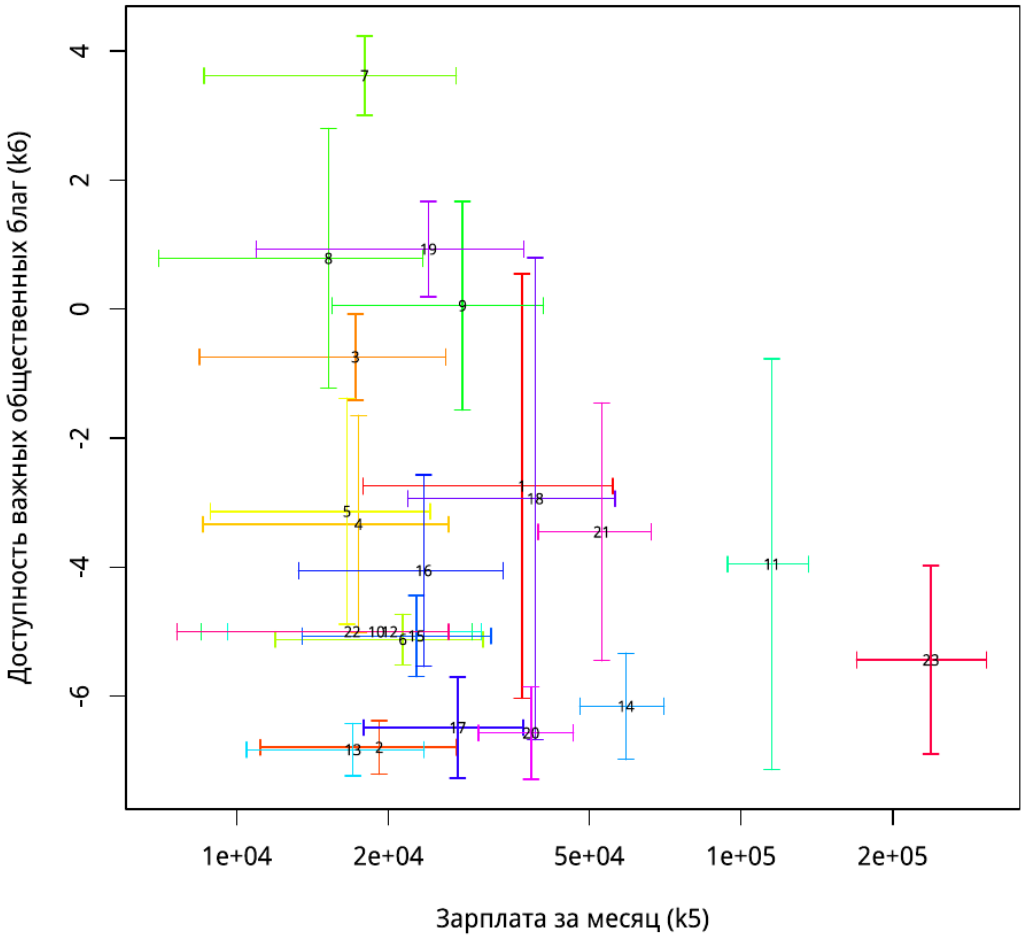
Результаты

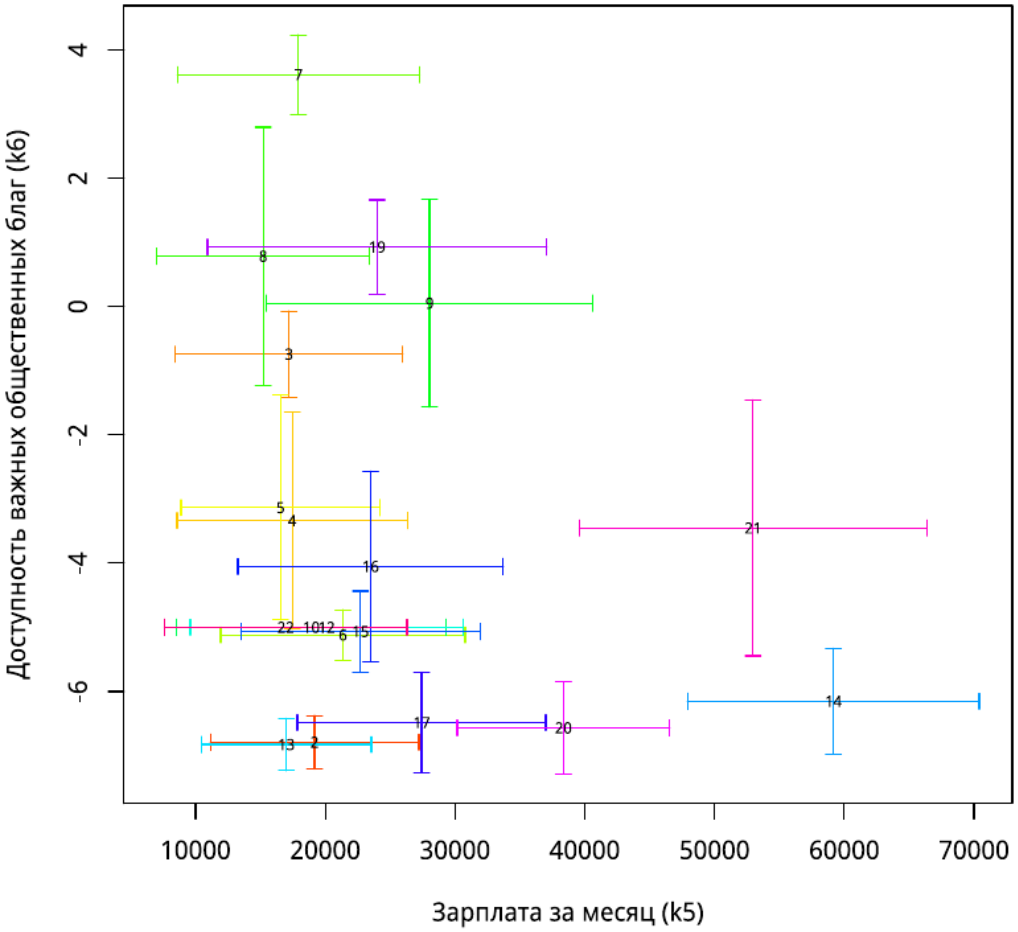

Зарплата за месяц (k5)
1e+04
2e+04
5e+04
1e+05
2e+05
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
Близость к предпринимательству (k2)

Результаты

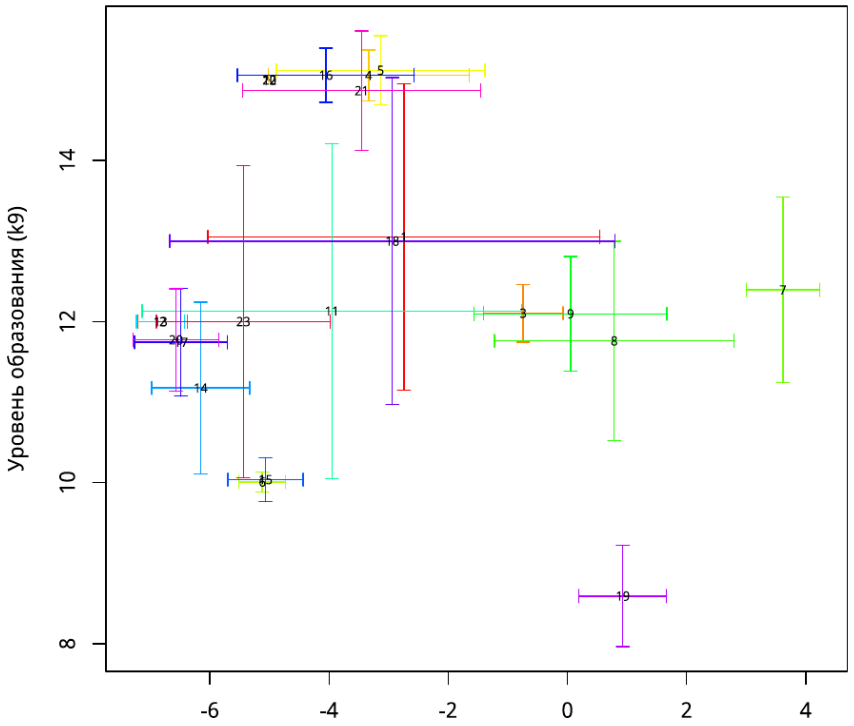

Субъективная оценка своего общественного положения(k7)
5
10
15
20
1e+04
2e+04
5e+04
1e+05
2e+05
Месячный доход(k5)
17
16
9
18
14
11
23
1
12
13
15
21
5
7
19
20
2
10
6
4
8

Результаты

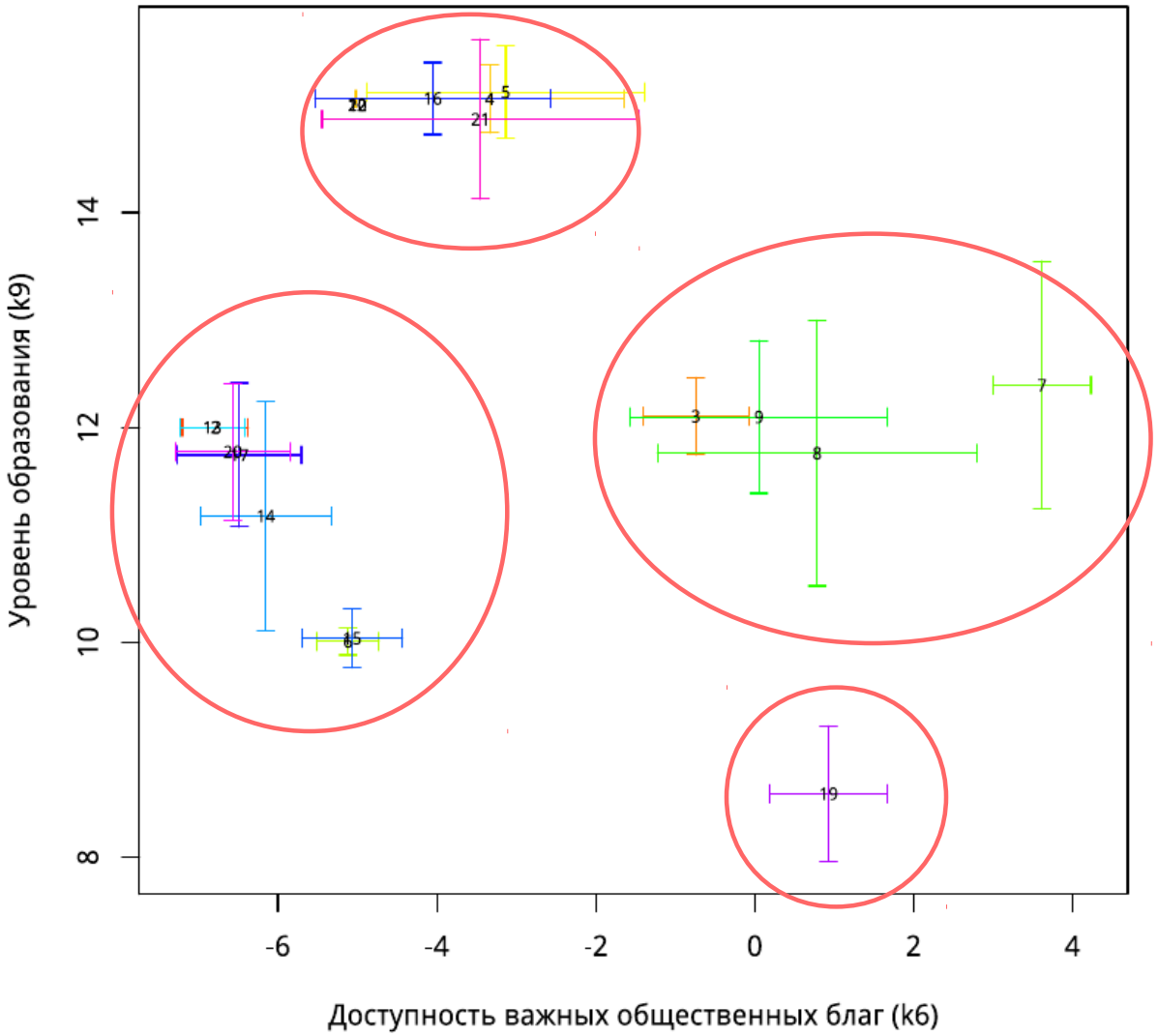

Доступность важных общественных благ (k6)
4
2
0
-2
-4
-6
1e+04
2e+04
5e+04
1e+05
2e+05
Зарплата за месяц (k5)


Доступность важных общественных благ (k6)
4
2
0
-2
-4
-6
10000
20000
30000
40000
50000
60000
70000
Зарплата за месяц (k5)

Результаты



Все 23 кластера



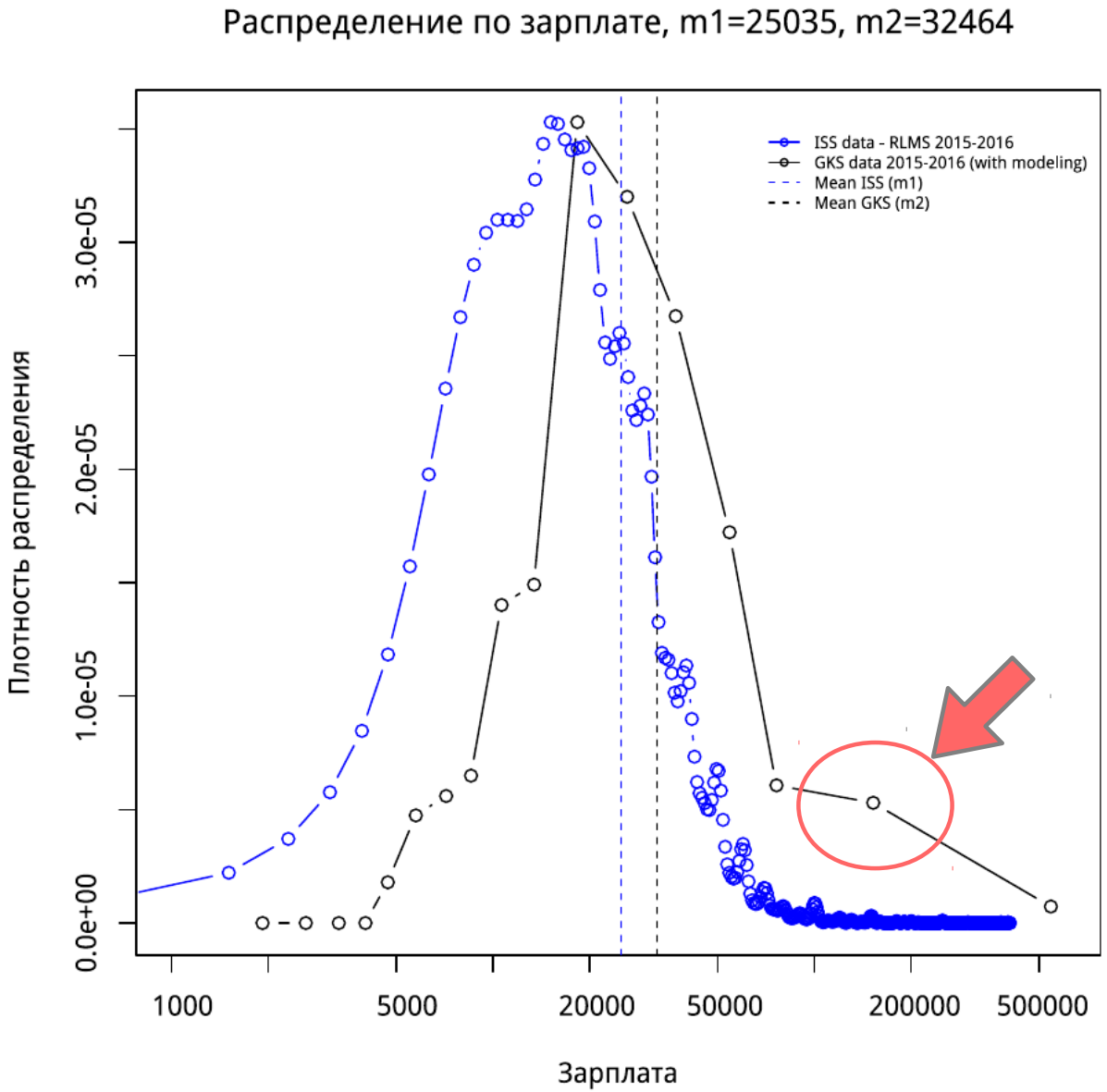
Без кластеров 11,23,1,18

Сравнение распределения зарплат в данных RLMS с данными Росстата за те же годы

Модель для плотности распределения данных Росстата за 2015-2016 гг.

$$d_{2015,2016} = \frac{3}{4} d_{2015} + \frac{1}{4} d_{2017}$$

Можно видеть необычно завышенное значение в области больших зарплат ~ 175 т.р.



Спасибо за внимание!