# ФАКТОРНЫЙ АНАЛИЗА КАК ИНСТРУМЕНТ ВИЗУАЛИЗАЦИИ МНОГОМЕРНЫХ ДАННЫХ. R-ТЕХНИКА. АЛГОРИТМ ХОТЕЛЛИНГА

**Д.В. Кутузова , В.А. Ляхина**
Научный руководитель А.Ю. Трусова
Самарский национальный исследовательский университет
имени академика С.П. Королева

Массивы многомерных данных сложны для восприятия и поэтому существует определенная методика, которая позволяет анализировать многомерные данные. В работе решались следующие задачи: изучение экономической сферы и уровня жизни граждан в доковидный и послековидный периоды в различных регионах России; изучение математического инструмента факторного анализа, алгоритм Хотеллинга; выделение двух главных факторов; сжатие массива 11 на 7 до размерности 11 на 2.

Шаги алгоритмаХоттелинга [1]:

1. Создание матрицы исходных данных, расчет матрицы корреляции R
2. Расчет общностей $h_j$ (редуцированная матрица с общностями на главной диагонали) методом максимальной корреляции
3. Пересчет$R_h$ - редуцированной матрицы
4. Возведение редуцированной матрицы в степень $R_h^2 \to R_h^4 \to \cdots \to R_h^{16}$ процедура заканчивается, когда0,01<d<0,02

(1) $d = \left| a^{(n)} - a^{(n=1)} \right|$

5. Выделение главного фактора:
   - 5.1) Расчет вектора бета
   - (2) $\beta_1 = R_H \quad \alpha_1^{(n)}$
   - 5.2) Расчет компоненты матрицы факторного отображения
   - (3) $A = \dfrac{U_1 \quad \sqrt{\lambda_1}}{\left( \quad U_{1i}^2 \right)^{1/2}}$
6. Расчет восстановленной матрицы корреляции $R_h^+$ и матрицы первых остаточных коэффициентов

(4) $R_1 = R_h - R_h^+$

7. Последующие итерации аналогично предыдущим, только вычисления производятся на данных матрицы остатков$R_1$, также рассчитывается остаточная матрица

(5) $R_2 = R_1 - A_2 \quad A_2^{'}$

В исследовании изучались следующие показатели:

$X_1$ - Среднемесячная номинальная начисленная з/п работников по полному кругу организаций, тыс. руб.

$X_2$ - Стоимость условного (минимального) набора продуктов питания, тыс. руб.

$X_3$ - Задолженность по кредитам в тыс рублях, предоставленными кредитными организациями физическим лицам

$X_4$ - Стоимость турпакетов, реализованных населению гражданам России по территории России, тыс. руб.

$X_5$ - Стоимость турпакетов, реализованных населению гражданам России по другим странам, тыс. руб

$X_6$ - Количество соглашений по экспорту, тыс. руб

$X_7$ - Количество соглашений по импорту, тыс. руб

В ходе исследования были сформированы массивы данных, используя статистику Росстата за 2019/2021 годы [2].

Анализируя матрицу факторного отображения, формируется первый латентный фактор.Он включает в себя показатели F1={$X_3$, $X_4$, $X_5$, $X_6$, $X_7$} - Фактор торгово-экономической сферы. Второй латентный фактор это F2={$X_1$, $X_2$} - Фактор уровня жизни.

На рисунке 1 визуализированы изучаемые показатели за 2019 год в пространстве латентных факторов.

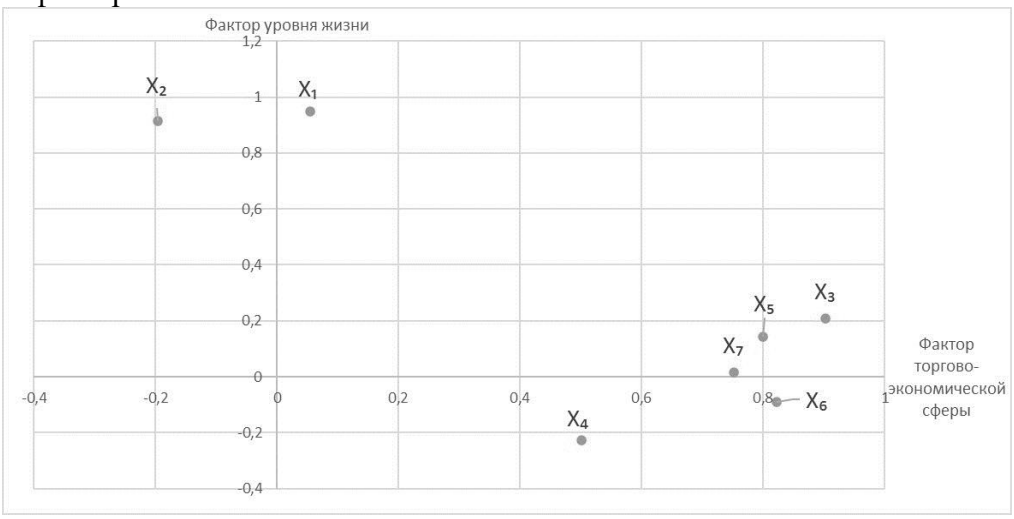


Рисунок 1 - Показатели в пространстве латентных факторов 2019 год.

Как можно заметить показатели $X_1$, $X_2$ локализованы у оси фактора уровня жизни, а $X_3$-$X_7$у оси фактора торгово-экономической сферы.

Используя соотношение 6, были рассчитаны координаты в латентном пространстве:

(6) $b = (A^T \ A)^{-1} \ A^T \ Y^T$

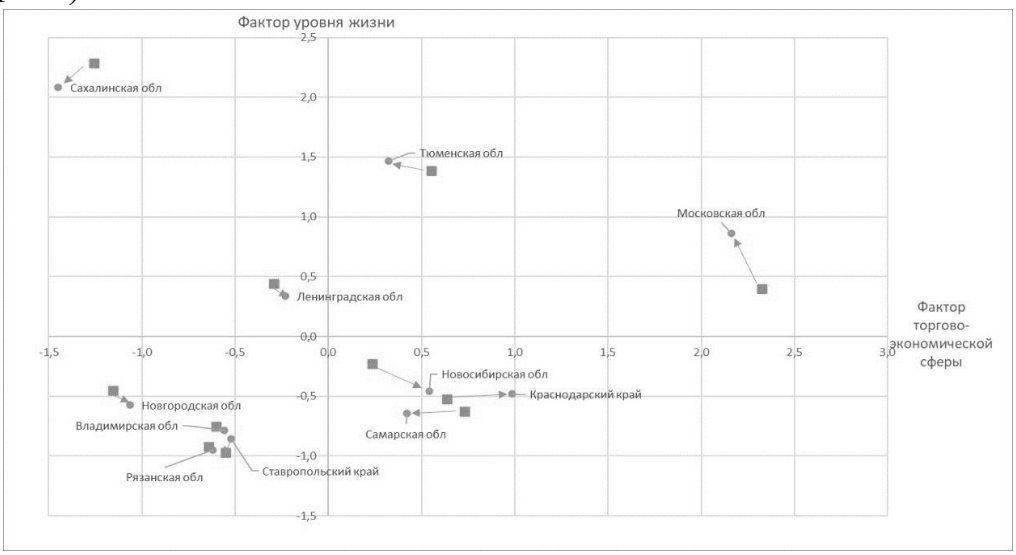


Рисунок 2 - Сравнение расположения объектов 2019/2021 года.

Рассчитаны компоненты латентных факторов в виде линейной комбинации изучаемых показателей:

Для 2019 года

(7) $$F_1 = 0{,}054\,x_1 + (-0{,}197)\,x_2 + 0{,}902\,x_3 + 0{,}501\,x_4 + 0{,}800\,x_5 + 0{,}822\,x_6 + 0{,}752\,x_7$$
$$F_2 = 0{,}948\,x_1 + 0{,}914\,x_2 + 0{,}209\,x_3 + (-0{,}277)\,x_4 + 0{,}144\,x_5 + (-0{,}089)\,x_6 + 0{,}015\,x_7$$

Для 2021 года

(8) $$F_1 = -0{,}089\,x_1 + (-0{,}339)\,x_2 + 0{,}907\,x_3 + 0{,}496\,x_4 + 0{,}906\,x_5 + 0{,}546\,x_6 + 0{,}803\,x_7$$
$$F_2 = 0{,}953\,x_1 + 0{,}872\,x_2 + 0{,}330\,x_3 + (-0{,}214)\,x_4 + 0{,}190\,x_5 + (-0{,}166)\,x_6 + 0{,}119\,x_7$$

Рассчитанный коэффициент информативности составили 70% и 71%, что свидетельствует о достаточности двух выделенных факторов.

**Список использованных источников**

1. Сошникова Л.А и др. Многомерный статистический анализ в экономике: Учеб. пособие для вузов // Под ред. проф. В. Н. Тамашевича. – М.: ЮНИТИ-ДАНА, 1999. – 598 с.

2. Официальная статистика // Федеральная служба государственной статистики. URL:https://rosstat.gov.ru/folder/10705 (дата обращения: 11.09.2023)

# КЛАСТЕРНЫЙ АНАЛИЗ. ПОКАЗАТЕЛИ СМЕРТНОСТИ В СТРАНАХ ЕС И СНГ

**А.В. Непогожева**

Научный руководитель А.И. Ильина

Самарский национальный исследовательский университет имени академика С.П. Королева

Актуальность: кластерный анализ позволяет сегментировать данные по схожести объектов для более детального изучения.

Научная новизна: применение методов кластеризации без обучения для визуализации многомерных данных показателей смертности.

Практическая значимость: по результатам кластеризации полученные группы объектов могут быть рекомендованы в использовании при разработке стратегий развития здравоохранения в различных странах мира.

Цель исследования: образование групп схожих между собой объектов по показателям смертности от различных болезней.

Задачи:

1. Изучить проблему смертности населения в странах ЕС и СНГ.
2. Ознакомиться с математическим инструментарием кластерного анализа и выбрать оптимальные алгоритмы.
3. Провести процедуры методов кластерного анализа, разделить страны на кластеры, визуализировать данные.

Кластерный анализ – это совокупность методов, позволяющих классифицировать многомерные наблюдения, каждое из которых описывается набором исходных переменных $X_1, X_2, ..., X_m$ [1].

В качестве доступного инструментария в работе использовался метод Уорда (иерархический метод) и метод поиска сгущений тип «форель» (итеративный метод).

Алгоритм:

1. Нормирование переменных $z_{ij}$.

(1) $z_{ij} = \frac{x_{ij}}{\overline{x_j}}$

2. Расстояние между объектами $d_{ij}$.

(2) $d_{ij} = \sqrt{\sum_{k=1}^{m} \left(x_{ik} - x_{jk}\right)^2}$

3. По методу
   a. Уорда: сумма квадратов отклонений $V_k$.