**Индексы по отдельным доменам**

**Введение и исходные данные**

Исходным элементом нашего исследования является набор данных, организованный в виде словаря, где ключом является год, а значением – DataFrame с данными для этого года. Каждый такой DataFrame имеет следующую структуру:

* **Строки** – регионы (названия субъектов Российской Федерации);
* **Столбцы** – фиксированные переменные, включающие:
  + object\_name: название субъекта,
  + object\_level: уровень объекта (например, регион, федеральный округ или страна; в дальнейшем остаются только регионы),
  + year: год (фактически повторяет ключ словаря),
  + далее следуют характерные для данного домена признаки. Для домена «Население» это могут быть коэффициент естественного прироста, ожидаемая продолжительность жизни и т.д., для домена «Финансы» – показатели по банковским счетам, задолженности, числу действующих кредитных организаций и т.п.

Получив этот словарь, который далее мы будем именовать доменом, мы переходим к построению интегрального доменного индекса регионов (DIR).

**Требования к свойствам индекса**

Мы считаем, что итоговый индекс должен удовлетворять следующим свойствам:

1. **Сопоставимость одного региона в разные периоды времени.**  
   Это позволяет отслеживать динамику развития региона.
2. **Сопоставимость двух любых регионов в рамках одного периода.**  
   Индексы разных регионов, рассчитанные в один год, должны быть сравнимы между собой.
3. **Сопоставимость двух любых регионов, рассчитанных в разные периоды.**  
   Такой подход обеспечивает возможность межрегионального сравнения вне зависимости от года расчёта.

Если все три условия выполнены, мы можем считать, что индекс является справедливым и информативным для оценки состояния региона.

**Методология формирования индекса**

**Приведение данных к единому масштабу (нормализация)**

Поскольку данные для каждого признака могут иметь различные шкалы и размерности, для корректного агрегирования признаков их необходимо привести к единому масштабу. Мы используем **мин-макс нормализацию**.  
Применяется два подхода:

* **Нормализация по текущему году (Current Normalization).**  
  Для каждого года отдельно для каждого признака вычисляются минимальное и максимальное значения, и затем данные нормируются в рамках данного года. Такой подход отражает положение региона относительно других регионов именно в этот период.
* **Нормализация относительно базового года (Base Normalization).**  
  Для каждого признака минимальное и максимальное значения определяются на основе заранее выбранного базового года (например, 2012 или 2010). Этот подход позволяет оценивать динамику региона относительно его исходного состояния, зафиксированного в базовом году.

В результате мы получаем два набора нормализованных данных:

* **Current Index** – набор нормированных значений, полученных с использованием текущих минимальных и максимальных значений каждого признака.
* **Base Index** – набор нормированных значений, полученных с использованием минимальных и максимальных значений признаков, рассчитанных по базовому году.

**Комбинирование нормализованных индексов**

После нормализации мы строим два индекса:

* **Current Index:** отражает текущее состояние региона, учитывая относительное положение в каждом конкретном году.
* **Base Index:** отражает состояние региона относительно базового года, позволяя оценить его эволюцию со временем.

Чтобы объединить оба подхода, итоговый доменный индекс (DIR) рассчитывается как линейная комбинация этих двух индексов:



где α – параметр, задающий степень влияния базового состояния на итоговый индекс. Такой подход позволяет гибко регулировать вклад исторического базового уровня и текущих изменений в итоговую оценку.

**Определение весов для нормированных признаков**

При построении как Current Index, так и Base Index итоговый индекс домена рассматривается как линейная комбинация нормализованных признаков:

Изображение выглядит как Шрифт, текст, белый, диаграмма

Контент, сгенерированный ИИ, может содержать ошибки.

где xi​ – нормализованное значение признака, а wi– вес, отражающий его информативность.

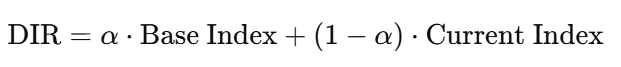
В нашем подходе для определения весов мы рассматриваем три метода:

1. **Метод главных компонент (PCA).**
   * PCA — это статистический метод, который позволяет найти линейные комбинации исходных признаков, так что первая главная компонента объясняет максимальную дисперсию данных. То есть, признаки, которые наиболее сильно варьируются и лучше различают регионы, получат больший вес.
   * Метод помогает уменьшить размерность исходного набора признаков, выявить основные факторы, влияющие на вариативность данных, и назначить веса признакам на основании их вклада в общую дисперсию. Это делает итоговый индекс более информативным и устойчивым к шуму в данных.
2. **Веса с обратной дисперсией.**  
   При этом методе признаки с более высокой дисперсией получают меньший вес, так как высокая изменчивость может говорить о нестабильности показателя.
3. **Равные веса.**  
   Каждый признак вносит одинаковый вклад в итоговый индекс. Этот метод прост, но не учитывает различия в информативности отдельных показателей.

В основном мы используем метод главных компонент (PCA), поскольку он позволяет автоматически определить наиболее значимые признаки с точки зрения формального определения степени информативности. Альтернативные методы (обратная дисперсия и равные веса) доступны для демонстрации и проверки чувствительности итогового индекса к выбору весов.

**Итоговая процедура формирования индекса DIR**

1. **Сбор данных:**  
   Исходный домен представлен в виде словаря, где ключ – год, а значение – DataFrame с региональными данными.
2. **Предобработка:**
   * Фильтрация данных: остаются только регионы.
   * Удаление пропусков: данные очищаются от недостающих значений.
   * Приведение к единой структуре.
3. **Нормализация:**  
   Применяются два подхода нормализации:
   * Нормализация по текущему году → Current Index.
   * Нормализация относительно базового года → Base Index.
4. **Определение весов признаков:**  
   С помощью метода главных компонент (PCA) (или альтернативных методов) вычисляются веса для каждого нормированного признака. Таким образом, для каждого домена получаются два вектора весов:
   * Вектор весов для Current Index.
   * Вектор весов для Base Index.
5. **Агрегация и построение итогового индекса:**  
   Итоговый индекс домена (DIR) рассчитывается как взвешенная сумма двух индексов:

где α — коэффициент, задающий вклад базового состояния (по умолчанию в текущем проекте принимается значение 0,9).

1. **Анализ и визуализация:**  
   После формирования DIR строятся графики динамики индекса, ранжирование регионов, а также анализ влияния отдельных признаков на итоговый индекс.

**Виджеты**

Рассмотрим примеры полученных результатов, которые оформлены в формате «системы мониторинга» с использованием виджетов. Сам прототип представлен следующим образом:

Изображение выглядит как текст, линия, Шрифт, снимок экрана

Контент, сгенерированный ИИ, может содержать ошибки.

Предположим, мы хотим исследовать степень развития определенного домена, например, здравоохранение. Выбираем из доступного перечня:

Изображение выглядит как текст, снимок экрана, Шрифт, число

Контент, сгенерированный ИИ, может содержать ошибки.

И загружаем: Изображение выглядит как текст, Шрифт, линия, снимок экрана

Контент, сгенерированный ИИ, может содержать ошибки.

Теперь рассмотрим опцию интерактивный анализ. В данном виджете можно выбрать конкретный регион, либо смотреть динамику показателей по всей стране в целом. Пусть выбран регион Нижегородская область. Кроме того, можно выбрать доступный и инетересующий нас показатель:

Изображение выглядит как текст, Шрифт, линия, снимок экрана

Контент, сгенерированный ИИ, может содержать ошибки.

Получаем следующий визуал:

Изображение выглядит как линия, текст, График, диаграмма

Контент, сгенерированный ИИ, может содержать ошибки.

Изображение выглядит как текст, диаграмма, График, линия

Контент, сгенерированный ИИ, может содержать ошибки.

То есть имеем динамику с линией тренда на основе классической регрсессионой модели, а также темпы прироста конкрентного показателя.

Сам же доменный индекс региона (DIR) доступен к просмотру в поле ниже:

Изображение выглядит как текст, линия, снимок экрана

Контент, сгенерированный ИИ, может содержать ошибки.

Выбираем, к примеру, Нижегородскую область и получаем следующий график:

Изображение выглядит как текст, линия, диаграмма, График

Контент, сгенерированный ИИ, может содержать ошибки.

Как ранее отмечалась сам DIR строиться на оснве линейной комбианции с коэффициентов α из двух вспомогательных индексов: Base index (степень развития региона с учетом базового года) и Current Index (степень развития региона с учетом ежегодных текущих показателей). Итоговая красная линия (DIR) может быть использована, например, чтобы установить наиболее похожие регионы с точки зрения развития отрасли здравоохранения. К примеру:

Изображение выглядит как текст, диаграмма, линия, снимок экрана

Контент, сгенерированный ИИ, может содержать ошибки.

Расстояние близости определяется примитивно без учета временной принадлежности, но с использованием единичной нормы последовательности индексов. Кроме того, можно посмотреть «рейтинг регионов» с учетом динамики и «антирейтинг». Изображение выглядит как текст, снимок экрана, Параллельный, линия

Контент, сгенерированный ИИ, может содержать ошибки.

Изображение выглядит как текст, снимок экрана, Параллельный, Шрифт

Контент, сгенерированный ИИ, может содержать ошибки.

Важно отменить, что построенный показатель именно что учитывает относительную степень развития региона по сравнению с его развитием в прошлом и развитию других схожих регионов. Чтобы формально понять как был сформирован индекс и что было наиболее значимо с точки зрения вклада конкретных признаков, можно ориентироваться на таблицу весов показателей для кажого года и каждого региона. На примере Нижегородской области и 2012 года имеем следующую таблицу:

Изображение выглядит как текст, Шрифт, число, снимок экрана

Контент, сгенерированный ИИ, может содержать ошибки.

Сразу отметим, что влкда мы определяем как отношение двух чисел. Первое - вес признака умноженный на показатель признака. Второе – значение индекса. На приведенном выше примере нетрудно заметить, что сущесвтеный вклад в формирование индекса в нес такой дестимулирующий показатель, как число аборотов с значением в 27.

**Интегральные индексы**

**Переход к SIIRD (Социальный Интегральный Индекс Регионального Развития)**

Итак, после формирования индексов DIR мы имеем доступ к следующим доменам:  
Изображение выглядит как текст, снимок экрана, Шрифт

Контент, сгенерированный ИИ, может содержать ошибки.

Это означает, что мы имеем формальную оценку развития определенного домена для определенного региона и определенного года. Имея такой набор данных можно продолжить направление агрегации в формирование социального интегрального индекса регионального развития (SIIRD). Для этого мы предположим, что SIIRD может быть сформирован как линейная комбинация индексов доменов (DIR). Однако, учитывая такой способ формирования, нам необходимо также каким-то образом задать веса уже для каждого домена. Руководствуясь доступной практикой, принято решение также использовать метод главных компонент для агрегации разных областей. Кроме того, для стабильности весов индексы каждого домена отнормированы минмакснормализацией, где минимальные и максимальные значения были получены в рамках глобального представления данных. То есть для конкретного признака учитывались значения за все года, а не только за один определенный.

Таким образом, полученные веса имеют следующее распределение:

Изображение выглядит как текст, снимок экрана, диаграмма, линия

Контент, сгенерированный ИИ, может содержать ошибки.

И итоговый сформированный индекс (SIIRD) получается исходя из формулы:

Изображение выглядит как текст, снимок экрана, Шрифт, алгебра

Контент, сгенерированный ИИ, может содержать ошибки.

Продемонструрем полученные результаты индекса в динмике, например, для таких регионов: Московская область, Нижегородская область, Республика Татарстан и Саратовская область.

Изображение выглядит как текст, диаграмма, График, линия

Контент, сгенерированный ИИ, может содержать ошибки.

Изображение выглядит как текст, диаграмма, снимок экрана, линия

Контент, сгенерированный ИИ, может содержать ошибки.

Таким образом, полученный SIIRD обладает рядом ключевых преимуществ:

* Сопоставимость во времени и между регионами. Нормализация данных и единый алгоритм расчёта делают индексы сравнимыми как для одного региона в разные периоды, так и между различными регионами за один и тот же период.
* Объективность оценки. Применение PCA для определения весов позволяет автоматически выявлять наиболее значимые домены, минимизируя субъективность в выборе параметров. Это делает итоговый индекс более объективным и надежным.
* Комплексный подход. SIIRD агрегирует несколько доменных индексов, каждая из которых отражает отдельный аспект социально-экономического развития региона. Такой интегральный подход обеспечивает всестороннюю оценку состояния региона.
* Интерпретируемость результатов. Весовое распределение, полученное с помощью метода главных компонент, позволяет определить, какие домены вносят наибольший вклад в общее состояние региона. Это полезно для определения приоритетов в региональном развитии и разработки целевых программ поддержки.

**Формирование FIIRD**

Финансовый интегральный индекс регионального развития формируется исходя из индекса DIR для финансового домена, особенностью которого является относительно большее число признаков. Единственный нюанс заключается в том, что предложенный способ формирования этого индекса не гарантирует, что его значения заключены между 0 и 1. Тем не менее, подобное свойство никак не фигурировало до этого и не требовало отображения в определенный промежуток. Однако FIIRD тем и будет отличаться от финансового DIR, что значения FIIRD будут нормализованы по методу минмакснормализации. Таким образом, полученный индекс выглядит следующим образом на примере двух субъектов: Санкт-Петербург и Кировская область:

Изображение выглядит как текст, диаграмма, График, снимок экрана

Контент, сгенерированный ИИ, может содержать ошибки.

Изображение выглядит как текст, диаграмма, График, линия

Контент, сгенерированный ИИ, может содержать ошибки.

**Формирование SFIIRD**

Социально-Финансовый интегральный индекс регионального развития может быть сформирован как линейная комбинация двух полученных ранее индексов: SIIRD и FIIRD. В самом простом случае используется сочетание с коэффицентов альфа равным 0.5, то есть просто среднее арифметическое. Ниже несколько примеров:

Изображение выглядит как текст, диаграмма, линия, График

Контент, сгенерированный ИИ, может содержать ошибки.

Изображение выглядит как текст, диаграмма, линия, График

Контент, сгенерированный ИИ, может содержать ошибки.

Изображение выглядит как текст, диаграмма, линия, График

Контент, сгенерированный ИИ, может содержать ошибки.

Изображение выглядит как текст, диаграмма, линия, График

Контент, сгенерированный ИИ, может содержать ошибки.

Изображение выглядит как текст, диаграмма, линия, График

Контент, сгенерированный ИИ, может содержать ошибки.