МЕТОДЫ И ТЕХНОЛОГИИ АНАЛИЗА ДАННЫХ РОССТАТА

Смольникова И.А.* **(Россия, г. Москва)**

Аннотация. *В России появилось много больших таблиц открытых данных [1]–[2]. Стали доступны и аналитические программы. В предшествующей статье [3] дан обзор (бесплатно) доступных информационных бизнес – инструментов (ПО), проведена классификация, для аналогов – сравнение. Их возможности охватывают не только подготовку (очистку и восстановление), разведку (визуализацию и описательную статистику) данных, но сравнительные и интеллектуальные методы выявления скрытых зависимостей для анализа (вклада и прогноза) и уменьшения размерности (для компактности хранения). Ниже на социально – экономических данных Росстата показаны разные способы и технологии анализа. Методику можно использовать при обучении информационно-аналитическим технологиям (ИАТ) управления.*

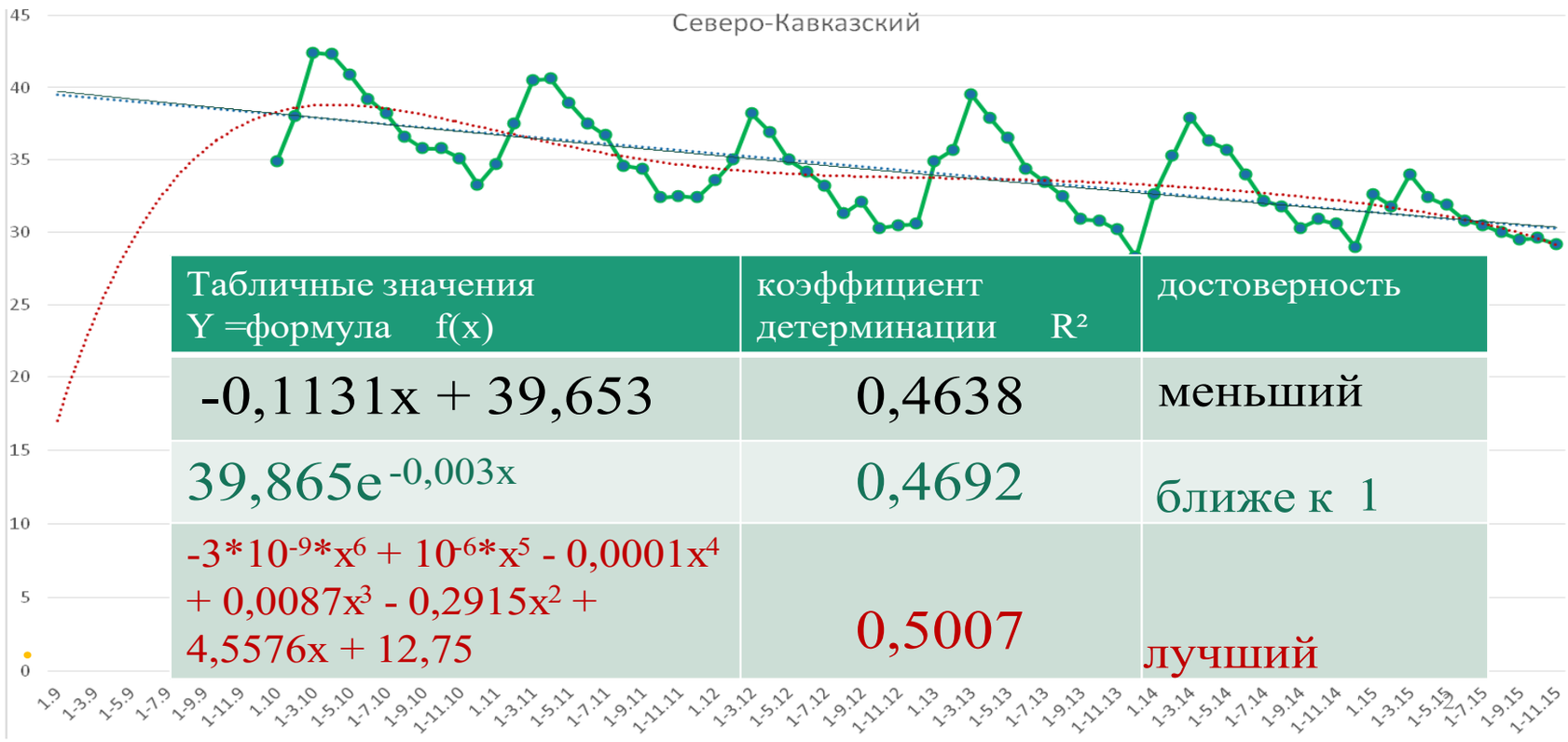
Сначала в развитых странах, потом и в России, в Интернете появилось много больших таблиц социально – экономических данных [1]. On-line можно искать и фильтровать нужные данные, даже графически визуализировать их малое количество, например, на информационной витрине ГАСУ [2]. Но это лишь один из способов 2-го этапа анализа **(разведки)**, который до сих пор преобладает.

В [3] даны все **5 этапов анализа** с их трудоемкостью. Продемонстрируем в MS Excel (модуль анализа), Statistica, Deductor (имеющихся на ФГУ) ход, модели и результаты аналитических исследований для 52-х показателей Росстата, собранных 85-ю субъектами РФ с 2009 года [1] по **этапам: 1**– I, **2**–II – IV, **3** –V или VI, **4** –VII, **5** – VIII.

I. Подготовка данных (проблемы и решения)

1. Неполные данные (особенно, в сфере культуры), остальные показатели:
   1) Астраханская и Тюменская области без округов – с декабря 2013 года,
   2) Северо-Кавказский округ – с 2010 года,
   3) Крым – с 2014 года (поквартально), с 2015 года – ежемесячно.
2. Удаление «полупустого» региона (Крымского округа).
3. Заполнение пропусков по трендам (без учета сезонности в Excel):

* *Смольникова Ирина Алексеевна*, кандидат физико-математических наук, доцент кафедры математических методов и информационных технологий в управлении, факультет государственного управления Московского государственного университета имени М.В.Ломоносова.



Табличные значения Y =формула f(x)	коэффициент детерминации R^2	достоверность
-0,1131x + 39,653	0,4638	меньший
$39{,}865e^{-0{,}003x}$	0,4692	ближе к 1
$-3*10^{-9}*x^6 + 10^{-6}*x^5 - 0{,}0001x^4 + 0{,}0087x^3 - 0{,}2915x^2 + 4{,}5576x + 12{,}75$	0,5007	лучший

Рис. 1. Приближение табличных данных 3-мя формулами

Видим – реалистичнее: внутренние пропуски – полиномиальный, назад – по среднему, а прогноз вперед – линейный (с учетом сезонных колебаний – в Statistica или Deductor).

4. Скрытые опечатки: вместо запятой – точка → не число, а текст; пустота → может игнорироваться в Excel.

5. Для гарантированного импорта первичных данных в ПО Statistica и Deductor – таблицы чисел должны быть в виде текстового файла с разделителем по столбцам.

II. Анализ временных рядов

В Excel построены графики с линейными трендами для округов (как и рисунок 1, рисунок 2 – тоже на примере накопительной в течение года доли **убыточных** организаций со с. 185–187 отчета Росстата за 84 месяца [1]):

Ри. 2. Графики 8-ми полных округов с уравнениями линейного тренда

Видим: пики – в унисон (накопление), почти все попарные корреляции строк > 0,5, что подтверждает **единую инвестиционную политику в России**.

Для схожих по динамике регионов (парная корреляция близка к 1) найдены формулы **регрессии**, например: Уральский округ = 1,117 * Приволжский округ + 0,84 ± 3,6. Достоверное приближение послужило основой для сэмплинга (п. VI ниже). От графической визуализации при переходе от 9 округов к 85-ти субъектам РФ вернулись к таблице с использованием формул для автоматизации выводов (см. модель III).

III. Модель динамики показателя представлена в таблице 5 в [3] (Excel-книга 11)

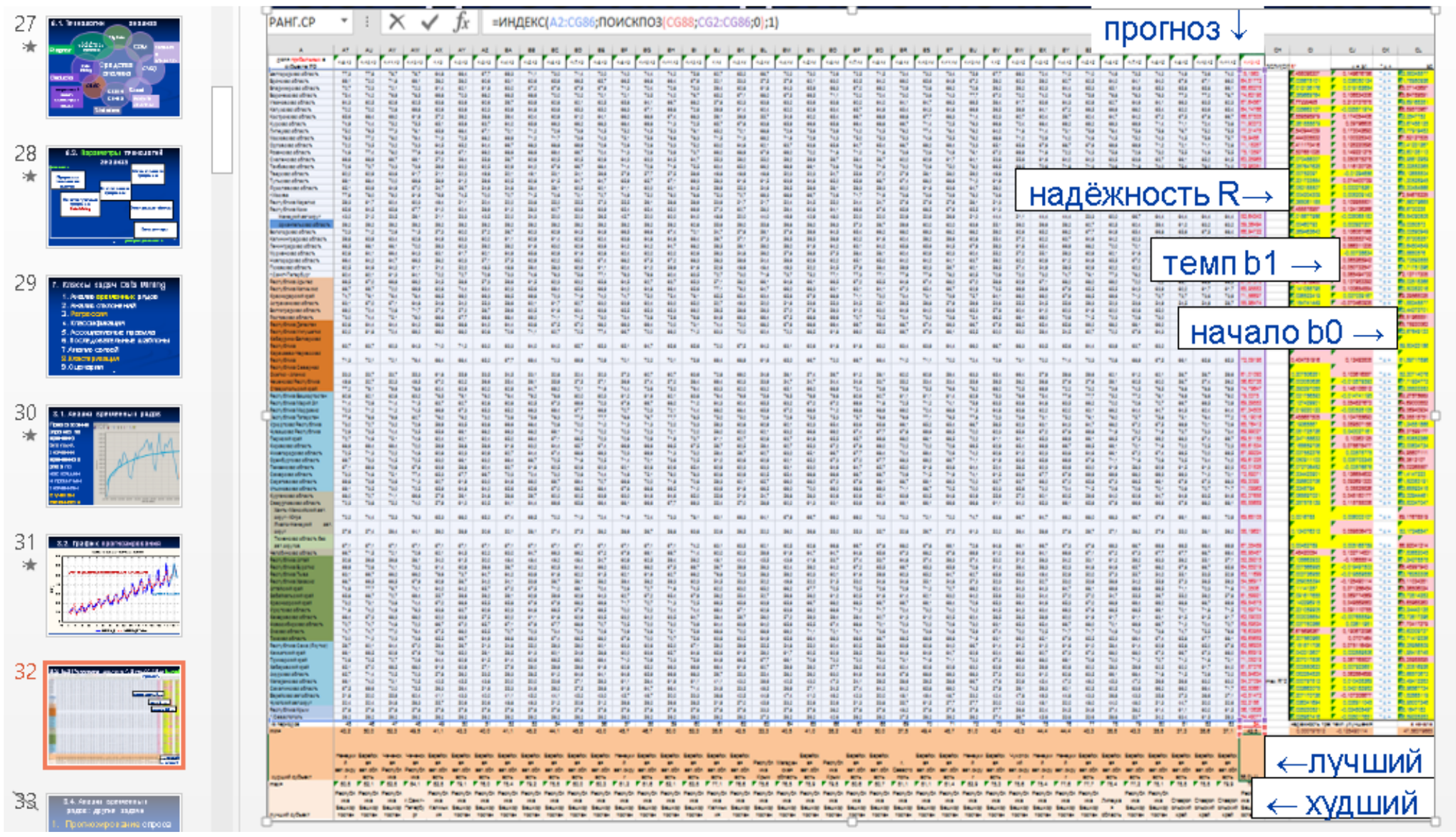


Рис. 3. Excel – модель мониторинга и прогноза значения показателя

Слева – регионы по округам, справа – рассчитываемые прогноз на следующий период, а также квадрат коэффициента корреляции R^2 его надежности, темп роста b1 (убыли) и начальное значение b0, внизу – минимальные и максимальные значения по периоду с указанием достигшего их региона (через функции ПОИСК ПОЗИЦИИ и ИНДЕКС – см. вверху в строке формул на рисунке 3) – лучшего и худшего региона в каждый отчетный период (внизу). Ниже (не видно) – сортировка значений темпа роста и прогноза посредством РАНГА с указанием достигшего их региона, а также определение регионов, близких (через разность соседних) по каждому параметру.

Так по наименьшей доле убыточных предприятий: до августа 2015г. и по прогнозу на 2016– лидер Башкортостан, хотя в сентябре – ноябре 2015 – Ставропольский край. Значит, в этих регионах **наилучшие условия для работы предприятий**.

IV. Моделирование динамики другого показателя по модели III:

Для однотипных меняющихся данных (например, Росстата) **моделирование** – единственный способ перейти от кустарных решений к промышленной работе. Однако сначала бывает нужно:

а) логическое восстановление отсутствующих данных по некоторым показателям – если:

1) регион не производит вид товаров (в Москве нет рыболовного промысла, а в Санкт-Петербурге – сельского хозяйства), то 0%

2) индекс не изменился, то 100%,

3) отчетность не сдавалась (Северный Кавказ, Крым), то по среднему из имеющихся.

б) преобразование – переход от номинала к доле (разделить на максимум и умножить на 100%) – для корректного сравнения данных по разным показателям в п.V.

Подставляя численные данные **любого из 51-го показателя с 2009 по 2015 год** в светлое рабочее поле таблицы в Excel (см. рис. 3 выше) справа и внизу сразу получаем результат (аналогичный описанному под рис. 3).

С магистрантами направления ГМУ проведено сравнение 85-ти регионов РФ по динамике каждого из 51-го показателя с 2009 по 2015 (и даже 2016) год.

Например, среди округов (без Крыма) **по среднедушевым денежным доходам**:

в январе 2009 г. Северо-Кавказский округ был беднейшим, а Уральский – богатейшим; по факту и прогнозу на январь 2017 г. аутсайдером стал Сибирский (худший по темпу роста, поэтому с августа 2014 г. попеременно с Северо-Кавказским становится минимальным); с февраля 2009 г. **бессменно лидирует Центральный**, хотя по темпу быстрее растет Дальневосточный округ.

V. Совместный анализ регионов по нескольким показателям (до 6):

1. Обозначим «0» – № субъекта. Рассмотрим, к примеру, **показатели *1 – 6***:

1) индекс промышленного производства (отношение к предыдущему периоду)

2) доля добычи полезных ископаемых

3) доля обрабатывающих производств

4) доля производства и распределения электроэнергии, газа и воды

5) доля отгруженных товаров собственного производства, выполненных работ и услуг, к max субъекту (Москва)

6) доля по вылову рыбы к max субъекту (Мурманск):

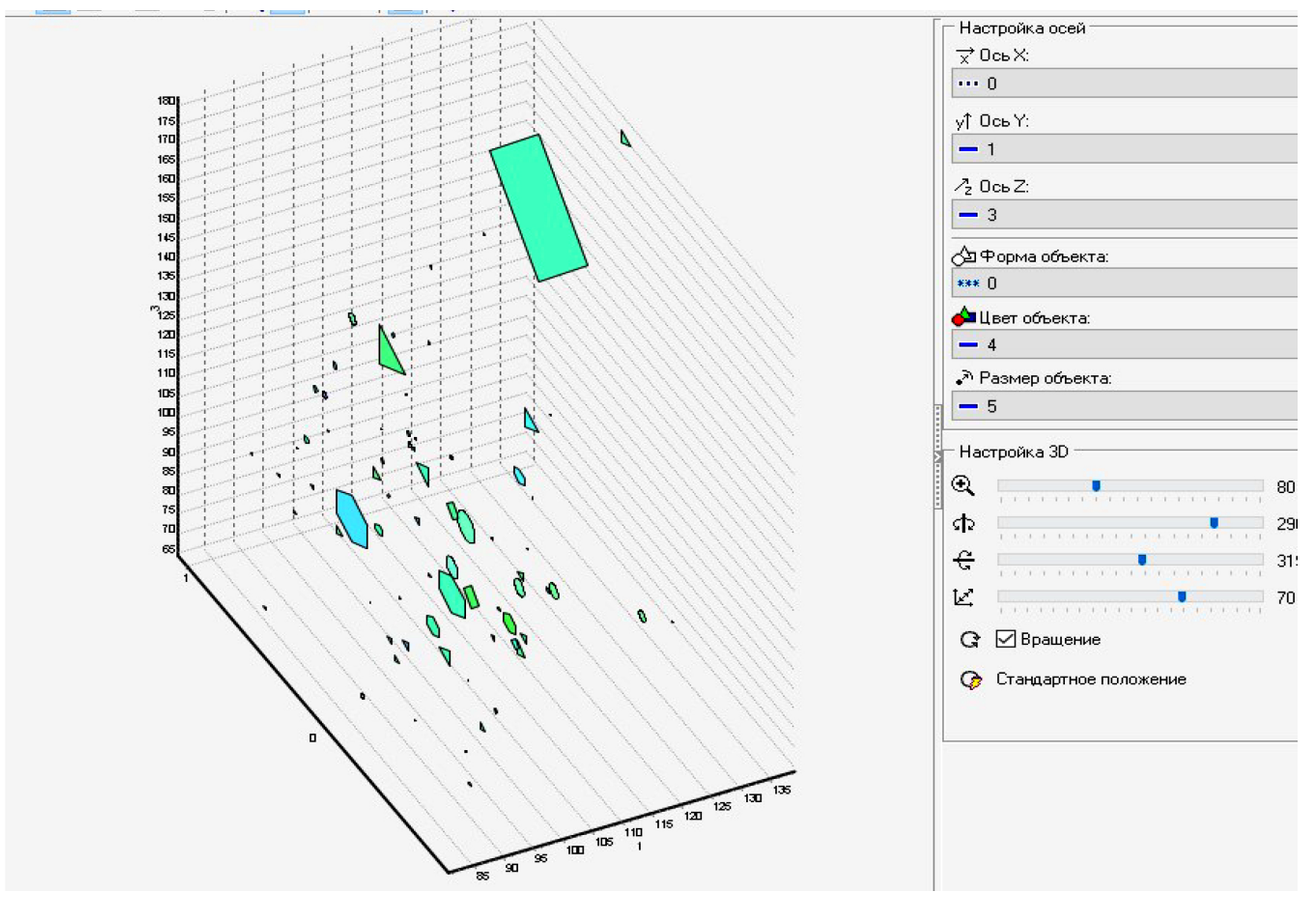


Рис. 4. Максимум 6 измерений: 3 пространственные оси, фигура, цвет и размер в Deductor

Для малого числа объектов (чтобы не загораживали друг друга) на диаграмме размещения видим выбросы. Их надо приблизить к остальным для обнаружения нетривиальных (в том числе, числовых) более надежных зависимостей показателей (см. п. 3 ниже) и скрытых переменных (факторов – см. п. 4

ниже). Классификация **подходов и задач углубленного анализа Big Data (BD)**, в т.ч., Business Intelligent (BI)-поддержки управления, даны в таблицах 1 и 2 статьи [3].

2. В Excel найдены попарные корреляции показателей для 85-ти регионов:

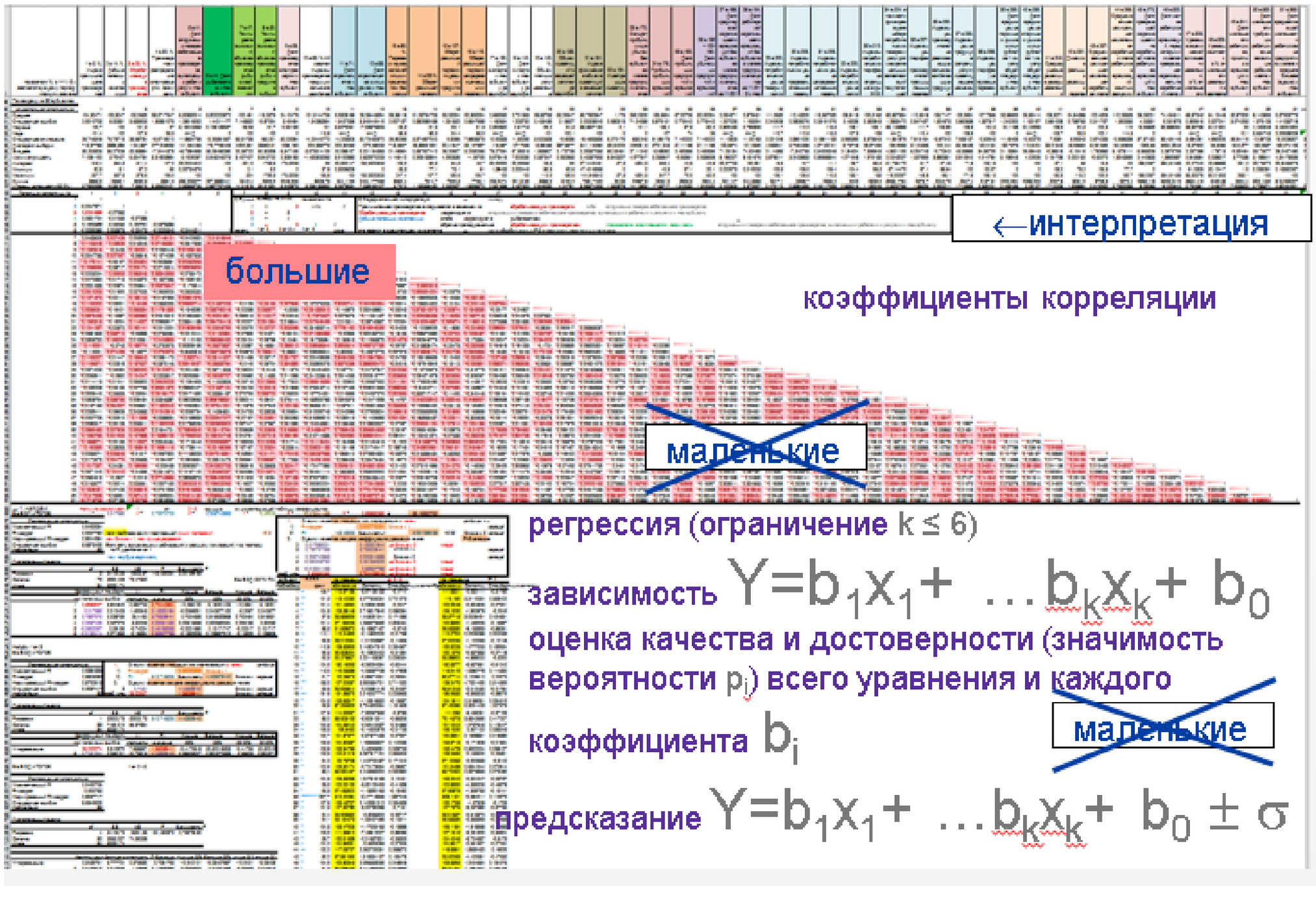


Рис. 5. Корреляционная и регрессионная модель

Выводы по знаку коэффициента корреляции **по показателям** со смысловой интерпретацией:

2~6, но ***1*** ⊥ ***2, 6*** и ***2*** ⊥ ***3, 4, 5*** («~» и «,»–одинаковый знак, «⊥»–разный) сырье на экспорт не дает вклада в отечественную промышленность). А по max модулю корреляции:

1 сильно коррелирует с ***3***, слабее – с ***5***, но ***3*** коррелирует с ***5*** →оставим б**о**льшую: ***1*** от ***3***.

3. Для корреляций, близких к 1, построены **регрессии** – формулы, с помощью которых:

а) найден вклад показателей в «1» = индекс промышленного производства

1 = 0,018 * ***2*** + **0,7597** * ***3*** + 0,207 * ***4*** – 0,06 * ***5*** – 0,013 * ***6*** + 2,599 ± 38,168

1 = **0,745** * ***3*** + 26,02 ± 52,91

1 = **0,757** * ***3*** + 0,205 * ***4*** – 0,064 * ***5*** + 5,0163 ± 38,144

б) последнее – лучшее (максимальный R^2), хотя погрешость (±) все равно велика.

Таким образом, наша промышленность на **76**% состоит из **обработки** тех малых количеств (**1,8%**) **полезных ископаемых**, что **остались в России**.

в) сравнены характеристики надежности в Deductor (81) и Excel (85) ® аналогичны.

4. Найдены скрытые зависимости **(факторы)** – линейные комбинации первичных показателей. От них найдены более надежные регрессии в Deductor и Excel (ниже):

Таблица 1

Сравнительная таблица характеристик регрессии для «***1***» от 4-х факторов

Источник	Сумма квадратов SS	Число степеней свободы, df	Средние квадраты, MS	F-критерий	Значимость
Регрессия	54718,12	4	13679,53	214,8777	0,0000
Ошибки	4838,31	76	63,66		
Сумма	59556,43	80			
Регрессия	54126,73	4	13531,68	181,480875	$2{,}7567*10^{-39}$
Остаток	5965,007	80	74,56		
Итого	60091,74	84			

Таблица 2

Сравнительная таблица характеристик коэффициентов регрессии для «1»

Deductor (81)	Нестандартизован-ные коэффициенты		Стандартиз-е коэффициенты	t-критерий	Значи-мость	Доверительный интервал (95%)	
	Значение	Ошибка				Значение	Ошибка
Константа	4,1481	6,3505		0,6532	0,5156	-8,5000	16,7962
«2» (X0)	0,0148	0,0114	0,0426	1,2979	0,1982	-0,0079	0,0376
«3» (X1)	0,7698	0,0266	0,9680	28,9564	0,0000	0,7169	0,8228
«4» (X2)	0,1918	0,0535	0,1177	3,5876	$5{,}882*10^{-4}$	0,0853	0,2982
«5» (X3)	-0,0914	0,0751	-0,0406	-1,2170	0,2274	-0,2409	0,0582

Excel (85)	Коэффи-циенты	Стандартная ошибка	t-статистика	P-Значение	Нижние 95%	Верхние 95%
Y-пересечение	2,5459791	6,777724	0,375639	0,7081793	-10,94212	16,034079
2	0,0179816	0,012328	1,45858	0,14859583	-0,006552	0,042515
3	0,759689	0,028585	26,57659	$1{,}8684*10^{-41}$	0,7028033	0,8165747
4	0,2067891	0,05673	3,64514	0,00047362	0,0938926	0,3196856
5	-0,0595566	0,080788	-0,73719	0,46316062	-0,22033	0,1012169

VI. Сэмплинг – уменьшение выборки субъектов до репрезентативной

1) по 1-му критерию удалось уменьшить с 85 до 48-ми со схожестью 99.7%,

2) стратифицированный – сразу по 4-м факторам: уменьшили с 85 до 42-х:

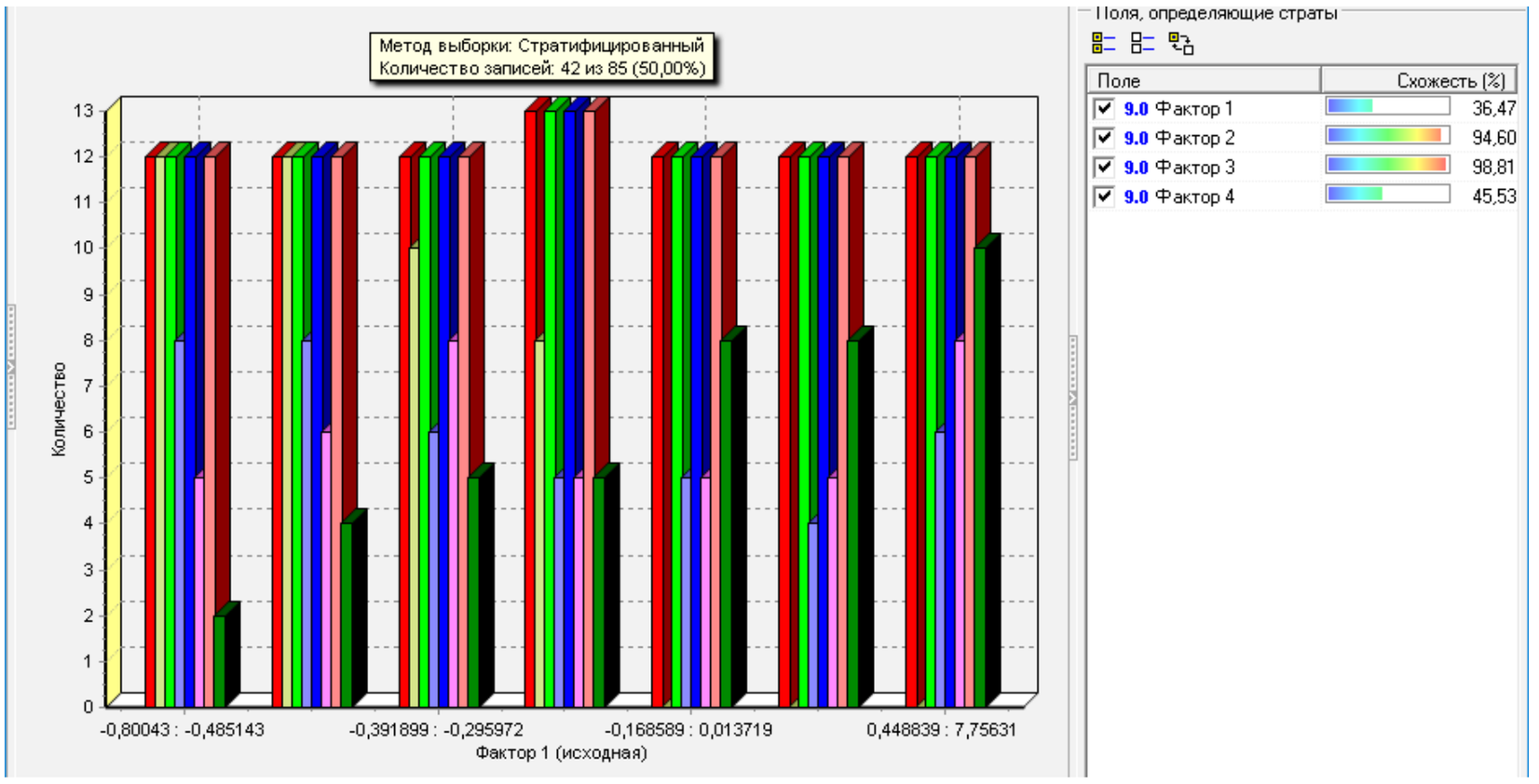


Рис. 6. Гистограмма 4-х факторов (для 85 и 42) по 7-ми диапазонам взвешенного критерия

VII. Кластеризация – объединение «похожих по критерию» объектов в сравнительно однородные группы, существенно отличающиеся от других групп

1. Deductor сам разбил РФ на 5 кластеров по экономике с помощью карт Кохонена:

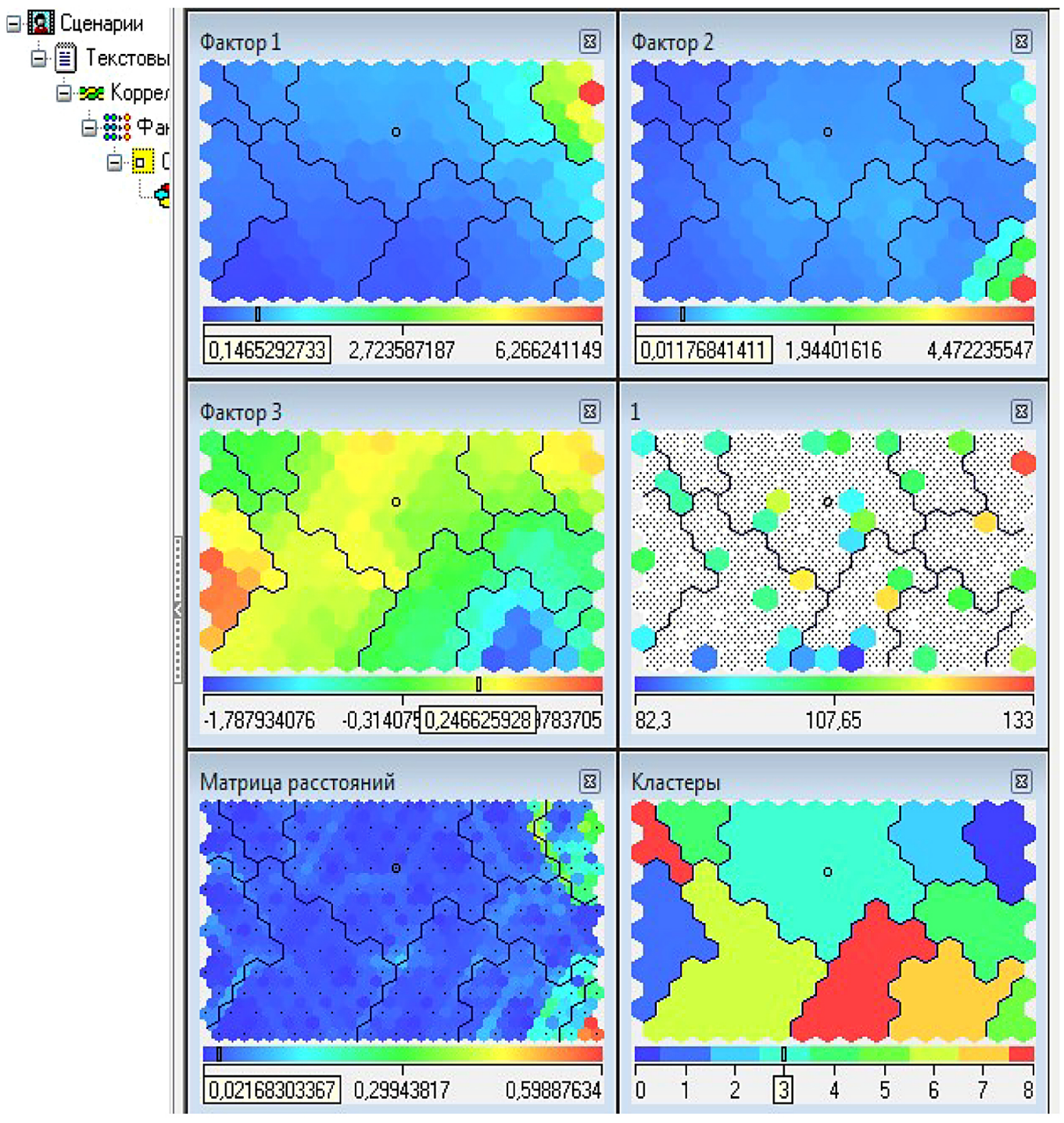


Рис. 7. Структура 5-ти экономических кластеров

По значению показателей их можно назвать:

0) застойные;
1) возрождающиеся большие регионы;
2) дотационные;
3) растущие;
4) возрождающиеся малые регионы.

21_OUT / 21_ERR / 21_ERR_1	Расстояние до центра кла...				
Номер кластера	0,012596	0,036862	0,046059	0,047394	0,052913
0					
1	2			2	2
2					
3		3	3		
4					
Итого:	2	3	3	2	2

Рис. 8. Структура 5-ти экономических кластеров

Видим, что застойных и дотационных – мало.

2. В Deductor задали 4 кластера по демографии (нумерация с 0 по 3):

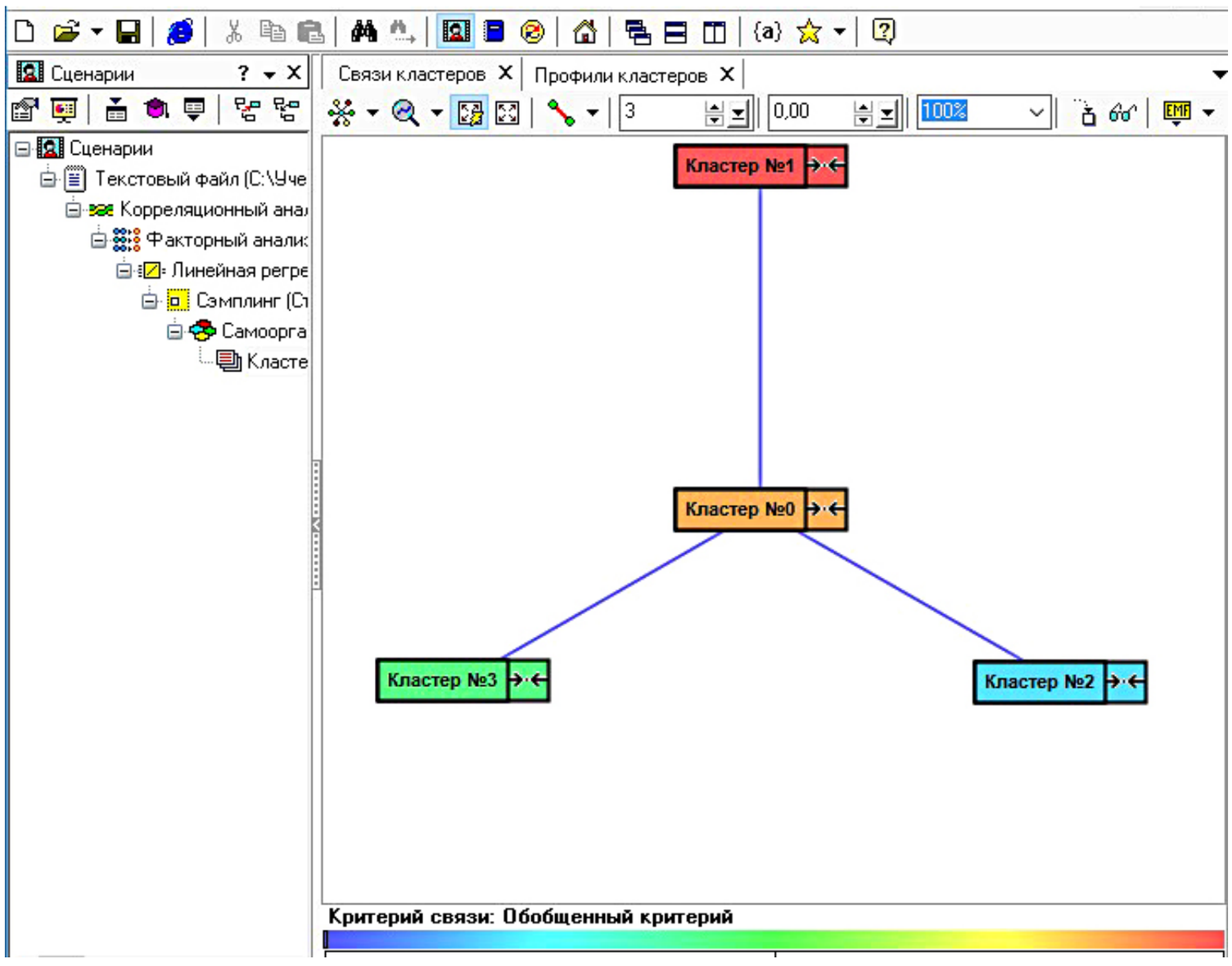


Рис. 9. Сильные (синие) связи 4-х демографических кластеров

Можно визуализировать характеристики кластеров и таблицы расстояний:

Что-если | Матрица сравнения | Профили кластеров | Обучающий набор | Диаграмма связей | Куб

Кластеры

Кол-во записей: 20

Поля	Показатели	Возрождающиеся 20 (22,5%)	Вымирающие 39 (43,8%)	Дотационные 13 (14,6%)	Вымирающие 17 (19,1%)	Итого
9.0 Среднегод	Значимость	90,4%	87,3%	99,5%	100,0%	100,0%
9.0 Число	Значимость	100,0%	100,0%	99,8%	71,8%	100,0%
9.0 Удельный	Значимость	95,5%	44,5%	100,0%	99,3%	100,0%
9.0 Число	Значимость	77,9%	100,0%	100,0%	77,9%	100,0%
9.0 Среднедуш	Значимость	97,4%	93,0%	50,7%	13,6%	97,4%
9.0 Изменение	Значимость	41,6%	31,7%	59,0%	0,7%	36,5%

Показатель	3
Среднее	78
Стандартно...	24
Стандартна...	25
Минимум	1,7
Максимум	50
Сумма	5,6
Сумма квад...	32
Кол-во пуст...	0

1 / 20

Регион	Численность населения (тыс. чел.)	Удельный вес городского населения %	Изменение численности населения	Число родившихся на 1000 населения.	Число умерших на 1000 .	Среднегодовая численность населения занятых в экономике (тыс. чел.)	Среднедушевые денежные доходы (руб./мес.)	Номер кластера	Расстояние д кластер
Республика Коми	1135	74,2	-1,3	8,5	10,7	495,3	1654,4	2	0,072333812
Ненецкий АО	46	59,2	-0,6	11,2	9,4	20,7	1268,8	2	0,21416035
Мурманская область	983	91,9	-1,7	7,6	10,1	422,5	1264,5	2	0,22732134
Республика Северная Осет	674	67,6	1,4	10,8	12,6	220,1	1162,4	2	0,17552675
Удмуртская Республика	1633	69,4	-0,2	9,7	12,7	744,7	1246,9	2	0,15501520
Томская область	1067	66,9	-0,4	9,1	12,4	455,6	1047,3	2	0,16760947
Тюменская область	3221	76,2	-0,2	10,3	8,5	1750	1480,1	2	0,26833765
Ханты-Мансийский АО	1368	91,2	-0,1	10,8	6,2	792,6	3266	2	0,34647985
Ямало-Ненецкий АО	495	82,7	-0,6	12,2	5,3	301,3	4398	2	0,47457968
Республика Бурятия	1035	59,6	-0,6	11	12,6	395,8	1421,7	2	0,20222299
Таймырский (Долгано-Нене	43	64,9	-1,4	10,3	8,7	22	1694,9	2	0,15349703
Республика Саха (Якутия)	977	63,9	-1,2	13	9,6	477,8	1717,5	2	0,21816988
Еврейская автономная обл	199	67,5	-1,8	9,3	13,6	73,4	2840,6	2	0,22834597
Чукотский автономный окр	72	67,9	-5,9	9	7,1	33,8	963,7	2	0,23671170
Приморский край	2174	78,3	-1	8	12,8	945,8	2450,8	2	0,19655993
Хабаровский край	1518	80,6	-1,1	7,9	13,1	665	1314	2	0,19052383
Амурская область	1006	65,3	-0,9	9,2	12,5	427,2	1621,3	2	0,13705237
Камчатская область	383	80,8	-1,8	9,1	10,3	177	1191	2	0,12922733
Магаданская область	233	92	-2,9	8,2	9,3	113,1	2416,6	2	0,22829564
Сахалинская область	598	86,7	-1,6	8,2	12,2	261,9	2423,2	2	0,19164902

Рис. 10. Профили 4-х демографических кластеров

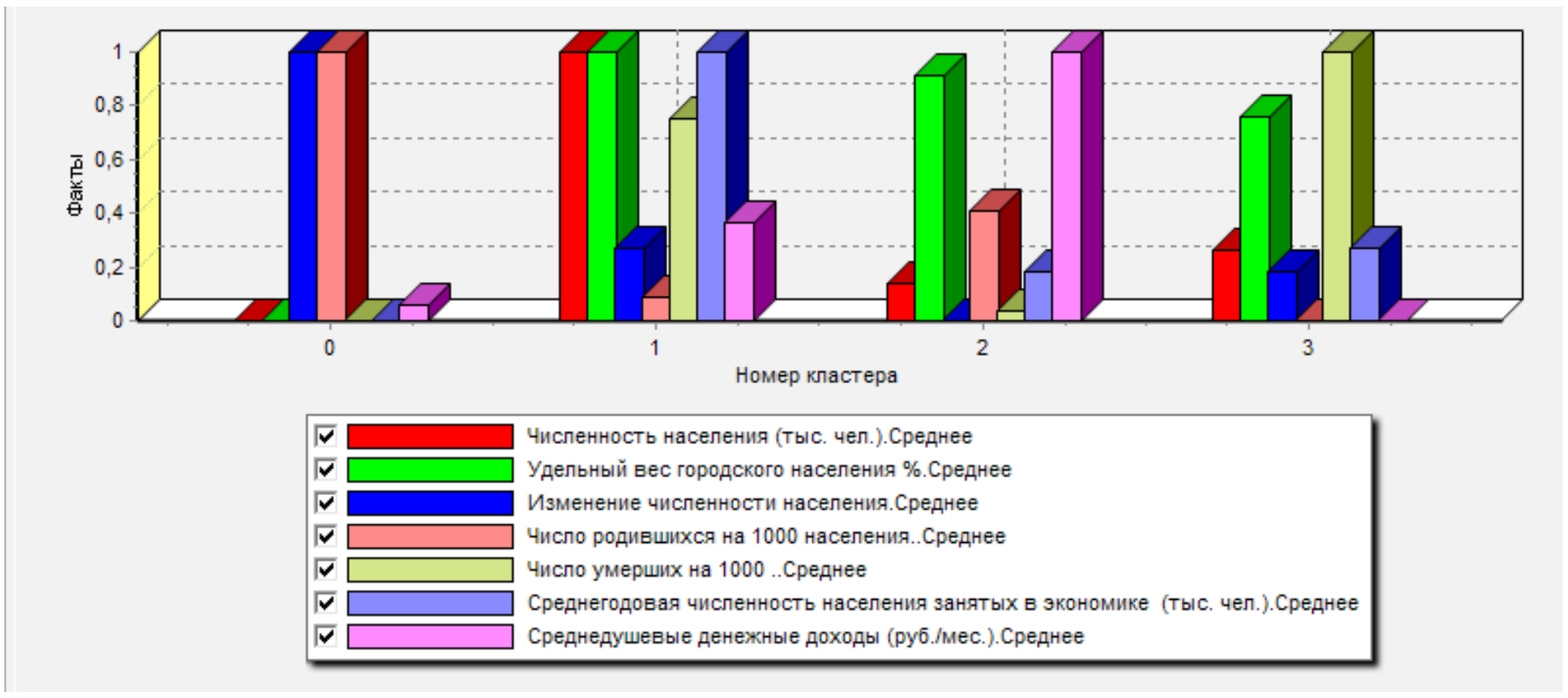


Рис. 11. Кросс-диаграмма 4-х демографических кластеров

Названия придумали по значениям смертности:

0) дотационные;

1) вымирающие большие;

2) возрождающиеся;

3) вымирающие малые регионы.

В кластере № 0 положение дел было хорошее. В остальных 3-х численность населения очень сильно падает. Поэтому с 2007 года в России **оснащают мед. учреждения и стимулируют рождаемость** по программе «Материнский капитал».

Кластеризация нужна для уменьшения размерности. Для каждой группы похожих регионов разрабатывают свою стратегию их развития, опробуют на пилотном регионе.

Операции VI–VII осуществлены посредством Data Mining: вместо неизвестных математических формул нейронные сети с обучением сами подбирают коэффициенты.

Итак, произведено прогнозирование экономических и группирование социальных показателей Росстата:

1) выборка данных (связывание данных);

2) очистка (заполнить пропуски, удалить аномалии);

3) трансформация (сгруппировать по близости, скользящее окно);

4) Data Mining – моделирование (линейная временн**а**я и множественная регрессия, нейронная сеть для группирования);

5) интерпретация результатов (ретро-прогноз, диаграмма рассеяния, распределение ошибки или таблица сопряженности для достоверности).

На основе отклонения прогноза от цели ОГВ принимают стратегию и разрабатывают программу мероприятий для достижения запланированных значений.

VIII. Стратегию развития вырабатывают в виде сценариев с использованием математических формул, например, в ПО Prognoz [3], *но в бесплатной версии это невозможно, а платную факультет не приобрел.*

Наконец, по названиям факторов руководству надо представить смысловую интерпретацию полученных знаний, дать управленцу практически полезные комплексные рекомендации. **Эффекты комплексного систематического анализа** – улучшение: планирования, экономии, интеграции, реализации программы, ее мониторинг и его оперативность для повышения качества жизни населения.

Управленцу без аналитической подготовки не то, что создать, трудно даже понять готовую модель, поэтому у него мало доверия к результатам и, как следствие, – отказ в применении модели. Для удобства в дорогих заказных разработках используют **цветовой индикатор – визуализатор результата** (комплексная **выходная** характеристика **= взвешенной сумме входов** = независимых показателей, где веса – например, по коэффициентам линейной регрессии) и гиперссылки для подробных комментариев («Интегрум»). В Excel подобное можно реализовать на отдельном листе.

Процессы создания (синтеза) социально-экономических моделей, их использование (моделирование) и сложности интерпретации результатов предполагают аналитическую (предметную, математическую и технологическую) квалификацию экспертов. Сейчас на эти должности берут программистов (для интеграции данных) и математиков (для анализа). Но при упрощении интерфейса аналитического ПО лучшие наши выпускники тоже справятся с этой деятельностью.

Авторские материалы [3] и **методику их использования в электронном виде** можно скопировать с https://sites.google.com/site/2018fguiatu/задания. Методика апробирована на магистрах различных групп **факультета государственного управления (ФГУ)**, показала возрастание интереса к исследова-

нию и повышение качества анализа данных. Для сильных студентов – не только работа по инструкции, но и создание своих **моделей** поддержки и принятия управленческих решений.

Это научит управленцев дифференцировать и обосновывать принимаемые решения. Накапливаемая **коллекция моделей** полезна как студентам, так и аналитикам.

Тенденции развития рынка и **перспективы Business Intelligent** перечислены в [3]. Они убедительно призывают к расширению и углублению подготовки по ИАТУ.

Это в дальнейшем поможет развитию аналитической составляющей и модельного ряда ИСУ, совершенствованию порталов ОГВ и сайтов органов местного самоуправления **для актуальной поддержки** экономики и населения.

Список литературы

[1] Центральная база статистических данных (ЦБСД) на Едином Интернет-портале Росстата. URL: http://cbsd.gks.ru (30.05.2015).

[2] Государственная автоматизированная информационная система «Управление» (ГАСУ) URL: http://gasu.gov.ru (30.05.2017).

[3] *Смольникова И.А.* Методы и технологии анализа данных // Государственное управление в XXI веке: Материалы 14-й Международной конференции факультета государственного управления МГУ имени М.В. Ломоносова, секц. 2. Инновации, 2016. – С. 108–118 URL: http://www.spa.msu.ru/uploads/files/books/publikazija.pdf