КОМПЛЕКС ЭКОНОМИЧЕСКОЙ ПОЛИТИКИ
И ИМУЩЕСТВЕННО-ЗЕМЕЛЬНЫХ ОТНОШЕНИЙ
ПРАВИТЕЛЬСТВА МОСКВЫ

Аналитический
центр Москвы

Зарубежный опыт применения методов Data mining в официальной статистике и направления их развития в деятельности Росстата

Зарова Е.В., д.э.н., проф.
ГБУ Аналитический центр города Москвы

24 ноября 2020

План выступления

1. **Понятие интеллектуального анализа данных (Data mining). Общие и отличительные характеристики понятий DM, статистика, Machine Learning**
2. **Зарубежный опыт применения методов DM в официальной статистике**
3. **Экспериментальные работы по интегрированию и анализу микроданных выборочных обследований с применением методов DM (ГБУ Аналитический центр города Москвы)**
4. **Предложения по возможным направлениям внедрения методов DM в практику Росстата**
5. **Выводы и рекомендации**

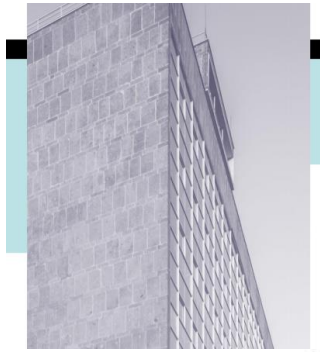
Определение DM

Интеллектуальный анализ данных (Data mining)

!!!!! НЕ предопределенных заранее сформированными гипотезами

Извлечение ценных полезных *«паттернов»* (внутренних структур, взаимосвязей переменных)

из массивов данных большого объема, неструктурированных или слабоструктурированных, а также составленных из данных разнотипных информационных источников



СТРАТЕГИЯ РАЗВИТИЯ

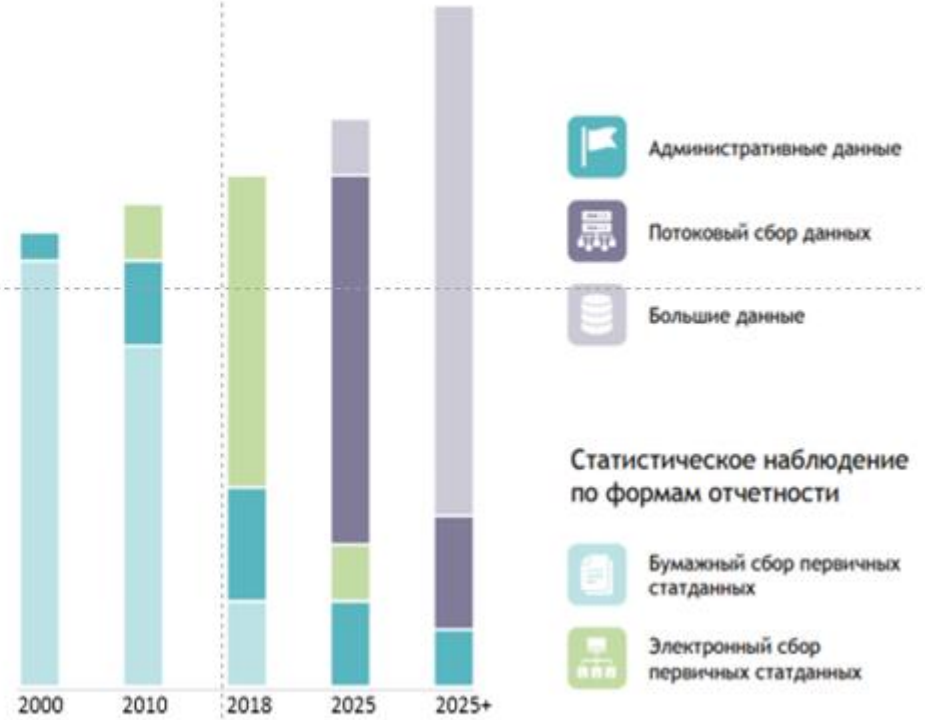
Федеральной службы государственной статистики до 2024 года

Изменение парадигмы статистического наблюдения

Переход от традиционной отчетности к новым источникам данных

Создание единой платформы и единой методологии

Вызовы российской статистики



Изменение парадигмы статистического наблюдения, предусмотренное Стратегией развития Федеральной службы государственной статистики, предопределяет необходимость разработки и внедрения методов интеллектуального анализа данных (Data Mining), и в том числе методов машинного обучения (Machine Learning), в процессах сбора, обработки и анализа данных

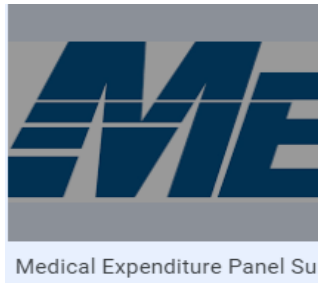
DM в практике национальных статистических офисов

Первая публикация о применении DM в официальной статистике – 1996*

Применение *деревьев классификации* для определения коэффициентов взвешивания единиц, содержащих неответы по классификационным признакам

2007 CENSUS OF AGRICULTURE

***Деревья классификации* для разделения респондентов на группы для корректировка весов с целью компенсации потери данных при неответах**

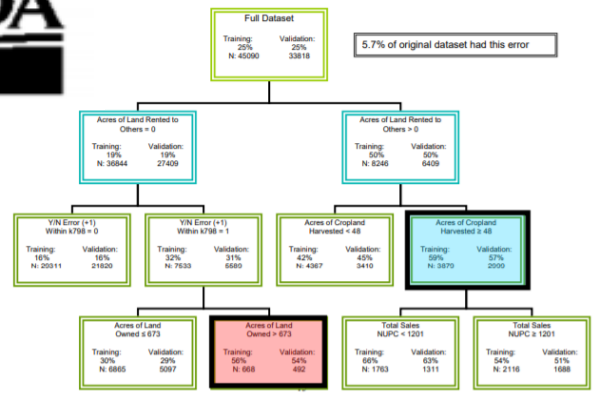


United States™ Census 2010

***Деревья регрессии + логистическая регрессии* для выработки классификации населения по рассовой принадлежности**

2009 – применение *деревьев решений* для классификации ошибок наблюдения площади посевных площадей

USDA



Примеры опыта Евростата и стран Евросоюза

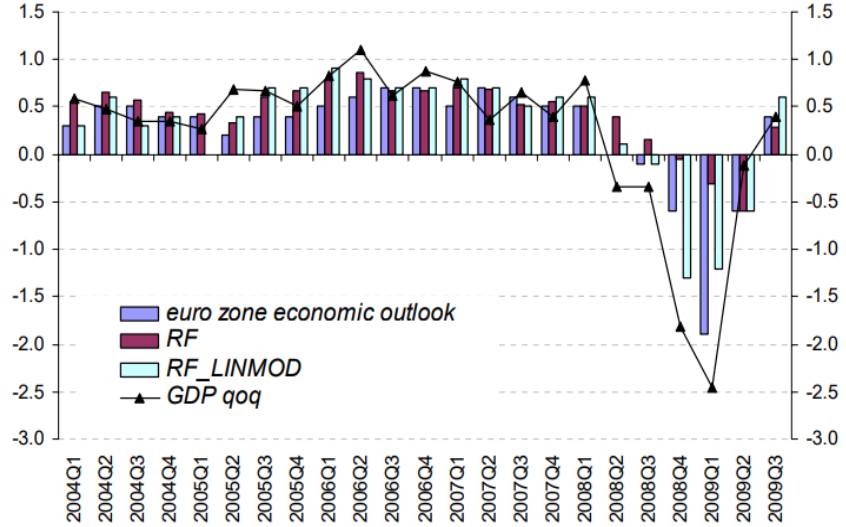
ISSN 1977-3331
EWP 2011/002
Euroindicators working papers

2012

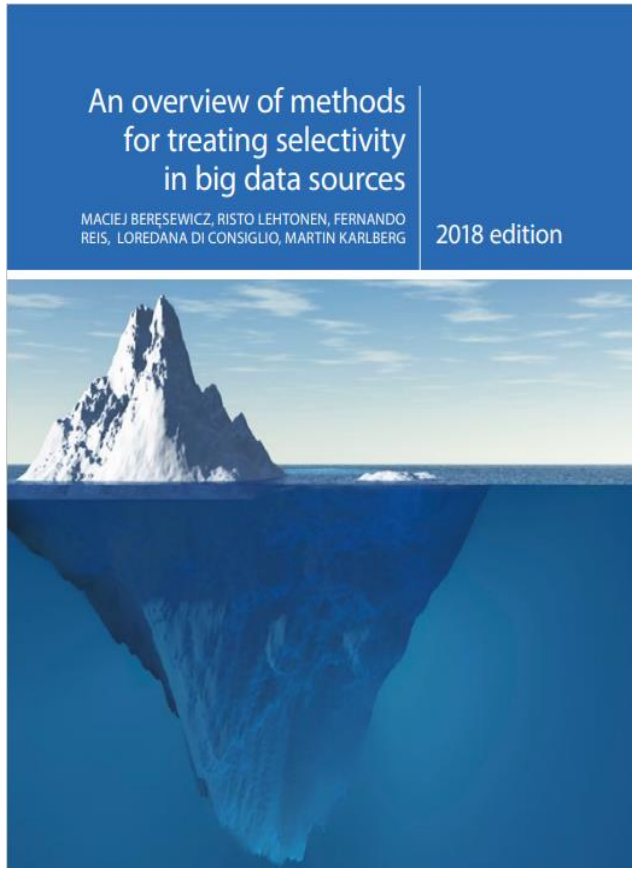
Euro area GDP forecasting using large survey datasets. A random forest approach

Источник данных:
Joint Harmonised European Union Business and Consumer Surveys (Совместные согласованные опросы бизнеса и потребителей Европейского союза) – квартальные данные по 5-ти секторам экономики

Оценка квартальных значений ВВП:
метод Монте Карло+деревья регрессии



Методические рекомендации Евростата



НОВЫЕ ИСТОЧНИКИ СТАТИСТИЧЕСКИХ ДАННЫХ

- Данные мобильной сети
- Социальные сети в Интернете - Twitter
- Данные о веб-активности –
  - Google Trends
  - Использование Википедии
- Административные данные

МЕТОДЫ ОБРАБОТКИ И АНАЛИЗА ДАННЫХ

МЕТОДЫ DM для предварительной обработки и структурирования данных

- К-ближайших соседей (kNN)
- Искусственные нейронные сети
- Деревья регрессии и классификации

МЕТОДЫ ИНТЕГРИРОВАНИЯ ДАННЫХ

- TEXT-MINING
- STATISTICAL MATCHING

Примеры методических рекомендаций по применению методов DM к большим и сложносоставным данным

Analysis of the most recent modelling techniques for big data with particular attention to Bayesian ones

GEORGE KAPETANIOS, MASSIMILIANO MARCELLINO, KATERINA PETROVA

2018 edition

STATISTICAL WORKING PAPERS | eurosta

Big Data and Macroeconomic Nowcasting: from data access to modelling

EMANUELE BALDACCI, DARIO BUONO, GEORGE KAPETANIOS, STEPHAN KRISCHE, MASSIMILIANO MARCELLINO, GIAN LUIGI MAZZI, FOTIS PAPAILIAS

2016 edition

STATISTICAL BOOKS | eurostat

ISSN 1977-0375

eurostat
Methodologies and Working papers

MEDSTAT II: Asymmetry in foreign trade statistics in Mediterranean partner countries

09 edition

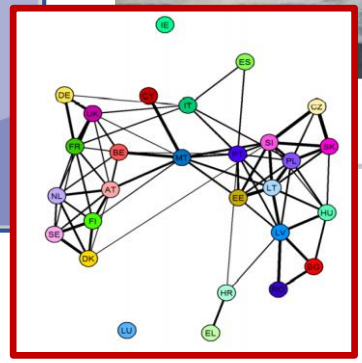
eurostat
EUROPEAN COMMISSION

Labour market attractiveness in the EU

JOAO SOLLARI LOPES, SONIA QUARESMA AND MARCO MOURA

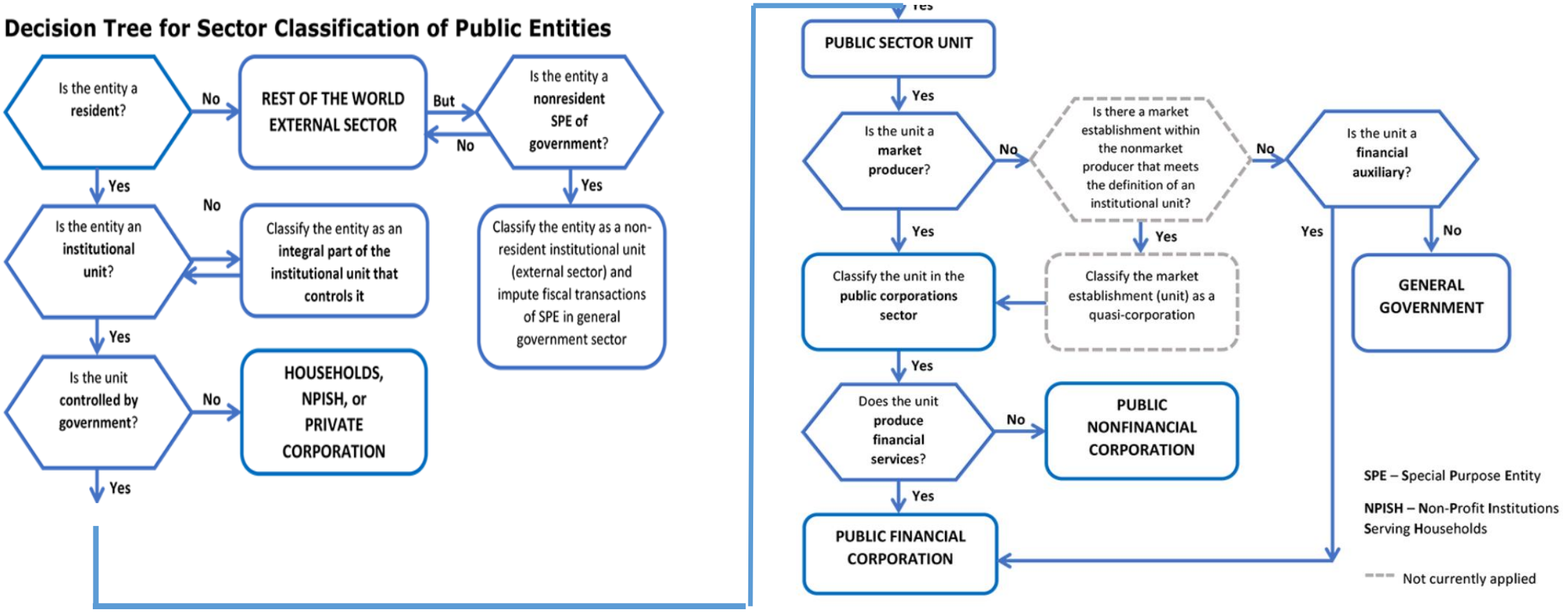
2018 edition

STATISTICAL WORKING PAPERS | eurostat



Statistics Canada

Дерево решений для классификации общественных организаций по секторам экономики



Национальный статистический офис Великобритании – Центр науки о данных

МЕТОДЫ DM

Model (feature set)

Linear regression

Random forest

Ada boost

XGBoost

Neural network

Разработка и внедрение новых методов и источников данных для целей официальной статистики по 5-ти проектам:

- **Эволюционирующая экономика**
- **Городская и сельская экономика**
- **Социальная сфера**
- **Устойчивость**
- **Великобритания в глобальном контексте**

Office for National Statistics

Data Science Campus

1.

ПРОЕКТЫ

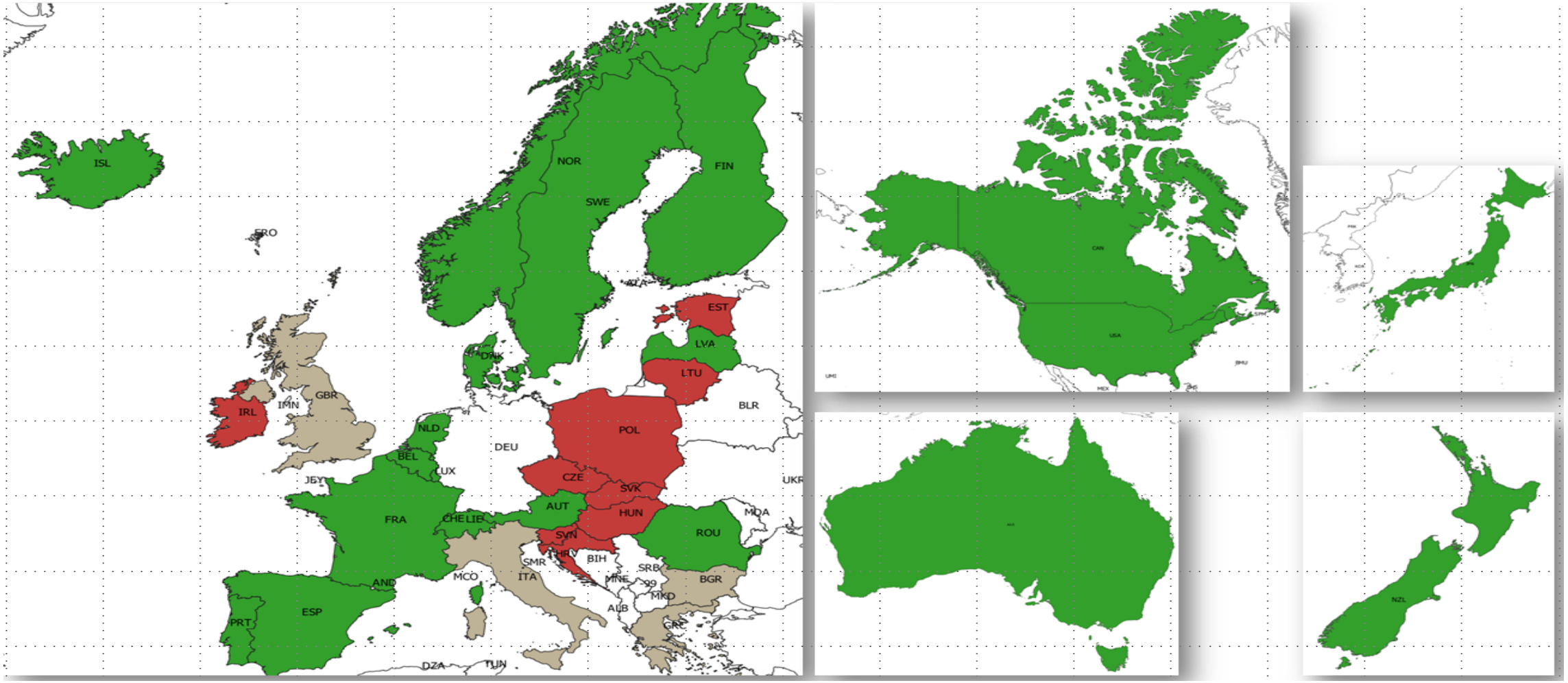
- **Оценка калорийности питания**
- **Новые и более быстрые индикаторы, систематические ошибки и аномалии в данных Управления по налогам и таможенным сборам по налогу на добавленную стоимость (НДС)**
- **Автоматизированная генерация и систематизация данных отчетов**
- **Улучшение статистики природного капитала, включая оценки естественных земель или зеленых насаждений**
- **Патентная статистика**
- **Оценка жилищных условий и энергоэффективности**
- **Классификация финансовых услуг**
- **Оценка зеленых насаждений в городской среде**

“Кампус реализует проекты для Управления национальной статистики (ONS) Великобритании С ПРИМЕНЕНИЕМ СОВРЕМЕННЫХ МЕТОДОВ MACHINE LEARNING и альтернативных источников данных”

https://datasciencecampus.ons.gov.uk/projects/

Применение методов Data mining национальными статистическими офисами (НСО):

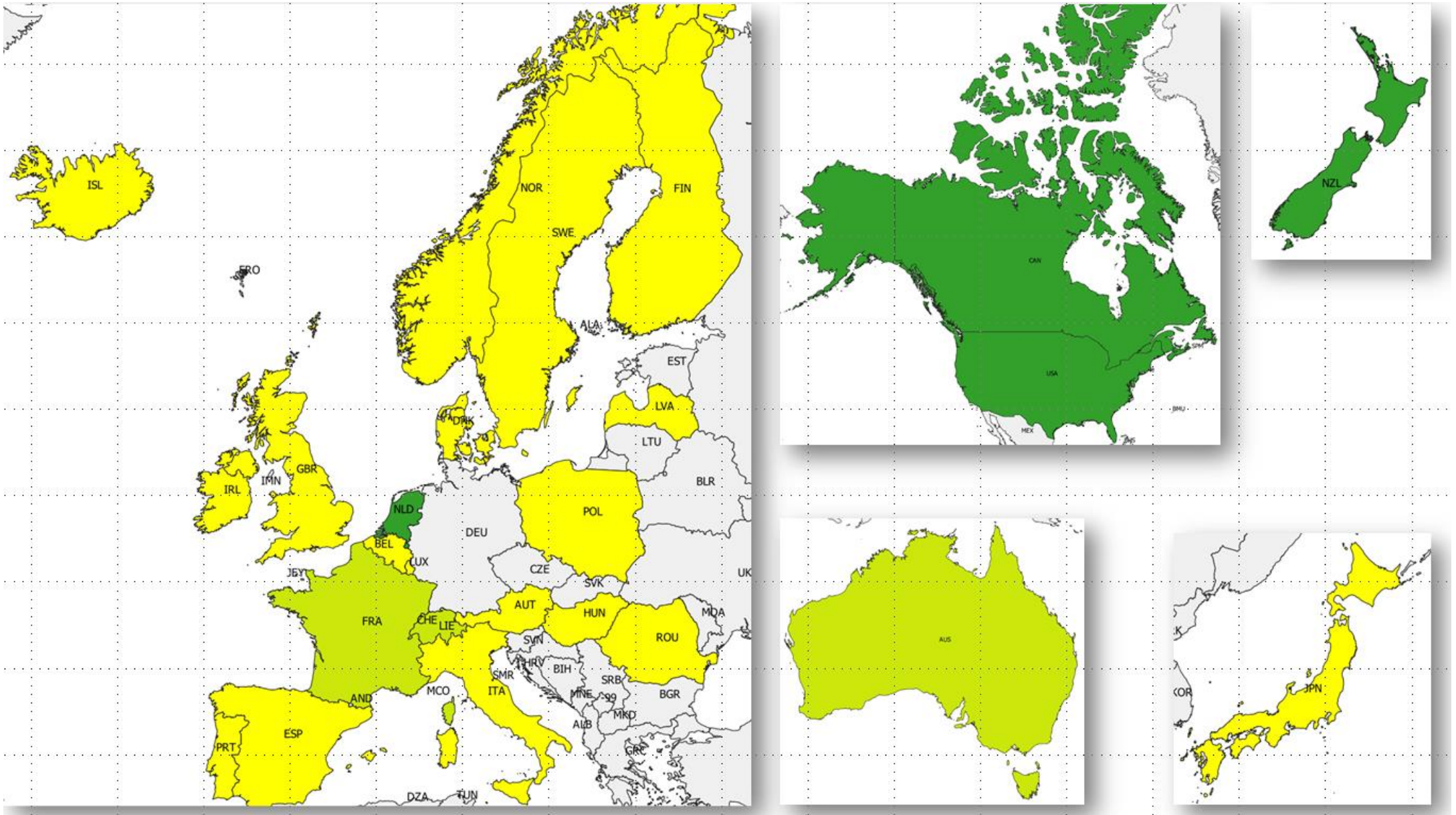
\+ – Нет ответа



Federal Statistical Office of Germany, 2018

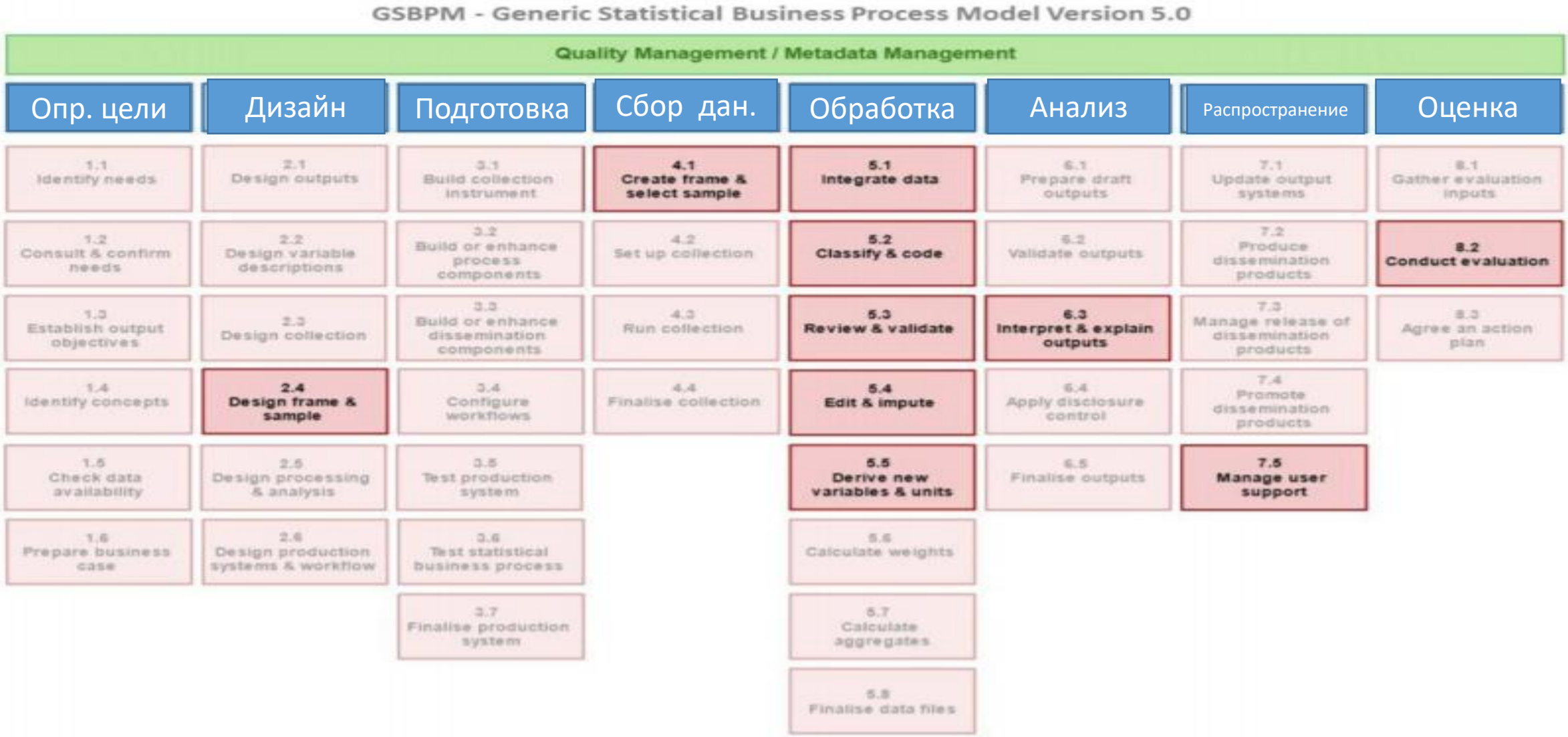
Число работ в НСО с применением методов DM:

1-6	6-10	Св. 10

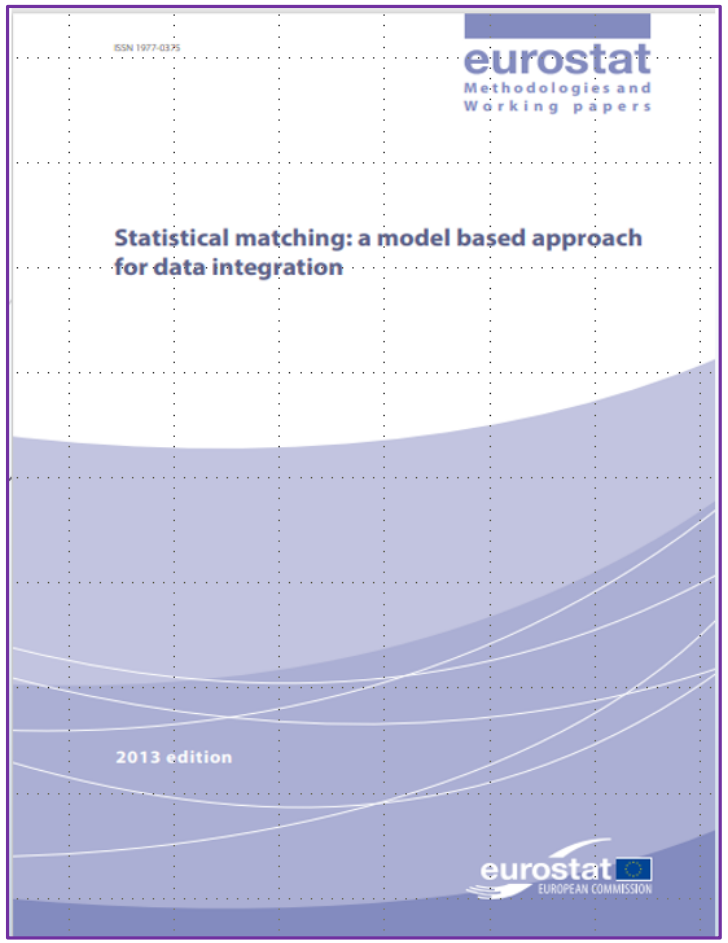


UNECE Методы Machine Learning по этапам GSBPM

(Типовая модель статистической информации)



Методы DM в интегрировании *микроданных* выборочных обследований (Пр. 1)



EU – SILC
Обследование доходов и условий жизни

SM

EQLS
Европейское исследование качества жизни

SM

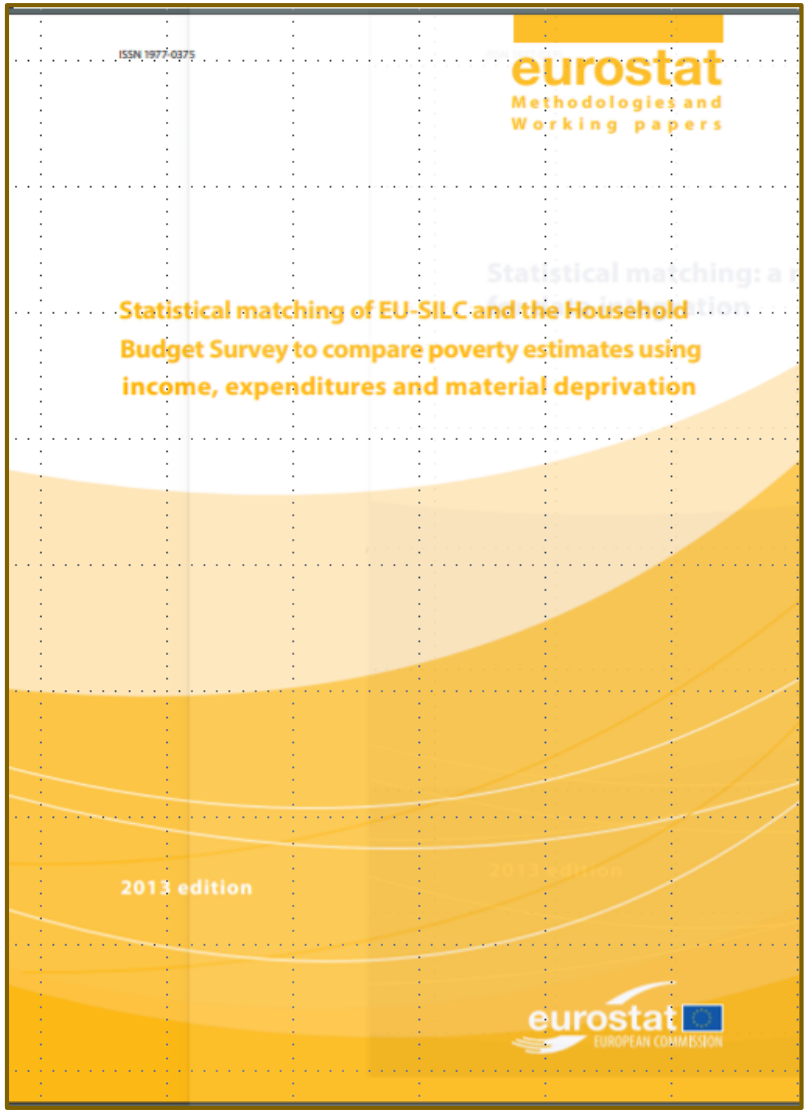
LFS
Обследование рабочей силы

NND.hotdeck( ),

***SM – Statistical matching* –**
методы интегрирования массивов микроданных путем вменения значений признаков единиц массива-донора единицам массива-реципиента.

Применяется семейство методов Hot Deck, основанных на методах кластерного анализа

Методы DM в интегрировании *микроданных* выборочных обследований (Пр. 2)



EU – SILC
Обследование доходов и условий жизни

SM

HBS
Обследование бюджетов домашних хозяйств

Цель – получение *синтетического набора микроданных* для расчета показателей доходов, материальной депривации, расходов населения
Методология STATMATCH

Опыт интегрирования микроданных ОДН и ОРС по городу Москве (2019)

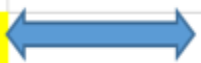
Обследование доходов населения и УСП (ОДН) – Массив – донор (8 182)	
Общие переменные	**Специфические переменные (в том числе целевые переменные)**
X	**Y**

Обследование рабочей силы (ОРС) – Массив – реципиент (43 665)	
Общие переменные	**Специфические переменные**
X	**Z**

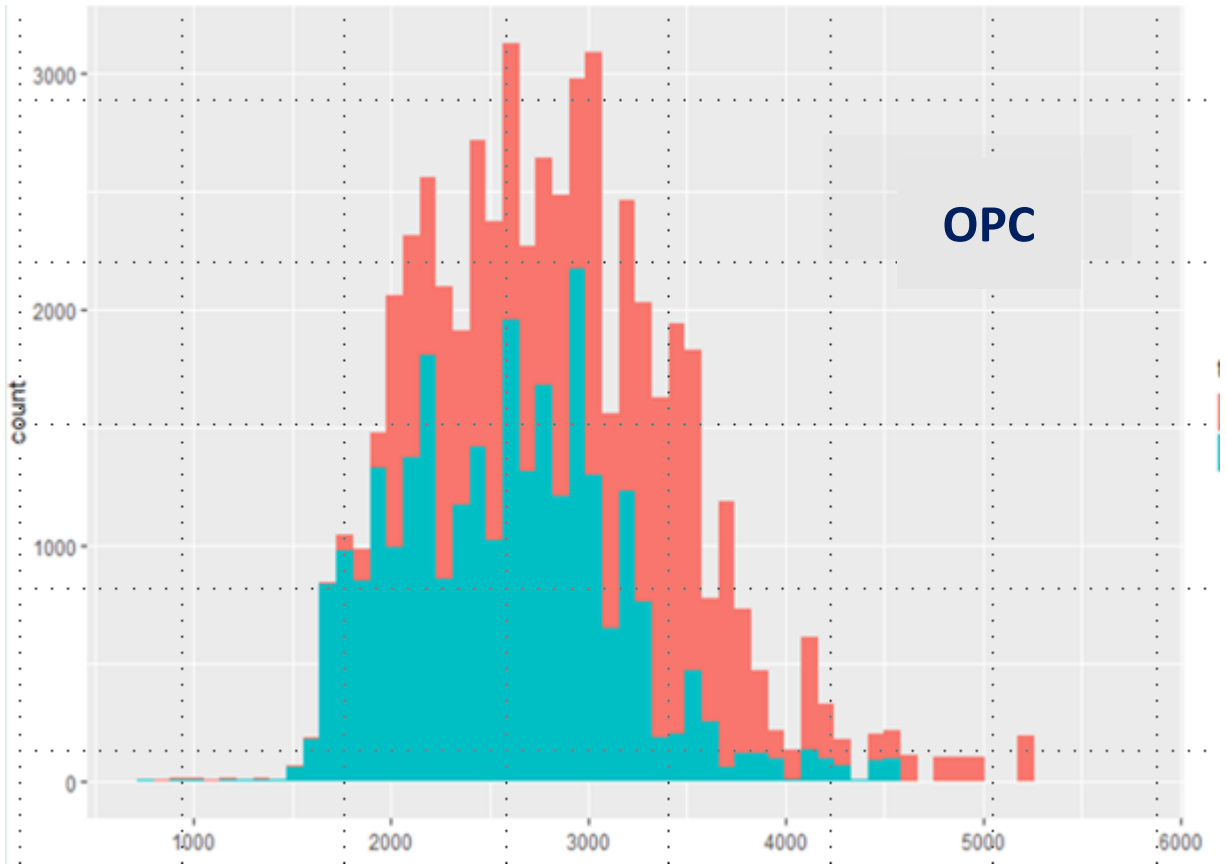
Синтетический массив микроданных (ОДН – LFS)		
X	**Y**	**Z**

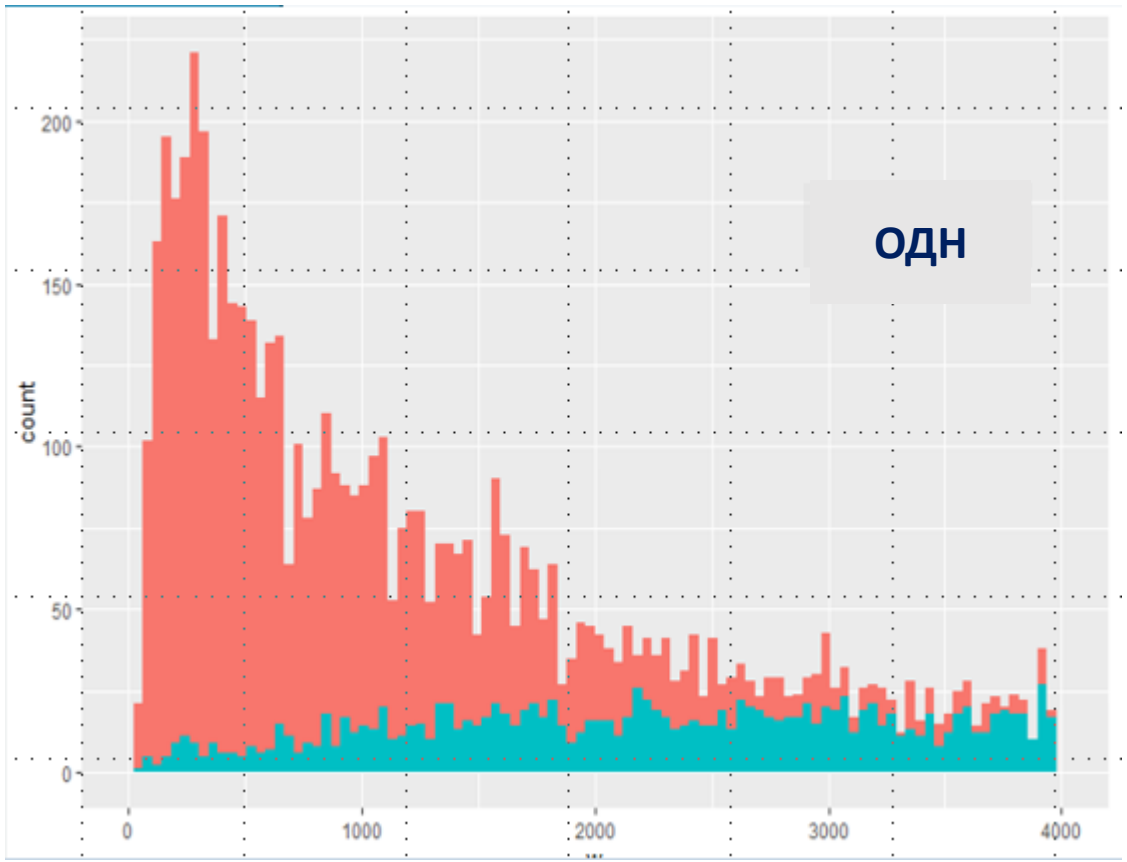
Гармонизация переменных ОРС и ОДН с целью обеспечения сопоставимости кодов

	A	B	C	D	G	H	I
1	Исходные коды	Сопоставимые коды	Гармонизируемые переменные ОДН		Гармонизируемые переменные ОРС	Исходные коды	Сопоставимые коды
2							
3							
4			ОДН		ОРС		
5			H01_01	⟷	nas_pol		
6	1	1	Мужской		Мужской	1	1
7	2	2	Женский		Женский	2	2
8			H01_04	⟷	NASBRACH		
9					Ваше состояние в браке?.		
10	1	1	состоит в зарегистрированном браке		Состою в зарегистрированном браке	1	1
11	2	2	состоит в незарегистрированном браке		Состою в незарегистрированном браке	2	2
12	3	3	вдовец/вдова		Вдовец, вдова	3	3
13	4	4	разведен (а)		Разведен(а)	4	4
14	5	5	разошелся (лась)		Разошелся(лась)	5	5
15	6	6	никогда не состоял(а) в браке		Никогда не состоял(а) в браке	6	6
16							
17							
18			I01_10	⟷	nasobraz		
19	1	1	кадры высшей квалификации (послевузовское) - аспирантура, доктора		Какое образование Вы имеете?		
20	2	1	высшее - специалитет, магистратура		Высшее профессиональное	1	1
21	3	1	высшее - бакалавриат		Неполное высшее профессиональное (неполное высшее)	2	2
22	4	2	неполное высшее (незаконченное высшее) – оконченные 3 курса и бол		Среднее профессиональное	3	3
23	5	3	среднее профессиональное по программе подготовки специалистов с		Начальное профессиональное	4	3
24	6	3	среднее профессиональное по программе подготовки квалифицирова		Среднее (полное) общее	5	5
25	7	5	среднее общее (среднее полное общее)		Основное общее	6	6

26	8	6	основное общее (неполное среднее)		Начальное общее (начальное)	7	7
27	9	7	не имеете основного общего		2017 послевузовское	8	1
28					2017 высшее-бакалавриат	9	1
29			I03_07				
30	1	1	наемный работник за заработную плату или вознаграждение деньгам		V_OSNRB		
31	2	1	ученик на производстве, стажер, практикант		на собственном предприятии или в собственном деле для по	2	2
32	3	2	владелец (совладелец) собственного предприятия (собственного дел		в качестве наемного работника за заработную плату деньгам	1	1
33	4	3	член производственного кооператива (артели, партнерства с другим		в качестве члена производст-венного кооператива (артели),	3	3
34	5	5	помогающий на предприятии или в собственном деле, принадлежащих		в качестве помогающего без оплаты на предприятии, принад	5	5
35	6	2	индивидуальный предприниматель по договору гражданско-правовог				
36							
37							
38			I03_04		V_OSNZAN		
39							
40							
41	1	1	на предприятии, в организации (или обособленном подразделении ор		на предприятии, в учреждении, организации,	1	1
42	2	4	на предприятии индивидуального предпринимателя или у лиц, осущес		в фермерском хозяйстве,	2	2
43	3	2	в фермерском хозяйстве		предпринимательская деятельность без образования юридич	3	3
44	4	3	в сфере предпринимательской деятельности без образования юридич		на индивидуальной основе,	5	5
45	5	4	по найму у физических лиц, индивидуальных предпринимателей		по найму у физических лиц, индивидуальных предпринимате	4	4
46	6	5	на индивидуальной основе (на основе самостоятельной занятости)				
47	7	5	в собственном домашнем хозяйстве по производству продукции сель				

Гармонизация распределений весовых коэффициентов



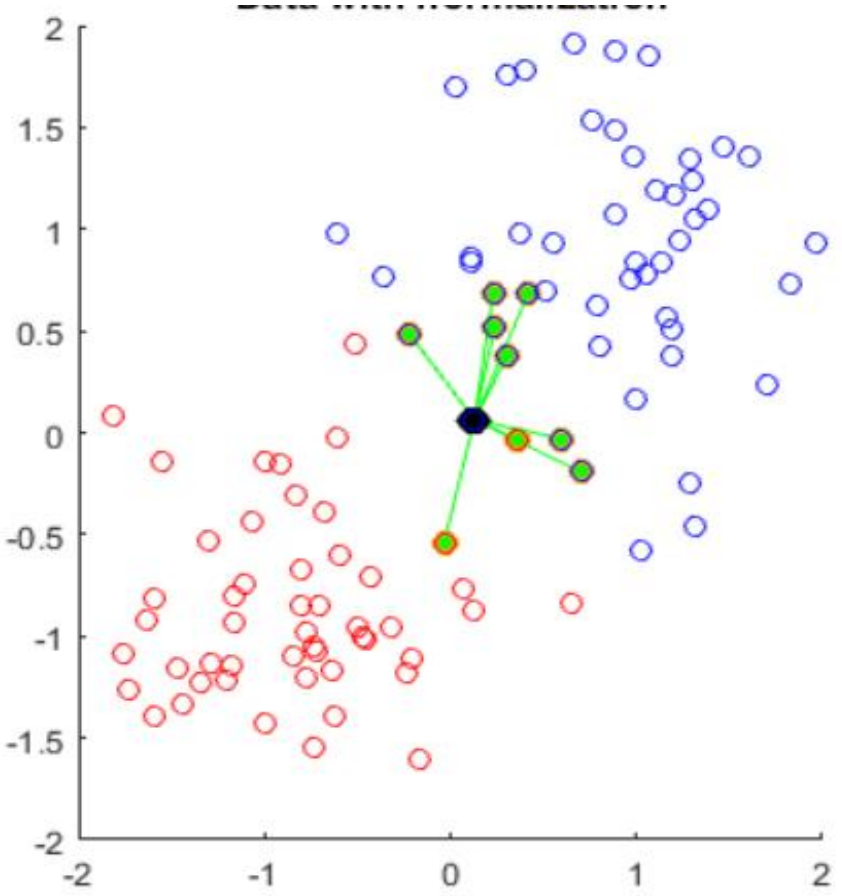


- исходные веса

- согласованные веса

Методы интегрирования массивов микроданных ОДН и ОРС (R, пакет StatMatch)

NND.hotdeck( )



поиск **«ближайшего соседа»** для каждой единицы массива-реципиента в массиве-доноре

RANDwNND.hotdeck( )

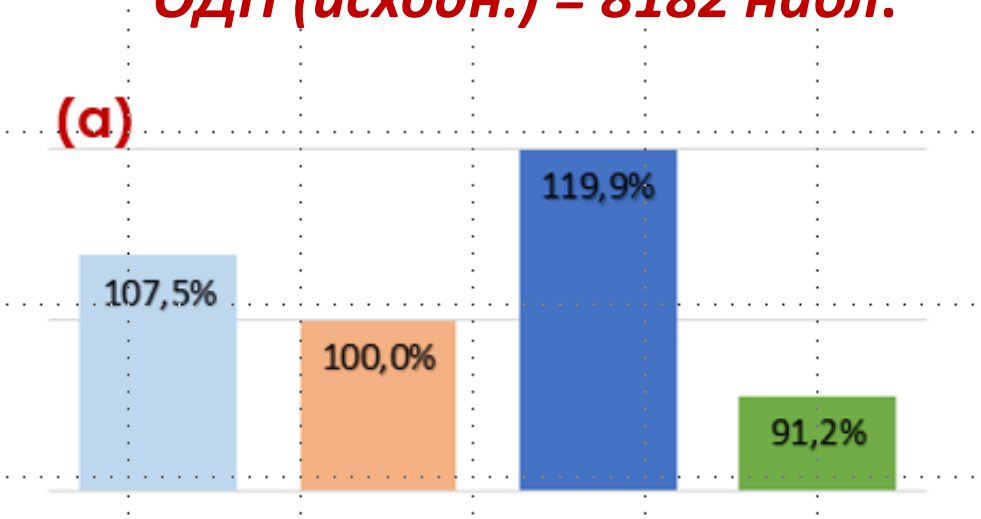
для каждой единицы массива-реципиента определяется **подмножество ближайших «доноров»,** а затем случайным образом выбирается донорская единица.

rankNND.hotdeck()

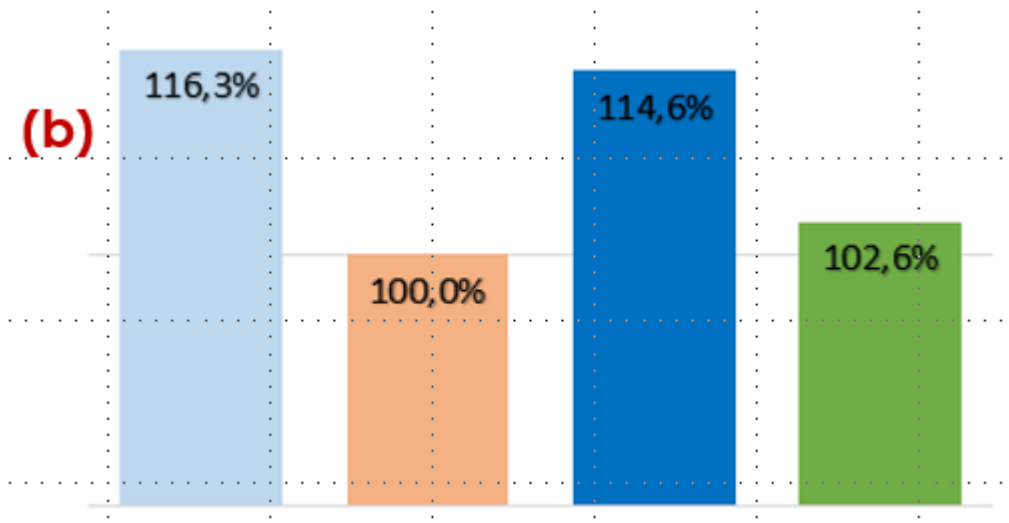
ближайший **«донор»** выбирается с учетом расстояния до ближайшей точки **эмпирической кумулятивной функции распределения**

Сравнительные результаты оценки среднедушевого денежного дохода, Москва, 2019

ОДН (исходн.) = 8182 набл.



ОДН + ОРС (интегрир.)= 43 665 набл.



Средний денежный доход на 1 члена ДХ (100 %)

Средний совокупный доход на 1 члена ДХ, %

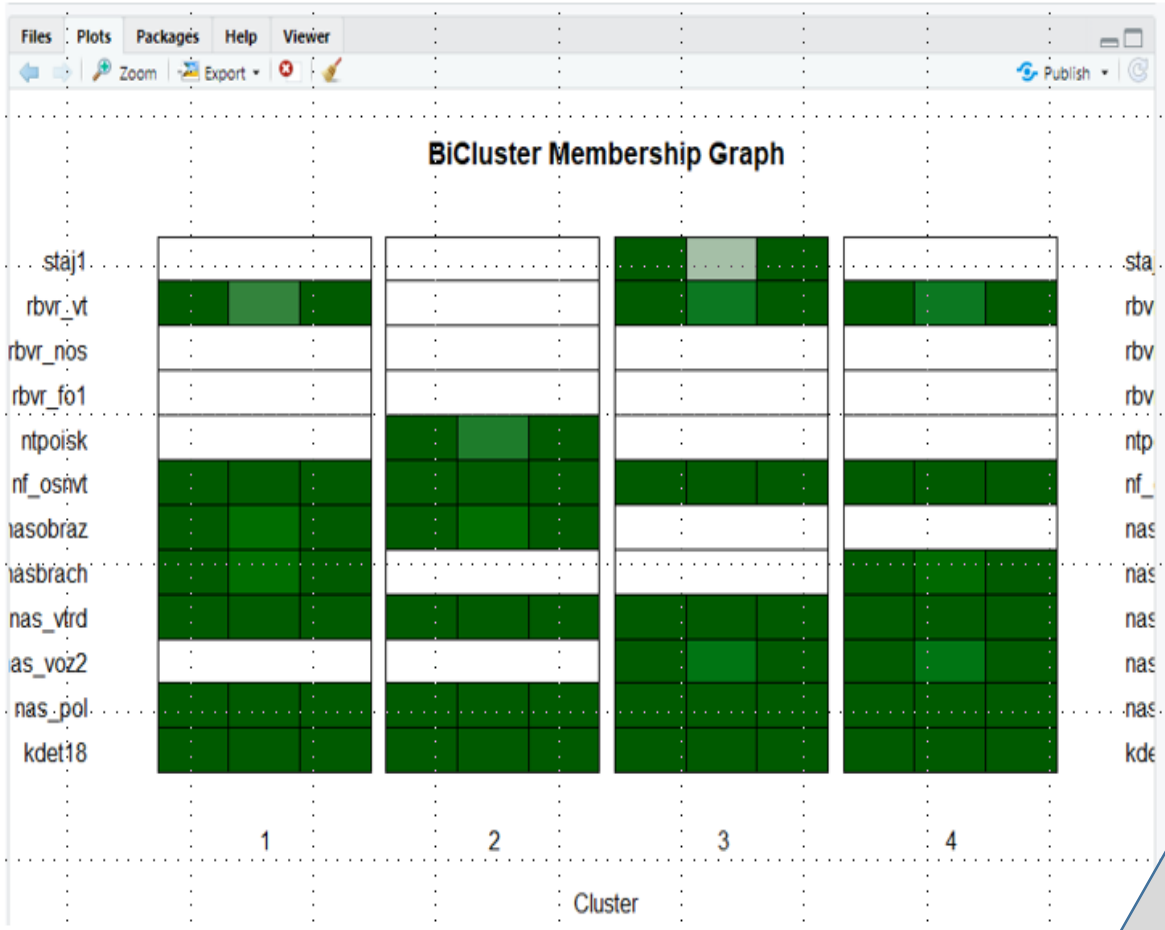
Средний располагаемый доход на 1 члена ДХ, %

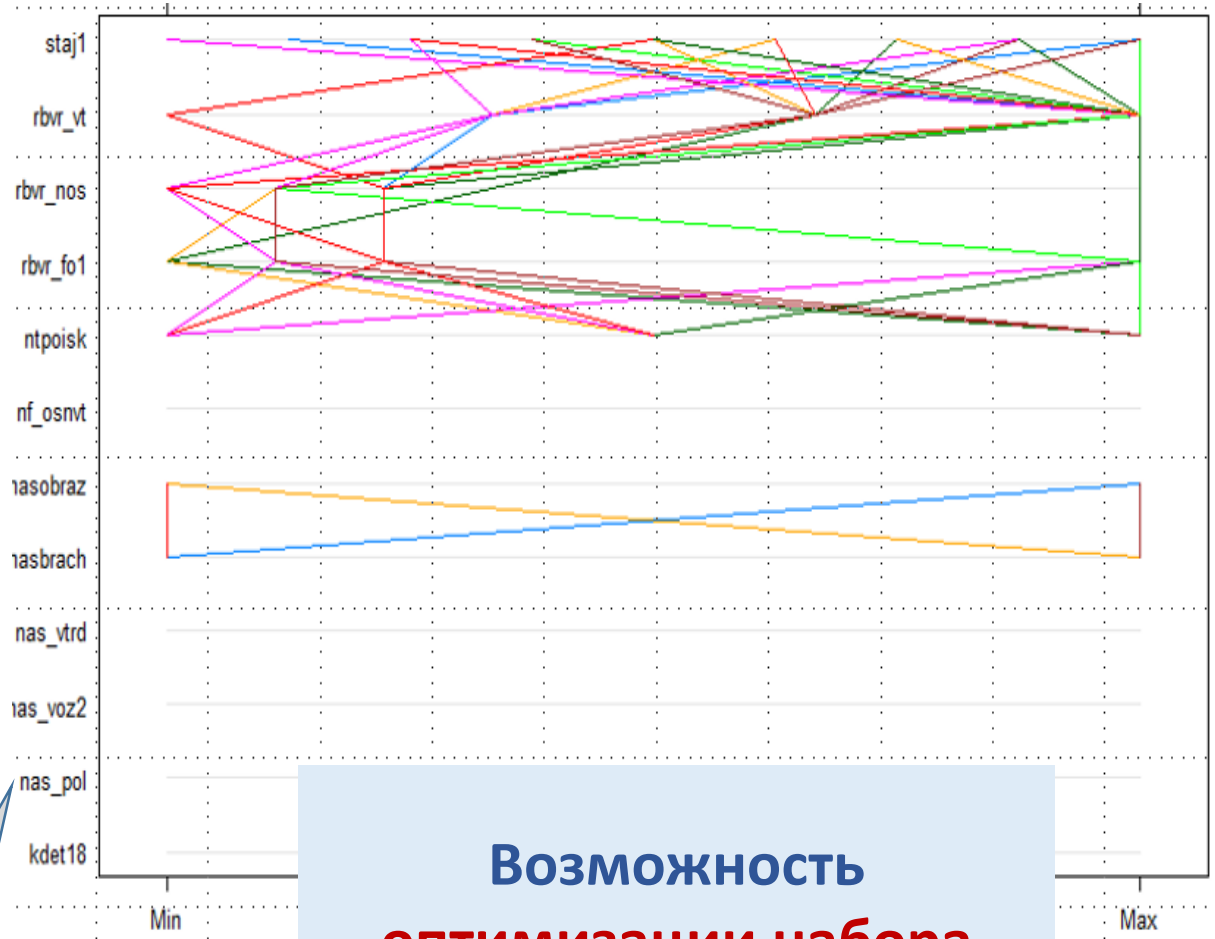
Средний доход от трудовой деятельности на 1 занятого, %

Экспериментальные оценки уровня бедности *по интегрированным микроданным ОДН и ОРС*

Уровень бедности, %	2017	2018	2019 предварительные данные
Российская Федерация			
Данные Росстата	12,9	12,6	12,3
Оценка по интегрированным микроданным ОДН и ОРС'	8,4	7,3	7,2
город Москва			
Данные Росстата	7,5	6,8	6,6
Оценка по интегрированным микроданным ОДН и ОРС'	6,9	6,2	6,1

Экспериментальное применение методов бикластерного анализа на микроданных ОРС.

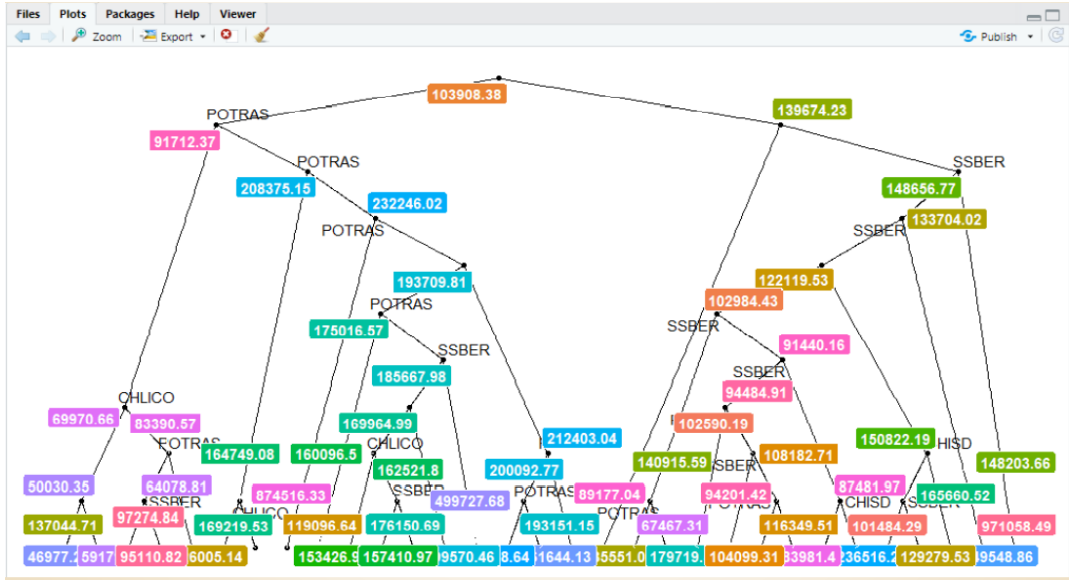


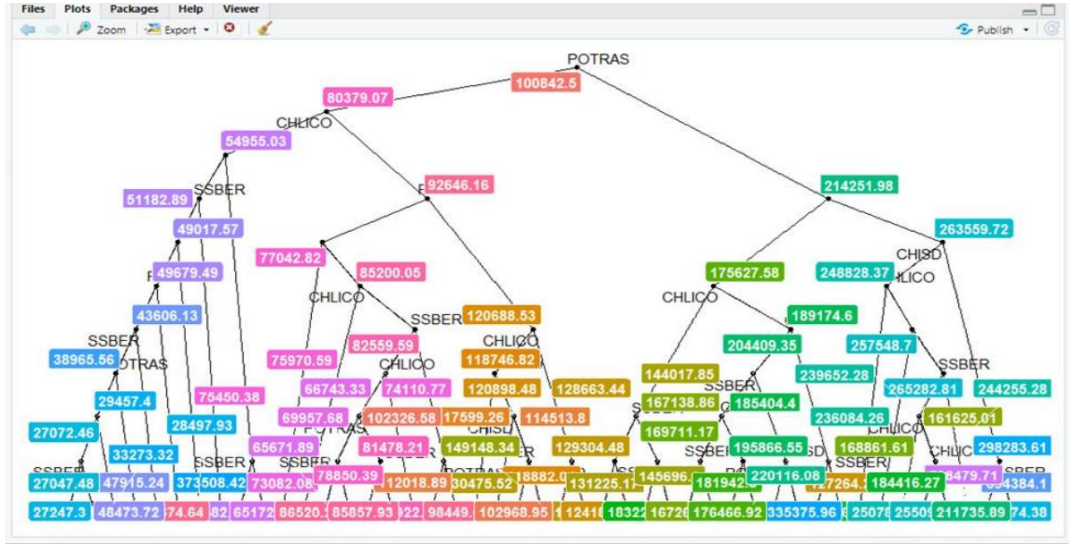


Переменные ОРС

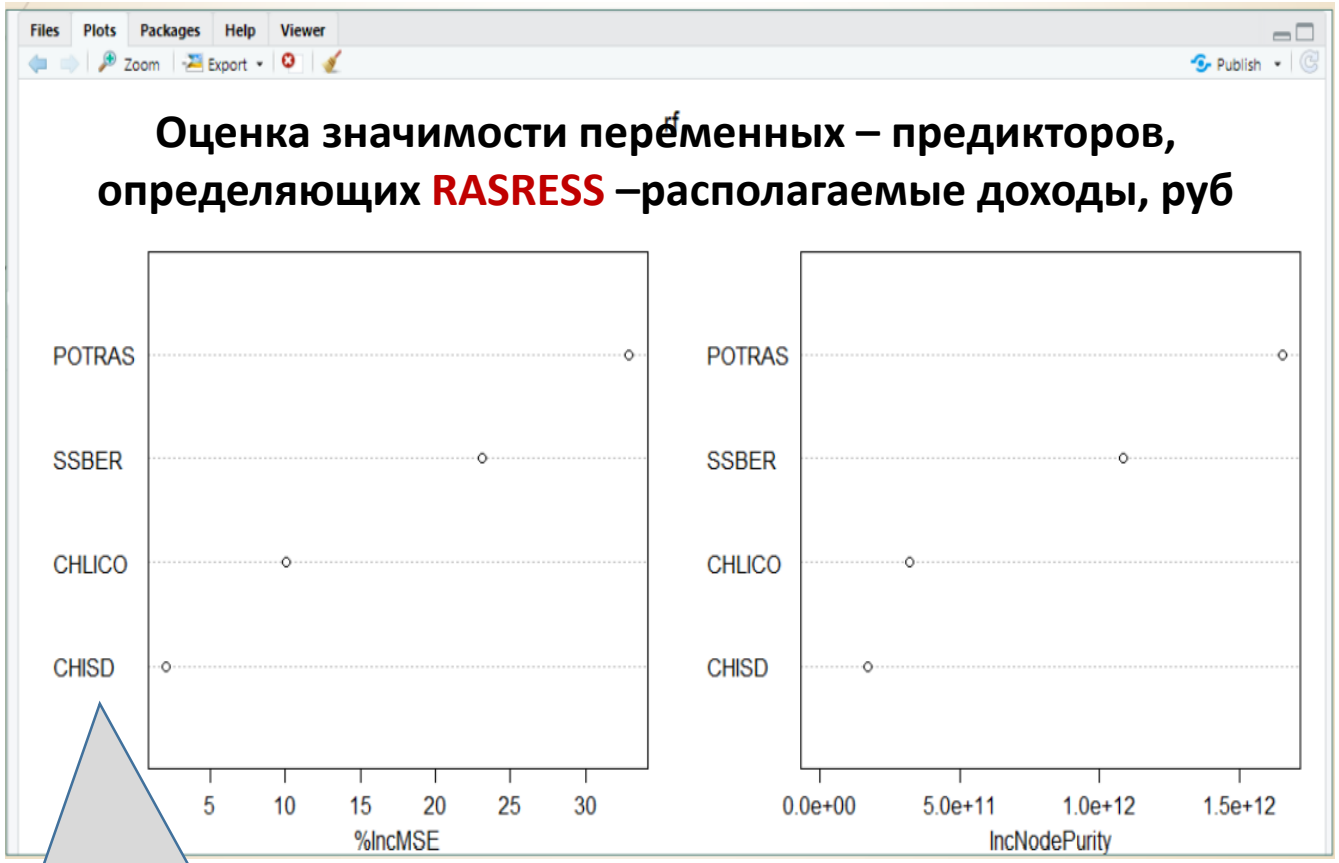
Возможность оптимизации набора наблюдаемых признаков при проведении ОРС

Методы «Случайный лес» -ансамбль деревьев регрессии - применительно к микроданным ОБДХ позволяют выявлять скрытые иерархические связи показателей





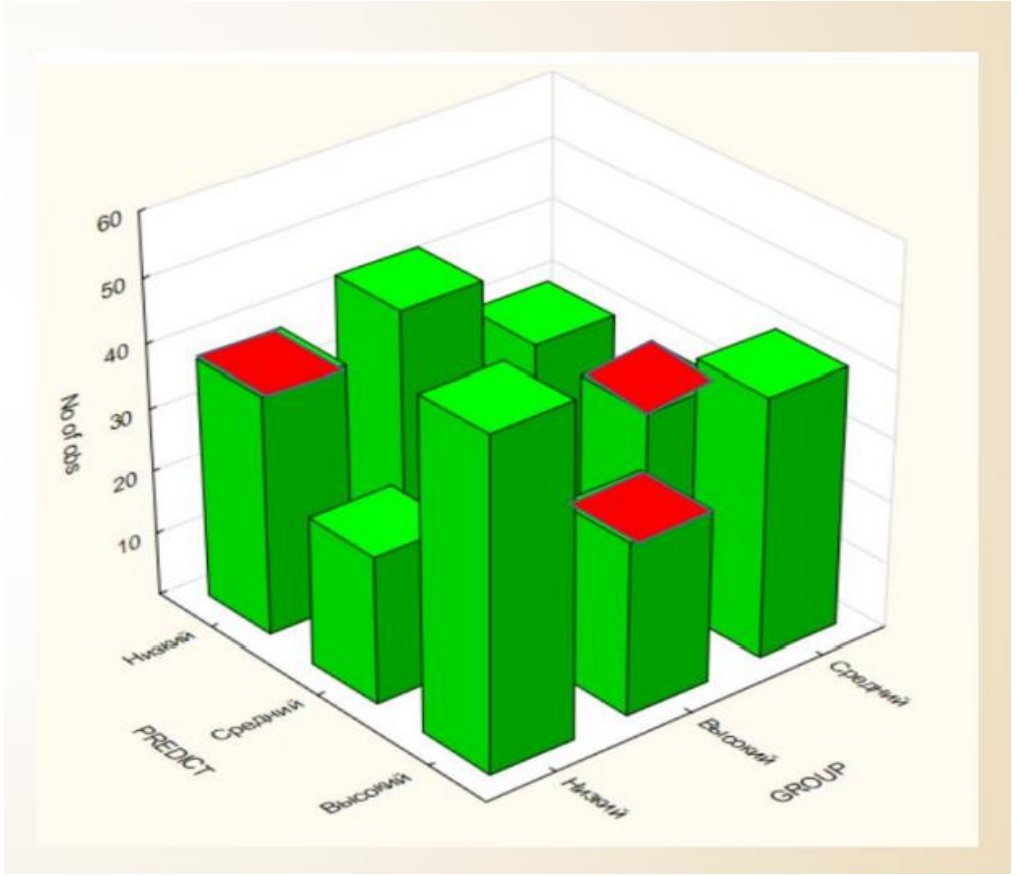
Оценка значимости переменных – предикторов, определяющих RASRESS –располагаемые доходы, руб



Обозначение переменных ОБДХ

Построение ансамблей *деревьев классификации* по микроданным ОБДХ позволяет уточнить исходную гипотетическую структуру данных

Число деревьев по 3-м изначально выделенным классам респондентов (по переменной RASRESS)



Сравнение исходной классификации и результатов применения метода «Случайный лес»

Апробация предложений

Предложения по возможным направлениям внедрения методов DM в практику Росстата:

- Интегрирование микроданных ОРС и ОДН с целью уточнения оценок доходов занятого населения на основе применения методов кластеризации HOT DECK
- Применение методов «Случайный лес» для выявления скрытых структур и взаимосвязей переменных в массивах микроданных ОБДХ
- Оптимизация наблюдаемых переменных и выявление скрытых структур с применением методов бикластерного анализа к массивам микроданных ОРС

Выводы и рекомендации:

1. **Применение методов интеллектуального анализа данных (DM) в официальной статистике обусловлено общемировым трендом вовлечения в производственный процесс официальной статистики новых источников данных, включающих большие, неструктурированные данные, потоковые данные, сложносоставные данные из источников различных типов**
2. **Методы DM позволяют повышать эффективность и качество статистического наблюдения, получать новые количественные и качественные характеристики наблюдаемых официальной статистикой явлений, не «регламентируемые» заранее предопределенными гипотезами (о составе элементов структуры, взаимосвязи показателей и т.п.)**
3. **Согласно опубликованным источникам в практике официальной статистики многих стран методы DM получают ускоренное развитие и применение, прежде всего, для целей предварительной обработки собираемых данных и их анализа, выработки альтернативных и экспериментальных оценок**
4. **Работы, выполняемые в ГБУ АЦ города Москвы, доказывают заинтересованность пользователей и возможность внедрения методов DM в практику Росстата для целей получения новых качественных статистических данных**
5. **Необходимо развивать сотрудничество между органами государственной статистики, научными и экспертными организациями по развитию методов DM и в целом науки о данных**

Спасибо за внимание!