



Институт системной социологии

# Использование кластерного анализа данных RLMS-HSE: новый подход к исследованию экономического положения и самооценки социальных групп

Алексеев Алексей

физический факультет МГУ им. М.В. Ломоносова

Евстюшева Екатерина

факультет архивного дела РГГУ

Алексеев Иван

ЦНИИОИЗ

Апрельская конференция ВШЭ

11 апреля 2018 г.



# План доклада

- Введение. Различные подходы к стратификации
- Выбор критериев стратификации
- Выбор вопросов и методов количественной оценки
- База данных RLMS-HSE — общая информация
- Выбор вопросов и способа количественной оценки критериев кластеризации
- Описание вычислительной части
  - Селекция данных
  - Нормализация данных
  - Предварительный анализ данных (корреляция столбцов данных)
  - Метод кластеризации K-средних
  - Поиск наилучшего количества кластеров (индекс Calinski-Harabasz)
- Результаты
- Валидация устойчивости определения кластеров
- Дальнейшие планы по развитию данного подхода

# Введение. Различные подходы к стратификации



**В.И.Ленин**

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



**Джон Голдторп**

Новатор многомерной  
стратификации, фактор  
степени контроля (власти)



**Н.Е. Тихонова**

Теоретик ресурсного  
подхода к стратификации



**Пьер Бурдьё**

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

## Выбор вопросов и метод количественной оценки

ID критерия	Критерий	Метод количественной оценки критерия	Допущения и погрешности метода
k2	Отношение к средствам производства	Близость к предпринимательству	Сложно получить данные о собственности, человек может быть иметь существенные активы, но не считать себя предпринимателем
k3	Роль в организации труда	Количество подчинённых	Положение в системе управления организации определяется не только количеством подчинённых, но и размером и внутренней структурой организации
k5	Размер получаемой части общественного богатства	Зарплата за месяц	Зарплата не всегда составляет основную часть дохода
k6	Доступ к ресурсам общества	Доступность важных общественных благ	Выбор вида материальных благ ограничен данными опроса RLMS
k7	Субъективная оценка общественного положения	Субъективная оценка своего общественного положения	Субъективная самооценка опирается на социальное окружение, поэтому является относительной величиной
k9	Уровень и специфика образования	Уровень образования	Специфику образования, например, гуманитарное или естественно-научное направление, мы не учитываем

# Общая информация про RLMS



Лонгитюдное обследование домохозяйств РМЭЗ НИУ ВШЭ (RLMS-HSE) представляет собой серию ежегодных общенациональных репрезентативных опросов на базе вероятностной стратифицированной многоступенчатой территориальной выборки.

В единую базу объединены **наблюдения за 22 года** проведения обследования, начиная с 1994 г.

Собрана информация о структуре доходов и расходов, материальном благосостоянии, работе и миграционном поведении, здоровье и структуре питания, об образовательном поведении и досуге т.д. Взятая в целом, она дает уникальную возможность получить достоверную и многомерную картину жизни населения страны.

Ссылка: <https://www.hse.ru/rlms/>

База является открытой

# Выбор вопросов из опросника RLMS

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?
l2.2	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

## 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

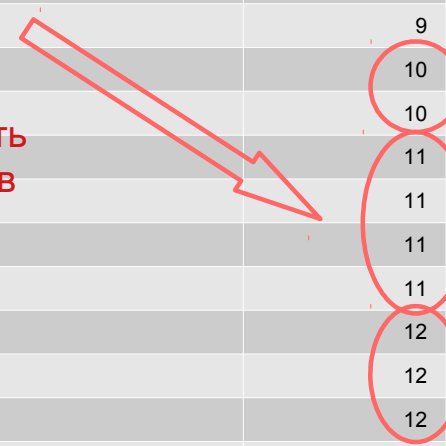


# Способ количественной оценки — уровень образования

	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

Величины выбраны для  
интуитивно-понятного  
определения уровня  
образования.

Одному значению  
может соответствовать  
несколько вариантов в  
опроснике.





## Селекция данных

- Отобраны данные 2015 и 2016 годов (24 и 25 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

## 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

Исходный набор данных

Каждый элемент столбца преобразуется по формуле

$$X' = (x - m_i) / s_{di}$$

где  $m_i$  – среднее арифметическое по  $i$ -му столбцу,  $s_{di}$  – 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

Нормированный набор данных

# Предварительный анализ данных

## Выбранные критерии

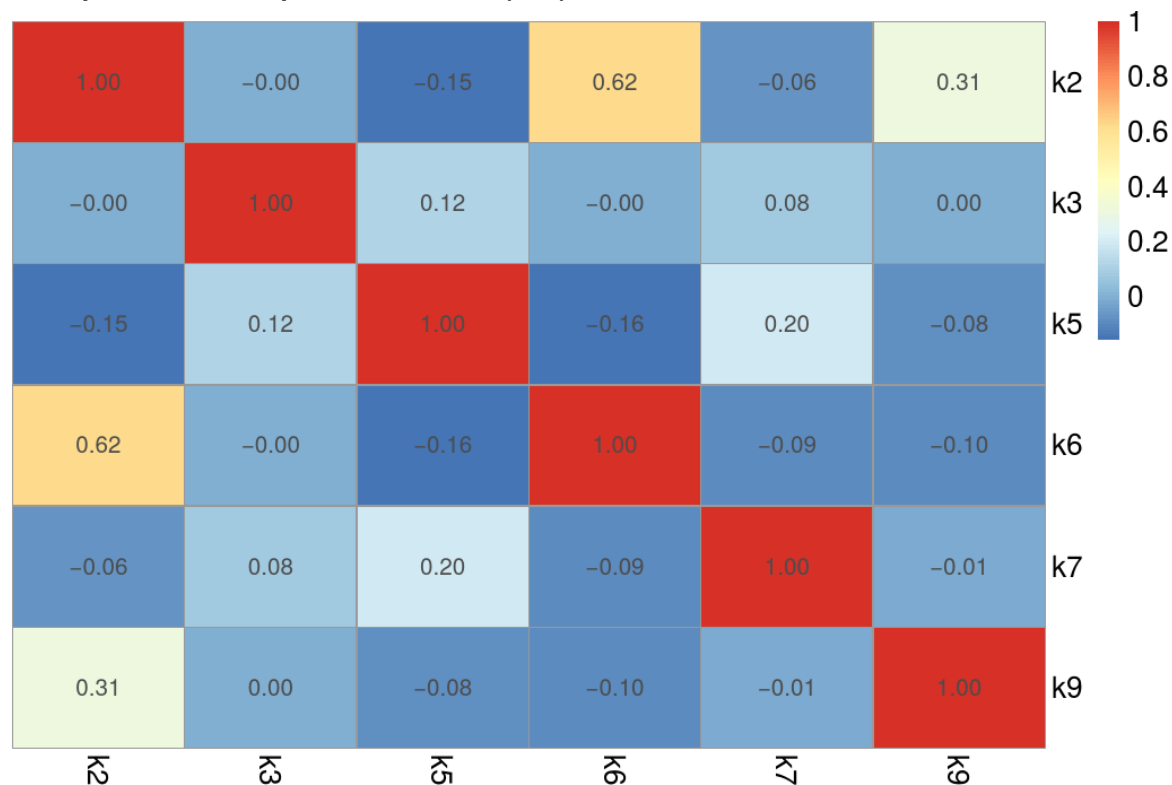
- Близость к предпринимательству (k2)
- Количество подчинённых (k3)
- Зарплата за месяц (k5)
- Доступность важных общественных благ (k6)
- Субъективная оценка своего общественного положения (k7)
- Уровень образования (k9)

## Положительная корреляция

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

## Отрицательная корреляция

- k5-k2
- k5-k6 (!)



Коэффициенты корреляции столбцов данных

Вывод — объем получаемых общественных благ связан с доходом, а не с зарплатой, однако доход в рамках нашего подхода мы пока не умеем оценивать.

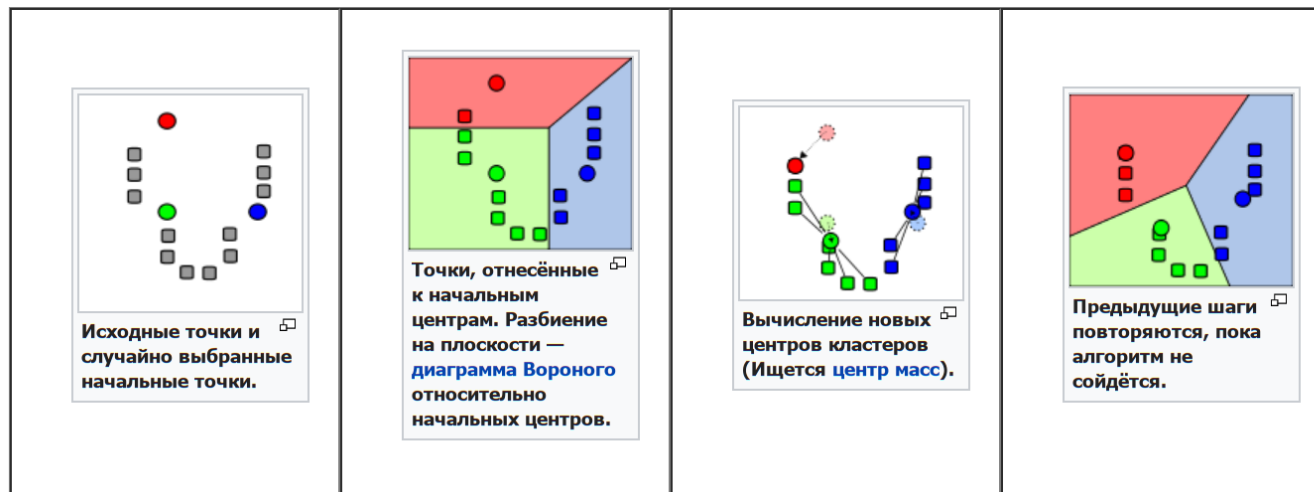
# Метод кластеризации К-средних

Метод k-средних (англ. k-means) — один из самых популярных методов кластеризации. Был изобретён в 1950-х годах математиком Гуго Штейнгаузом и почти одновременно Стюартом Ллойдом [1,2]. Особую популярность он приобрёл после работы Маккуина [3].

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,  $i = 1, 2, \dots, k$  и  $\mu_i$  — центры масс векторов  $x_j \in S_i$ .

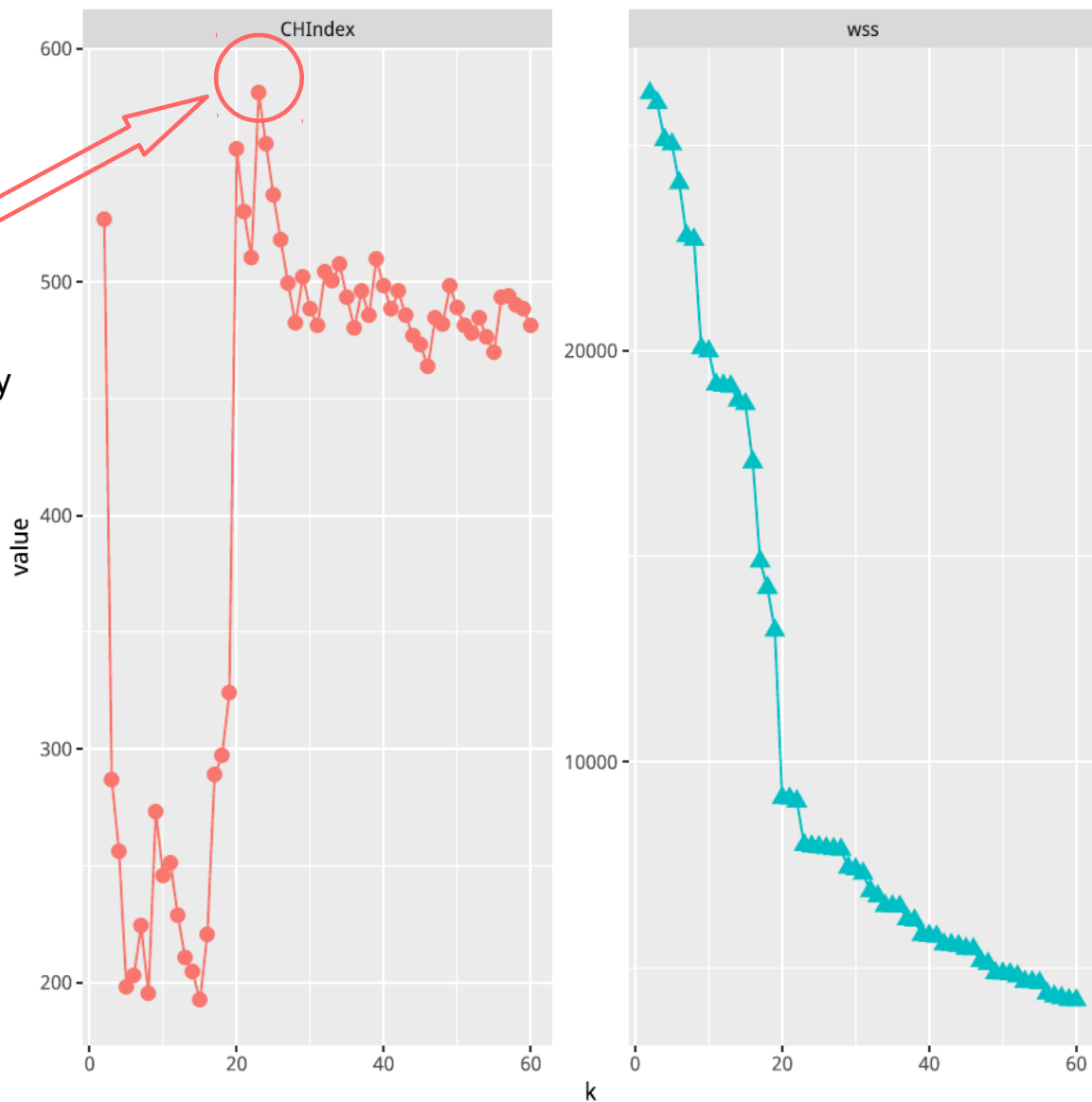


Источники:

1. Steinhaus H. (1956). Sur la division des corps materiels en parties. Bull. Acad. Polon. Sci., C1. III vol IV: 801—804.
2. Lloyd S. (1957). Least square quantization in PCM's. Bell Telephone Laboratories Paper.
3. 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.

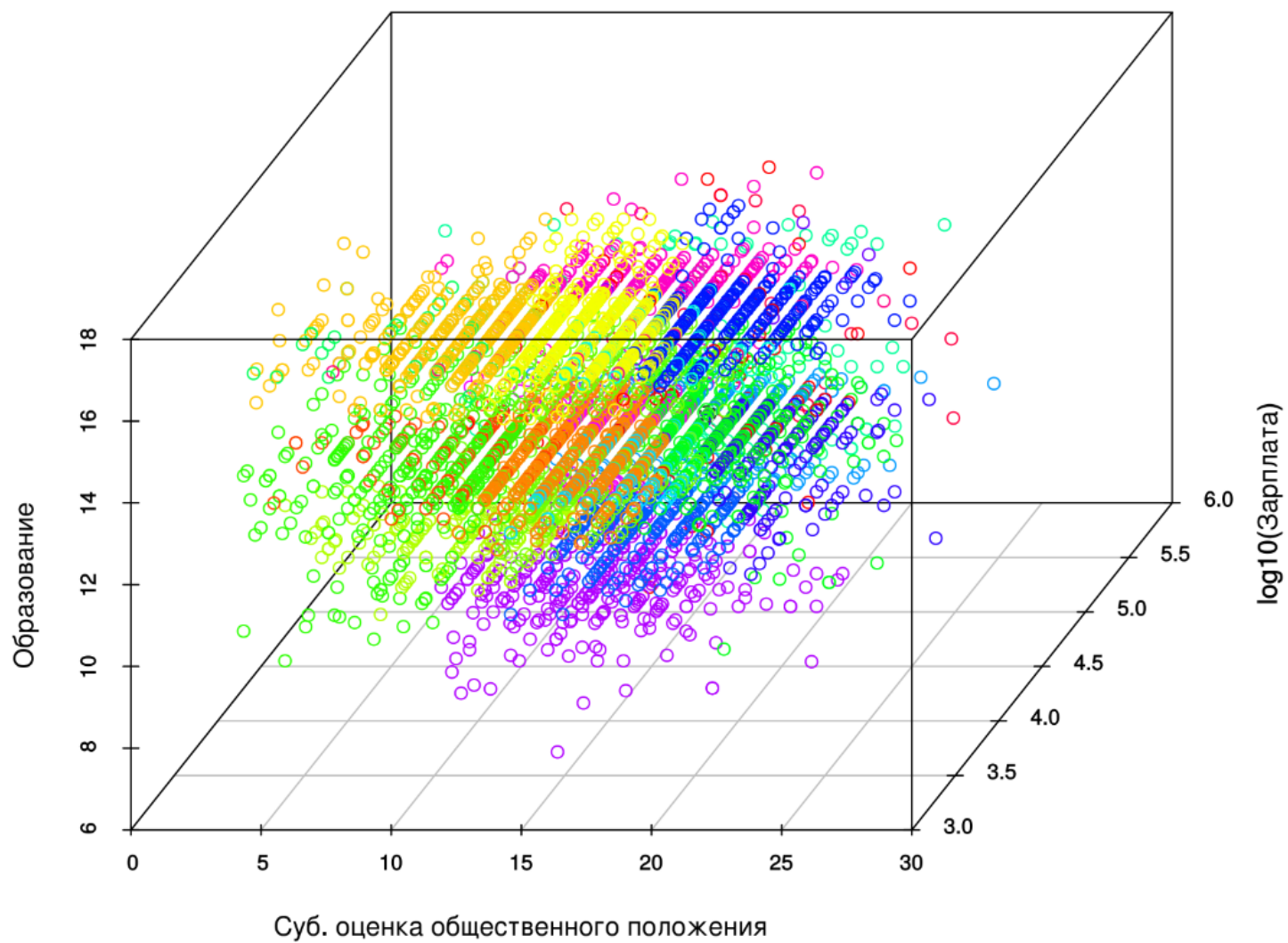
Максимум  
соответствует  
наилучшему количеству  
кластеров



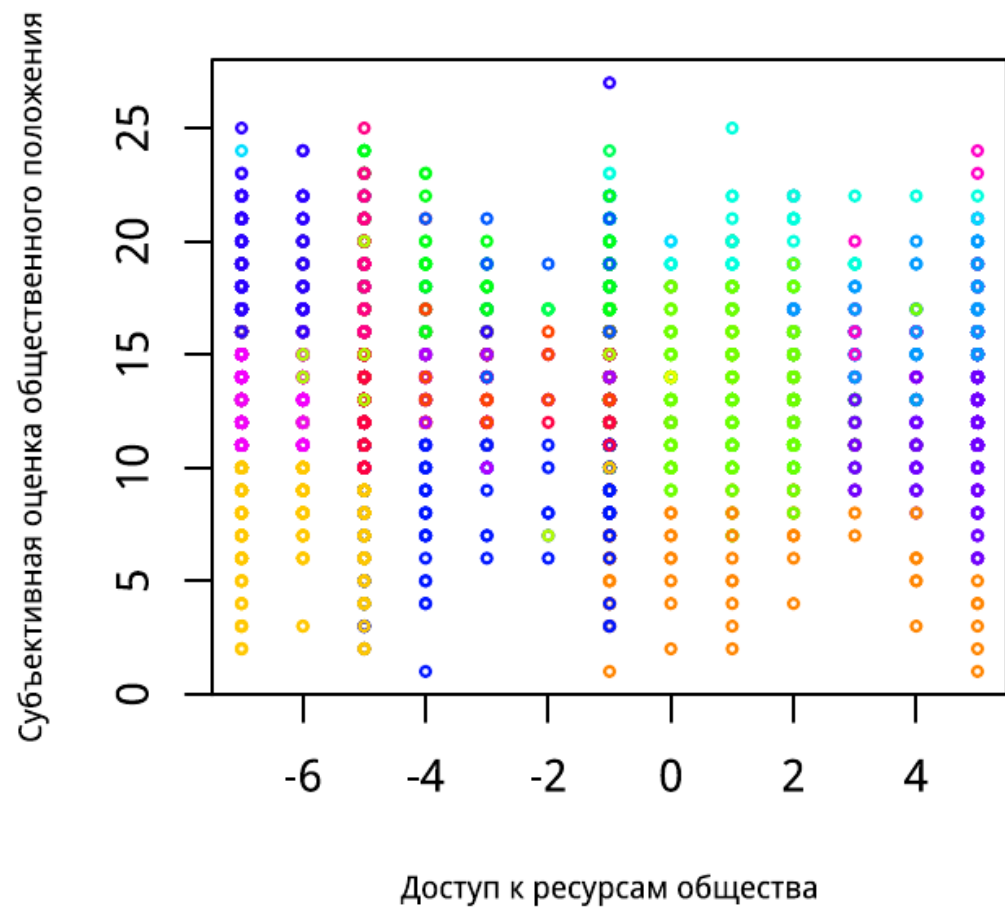
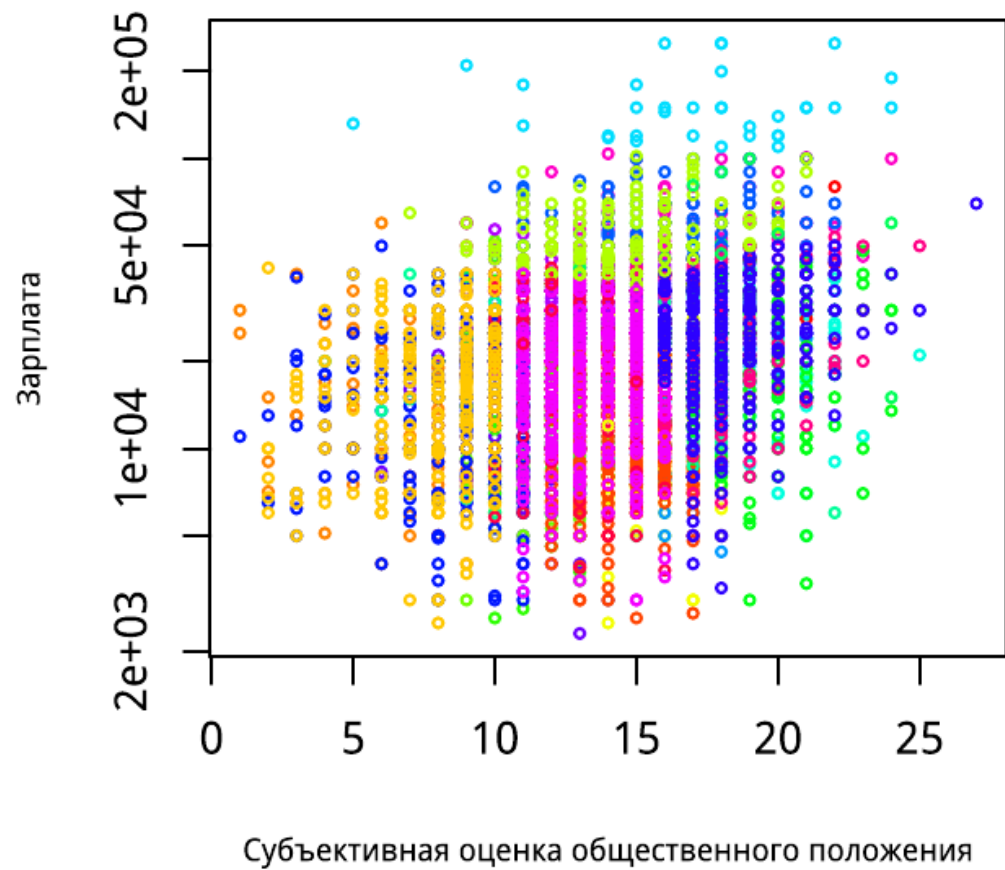
- CH-index достигает максимума при  $k=23$

# Результаты

Результат кластеризации в 3D

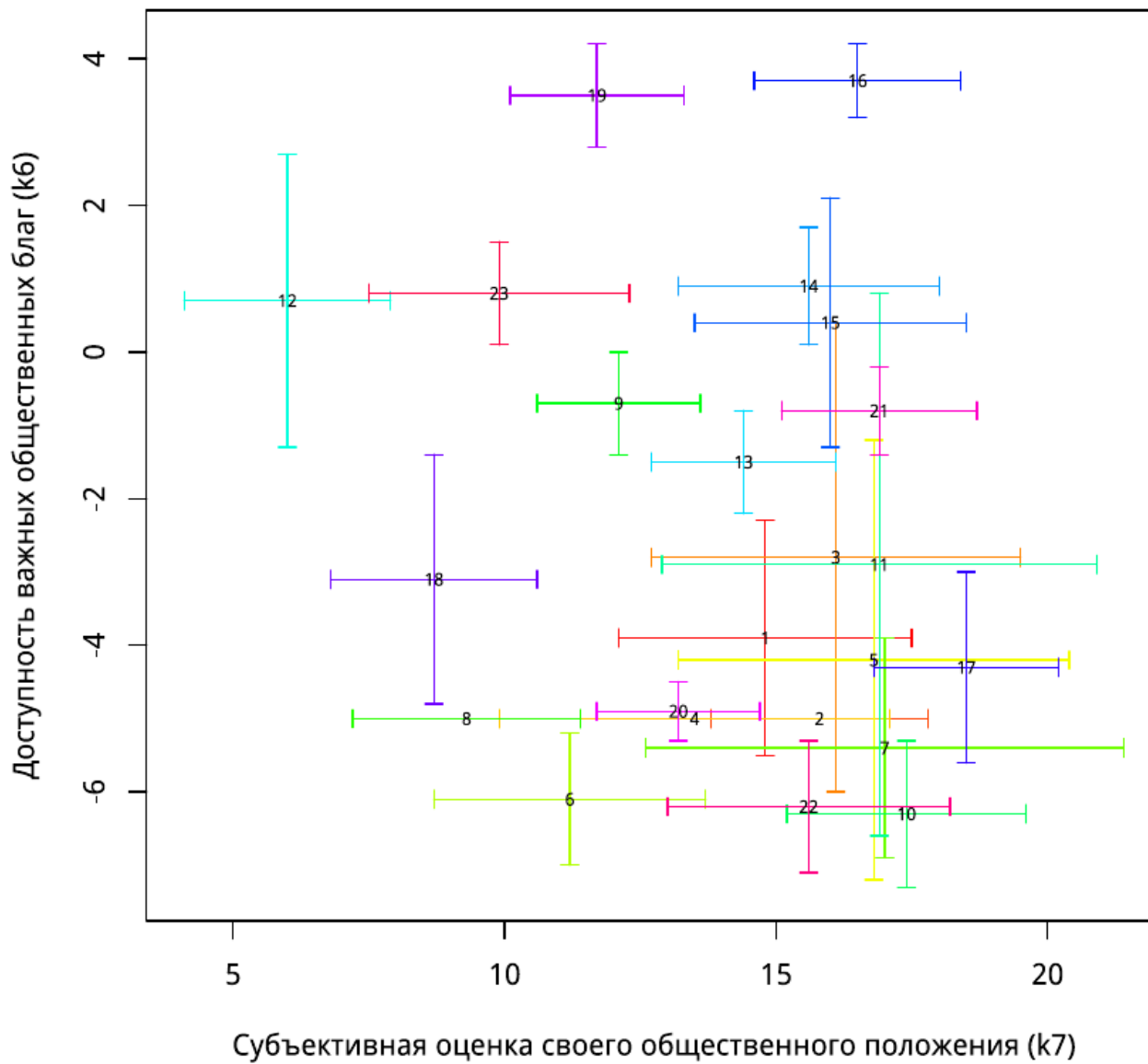


## Результаты - 2D

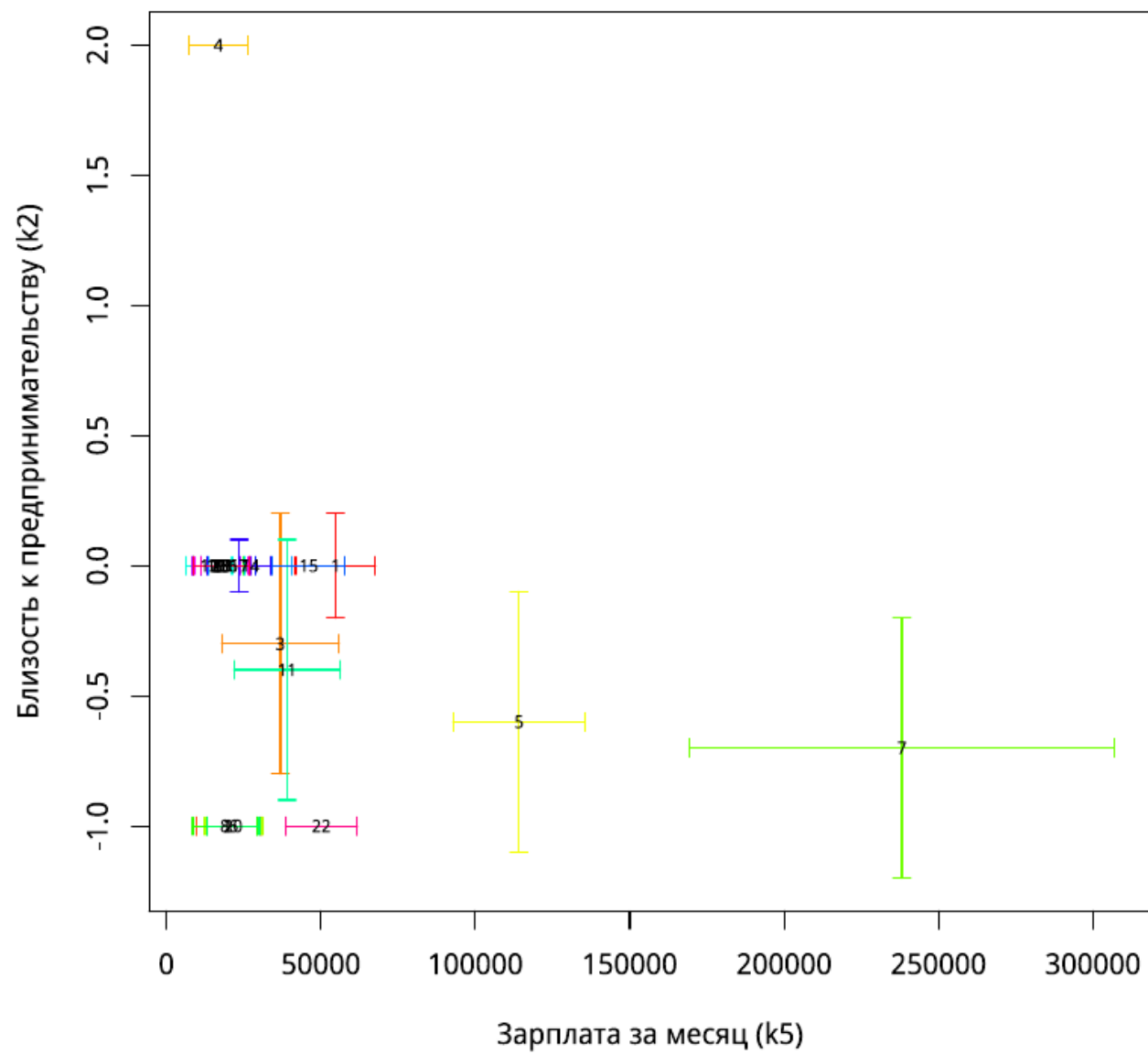




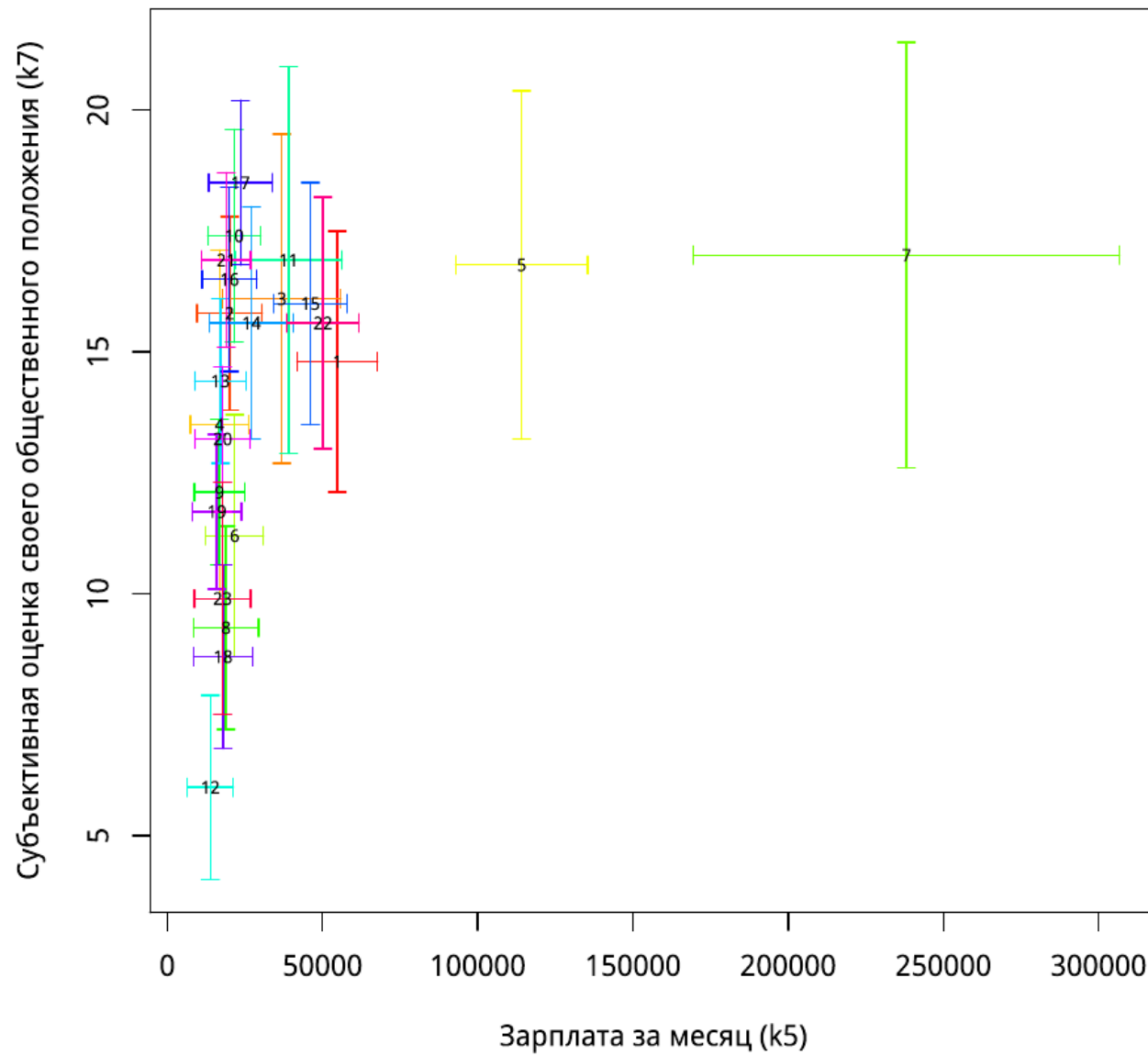
# Результаты



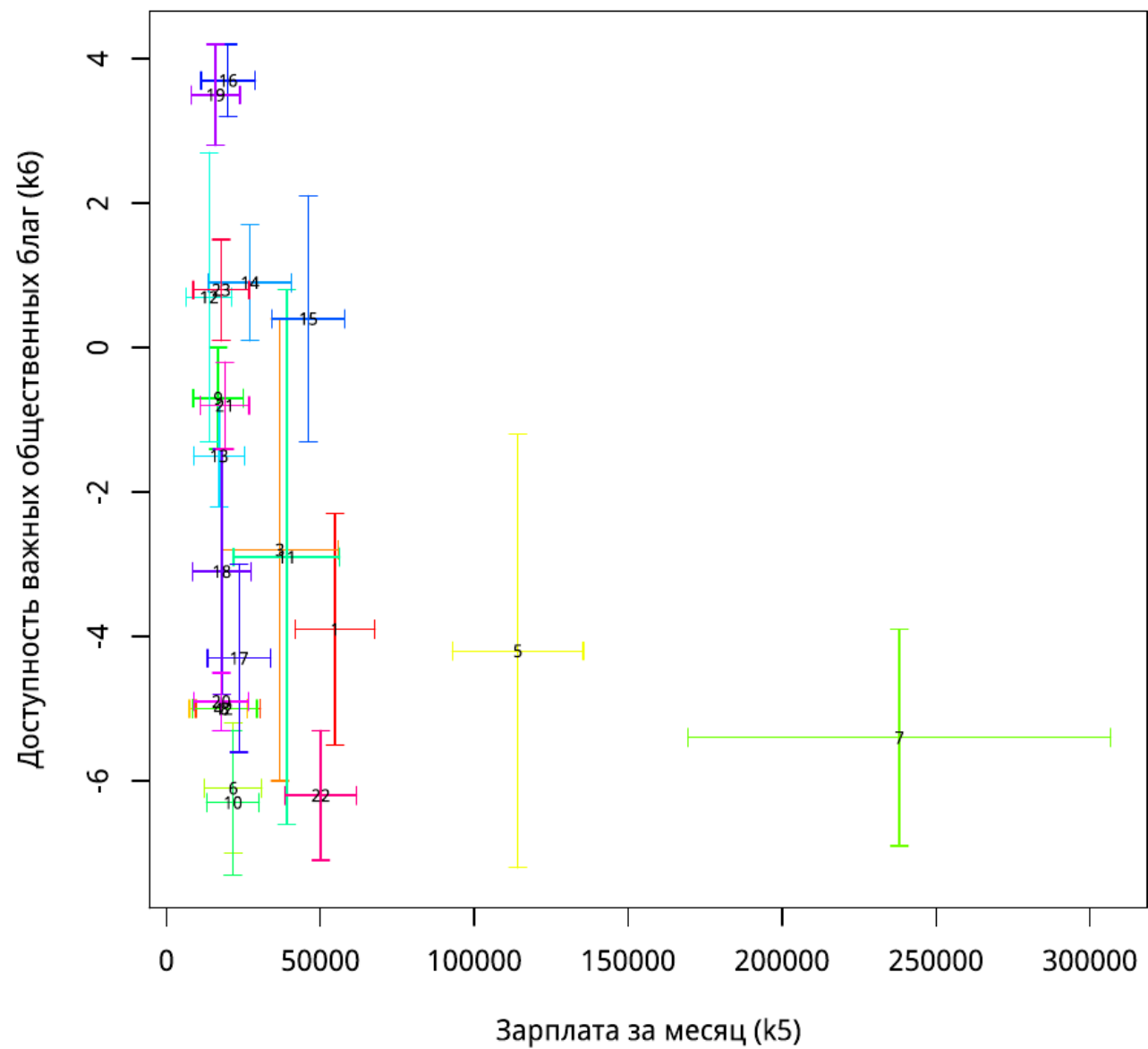
# Результаты



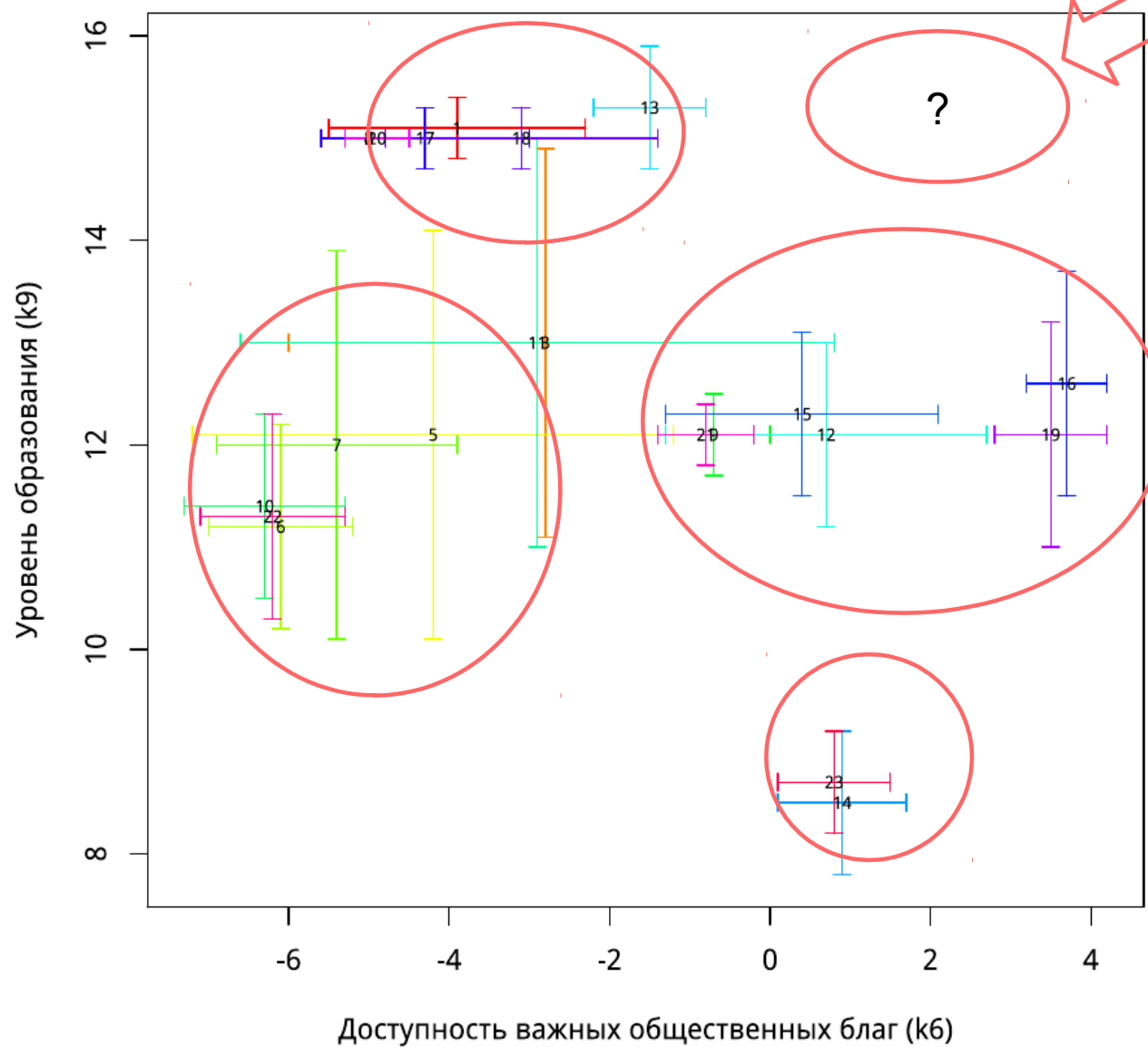
# Результаты



# Результаты

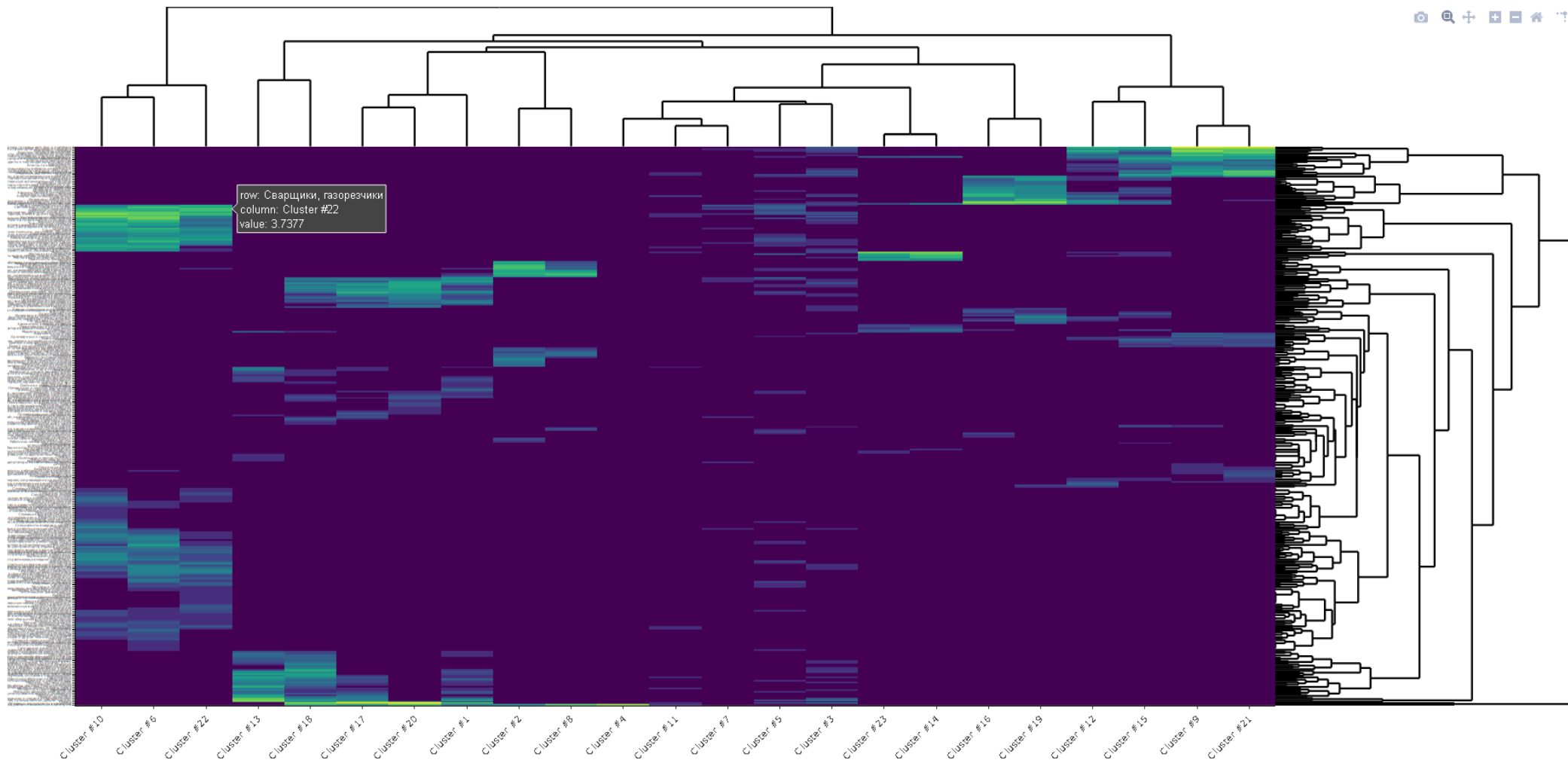


# Результаты



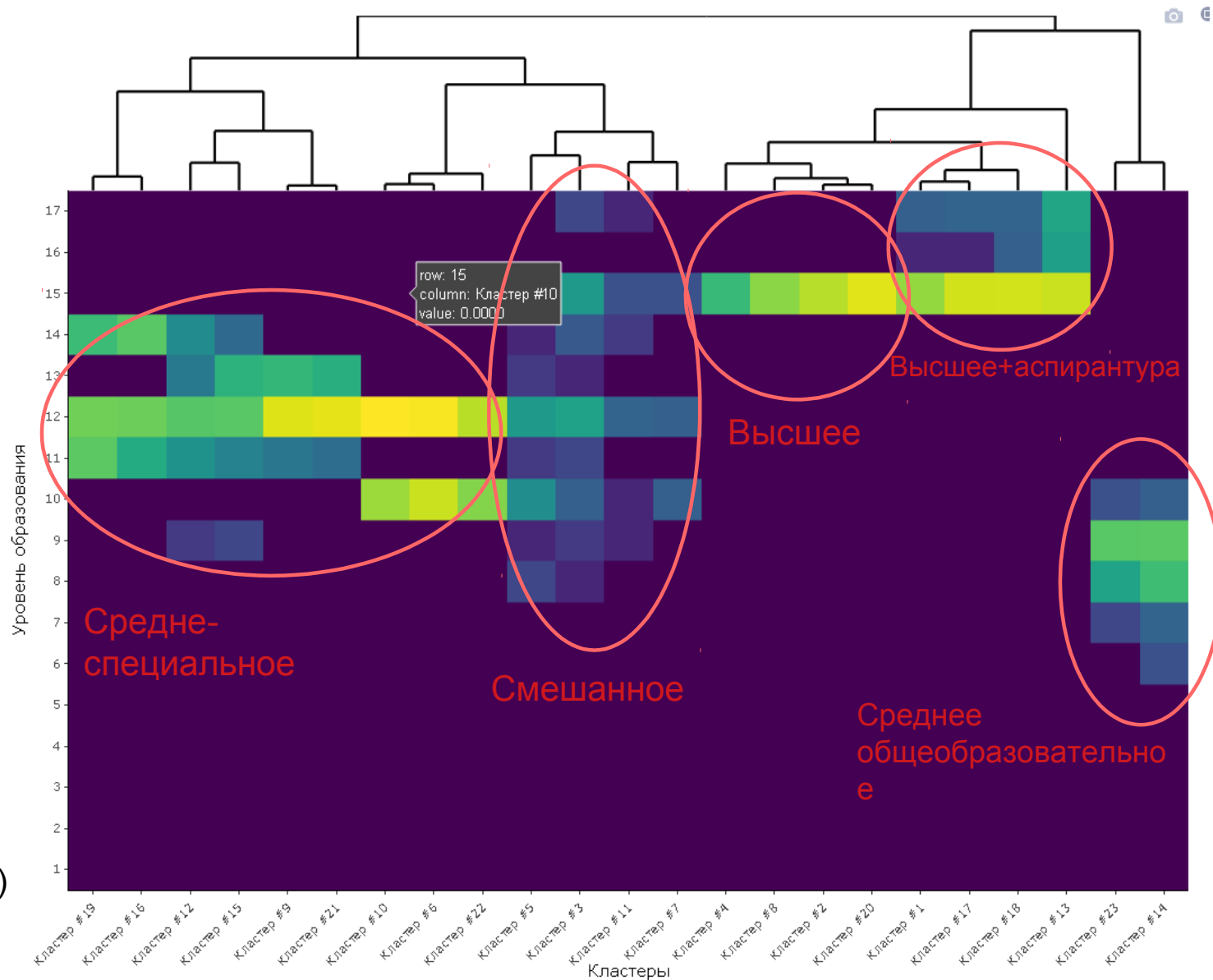
Не охвачены  
опросом, либо не  
включены в  
обработку, поскольку  
не указана зарплата

## Профессиональный состав кластеров (см. интерактивный график)



При построении тепловой диаграммы использован метод hclust (иерархическая кластеризация) - для строк и столбцов. Данный алгоритм реализован в R (пакет stats). Для построения интерактивной диаграммы использован пакет heatmaply.

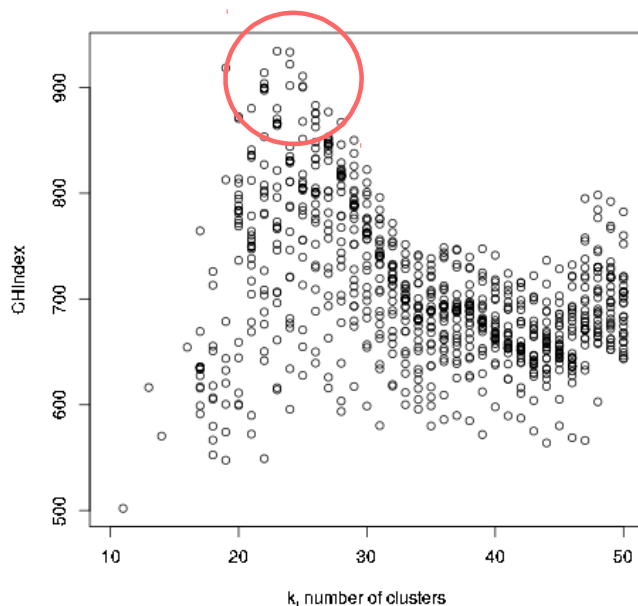
## Распределение участников кластеров по уровню образования



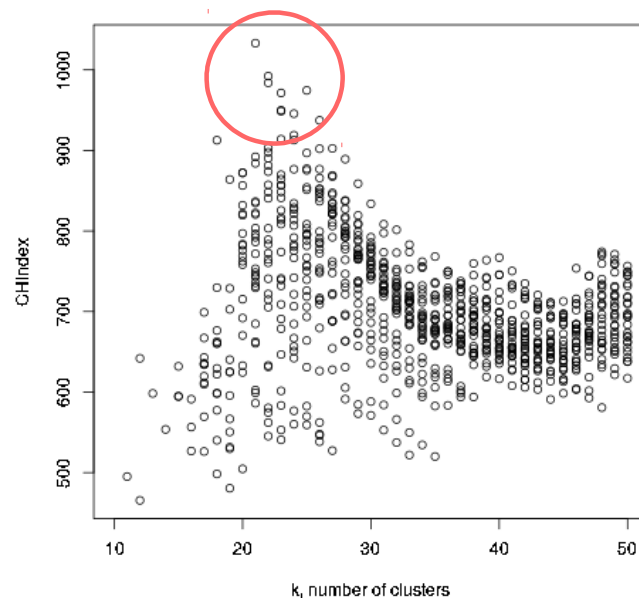
Hclust - по столбцам (кластеры упорядочены по схожести профиля распределения по образованию)

# Валидация определения количества кластеров

P=0.1%



P=0.2%



P=0.5%

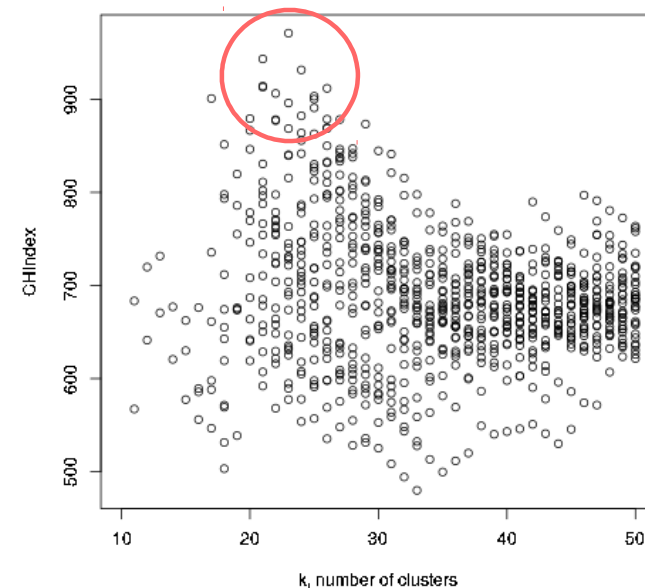


График величины СН-индекса от количества кластеров (k) для различного процента исключения строк данных (P). Исходное количество строк 8357. Количество расчётов каждого варианта (k и P) — 30.

Максимум СН-индекса достигается при k=**23** (P=0.1), k=**21** (P=0.2), k=**23** (P=0.5)





## Дальнейшие планы по развитию данного подхода

- Использование нейронных сетей для **предсказания зарплаты** для участников опроса, которые не указали её в опросе
- Разработка подход (модель) для **оценки дохода индивида** на основании зарплаты, оценки доступа к ресурсам и других данных опросника
- Использование данных **RLMS других лет** (более ранних)
- Сравнение результатов для RLMS – данные для **индивидов и домохозяйств**
- Использование данных **Мониторинга Института социологии РАН**, сравнение результатов кластеризации с результатами с использованием RLMS
- Применения нашего подхода для данных по **другим странам**.  
Например, для США имеется база [General Social Survey \(GSS\)](#)
- Уточнение определения и кластерного состава **«среднего класса»** (из каких кластеров он может состоять, по каким параметрам производить выделение этой группы)
- Публикация исходных данных и скриптов на **открытом репозитории GitHub**




# Литература

1. Goldthorpe, J.H. Social Mobility and Class Structure in Modern Britain. - Oxford: Clarendon Press; 1987. - 398 p.
2. Goldthorpe, J.H. The economic basis of social class. - London: Centre for Analysis of Social Exclusion; 2004. - 36 p.
3. Аникин В.А, Тихонова Н.Е. Профессиональный портрет среднего класса и особенности его эволюции // В кн.: Средний класс в современной России. Опыт многолетних исследований / Под ред. М.К. Горшкова и Н.Е. Тихоновой. М.: Весь Мир, 2016. Гл. 3. С. 58–79.
4. Бобровский О.В. Влияние материального положения населения на уровень социальной напряженности в регионе // Известия Тульского государственного университета. Гуманитарные науки, 2016. - С. 80-85.
5. Бобровский О.В. Влияние трансформации социальной структуры на уровень социальной напряженности в современном российском обществе // Теория и практика общественного развития, 2015.
6. Бобровский О.В. Региональные особенности формирования социальной структуры современного российского общества. - Тула, 2013. - 198 с.
7. Бурдые П. Социальное пространство: поля и практики: Пер. с фр. / Сост., общ. ред. пер. и послесл. Н.А. Шматко. – М., Институт экспериментальной социологии, 2005. – 576 с.
8. Ленин В.И. Полное собрание сочинений. – Т. 39. – М.: 1969.
9. Тихонова Н.Е. Ресурсный подход как новая теоретическая парадигма в стратификационных исследованиях // Социологические исследования, 2006, № 9.
10. Тихонова Н.Е. Социальная структура России: теории и реальность. - М.: Новый хронограф, 2014. - 408 с.



Спасибо за внимание!



Давайте посмотрим  
интерактивные диаграммы!