Методы автоматизации сбора и обработки статистических данных

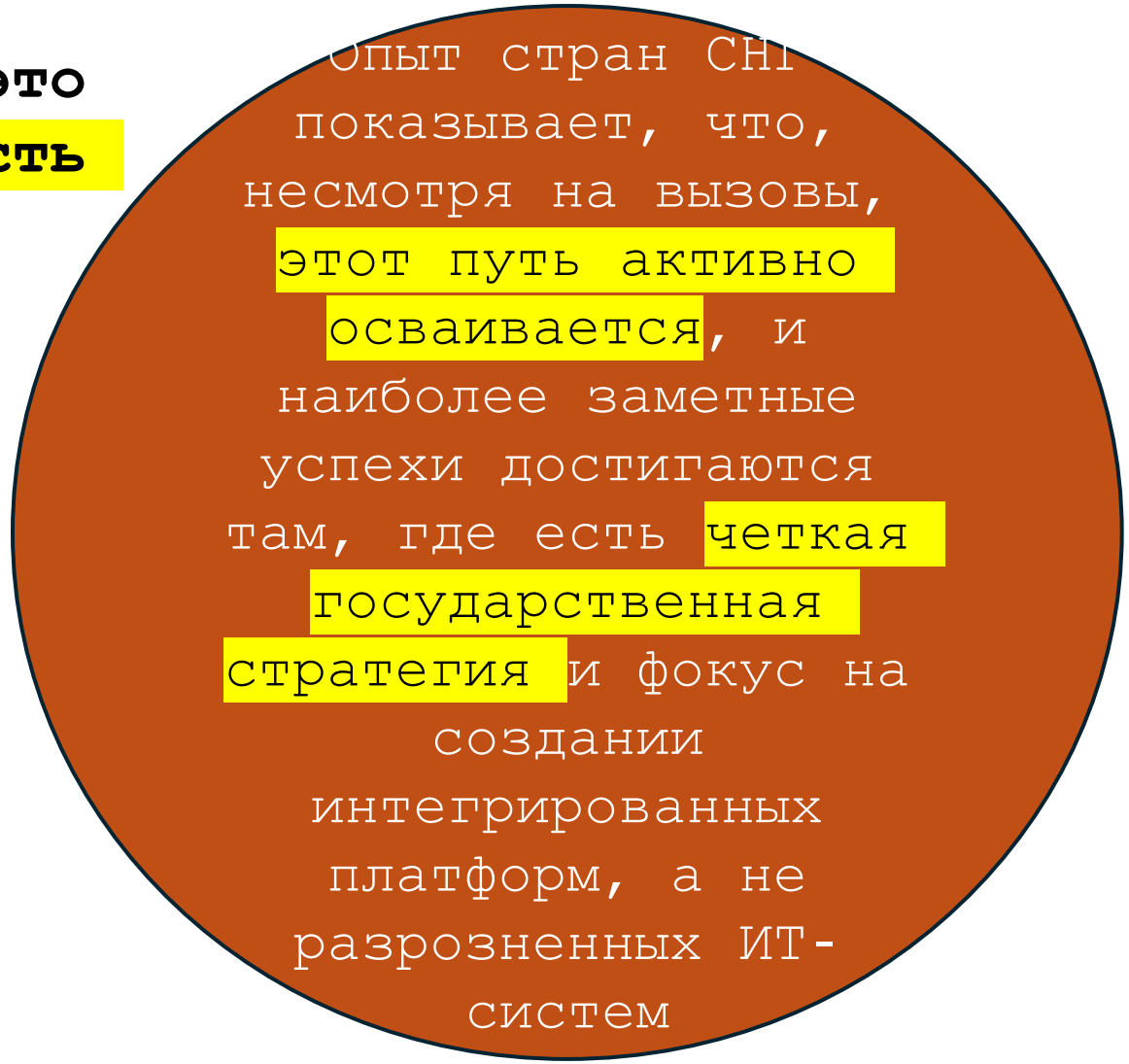
Лекции 2 и 3

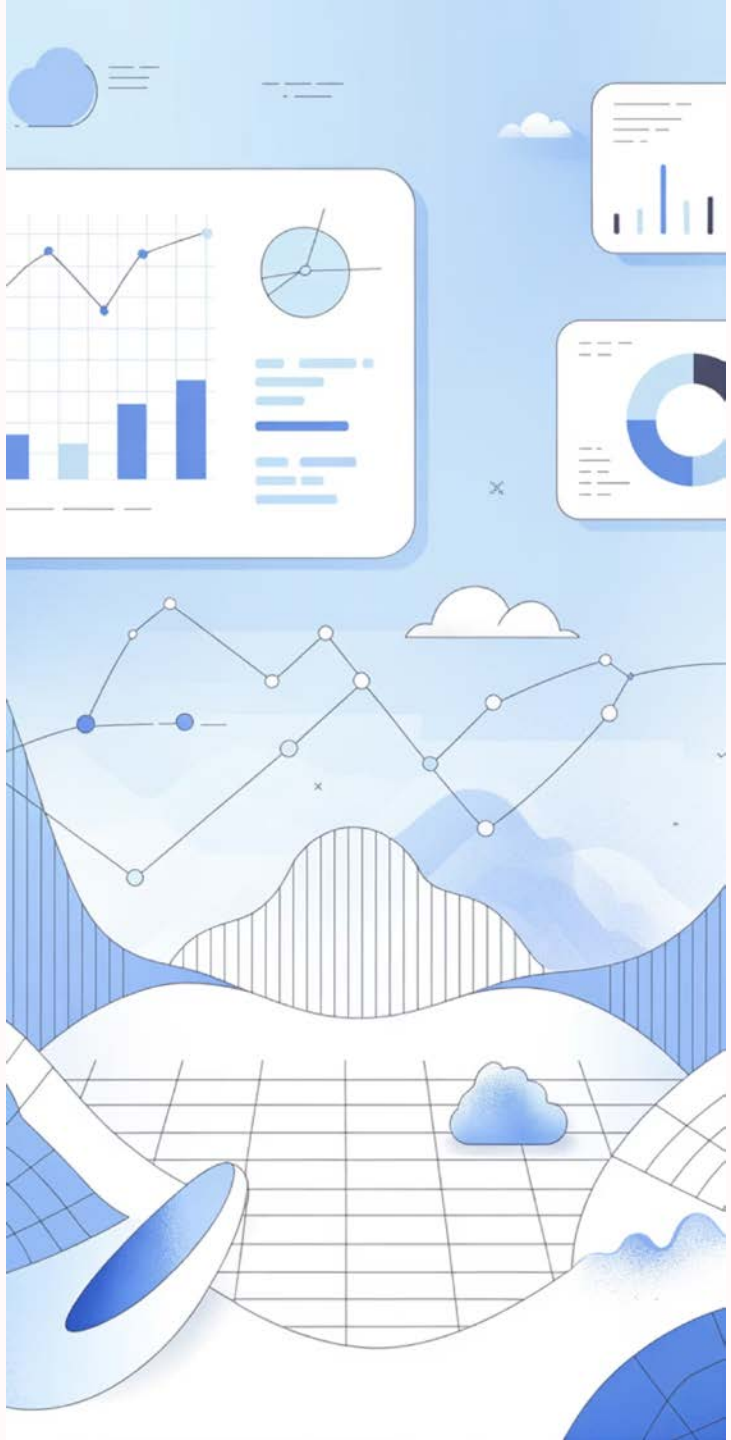
Зарова Елена Викторовна, доктор экономических наук. профессор

План лекции

- **Использование цифровых платформ, облачных решений и ИИ для автоматизации сбора и обработки данных *(лекция 3)***
- Практические примеры: автоматизация переписей, обработка административных данных
- Теоретические основы: обработка больших данных, обезличивание и защита информации

Использование цифровых платформ, облачных решений и ИИ для автоматизации сбора данных – это не просто тренд, а необходимость для повышения эффективности государственного управления и конкурентоспособности бизнеса





Применение МО и ИИ на этапах Типовой модели производства статистической информации

Уровни 1 и 2: принципы, задачи, опыт официальной статистики стран СНГ

Введение: Значение МО и ИИ в современной официальной статистике

Рост рынка ИИ

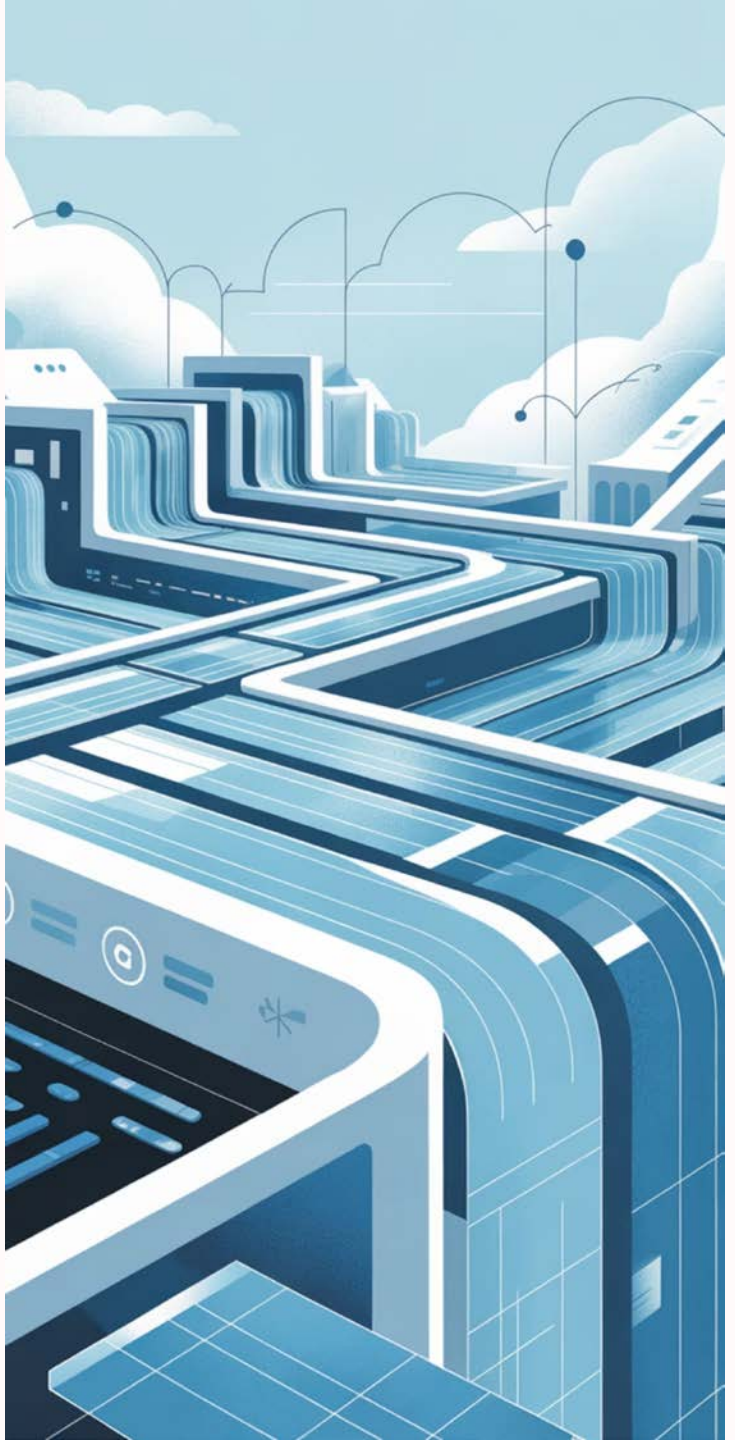
Рынок искусственного интеллекта в России достигнет 300 млрд рублей в 2024 году, становясь мощным драйвером цифровой трансформации экономики

Повышение качества

МО и ИИ позволяют существенно повысить качество, скорость обработки и точность статистических данных

Инновации СНГ

Страны СНГ активно внедряют передовые технологии в национальных статистических службах



Типовая модель производства статистической информации

Уровни 1 и 2

Первый уровень

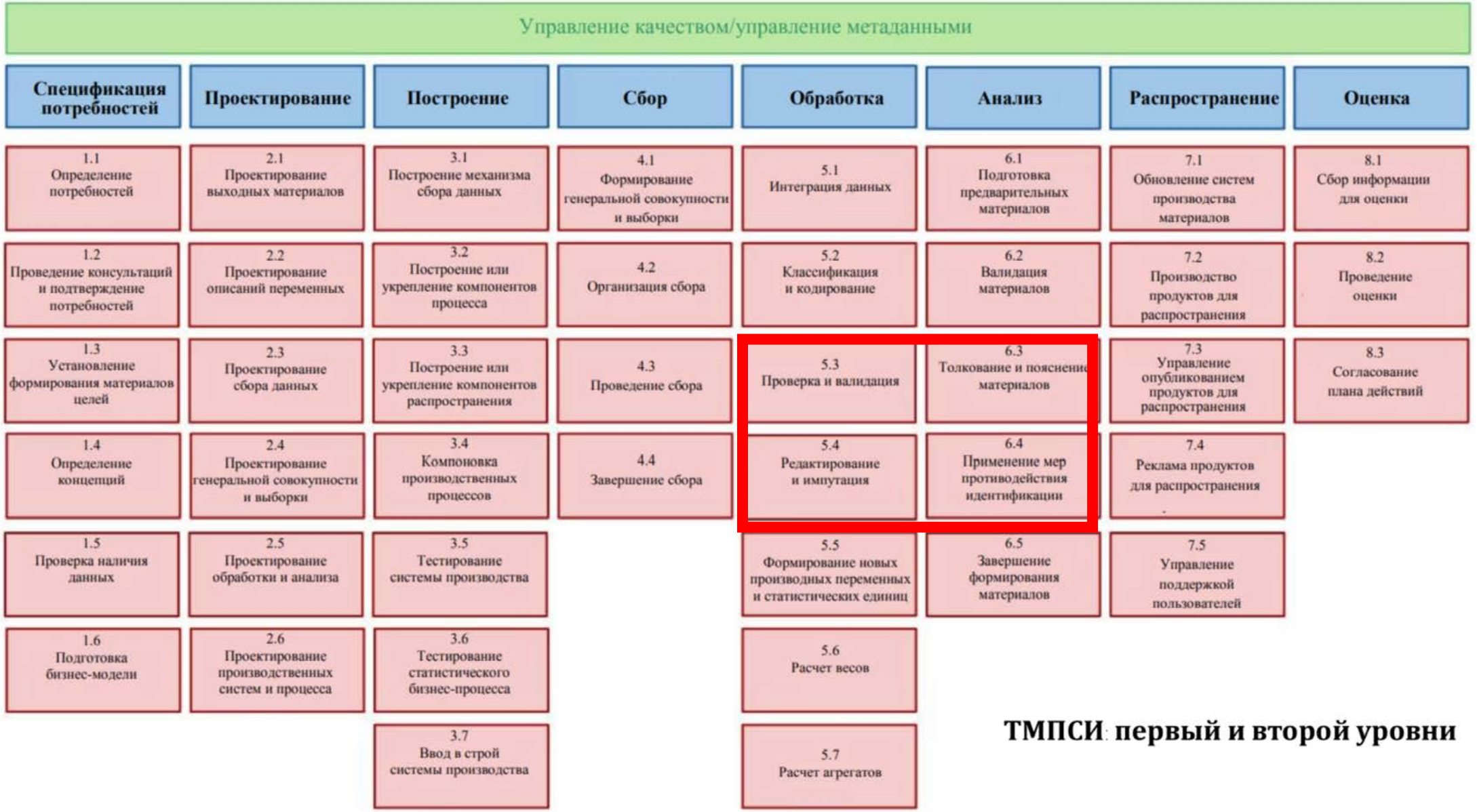
Сбор и первичная обработка данных — ключевая база для всей статистической системы, требующая особого внимания к качеству

Второй уровень

Интеграция, верификация и подготовка агрегированных данных для последующего анализа

Роль технологий

МО и ИИ автоматизируют рутинные операции, выявляют аномалии и улучшают качество данных на всех этапах



ТМПСИ: первый и второй уровни

ИИ и МО в стратегиях развития государственной статистики стран СНГ

Страна	Наличие и ссылка на стратегию	Год утверждения	Упоминание ИИ/МО в стратегии
Азербайджан	**Да.** Государственная программа по развитию статистики в Азербайджанской Республике на 2018-2025 годы	2018	**Косвенное указание.** Упоминается развитие IT-инфраструктуры и использование "альтернативных источников данных", но прямого указания на ИИ/МО нет.
Армения	**Да.** Стратегическая программа развития статистики Армении на 2022-2026 гг.	2022	**Прямое указание.** Содержит задачу "исследовать и внедрять передовые технологии, [illegible]е как искусственный интеллект и машинное обучение, для [illegible]анных и

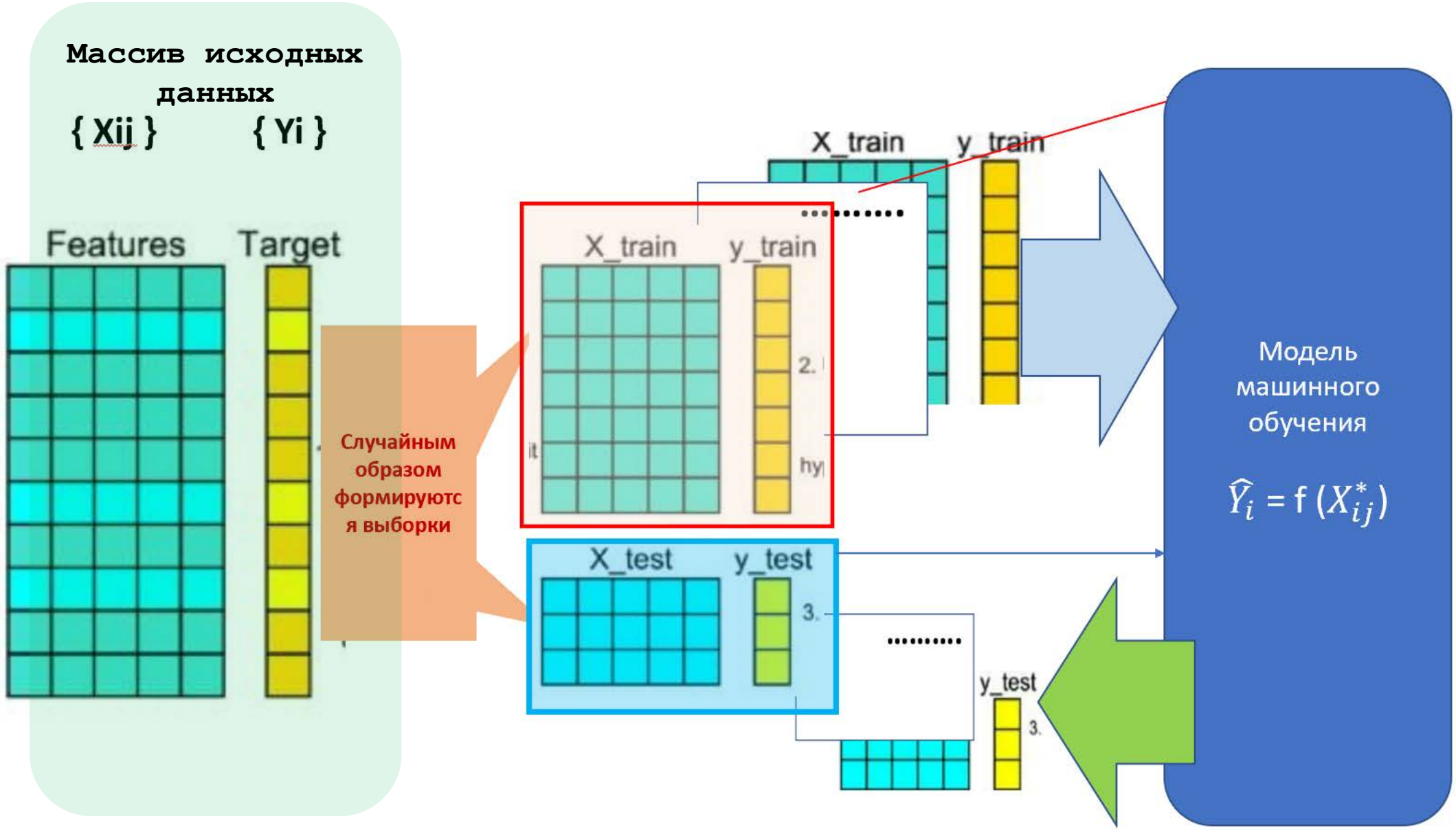
Сообщение для DeepSeek

Требует проверки

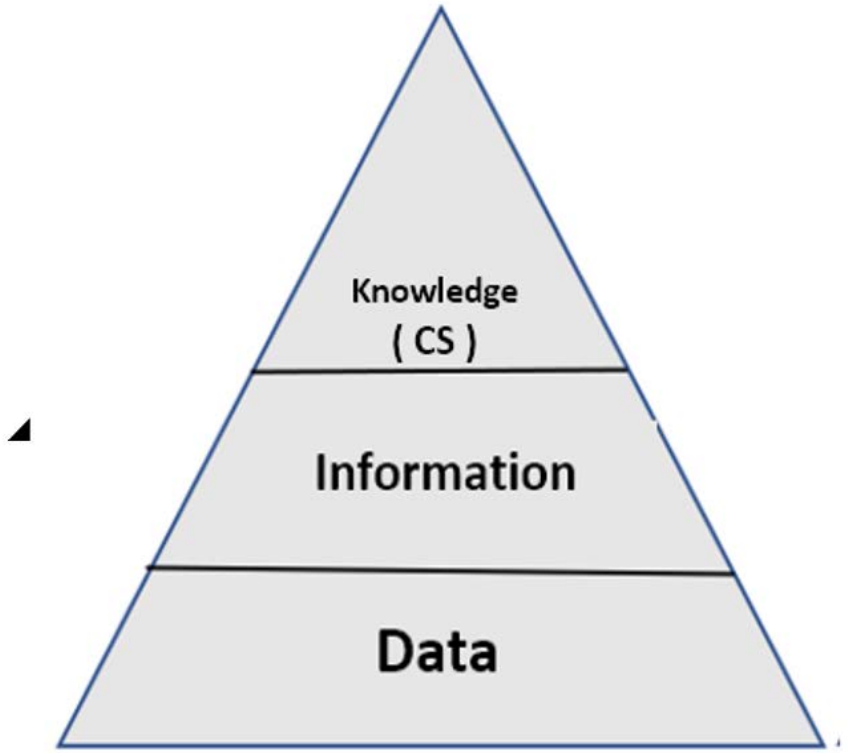
Беларусь	**Да.** Концепция развития государственной статистики до 2027 года	2020	**Косвенное указание.** Упоминается "анализ больших данных" и "совершенствование ИТ-инфраструктуры", но прямого указания на ИИ/МО нет.
Казахстан	**Да.** Концепция развития государственной статистики до 2026 года	2021	**Прямое указание.** Планируется "внедрение методов больших данных, машинного обучения и искусственного интеллекта для повышения оперативности, точности и эффективности статистических процессов".
Кыргызстан	**Да.** Стратегия развития государственной статистики Кыргызской Республики на 2019-2030 годы	2019	**Прямое указание.** Запланировано "использование методов интеллектуального анализа данных, машинного обучения и искусственного интеллекта для обработки больших ма ов информации".

Молдова	**Да.** Стратегия развития национальной статистической системы на 2021-2025 годы	2021	"внедрении новых технологий, таких как искусственный интеллект и машинное обучение, для анализа больших данных и повышения эффективности производства статистики".
Россия	**Да.** Концепция развития государственной статистики на период до 2024 года и дальнейшую перспективу до 2030 года	2021	**Прямое указание.** Включено "использование методов машинного обучения и искусственного интеллекта для анализа данных, в том числе больших данных, и автоматизации процессов".
Узбекистан	**Да.** Стратегия развития государственной статистики на 2020-2025 годы	2020	**Прямое указание.** Включено "внедрение передовых методов, включая машинное обучение и искусственный интеллект, для анализа больших данных и повышения качества статистической информации".

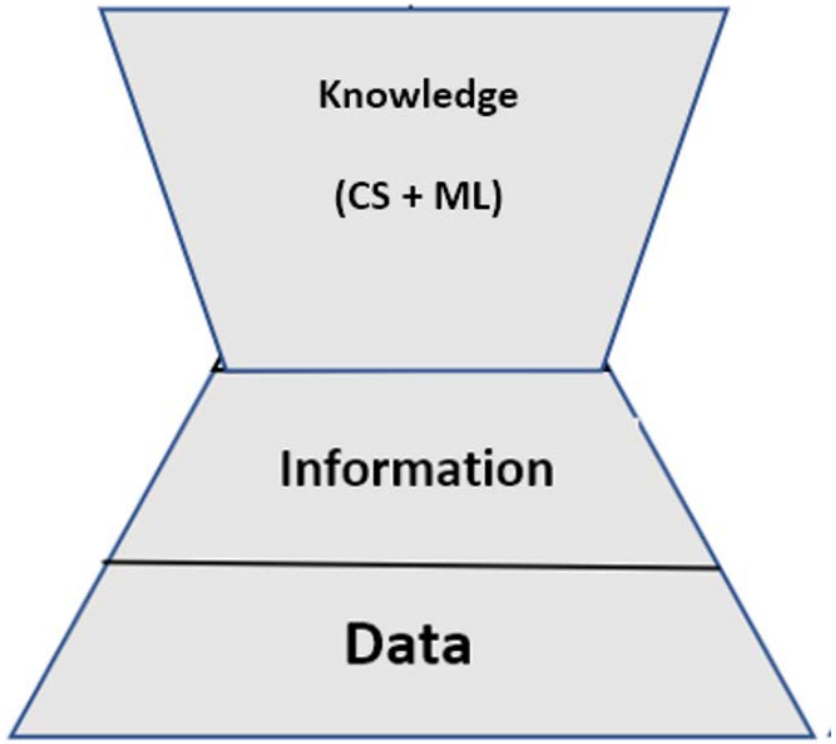
Идея метода машинного обучения



“Data-information-knowledge pyramid”



“a”

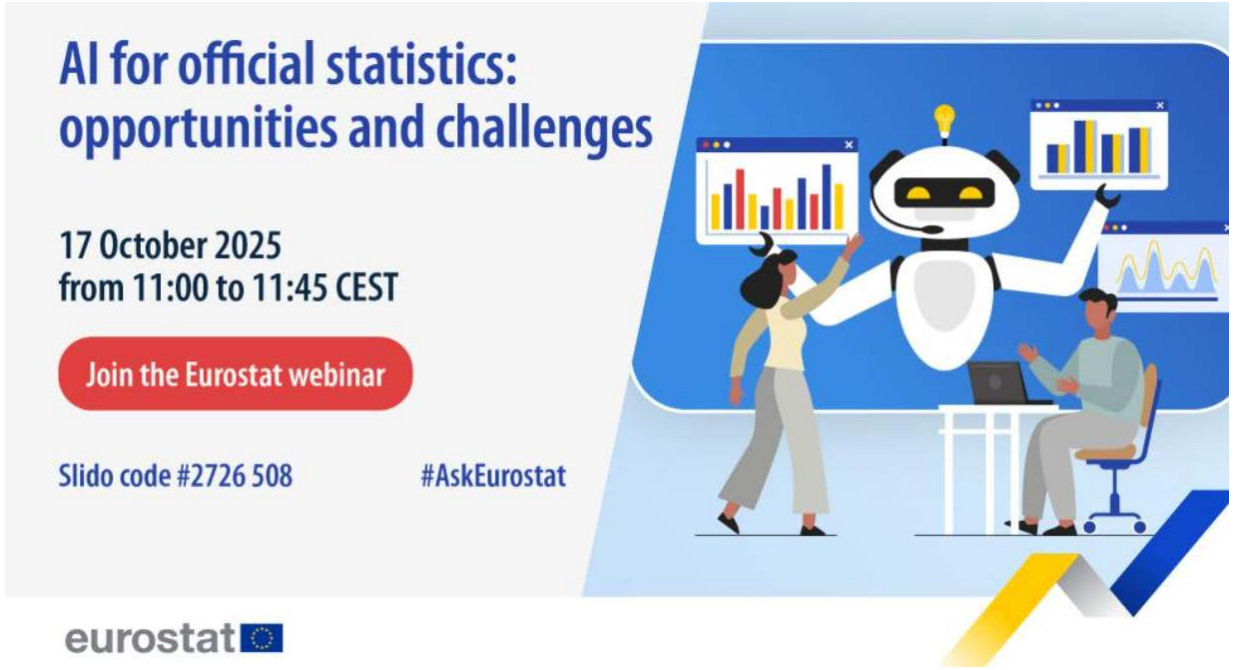


“b”

(*proposed by the author*)

* Conventional statistics (CS)

Семинар по ИИ в официальной статистике



https://unstats.un.org/bigdata

DM в руководствах EUROSTAT

Методы интеллектуального анализа данных (социальные нейронные сети, кластерный анализ, анализ выбора модели и взвешенный сетевой корреляционный анализ, чтобы установить связь между характеристиками рынка труда ЕС и индикаторами рынка труда.

Figure 1: Country social network using Labour Market Attractiveness set

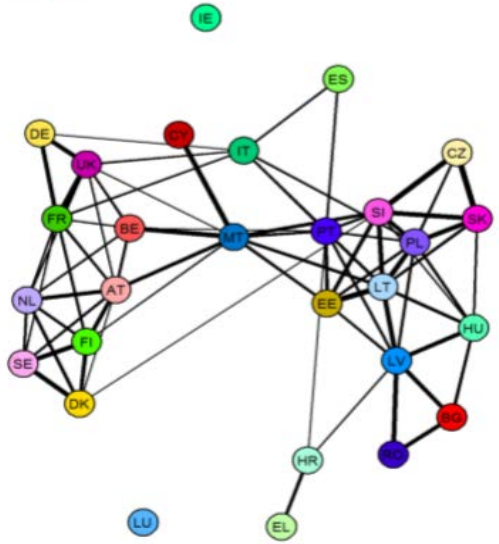


Figure 2: Cluster analysis using Labour Market Attractiveness set. Average silhouette width for all possible number of clusters (left). Silhouette width of regions considered for 10 clusters (right)

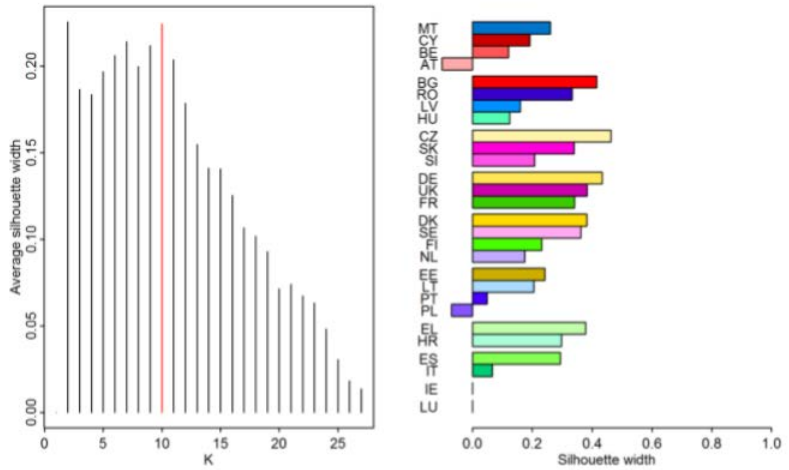


Table 3: Description of clusters

cluster	NUTS(a)	n	sep
AT-BE-CY-MT	MT, CY, BE, AT	4	2.66
BG-HU-LV-RO	BG, RO, LV, HU	4	2.49
CZ-SK-SI	CZ, SK, SI	3	2.31
DE-FR-UK	DE, UK, FR	3	3.19
DK-FI-NL-SE	DK, SE, FI, NL	4	2.66
EE-LT-PL-PT	EE, LT, PT, PL	4	2.31
EL-HR	EL, HR	2	3.53
ES-IT	ES, IT	2	3.10
IE	IE	1	3.75
LU	LU	1	4.68

(a) Order of countries reflects belongingness to cluster (i.e. silhouette width).

Многофакторные регрессионные модели

Table 5: Summary Table for best multivariate model for Skills Mismatch

Variables	Estimate	Std. Error	t-value	Pr(>|t-value|)
(Intercept)	100.60	8.694	11.570	<0.001
emp_Y15–24	-3.18	0.356	-8.920	<0.001
emp_Y15–24_NaceM-N	6.56	1.000	6.560	0.001

skills_mismatch ~ emp_Y15–24 + emp_Y15–25_NaceM–N

RSS = 5.51 (5 d.f.)
R^2 = 0.94, R^2-adjusted = 0.92
F(2,5) = 43.76 (p-value < 0.001)

Note: 20 data entries and 34 variables were removed to keep only complete-cases; 10 variables highly correlated among them ($\rho > 0.90$) and 2 variables least correlated with dependent variable ($\rho < 0.06$) were removed to run analysis at the most with 30 variables.

Table 6: Summary Table for best multivariate model for Mobility

Variables	Estimate	Std. Error	t-value	Pr(>|t-value|)
(Intercept)	1.82	6.196	0.290	0.772
emp_Y25–64_NaceK	4.61	0.711	6.470	<0.001
emp_Y25–64_NaceL	10.04	2.892	3.470	0.002
emp_Y15–24_NaceB–E	-0.38	0.198	-1.930	0.067

mobility ~ emp_Y25–64_NaceK + emp_Y25–64_NaceL + emp_Y15–24_NaceB–E

RSS = 7.10 (21 d.f.)
R^2 = 0.79, R^2-adjusted = 0.76
F(3,21) = 25.85 (p-value < 0.001)

Note: 3 data entries and 34 variables were removed to keep only complete-cases; 1 variables highly correlated among them ($\rho > 0.90$) and 11 variables least correlated with dependent variable ($\rho < 0.15$) were removed to run analysis at the most with 30 variables.

Table 7: Summary Table for best multivariate model for Emigration

Variables	Estimate	Std. Error	t-value	Pr(>|t-value|)
(Intercept)	0.52	0.276	1.900	0.071
emp_Y25–64_NaceK	0.18	0.031	5.790	<0.001
emp_Y15–24_ED5–8	0.03	0.006	4.420	<0.001
emp_Y15–24_NaceO–Q	-0.03	0.011	-2.880	0.009
pop_Total	0.00	0.000	-2.500	0.020
emp_Y15–24_NaceB–E	-0.01	0.008	-1.870	0.075

Регрессия на главных компонентах

Table 8: Eigenvariables of Labour Market Attractiveness set

Eigenvariables	n	mean MM	Variables (>0.75 MM)
Unemployment	3	0.83	+: emp_Y[15–24]_Nace[G–I]; unemp_Y[15–24, GE25]. -: none.
Poverty	2	0.95	+: ARPR; AROPE. -: none.
Ageing Population	2	0.86	+: pop_Y[GE75]. -: pop_Y[15–24].
Education (Employed Adults)	4	0.71	+: emp_Y[25–64]_ED[3–4]. -: emp_Y[25–64]_ED[0–2]
Employment	4	0.80	+: emp_Y[15–24, 25–64] -: none.
Earnings	51	0.76	+: earn_OC[1–5, 7–9]_Nace[B–F]; earn_OC[1–5, 7–9]_Nace[G–N]; earn_OC[1–5, 9]_Nace[P–S]; emp_T[P]; emp_Y[25–64]_Nace[M_N, O–Q]; GDP; rooms_pp; training; -: mat_depriv.

n, number of variables grouped in Eigenvariable; MM, module membership calculated as the correlation to belonging Eigenvariable.

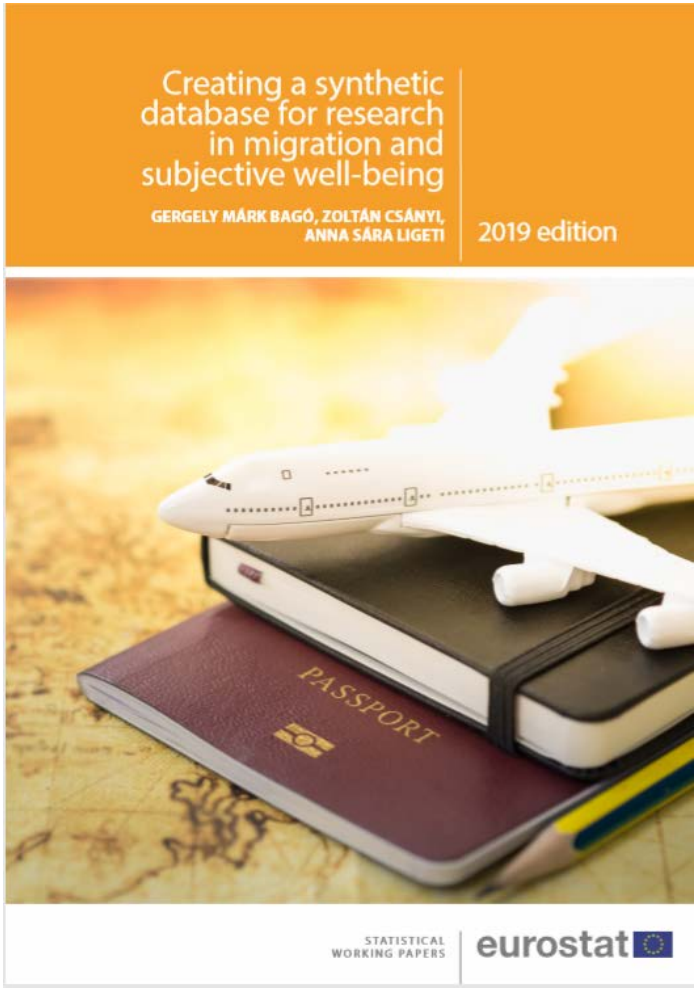
Note: 8 variables with more than 15% of missing data were removed. 2 variables did not group into any Eigenvariable.

Table 9: Correlation between labour market indicators and Eigenvariables

Eigenvariables	Skills Mismatch	Mobility	Emigration
Unemployed	0.36 (0.385)	-	-
Poverty	0.38 (0.352)	-	-
Ageing Population	-	-	-0.36 (0.063)
Education (Employed Adults)	-0.38 (0.352)	-	-0.50 (0.007)
Employed	-0.69 (0.058)	0.35 (0.082)	-
Earnings	-	0.59 (0.002)	-

[-] correlations between -0.30 and 0.30.

Note: Correlations calculated using Spearman correlation (p-values between brackets).



Statistical matching and hot deck variants applied

Объединение данных двух источников (статистики миграции и статистики благосостояния) STATMATCH

Table 2: Scheme of the experiment

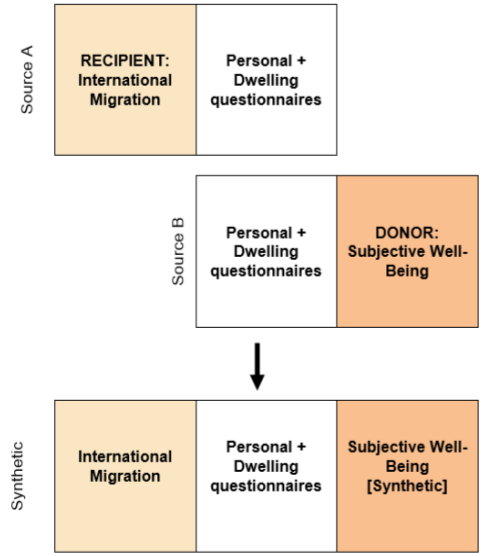
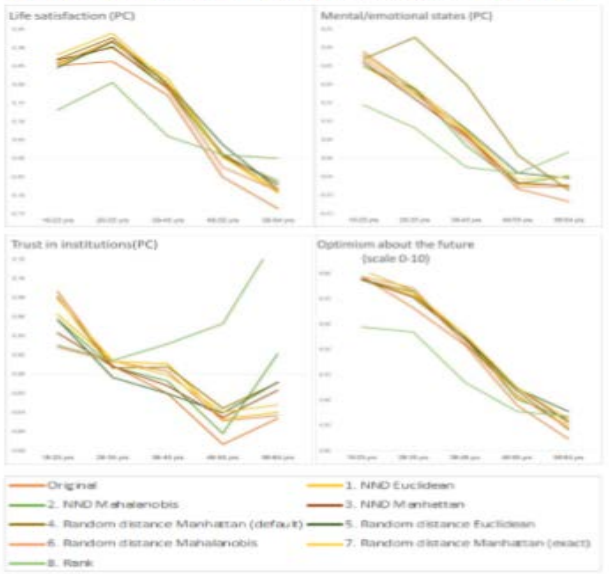
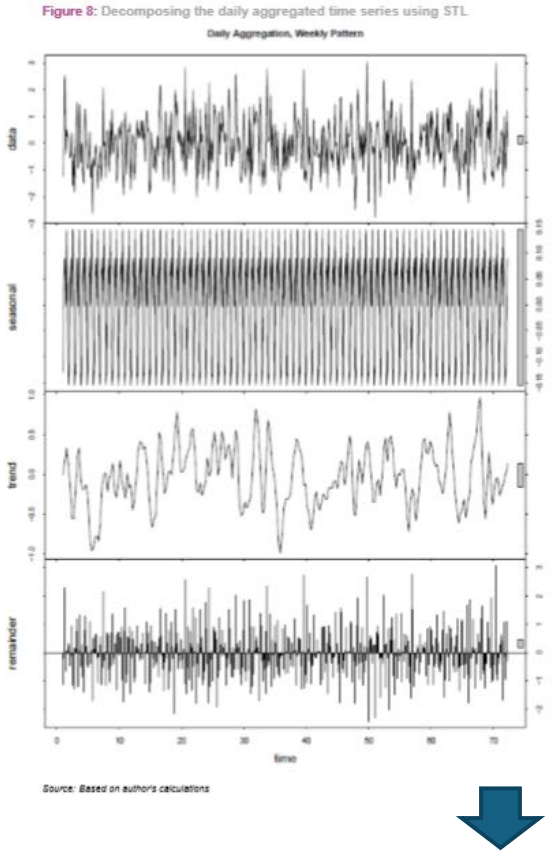


Table 3: Variants of hot deck methods, matching variables and donation classes used in the experiment

	Nearest neighbour distance hot deck			Random distance hot deck				Rank hot deck
	Method 1	Method 2	Method 3	Method 4	Method 5	Method 6	Method 7	Method 8
Distance Method	Euclidean	Mahalanobis	Manhattan	Default	Euclidean *	Mahalanobis*	Manhattan *	**
Constraint	Unconstrained							
Matching variables	Age, partner, underage child, type of settlement							Age
Class variables (donation classes)	Sex, Education							
Donor variables	Satisfaction (PC), Mental / emotional state (PC), Trust in institutions (PC), Optimism about the future							

Figure 1: Mean values of the original and the synthetic donor variables by age groups





Electronic payments data.

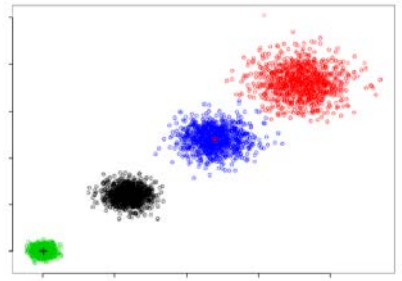
Mobile phone usage data.

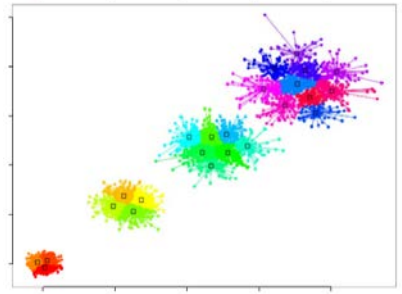
Sensor data.....

Satellite images data.....

Price data.....

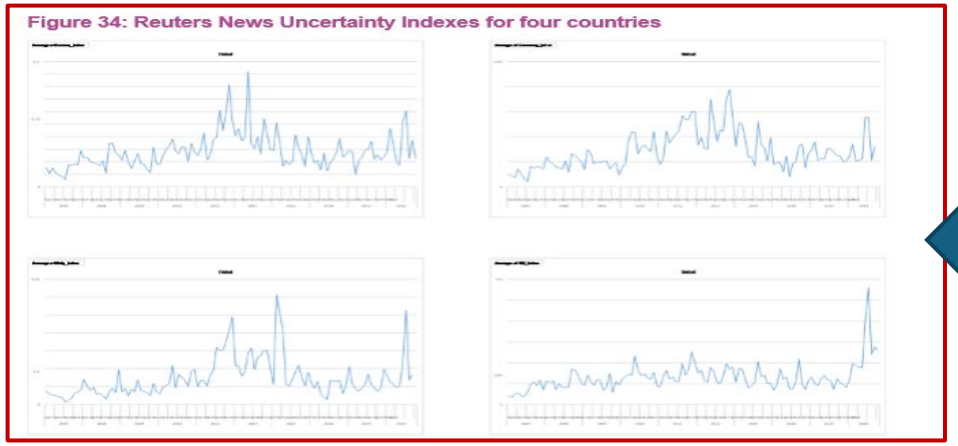
Textual data.....





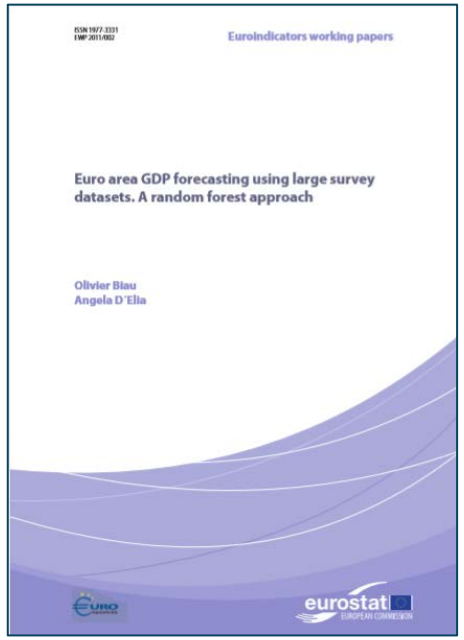
R packages

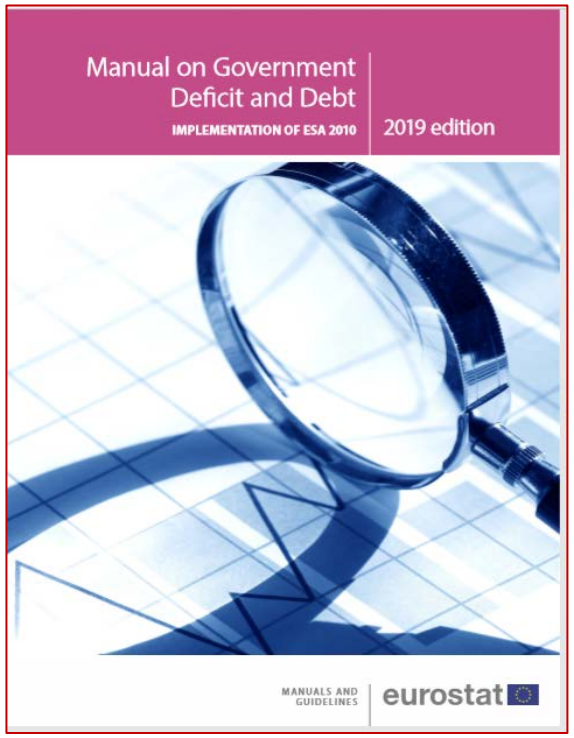
- "highfrequency" for intraday data,
- "ggmap" to plot the position of user 100 in the Sensor data analysis on Google maps,
- "gtrendsR" to download Google Trends series in R,
- "twitteR" to download Twitter series in R,
- "lubridate" to manipulate dates,
- "apcluster" for APC,
- functions from "clusterv" package,
- "rCUR" for CUR.

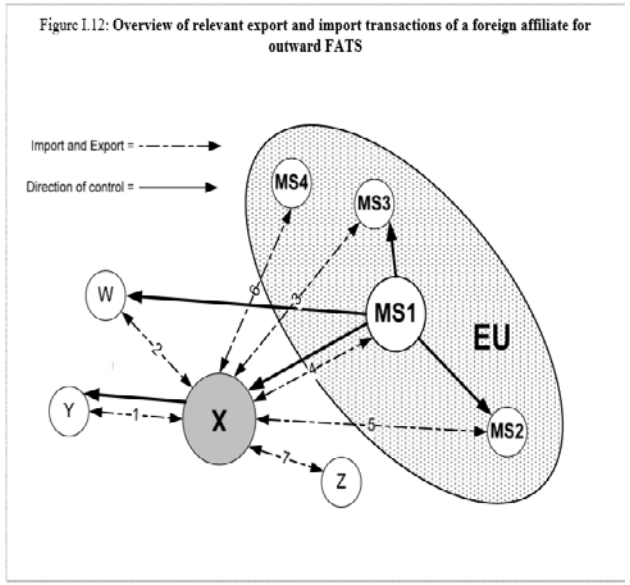


TEXT MINING

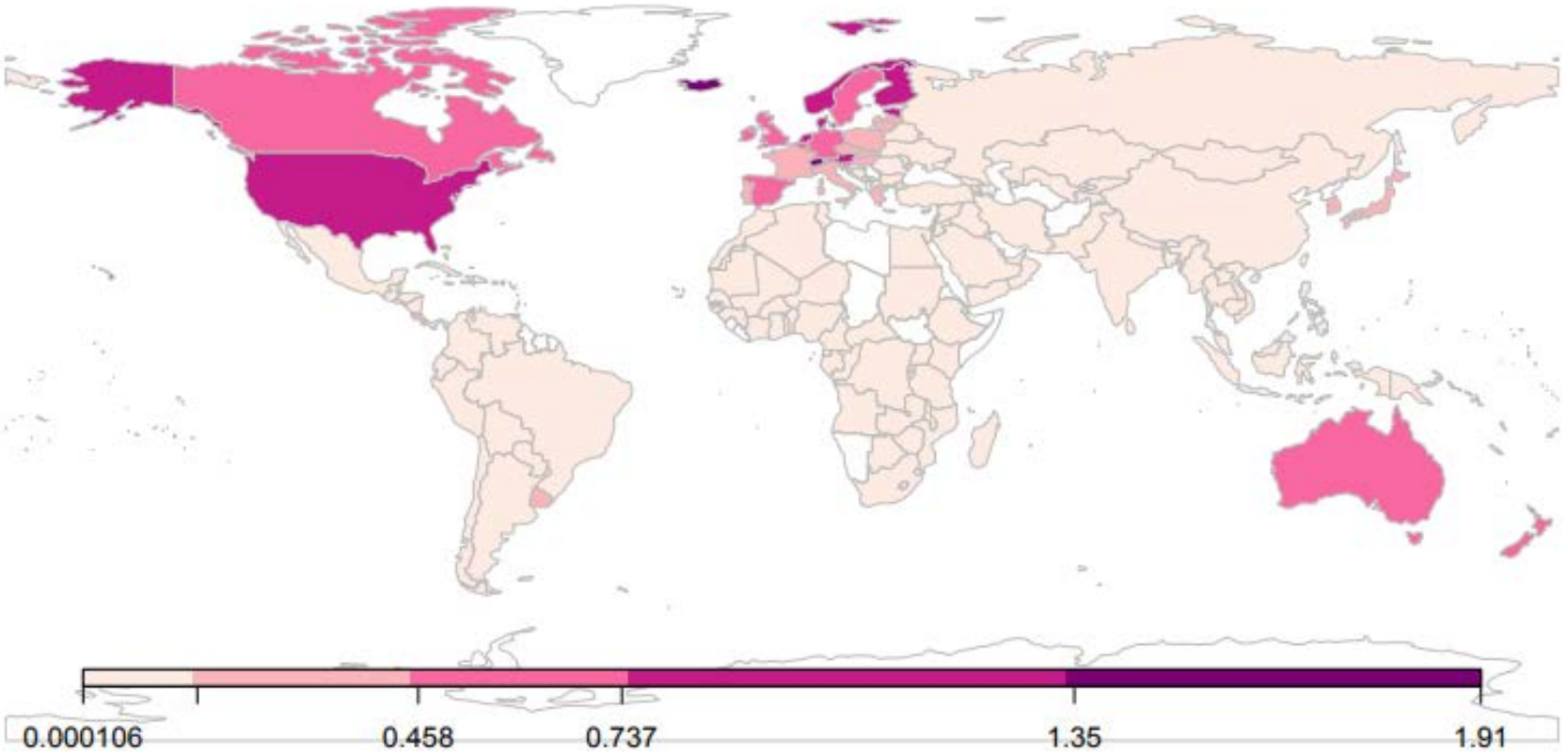
Выявление структурных «паттернов» и моделирование зависимостей с использованием технологий “random forest”







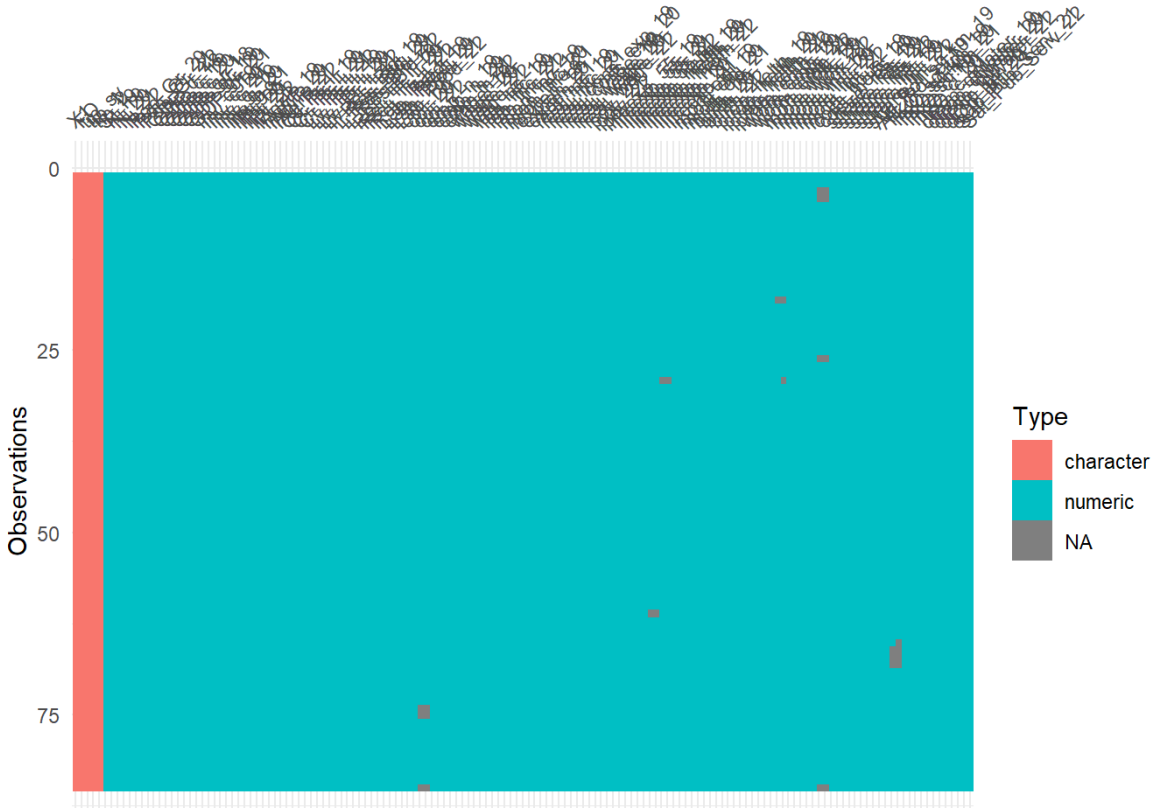
Использование пакетов R в официальной статистике

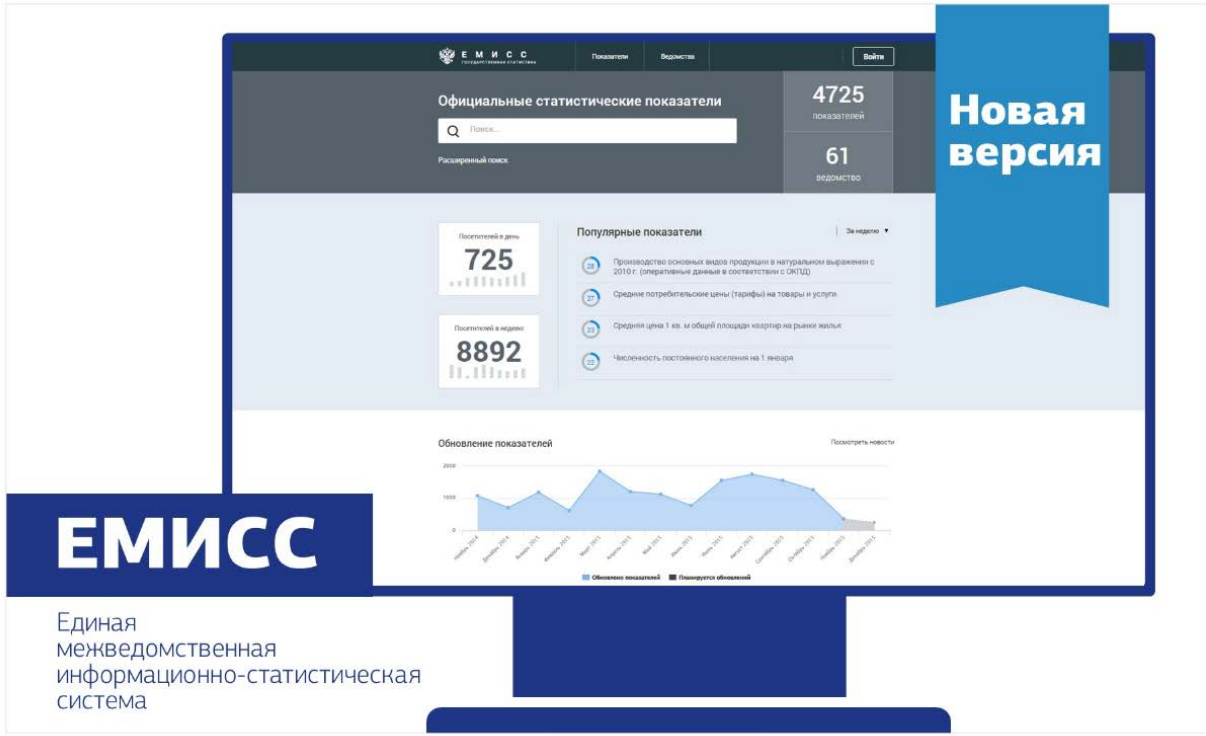


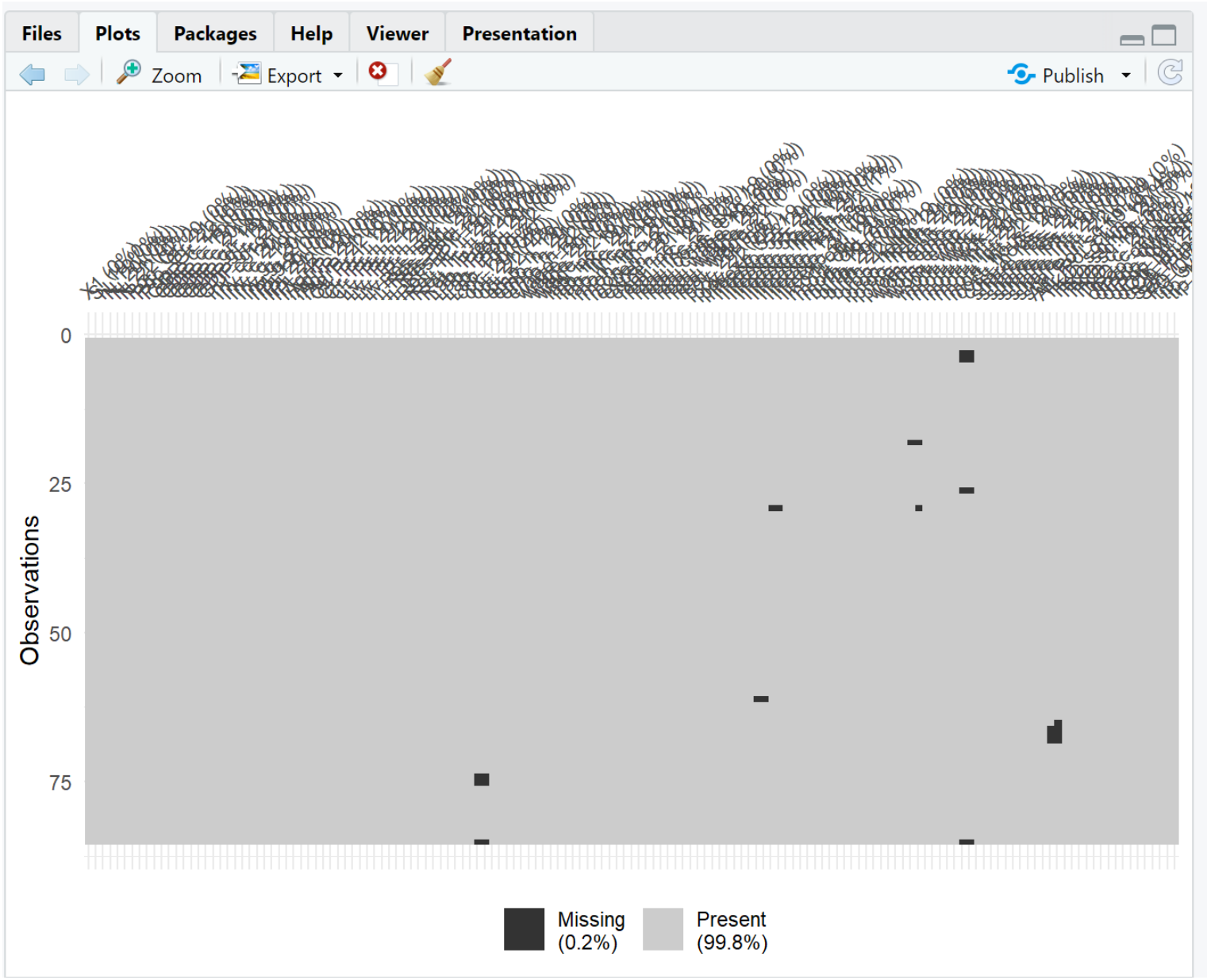
Загрузки пакетов R, включенных в представление задач CRAN по официальной статистике и методологии опросов по всему миру. Количество загрузок представлено на душу населения (то есть нормализовано путем деления на количество населения)

https://www.researchgate.net/publication/297656682_The_Software_Environment_R_for_Official_Statistics_and_Survey_Methodology

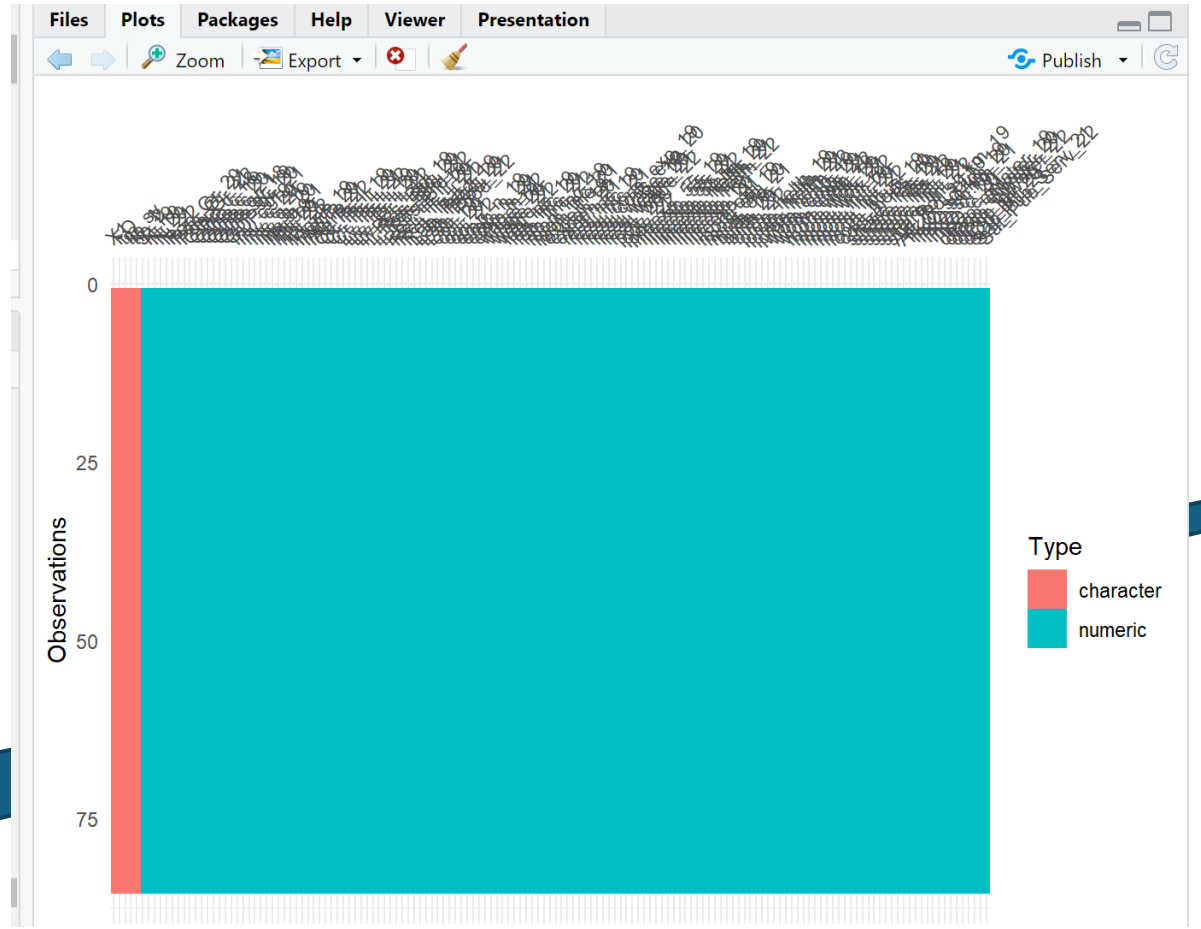
ПРИМЕР (1) на данных ЕМИСС. Региональные данные ЕМИСС -148 показателей, 2022 г.



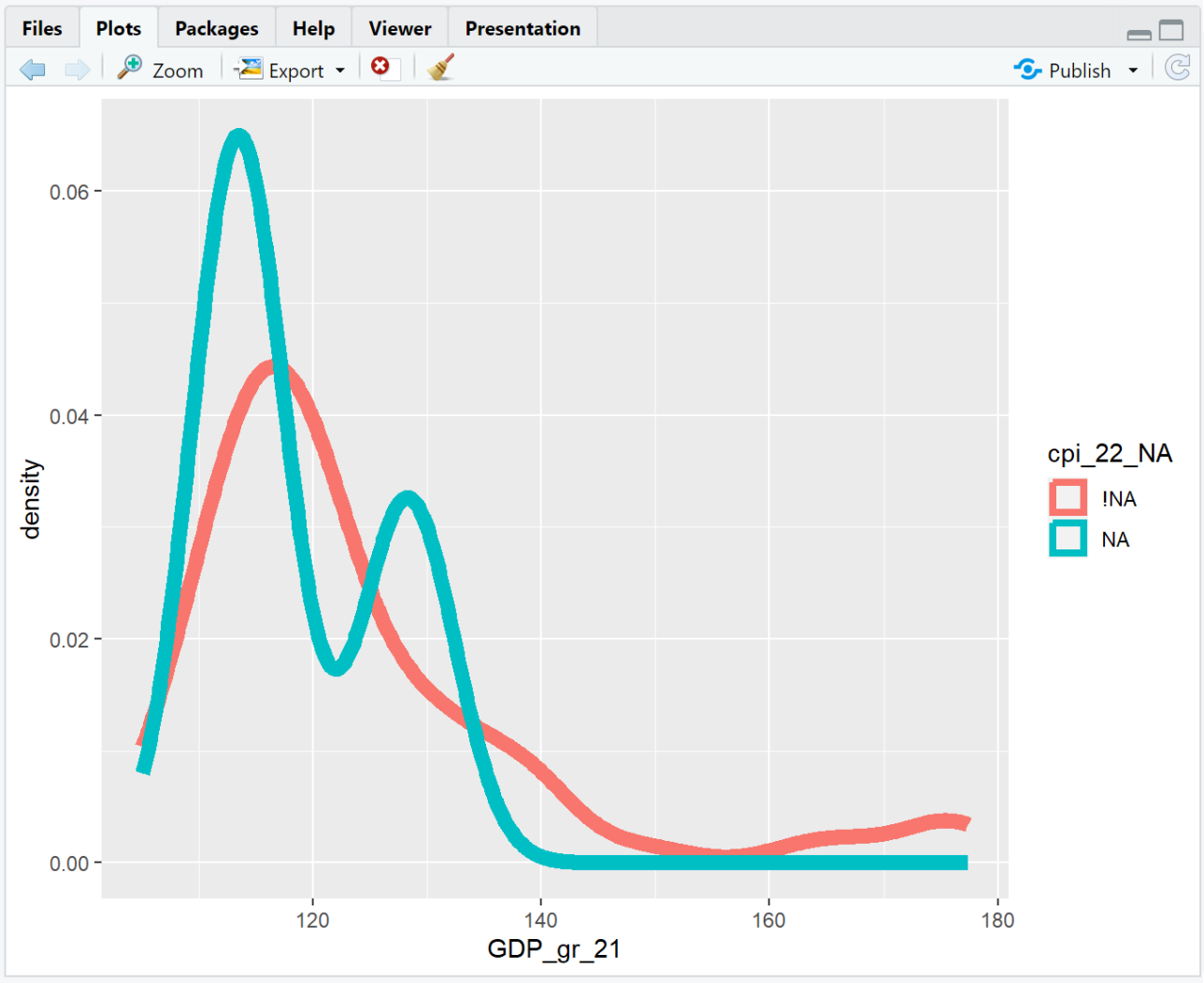


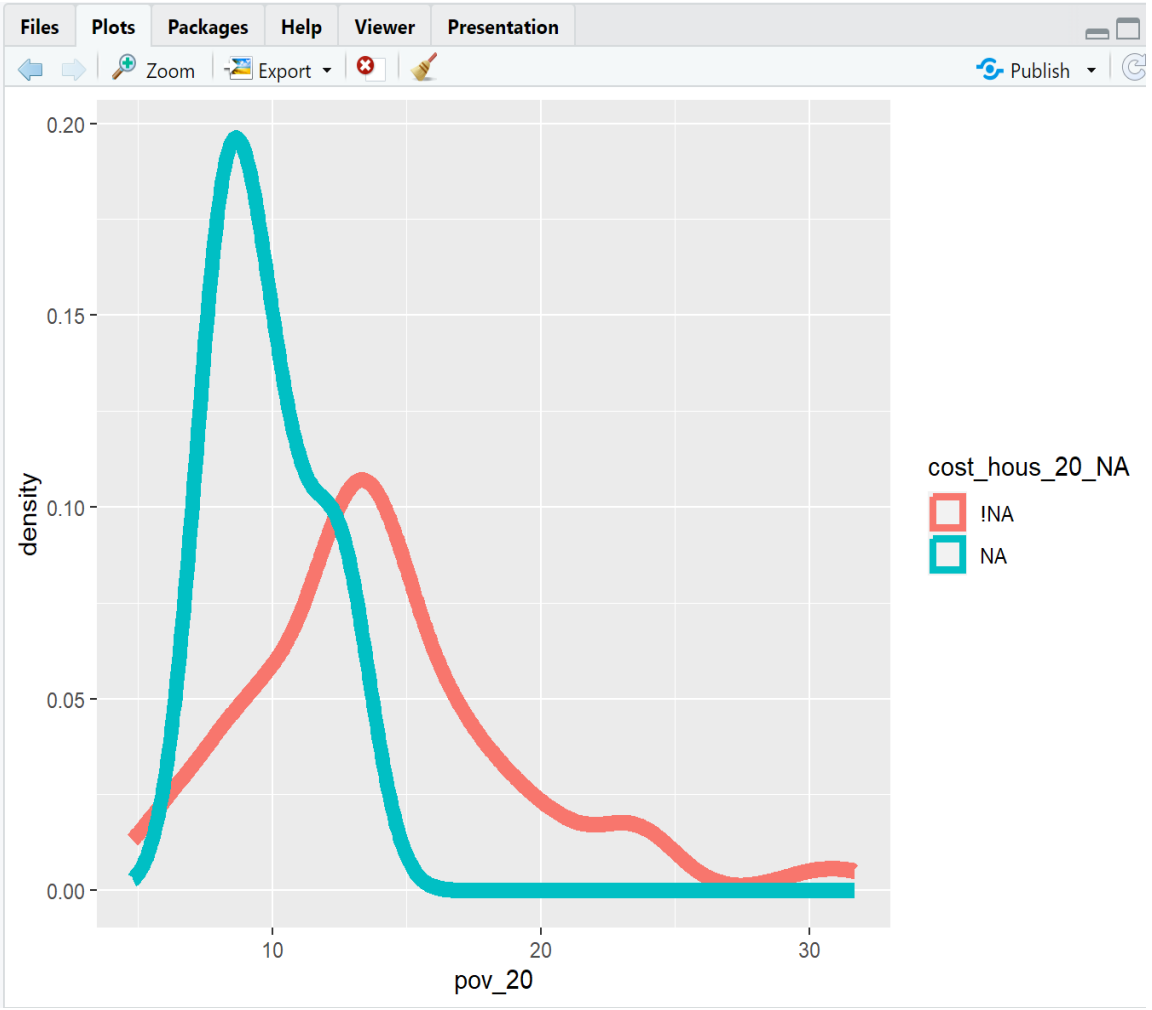

Files
Plots
Packages
Help
Viewer
Presentation
Zoom
Export
Publish
0
25
50
75
Observations
Missing
(0.2%)
Present
(99.8%)

Метод машинного обучения

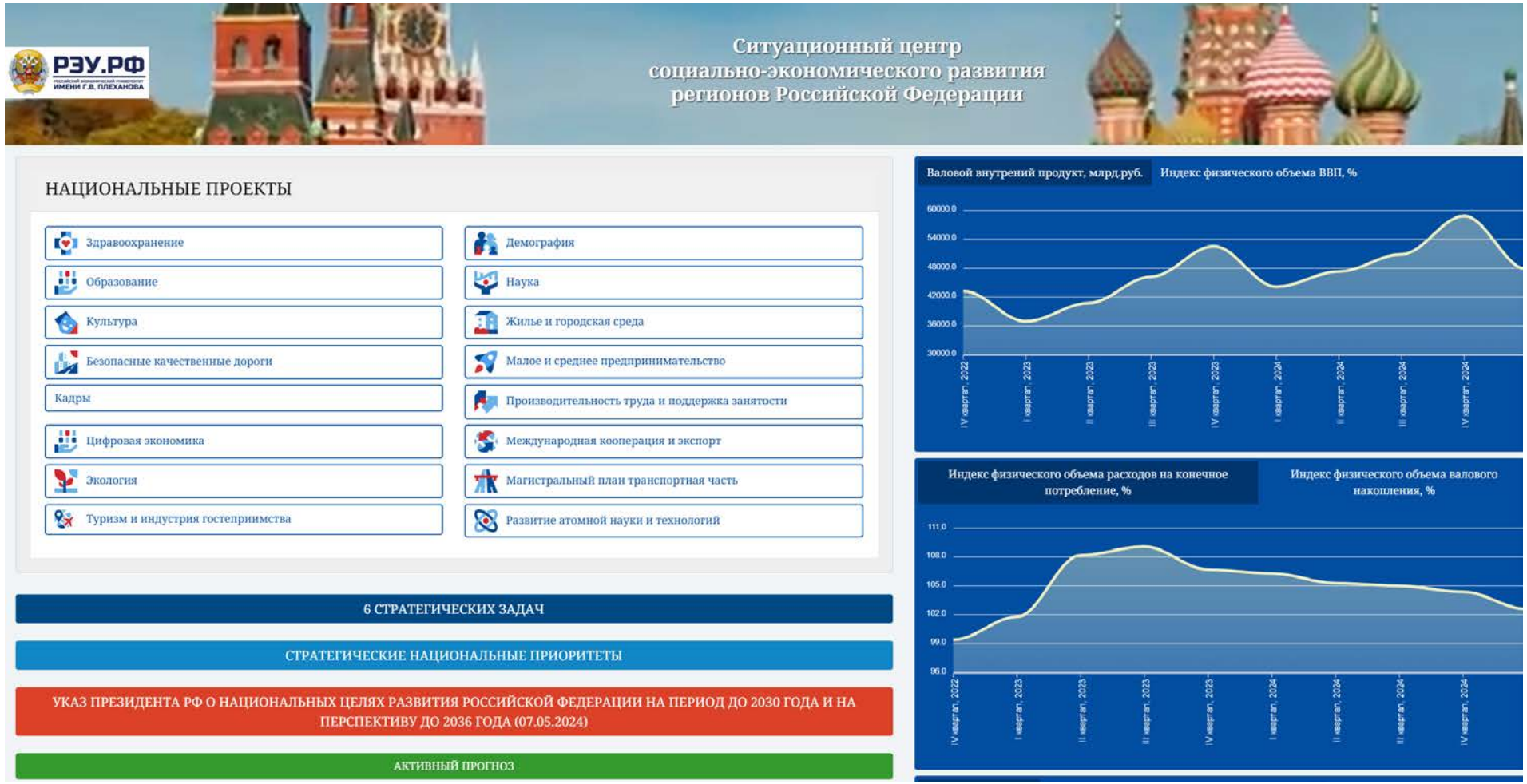


Технологии ИИ – предсказание пропусков данных



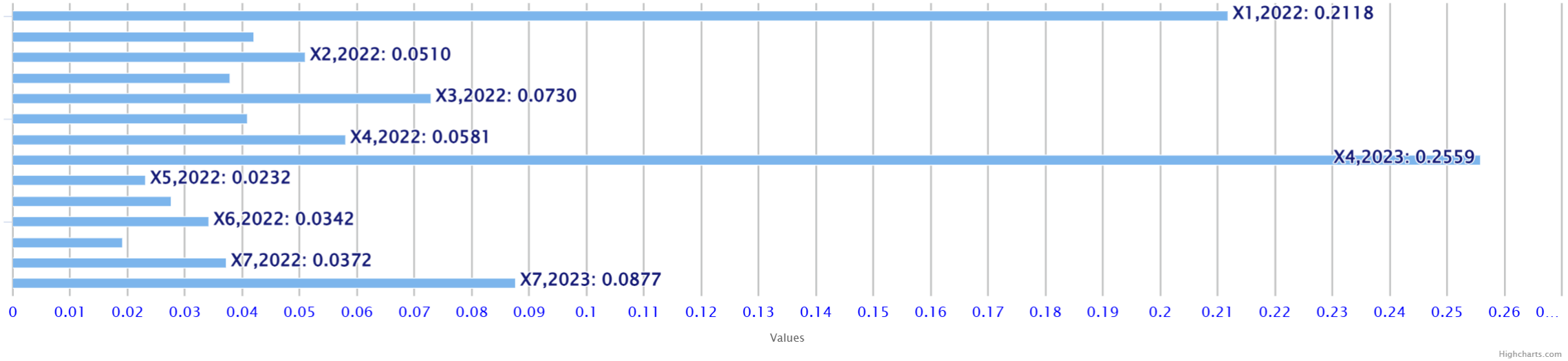


ПРИМЕР (2)платформа СЦ РЭУ им. Г.В. Плеханова



ЗНАЧИМОСТЬ ФАКТОРНЫХ ПЕРЕМЕННЫХ ДЛЯ ЦЕЛЕВОЙ ПЕРЕМЕННОЙ

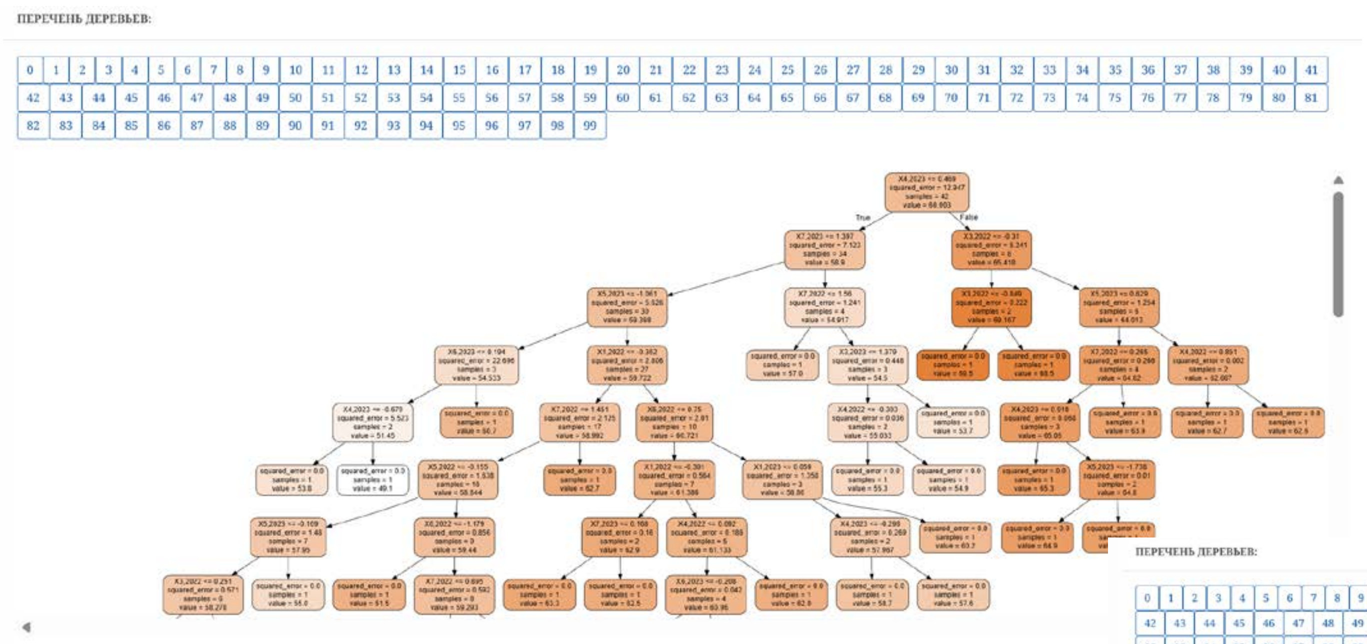
Труд / Уровень занятости населения, по данным выборочных обследований рабочей силы; в процентах

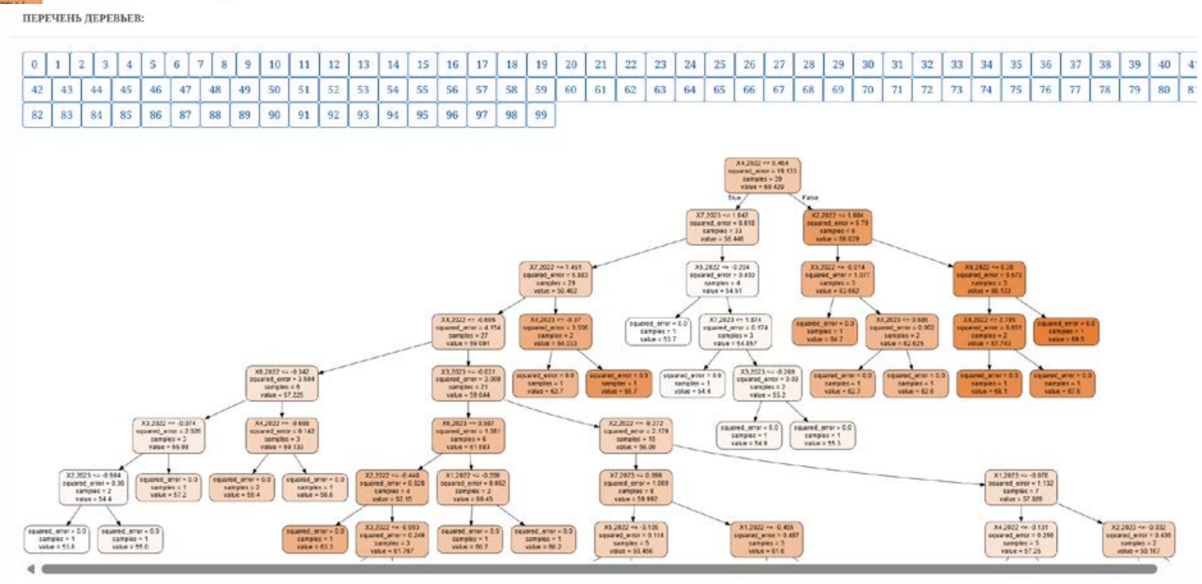


Описание	Переменная
Труд / Уровень занятости населения, по данным выборочных обследований рабочей силы; в процентах	**Y1**
Уровень жизни населения / Среднедушевые денежные доходы населения, в месяц; рублей	**X1**
Уровень жизни населения / Среднемесячная номинальная начисленная заработная плата работников организаций, рублей	**X2**
Реальные денежные доходы / Реальная начисленная заработная плата работников организаций, в процентах к предыдущему году	**X3**
Инвестиции / Инвестиции в основной капитал на душу населения, в фактически действовавших ценах; рублей	**X4**
Образование / Валовой коэффициент охвата дошкольным образованием, на конец года; в процентах от численности детей в возрасте 1-6 лет	**X5**
Численность зрителей театров и число посещений музеев на 1000 человек населения / Численность зрителей театров	**X6**
Здравоохранение / Заболеваемость на 1000 человек населения, зарегистрировано заболеваний у пациентов с диагнозом, установленным впервые в жиз	**X7**

КОЭФФИЦИЕНТ ДЕТЕРМИНАЦИИ

$R^2 = 0.9024$


ПЕРЕЧЕНЬ ДЕРЕВЬЕВ:
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41
42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81
82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99


ПЕРЕЧЕНЬ ДЕРЕВЬЕВ:
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40
42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80
82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99

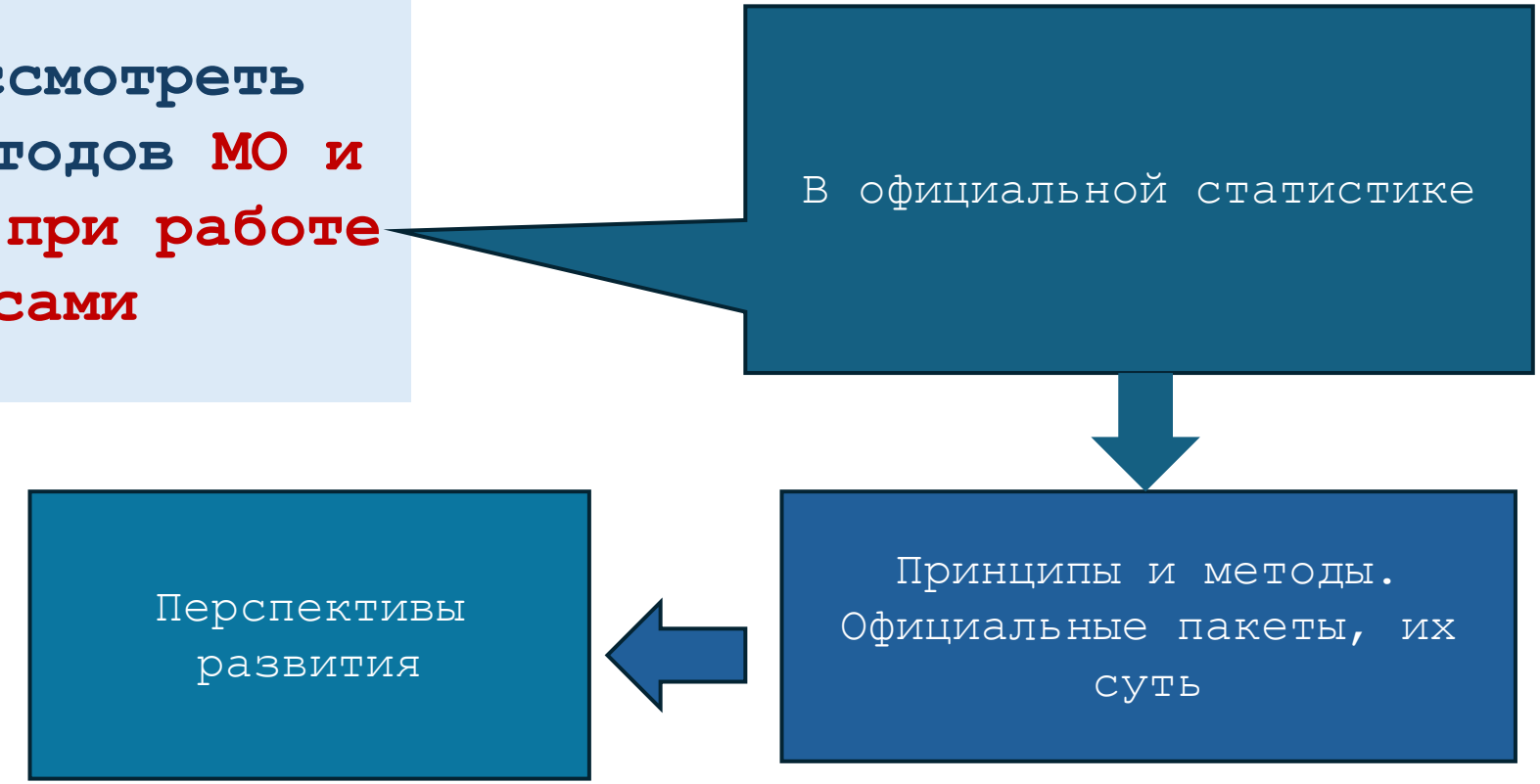
Выводы по разделу

1. **ИИ формирует потенциал аналитической функции официальной статистики, требования к развитию которой будут усиливаться со стороны пользователе**
2. **Потенциальным риском является «поведенческий дрейф», который возникает, когда модели ИИ корректируют свои выходные данные в соответствии с новой информацией. Это создает проблему, поскольку официальная статистика основана на достоверности, последовательности и сопоставимости.**
3. **Без сомнения, НСУ должны активно проверять системы искусственного интеллекта и снижать риски посредством тщательного тестирования и прозрачных операций.**
4. **Необходима нормативная проработка использования ИИ в официальной статистике.**
5. **Создание экосистемы данных с центральной интегрирующей ролью национального статистического офиса**

Методы МО в обработке первичных данных (выявление статистических выбросов)

Выявление выбросов является важной задачей производства статистической информации

Задача – рассмотреть применение методов МО и технологий ИИ при работе с выбросами



«Выбросы»

Выбросом (*outlier*) **считается значение в данных, которое находится далеко за пределами других наблюдений**

Экстремальные значения – это устранимые или неустранимые ошибки, возможные фиктивные значения

Статистический выброс – это наблюдение, которое существенно отклоняется от основной массы данных и расположено далеко от кривой плотности распределения вероятностей, к которой относятся основной объем данных. Формально, выброс – это точка данных с низкой вероятностью в данном распределении

Типы ошибок

Причины выбросов разнообразны:

- **Ошибки измерения**
- **Ошибки ввода данных**
- **Ошибки обработки данных**

ОШИБКИ РЕГИСТАЦИИ

- **Ошибки выборки**
- **Ошибки эксперимента**
- **Естественные выбросы**. ***Эти отклонения не являются ошибками, хотя и «выбиваются» на фоне остальных данных***

ОШИБКИ РЕПРЕЗЕНТАТИВНОСТИ

Сведения о применении пакетов машинного обучения в официальной статистике стран СНГ (пакеты R)

BY Белстат (Беларусь)

- `rpart` - для автоматической классификации видов экономической деятельности
- Базовые ML-пакеты в проектах по big data

KZ Казстат (Казахстан)

- `rpart` - для стратификации в выборочных обследованиях домохозяйств
- `e1071` - контроль качества данных переписи населения

RU Росстат (Россия)

- `e1071` - для обнаружения выбросов в экономической статистике
- `rpart` - для классификации предприятий по обороту
- `caret` - в исследовательских проектах по прогнозированию

UZ Узстат (Узбекистан)

- Ограниченное использование ML-пакетов
- `rpart` - в пилотных проектах по прогнозированию урожая

ROSA: выявление выбросов (одномерный и многомерный подходы, применение методов МО)

ROSA (R FOR OFFICIAL STATISTICS AND DATA ANALYSIS):

Создатель и целевая аудитория:

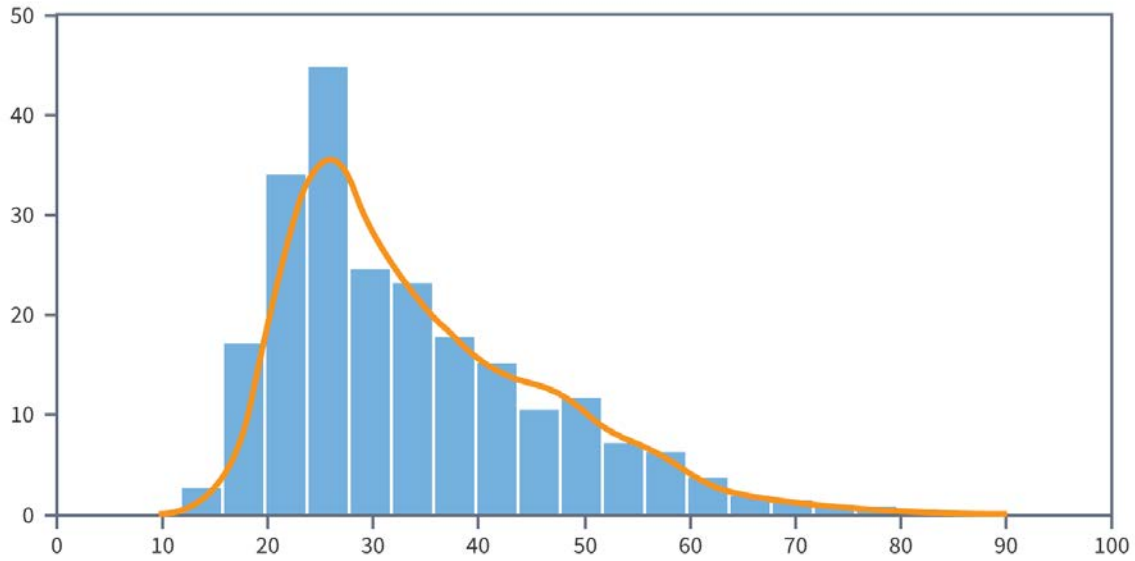
- **Разработчик:** Евростат (Eurostat), статистическая служба Европейского Союза.
- **Для кого:** Специально создан для национальных статистических офисов и официальных статистиков.

Основные функции пакета univOutl для одномерного выявления выбросов:

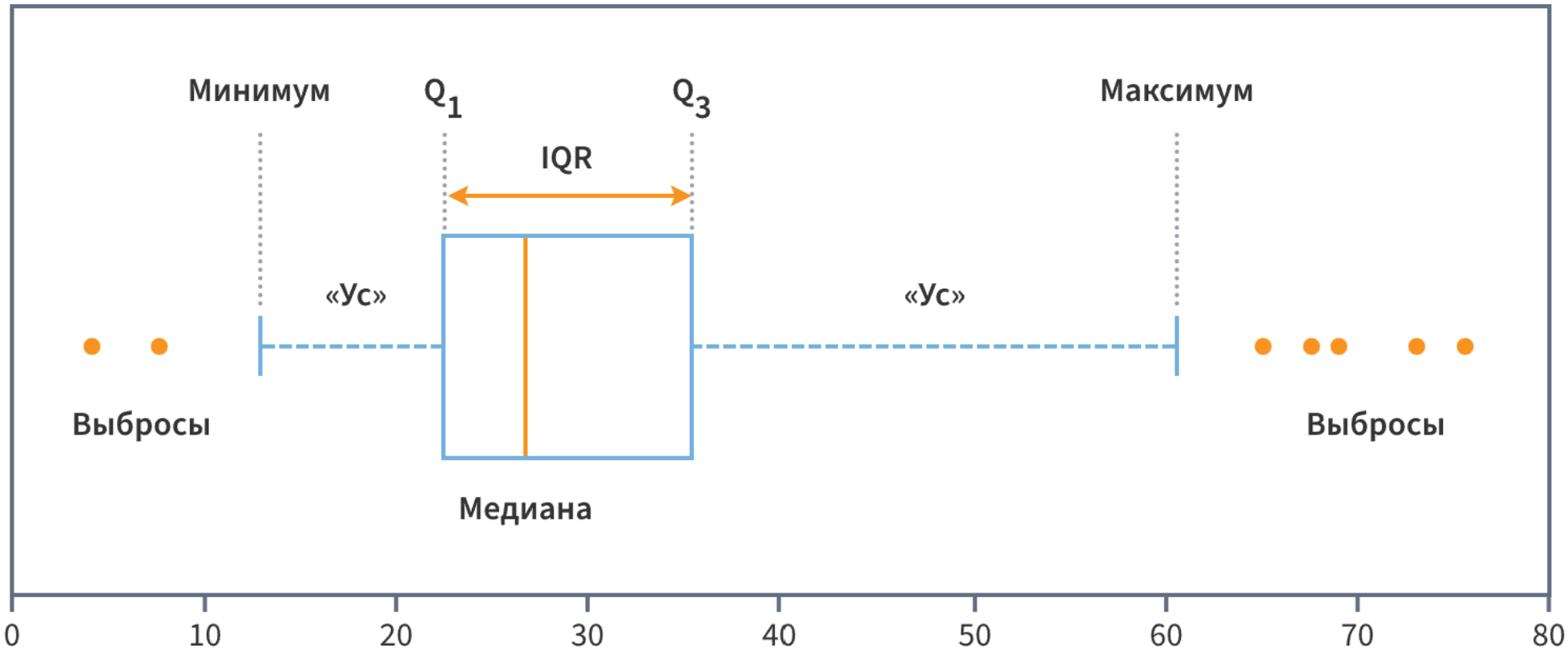
- LocScaleB(): Выявляет выбросы на основе робастного расположения и масштаба (аналог метода "медиана ± 3 MAD").
- QCD(): Использует робастный квартильный коэффициент дисперсии (QCD).
- adjbox(): Строит "скорректированные" диаграммы размаха (boxplot), которые лучше учитывают асимметрию данных.
- HDoutliers(): Алгоритм для обнаружения выбросов на основе теории больших отклонений (Heavy-Depth), эффективен для многомодальных распределений.

Практический пример

Асимметричное распределение



IQR – метод установления выбросов



Микроданные ОРС (фрагмент)

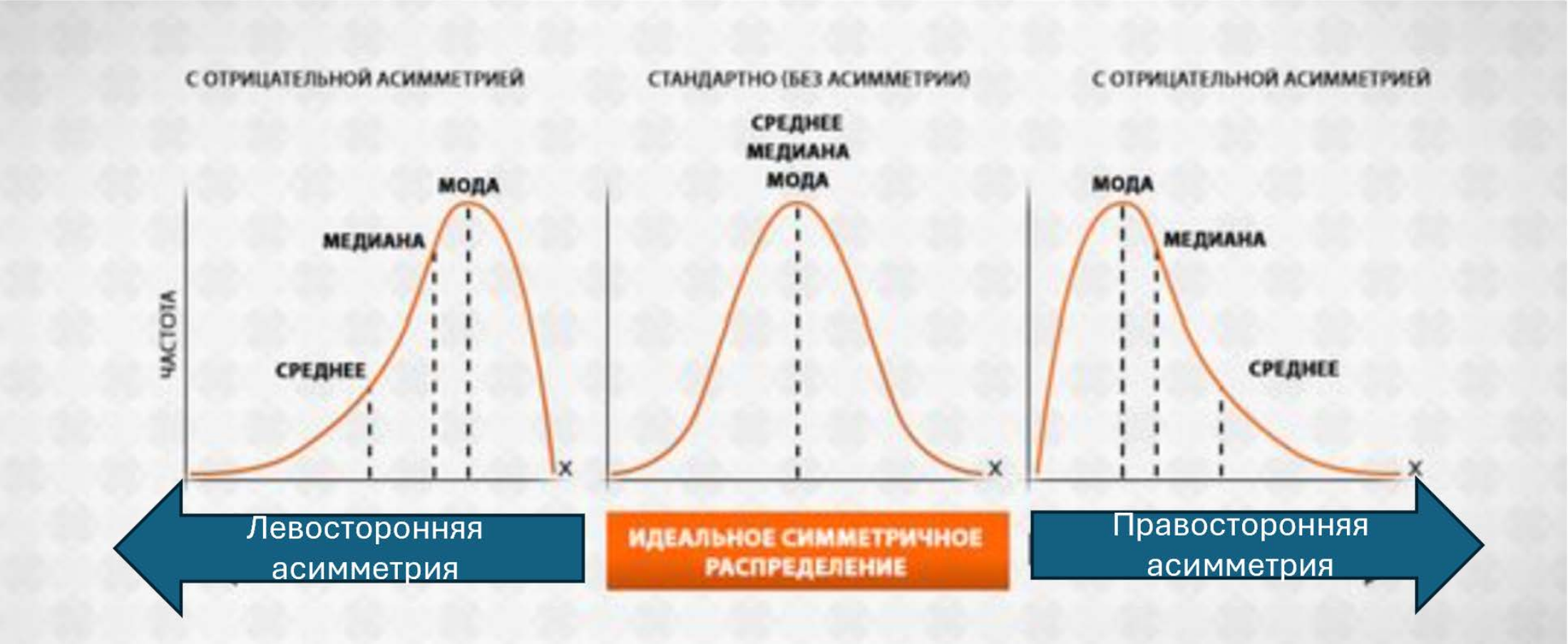
EA_TM_CTRAT.R | OUTLIERS_1.R | LFS_300 | DESCR. STAT..R | EDA_1.R | ROSSTAT_2025.R | 11.R | reg_1

Filter

	territ	posel	BB2	NAS_POL	NAS_VOZR	NASOBRAZ	INVALID	V_OSNZAN	KAT_TRUD1	V_OSNRB	OKZ_OSN1
1	40	1	4	1	51	10	0	1	2	1	
2	76	1	2	1	68	10	0	0	0	0	
3	40	1	4	2	20	5	0	0	0	0	
4	40	1	4	2	46	10	0	1	2	1	
5	45	1	1	2	73	8	0	0	0	0	
6	89	1	1	2	82	3	1	0	0	0	
7	88	2	2	1	68	4	0	0	0	0	
8	63	1	3	1	28	3	0	10	2	1	
9	63	1	2	1	22	5	0	0			
10	45	1	3	2	16	6	0	0			
11	52	2	3	2	48	5	0	1			
12	87	1	1	1	78	4	1	0			
13	87	1	3	1	29	11	0	1			
14	52	2	1	2	73	4	1	0			
15	63	1	2	2	75	5	0	0			

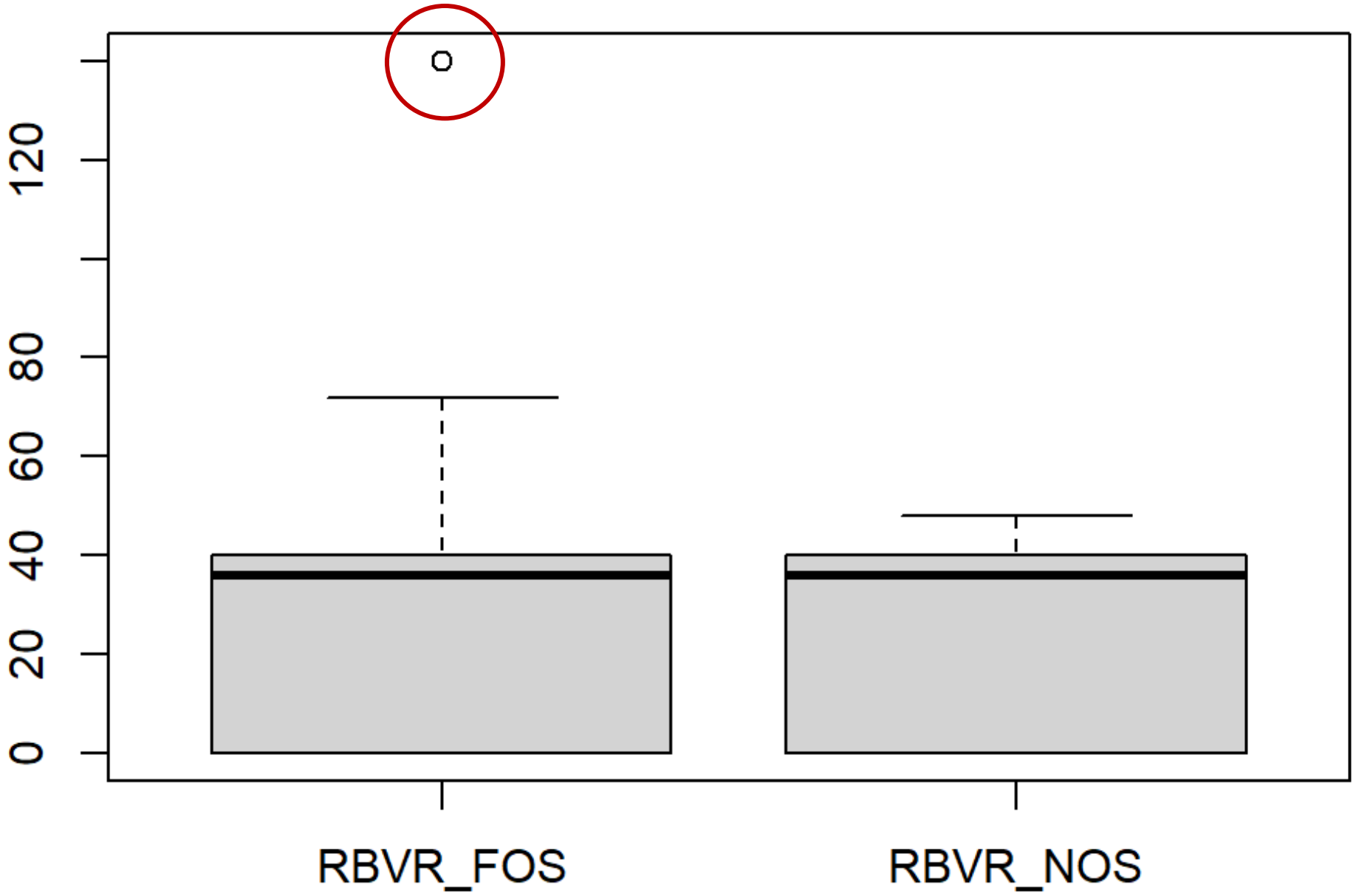
```
> summary(LFS_300[c('RBVR_FOS','RBVR_NOS')])
    RBVR_FOS         RBVR_NOS
 Min.   :  0.00   Min.   : 0.00
 1st Qu.:  0.00   1st Qu.: 0.00
 Median : 36.00   Median :36.00
 Mean   : 21.61   Mean   :21.37
 3rd Qu.: 40.00   3rd Qu.:40.00
 Max.   :140.00   Max.   :48.00
>
```

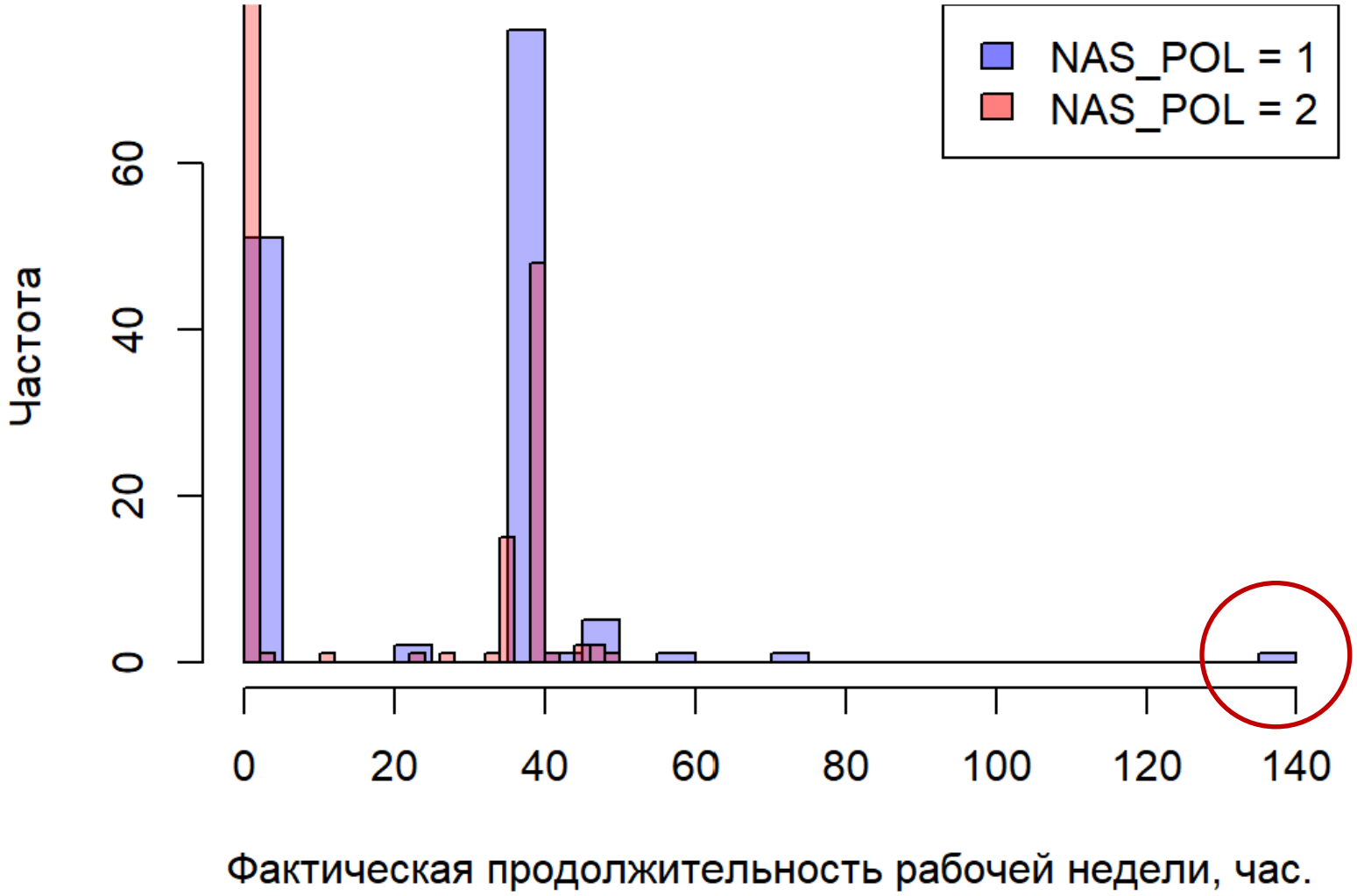
Справочно: типы асимметричных распределений



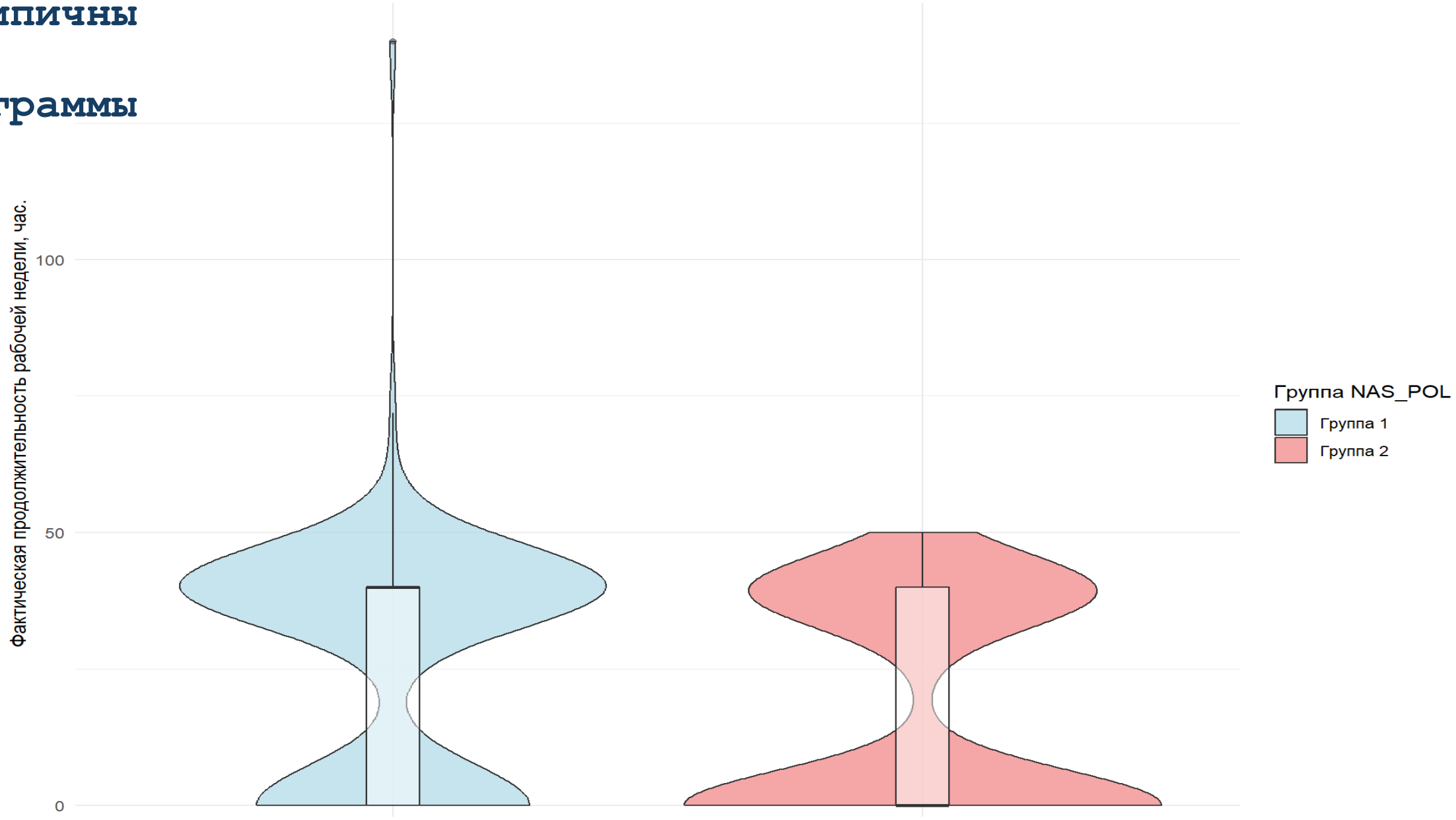
Описательная статистика

Статистическая характеристика	RBVR_FOS (час.) - фактическая продолжительность рабочей недели	RBVR_NOS (час.) - Нормальная продолжительность рабочей недели
Среднее (mean)	21.61	21.37
Станд. отклонение (sd)	21.34	19.68
Минимум	0	0
Максимум	140	48
Коэффициент вариации, %	**98,6**	**92,0**


0
20
40
60
80
120
RBVR_FOS
RBVR_NOS


NAS_POL = 1
NAS_POL = 2
Частота
0
20
40
60
0
20
40
60
80
100
120
140
Фактическая продолжительность рабочей недели, час.

Скрипичные диаграммы



Типы переменных в примере (*микроданные ОРС*) – для применения МО «без учителя» (кластерный анализ) с целью выявления статистических выбросов

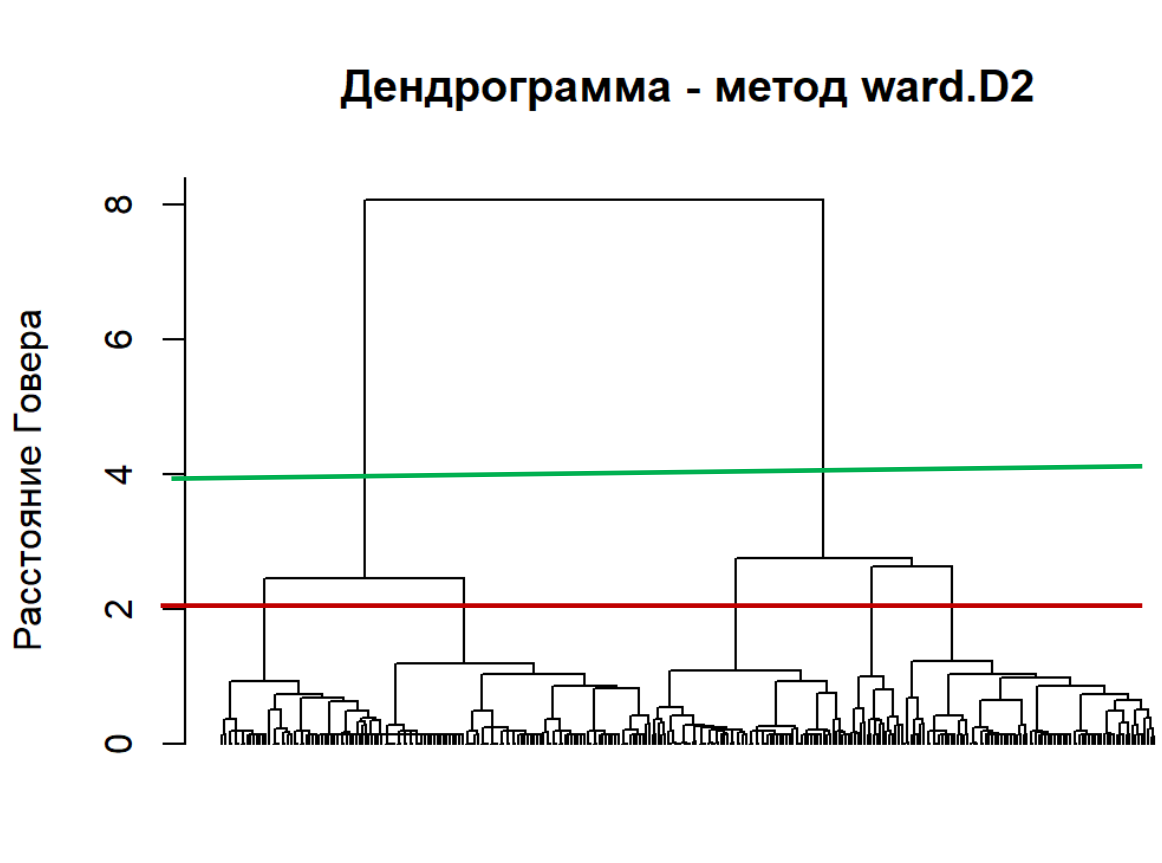
Количественные (числовой тип)

RBVR_FOS,
RBVR_NOS

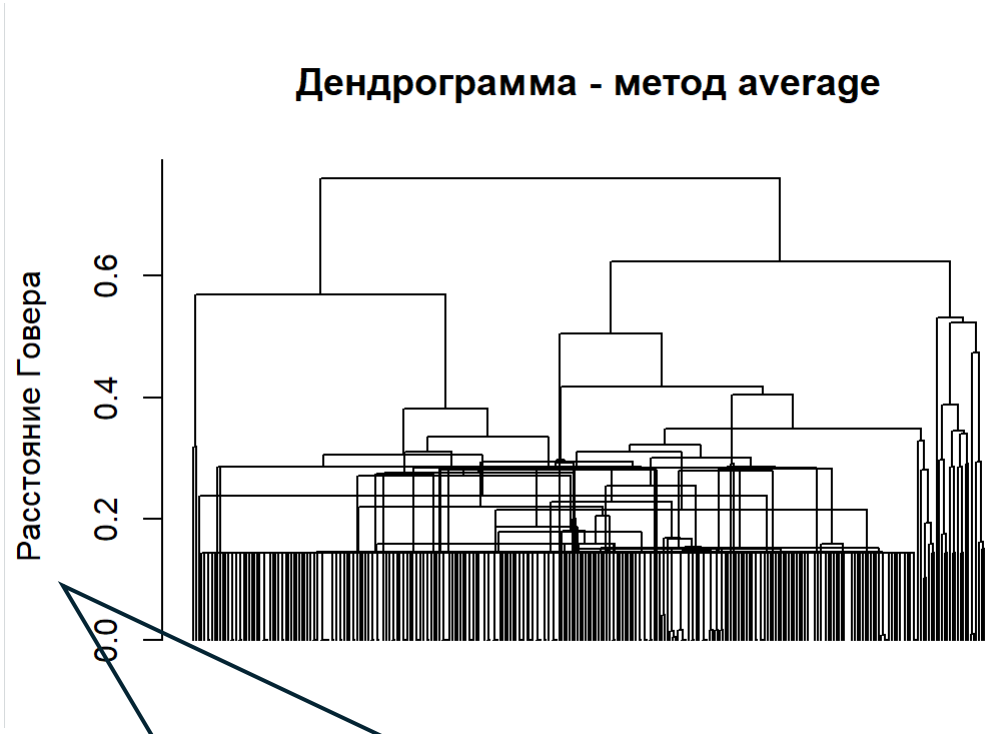
Категориальные (факторный тип)

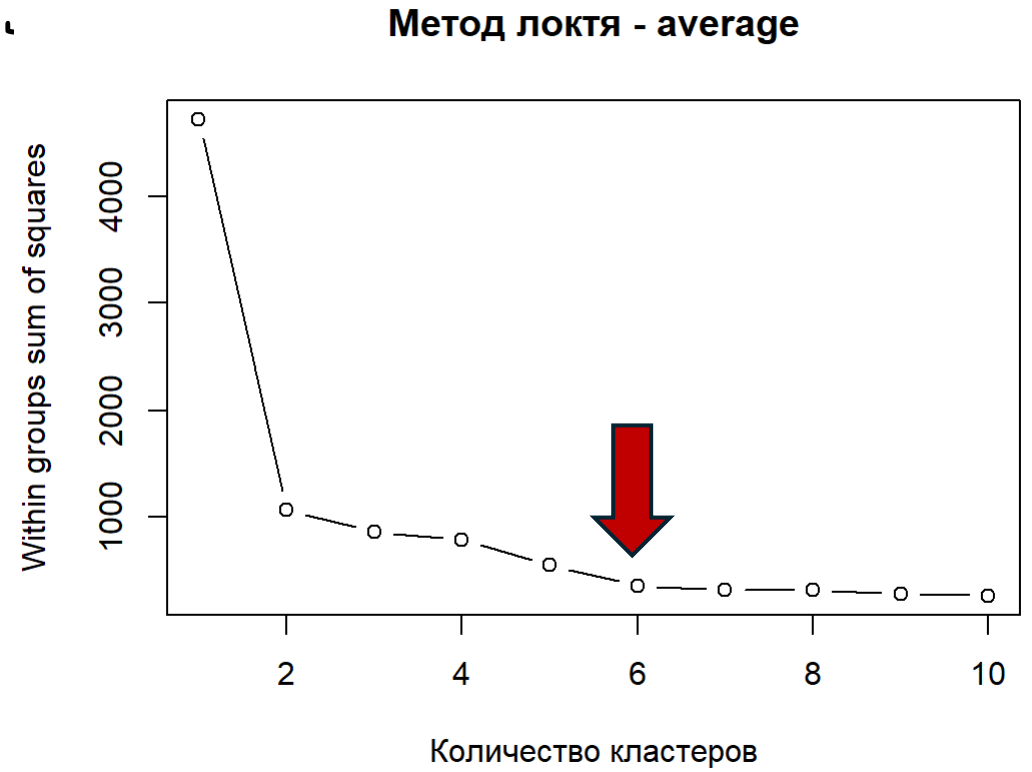
```
NAS_POL : 2 уровней
NAS_VOZR : 71 уровней
NASOBRAZ : 9 уровней
V_OSNZAN : 6 уровней
V_OSNRB : 4 уровней
```

Метод 1. МО «без учителя» - иерархическая кластеризация



Отсечение ранжированных высот на 90-м перцентил





Расстояние Говера* - это "универсальная метрика" для смешанных данных.
= "Умный способ измерить похожесть между объектами, когда ***данные разного типа (Приложение 1)***"

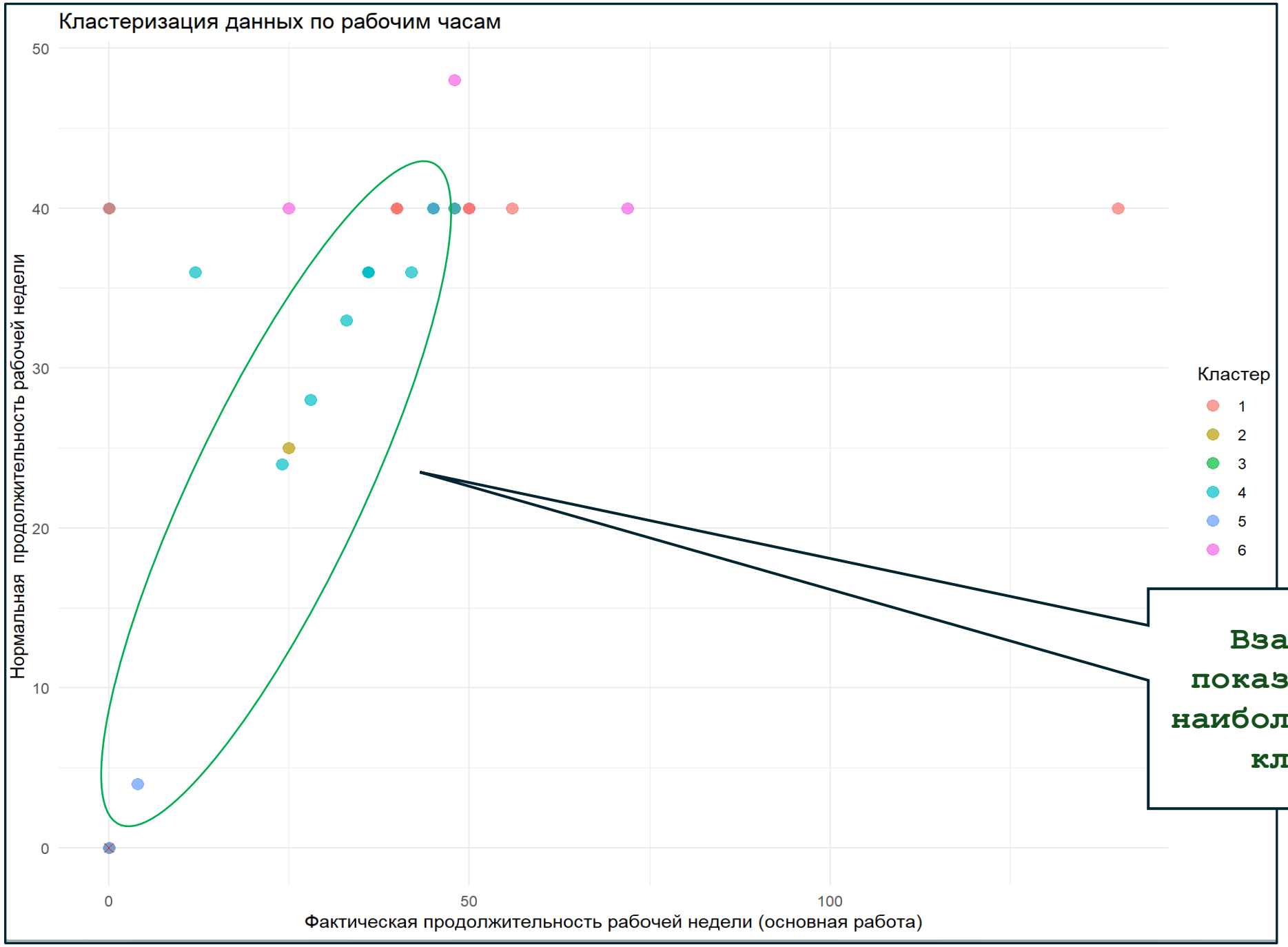
6 кластеров

Результат выделения статистических выбросов в микроданных ОРС (фрагмент) методом МО «без учителя»

cluster_k	n	mean_RBVR_FOS	mean_RBVR_NOS	sd_RBVR_FOS	sd_RBVR_NOS
1	81	40.9	40.0	13.1	0.444
2	51	0.490	0.490	3.50	3.50
3	64	37.7	38.4	6.69	2.97
4	81	0.0494	0.0494	0.444	0.444
3	17	42.8	40.5	9.24	1.94

Итого 294

Потенциальные «выбросы» 300-294 = 6



Взаимосвязь показателей для наиболее типичных кластеров

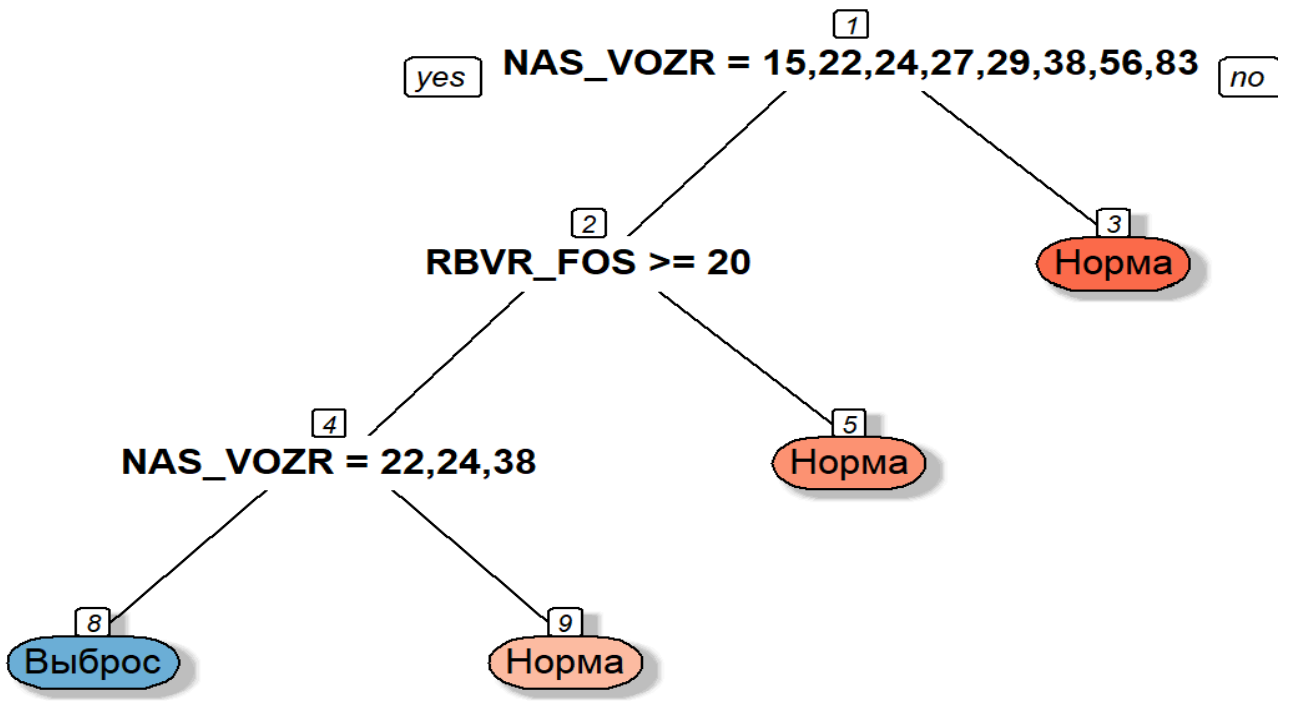
Итоговая таблица выбросов

ТАБЛИЦА ВЫБРОСОВ - Кластерный анализ - 2025-11-18

ID_наблюдения	Статус	Кластер	NAS_POL	NAS_VOZR	NASOBRAZ	V_OSNZAN	V_OSNRB	RBVR_FOS	RBVR_NOS
114	Выброс	1	2	47	3	1	1	24	24
18	Выброс	1	2	49	4	1	1	36	36
277	Выброс	1	2	46	4	1	1	48	40
27	Выброс	1	2	65	4	1	1	40	40
198	Выброс	1	2	29	4	1	1	40	40
106	Выброс	1	2	32	4	8	0	40	40
173	Выброс	2	1	44	5	5	0	25	40
98	Выброс	2	2	61	3	8	0	50	40
234	Выброс	1	2	33	4	8	0	40	40
105	Выброс	2	2	23	5	8	0	40	40
96	Выброс	2	2	45	5	8	0	40	40

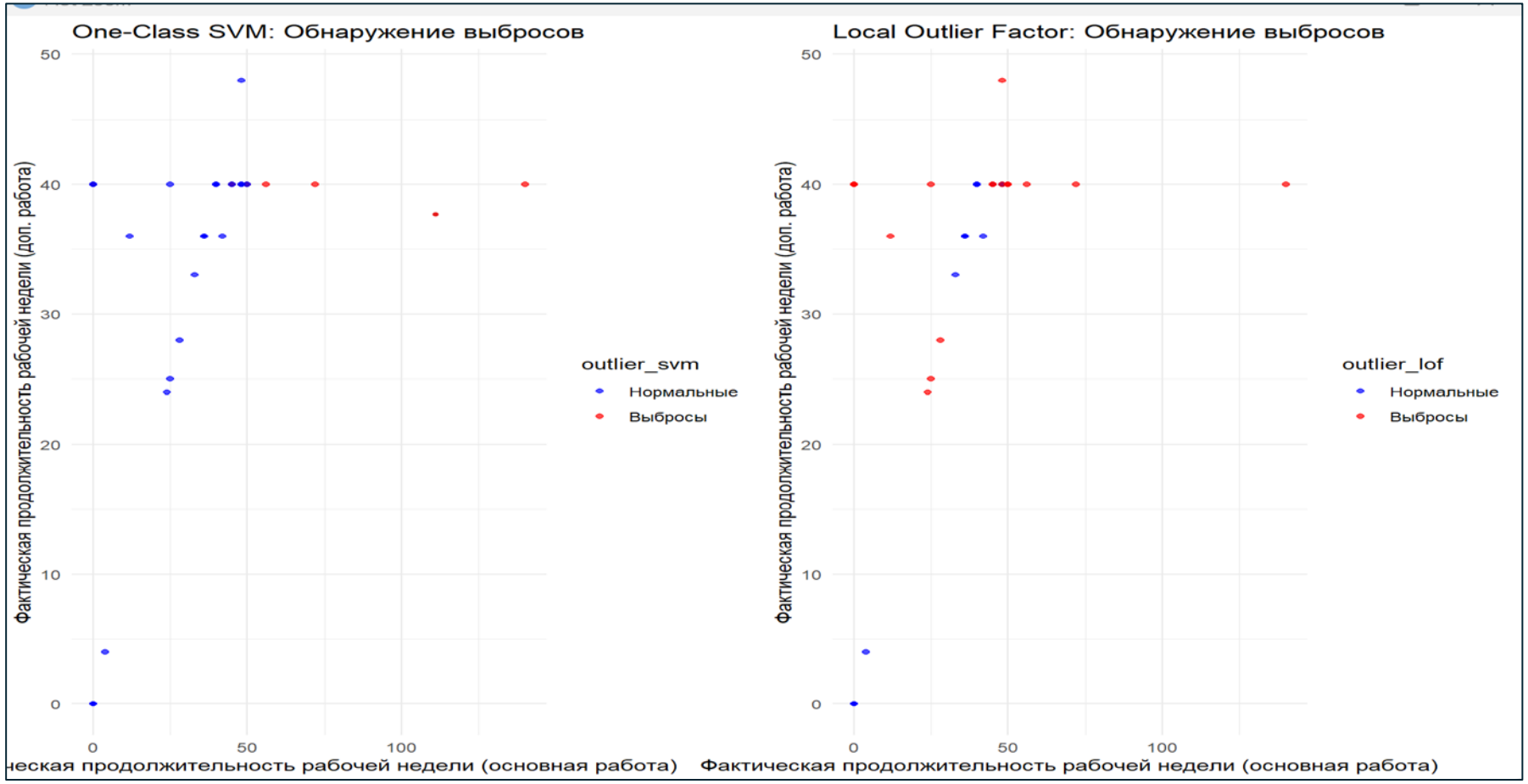
Метод 2. Выявление выбросов методом МО «с учителем»- деревья классификации (фрагмент)

Примечание 2.



Визуализация выбросов.

Примечание 3.



ПРИМЕЧАНИЯ

Примечание 1. Расстояние Говера

Расстояние Говера - это "универсальная метрика" для смешанных данных.

Если в данных имеются:

Числовые переменные (возраст, зарплата)

Категориальные (пол, профессия, цвет)

Расстояние Говера умеет корректно сравнивать объекты по всем этим типам переменных одновременно, приводя их к общей шкале.

Как работает:

- Для числовых переменных - использует нормализованную разницу
- Для категориальных - считает, совпадают значения или нет

Объединяет всё в одно итоговое расстояние

Расстояние Говера - это "универсальная метрика" для смешанных данных.

Примечание 2. Что значит «норма» и «выброс» в решении методом МО на основе деревьев классификации

Почему такое разделение?
Дерево решений автоматически выявило **естественные границы** в данных на основе:
Фактической продолжительности рабочей недели на основной работе (RBVR_FOS)
Возраста сотрудников (NAS_VOZR)
Возможно, других переменных
Критерии "выбросов" определяются статистически - это наблюдения, которые:
Значительно отличаются от большинства
Находятся в "хвостах" распределения
Имеют нестандартные сочетания характеристик
Такой подход позволяет **автоматически идентифицировать** случаи, требующие дополнительной проверки в статистическом обследовании.

Анализируя визуализацию дерева решений, можно определить логику разделения на **"Норма"** и **"Выброс"**:

"НОРМА" (Normal)

Это **типичные/стандартные наблюдения,** которые соответствуют основным паттернам данных

В контексте рабочего времени: сотрудники с **стандартной продолжительностью рабочей недели**

Например: большинство сотрудников с рабочим временем в типичном диапазоне (вероятно 35-50 часов)

"ВЫБРОС" (Abnormal)

Это **аномальные наблюдения,** которые значительно отклоняются от типичных паттернов

В контексте рабочего времени могут включать:

- **Слишком короткая** рабочая неделя (< определенного порога)
- **Слишком длинная** рабочая неделя (> определенного порога)
- **Необычные комбинации** основной и дополнительной работы

Примечание 3. Обнаружение выбросов: **One-Class SVM** и **Local Outlier Factor (LOF)**

1. One-Class SVM - обнаружено выбросов:

Модель обучается только на "нормальных" данных

Строит границу вокруг нормальных наблюдений

Все, что выходит за эту границу - считается выбросом

Аналогия: "Рисует забор вокруг типичных данных, кто за забором - выброс"

2. Local Outlier Factor (LOF) - обнаружено выбросов:

Сравнивает плотность точек вокруг каждого объекта

Если у точки соседи редкие/далекие - она выброс

Учитывает локальную структуру данных

Аналогия: "Ищет одинокие точки в малолюдных районах"

Различие: SVM ищет глобальные выбросы, LOF - локальные аномалии в контексте их окружения.