Санкт-Петербургский государственный университет

Янн Александр

Выпускная квалификационная работа

Уровень дохода по регионам Российской Федерации. Пространственный анализ.

Уровень образования:
Направление: 01.03.02 Прикладная математика и информатика
Основная образовательная программа: 01.03.02 Прикладная математика, фундаментальная информатика и программирование
кафедра математической теории игр и статистических решений

Научный руководитель:
кандидат физико-математических наук,
доцент кафедры математической теории игр и статистических решений
Панкартова Ярославна Борисовна.

Рецензент: кандидат физико-математических наук, доцент кафедры теории систем управления электрофизической аппаратурой
Гончарова Анастасия Борисовна.

Санкт-Петербург
2023

Оглавление

Введение

Пространственный анализ данных является важной областью статистики, которая позволяет исследовать взаимодействие между переменными в пространственном контексте. Один из способов изучить данное взаимодействие — это применить пространственную авторегрессионную модель, которая позволяет оценить влияние объясняющих переменных на зависимую переменную в пространственном контексте.

В данной работе рассматривается построение пространственной регрессионной модели, где зависимой переменной являются среднедушевые денежные доходы населения, а объясняющими - различные социально-экономические факторы, такие как уровень безработицы, наличие инфраструктуры и т.д.

Целью данной работы является исследование пространственной зависимости между регионами Российской Федерации и среднедушевыми денежными доходам населения, и установлению какие объясняющие факторы на нее влияют. Провести сравнительный анализ моделей с учетом пространственной зависимости и без нее. Для достижения этой цели будут использованы методы пространственной статистики и программное обеспечение для анализа пространственных данных.

Результаты данной работы могут быть полезны для различных социально-экономических исследований, а также для разработки политических программ на уровне регионов и государства в целом.

Построение пространственно-регрессионной модели для объяснения среднедушевых денежных доходов населения имеет ряд актуальных применений в современном мире. Ниже перечислены несколько причин, почему такая модель может быть полезной.

1. Политический анализ: среднедушевые денежные доходы населения часто используются в качестве индикатора уровня жизни населения. Поэтому пространственно-регрессионная модель может помочь

оценить, какие факторы влияют на уровень доходов и как эти факторы меняются в пространстве. Например, такая модель может помочь определить, какую роль играет местоположение на доходы населения.

2. Экономический анализ: пространственно-регрессионная модель может являться полезной для понимания, какие экономические факторы влияют на среднедушевые денежные доходы населения. Например, какую роль играет уровень безработицы, уровень образования или доступность капитала на доходы населения.
3. Социальный анализ: среднедушевые денежные доходы населения являются важным индикатором социальной справедливости.

Постановка задачи

Для того чтобы исследовать наличие пространственной зависимости между регионами Российской Федерации и среднедушевыми денежными доходами населения, и установить какие факторы могут влиять на нее, необходимо решить ряд задач:

1. Изучить теоретические основы построения регрессионных моделей с учетом пространственных данных: методы отбора параметров, способ построения матрицы весов, оценка параметров регрессии и способы оценки качества модели.
2. Собрать необходимые данные и сформировать выборку.
3. Найти и обработать ГИС данные субъектов Российской Федерации.
4. Проверить наличие пространственной автокорреляции.
5. Построить и оценить качество пространственной авторегрессии и множественной линейной регрессии.
6. Провести верификацию и сравнение построенных моделей.
7. Дать интерпретацию полученным результатам.

Обзор литературы

Для проведения данного исследования был изучен огромный список научной литературы и различных публикаций из архивов, журналов.

Для определения объясняющих переменных, влияющих на среднедушевые доходы, был изучен ряд экономической литературы:

1. М.И. Баканов, М.В. Мельник, А.Д. Шеремет; под ред. М.И. Баканова. — 5-е изд., перераб. и доп. — М.: Финансы и статистика, 2007. — 536с.
2. Абель Э., Бернанке Б., Макроэкономика. 5-е изд. СПб.: Питер, 2012. — 768 с

Для работы с географическими координатами и построения пространственной матицы весов посредством программных библиотек языка R было изучено пособие: Самсонов Т.Е. Визуализация и анализ географ. данных на языке R. Географический факультет МГУ, 2017. DOI: 10.5281/zenodo.901911.

Для проверки гипотез по средствам программных пакетов языка R, были дополнительно изучено: Методы прикладной статистики в R и Excel / В. М. Буре, Е. М. Парилина, А. А. Седаков. — 3-е изд., стер. — Санкт-Петербург : Лань, 2022. — 152 с. — ISBN 978-5-8114-2229-6.

Изучение устройства и работы пространственной авторегрессионной модели была произведена по материалам Luc Anselin. Spatial Econometrics: Methods and Models, 1988. https://doi.org/10.1007/978-94-015-7799-1.

Данные для построения моделей были собраны с официальных источников Федеральной службы государственной статистики. [7]

Географические данные административных районов Российской Федерации были взяты из официальной международной базы данных глобальных административных районов (GADM). [17]

Глава 1. Данные и актуальность

1.1. Географическая информационная система (ГИС)

В 21 веке количество информации и ее доступность значительно повысилась, как никогда раньше, в том числе географических данных, то есть данные, которые имеют пространственную привязку. Использование данных ГИС существенно обогащает возможности статистического анализа, так как благодаря этому можно относительно учитывать взаимное расположение объектов.

Данные ГИС отражают расположение объектов внутри некоторой территории. Специализированные методы и модели анализа территориально-распределенной информации позволяют учитывать изменчивость изучаемого процесса по территории или взаимосвязь значений показателей для соседних объектов или смежных областей.

Попытаемся формализовать ранее описанное на языке математических терминов. Существование пространственной автокорреляции ясно характеризует простую географическую истину, а именно первый закон географии Уолда Тоблера: «… все имеет отношение ко всему, но ближние вещи влияют сильнее, чем отдаленные» [13], иначе говоря, все связано со всем, но близкорасположенные объекты связаны сильнее.

1.2. Факторы

В этом пункте опишем факторы, которые использовались в исследовании. За объясняющие переменные возьмем [5],[12]:

1. Валовой региональный продукт (VRP) – обобщающий показатель экономической деятельности региона, характеризующий процесс производства товаров и услуг для конечного использования. Единица измерения – млрд. рублей.
2. Население (PEOPLE) – совокупность людей, живущих в регионе. Единица измерения – человек.

3. Располагаемые ресурсы хозяйств (DOMESTIC_SOURCES) – это совокупность денежных доходов домохозяйств, сумм израсходованных накоплений и привлеченных (заемных) средств и стоимости натуральных поступлений. Единица измерения – рублей.

4. Уровень безработицы (UNM) – отношение численности безработных определенной возрастной группы к численности экономически активного населения соответствующей возрастной группы. Единица измерения – проценты.

5. Индекс производительности труда (INDWORK) – это процент динамик производительности труда в сравнении с базовым значением. Единица измерения – % к предыдущему году

6. Дефицит денежного дохода (DEFICIT_INCOME) – сумма денежных средств, необходимая для доведения доходов населения с денежными доходами ниже прожиточного минимума до величины прожиточного минимума. Единица измерения – млрд. рублей.

7. Год (YEAR) – наблюдаемому признаку может присутствовать временная изменчивость.

Все перечисленные факторы [5] могут быть хорошими кандидатами для включения в пространственно-регрессионную модель для объяснения среднедушевых денежных доходов населения. Каждый из этих факторов может оказывать влияние на доходы населения в разной степени, и включение их в модель может помочь понять, какие из этих факторов являются наиболее значимыми.

Валовой региональный продукт отражает общее экономическое состояние региона и уровень его развития, что, в свою очередь, может повлиять на уровень доходов населения.

Население является важным фактором, поскольку большее количество людей может означать большее количество рабочих мест и, следовательно, более высокие доходы населения.

Располагаемые ресурсы домашних хозяйств могут включать в себя доходы от всех источников, а также сбережения, кредиты и другие финансовые ресурсы, которые могут влиять на уровень доходов.

Уровень безработицы связан с уровнем доходов, так как более высокий уровень безработицы может означать меньшее количество рабочих мест и, следовательно, более низкие доходы населения.

Индекс производительности труда отражает общий уровень производительности в регионе и, следовательно, влияет на уровень доходов населения.

Дефицит денежного дохода может означать, что доходы населения недостаточны для покрытия жизненных расходов и могут отразить социальную и экономическую несправедливость в регионе.

Год учитывает временные изменения, такие как экономические циклы или изменения в законодательстве, которые влияют на доходы населения.

Глава 2. Основные понятия и определения

2.1 Матрица весов

Определение 1.1 Матрица весов W – это матрица размером N x N, где N число регионов. На пересечении i-ой строки и j-го столбца стоит вес связи между i-ым и j-ым регионом [1].

В данной работе рассмотрим построение бинарной и нормированной матрицы весов между регионами Российской Федерации по правилу ферзя.

По правилу ферзя соседними будут считаться все пары территориальных единиц, имеющие хотя бы одну общую точку на границе.

В бинарной матрице весов каждая строка соответствует объекту, а каждый столбец - соседнему объекту, и элементы матрицы принимают значение 1, если объекты являются соседними, и 0 в противном случае.

$$Wbin_{ij} = \begin{cases} 1, & \text{имеет общую точку на границе,} \\ 0, & \text{не имеет общей точки границы,} \end{cases}$$

Рассмотрим также построение нормированной матрицы весов. В ней матрица веса всех соседей нормируется на их количество. Если у региона есть в наличии три соседних региона, то значения в строке этого региона будут 0.33 в столбцах, с которыми регион граничит, а регионы, с которыми этот регион не граничит будет стоять 0.

$$W_{ij} = \begin{cases} \frac{1}{n_i}, & \text{имеет общую точку на границе} \\ 0, & \text{не имеет общей точки границы} \end{cases}$$

где $n_i -$ количество граничащих регионов для i объекта.

2.2. Пространственная автокорреляции

Как уже было пояснено ранее, пространственная автокорреляция является характеристикой связи между объектами.

Определение 2.1 Для множества S, состоящего из n географических единиц, *пространственная автокорреляция* есть соотношение между

переменной, наблюдаемой в каждой из n единиц и мерой географической близости, определенной для всех n*(n−1) пар единиц из S [15].

Определение 2.2 Пространственная автокорреляция является мерой того, в какой степени расположенные вблизи друг от друга объекты характеризуются тенденцией иметь сходные значения по некоторому показателю[14].

Конечной целью исследований пространственной автокорреляции является построение статистической модели зависимости значения показателя в каждой единице от значений в соседних единицах и (опционально) неких факторов. Наличие статистически значимой пространственной автокорреляции говорит о влиянии процессов, обуславливающих кластеризацию значений в соседних территориальных единицах.

Определение 2.3 Индекс Морана – это мера пространственной автокорреляции, основанная на расположении объектов и их значений, позволяющая выявлять региональные кластеры.

Значение индекса:

«+1» - означает определенную прямую зависимость, то есть наличие кластеров

«0» - абсолютно случайно распределение

«-1» - означает обратную зависимость (пример, шахматная доска)

Индекс Морана осуществляет анализ пространственной автокорреляции и вычисляется по формуле [17]:

$$I = \frac{n \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\left[\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}\right] \left[\sum_{i=1}^{n} (y_i - \bar{y})^2\right]}$$

2.3. Множественная линейная регрессия

Множественная линейная регрессия — это модель зависимости переменной Y от одной или нескольких других переменных с линейной функцией зависимости.

Пусть дан вектор $y = \{y_1, y_2, \dots, y_n\}$ измеряемой объясняемой переменной, а также пусть дана матрица X = $\{x_{ij}\}$ состоящая из m значений и n зависимых переменных.

Модель линейной регрессии можно записать в таком виде:

$$y = a_1x_1 + a_2x_2 + a_3x_3 + a_nx_n + \mathrm{b} + \mathrm{e},$$

a = $(a_1, a_2, \dots, a_n)$ - вектор коэффициента регрессии,

b – смещение,

e – вектор случайных ошибок, независимо распределенных относительно среднего значения в нуле.

Модель множественной линейной регрессии, построенная по данным, полученным из регионов, может не так хорошо описывать зависимость между переменными, так как остатки множественной линейной регрессии могут демонстрировать пространственную зависимость, как правило свидетельствующую о наличии дополнительных неучтённых факторов. Это означает, что обычная модель линейной регрессии недостаточно хорошо объясняет зависимость.

2.4. Пространственная авторегрессия

Чтобы усовершенствовать нашу модель и учесть пространственную автокорреляцию добавим в модель компоненту пространственной авторегрессии (RWy). Теперь модель имеет такой исходный вид и называется пространственной регрессией [15], [2]:

$$\mathrm{y} = \mathrm{Xa} + \mathrm{R}W\mathrm{y} + \mathrm{e}$$

R – коэффициент регрессии, данный коэффициент характеризует пространственную автокорреляцию.

W – матрица весов.

Для нахождения R и a проведем ряд преобразований:

$$\mathrm{y} = \mathrm{Xa} + \mathrm{RWy} + \mathrm{e},$$

$$\mathrm{y} - \mathrm{RWy} = \mathrm{Xa} + \mathrm{e},$$

$$(\mathrm{I} - \mathrm{RW})\mathrm{y} = \mathrm{Xa} + \mathrm{e},$$

Пусть $(\mathrm{I}-\mathrm{RW})$ обратима, тогда

$\mathrm{y}=(\mathrm{I}-\mathrm{RW})^{-1}\mathrm{Xa}+(\mathrm{I}-\mathrm{RW})^{-1}e$ - идентична линейной регрессии. Минимизировав квадрат случайной компоненты, найдем коэффициенты a и R, а сама случайная компонента может быть представлена как $\mathrm{e}=\mathrm{y}-X\mathrm{a}-RWy$ [1]:

$$\sum_i \left(y_i - \sum_j \mathrm{a}_j x_{ij} - \mathrm{R} \sum_j w_{ij} y_j \right)^2 \to \min_{\mathrm{a}_j, R} e^2$$

2.5. Верификация модели

Проверка качества оцененной множественной регрессионной модели проводится по следующим направлениям:

Коэффициент детерминации - оценка тесноты связи рассматриваемого набора факторов с исследуемым признаком.

Чем ближе к 1 индекс множественной корреляции, тем теснее связь результативного признака со всем набором исследуемых факторов.

$$R^2 = \frac{RSS}{TSS} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})}$$

Среднее абсолютное отклонение (MAD) – это статистический показатель, который характеризует разброс значений набора данных относительно их среднего значения. Оно представляет собой среднюю арифметическую величину расстояний между каждым значением набора данных и средним значением этого набора данных. MAD используется в статистических исследованиях для оценки степени изменчивости данных и для сравнения разных наборов данных.

$$MAD = \frac{1}{n}\sum_{i=1}^{n}|e_i|$$

Средняя ошибка аппроксимации – среднее отклонение расчетных значений от фактических.

$$A = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| * 100\%$$

Значение средней ошибки аппроксимации до 10% свидетельствует о хорошо подобранной модели уравнения.

Проверка выполнимости предпосылок МНК (Теорема Гаусса-Маркова).

Теорема Гаусса-Маркова — это основная теорема статистической регрессии, которая устанавливает оптимальные свойства оценок моделей линейной регрессии. Суть теоремы заключается в том, что при определенных предположениях на данные модели регрессии, оценки параметров этих моделей, полученные методами наименьших квадратов, обладает наименьшей дисперсией, и то, что ее математическое ожидание равное истинному значению.

1. Во всех наблюдениях среднее значение (математическое ожидание) случайного возмущения равно нулю или иначе ошибки ε должны быть независимыми, одинаково распределенными и иметь нулевое среднее значение:

$$E(\varepsilon) = 0.$$

2. Дисперсия ошибок ε должна быть постоянной для всех наблюдений:

$$Var(\varepsilon) = \sigma^2.$$

Это свойство получило название свойства гомоскедастичности или однородности. В случае невыполнения данного условия говорят, что случайные возмущения в уравнениях наблюдения гетероскедастичные или неоднородные. Если случайные возмущения гетероскедастичные, то оценки параметров модели остаются несмещенными, но теряется эффективность оценки дисперсий параметров.

3. Отсутствие автокорреляции - ошибки в разных наблюдениях не связаны между собой. Независимость случайных возмущений. Другими словами, каким бы не оказалось значение случайного возмущения в первом наблюдении, оно никак не сказывается на значениях случайного возмущения в любом другом наблюдении.

$$\mathrm{Cov}(\varepsilon_i; \varepsilon_j) = 0 \; \forall \; i \neq j$$

4. Мультиколлинеарность – это линейная зависимость между двумя или несколькими факторными переменными в уравнении множественной регрессии. Если такая зависимость является функциональной, то говорят о полной мультиколлинеарности. Простейшим методом устранения мультиколлинеарности является исключение из модели одной или ряда коррелированных переменных.

Для выбора лучшей модели введем Информационный критерий Акаике. Информационный критерий Акаике (AIC) – это критерий, применяющийся для выбора лучшей из нескольких статистических моделей.

В случае сравнения моделей на выборках одинаковой длины, имеем такое выражение:

$$AIC = 2k + n(\ln(RSS)),$$

k - количество параметров статистической модели,

n – число наблюдений.

Считается, что наилучшей будет модель с наименьшим значением критерия AIC.

Глава 3. Практическая реализация и анализ

3.1. Отбор данных и построения таблицы

Для построения необходимой таблицы и ее дальнейшего анализа отберем необходимые нам данные с официального сайт Федеральной службы государственной статистики [5].

Предобработаем данные в нужном формате, перед этим не включая в таблицу такие регионы как остров Сахалин, Калининградскую область, Республику Крым, город федерального значения Севастополь из-за отсутствия общей территориальной границы с другими регионами. Также не включим города федерального значения Санкт-Петербург и Москва из-за возможных выбросов которые могут сказаться на качестве модели. Добавив столбец NAME_1 для дальнейшего объединения с таблицей координат.

	name	NAME_1	year	mid_soul_income	vrp	domestic_sources	deficit_income	indwork	people	unm
1	Белгородская область	Belgorod	2014	24750.00	619677.7	21684.8	0.6	102.8	1546022	4.0
2	Белгородская область	Belgorod	2015	28043.24	693379.4	23863.3	0.7	103.2	1549037	4.1
3	Белгородская область	Belgorod	2016	29798.58	778027.8	23894.2	0.7	102.3	1551500	4.0
4	Белгородская область	Belgorod	2017	30342.12	837306.8	27841.2	0.6	103.1	1551370	3.9
5	Белгородская область	Belgorod	2018	30778.00	911597.9	27576.9	0.6	102.8	1548647	4.0
6	Белгородская область	Belgorod	2019	32398.00	955329.2	34497.1	0.7	102.2	1548284	3.9
7	Белгородская область	Belgorod	2020	32884.00	997330.9	28796.0	0.6	101.1	1545205	4.9
8	Белгородская область	Belgorod	2021	35612.00	1354810.5	31872.0	0.6	101.8	1536588	4.2
9	Брянская область	Bryansk	2014	20594.45	242722.4	15585.3	1.4	106.4	1237769	5.0
10	Брянская область	Bryansk	2015	23428.12	271782.5	16842.0	1.5	104.3	1229340	4.6
11	Брянская область	Bryansk	2016	24005.63	316489.4	16591.0	1.6	101.4	1223135	4.6
12	Брянская область	Bryansk	2017	25106.61	341177.8	16852.8	1.6	105.6	1215756	4.4
13	Брянская область	Bryansk	2018	26658.00	367157.1	19037.8	1.5	104.9	1205584	3.9
14	Брянская область	Bryansk	2019	28422.00	399113.8	20014.1	1.5	105.0	1196339	3.8
15	Брянская область	Bryansk	2020	28636.00	414179.4	20724.4	1.5	104.2	1187587	4.0
16	Брянская область	Bryansk	2021	31608.00	468666.2	24157.2	1.4	101.1	1175726	3.4

Рис. 1 Данные

Далее, загрузим данные глобальных административных районов Российской Федерации из международной базы данных (GADM) [17] в формате .json.

Предварительно, также исключив ранее перечисленные регионы.

3.2. Построение матрицы весов

Загрузив географические координаты регионов и при помощи функции poly2nb() из программного пакета “spdep” построим матрицу весов по правилу ферзя. Для визуализации соседства регионов построим граф

соседства. В качестве вершин графа возьмем географические центры регионов.

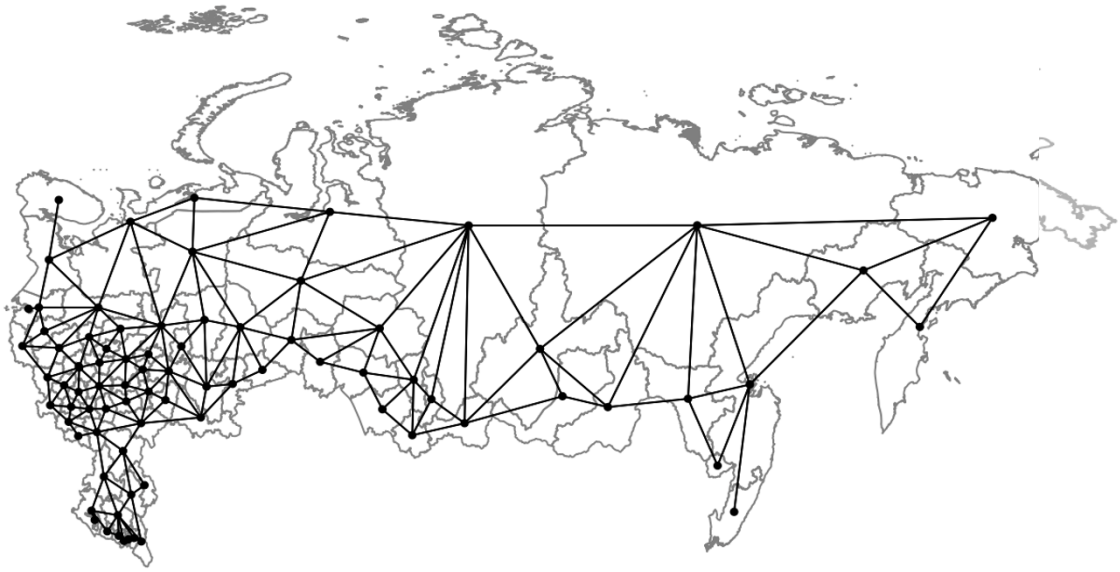

Рис. 2 Граф соседства по правилу ферзя

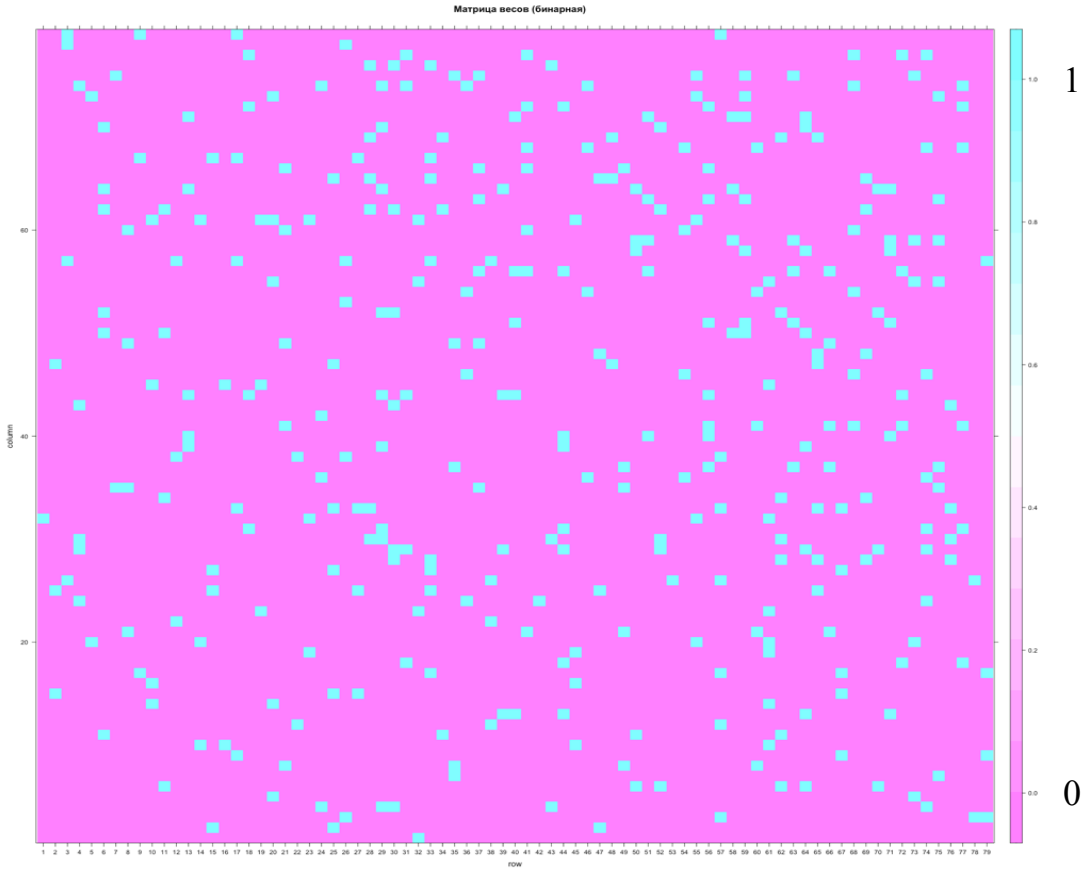

Рис. 3 Матрица весов

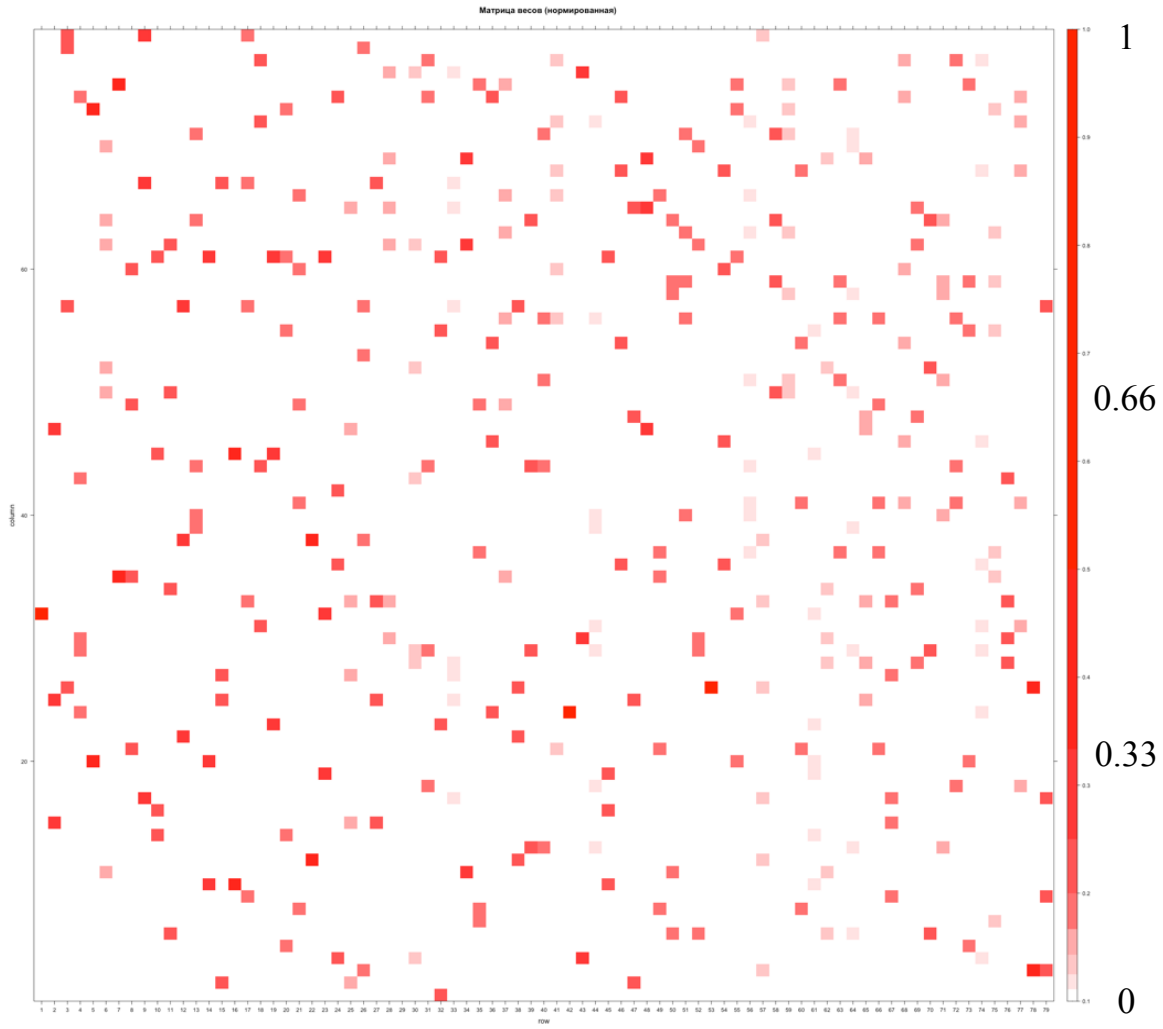

Рис. 4 Нормированная матрица весов

Объект списка соседей:

Количество регионов: 79

Количество ненулевых ссылок: 372

Процент ненулевых весов: 5,960583

Среднее количество ссылок: 4.708861

3.3. Анализ данных

Построение корреляционной матрицы из массива объясняемой и объясняющих переменных.

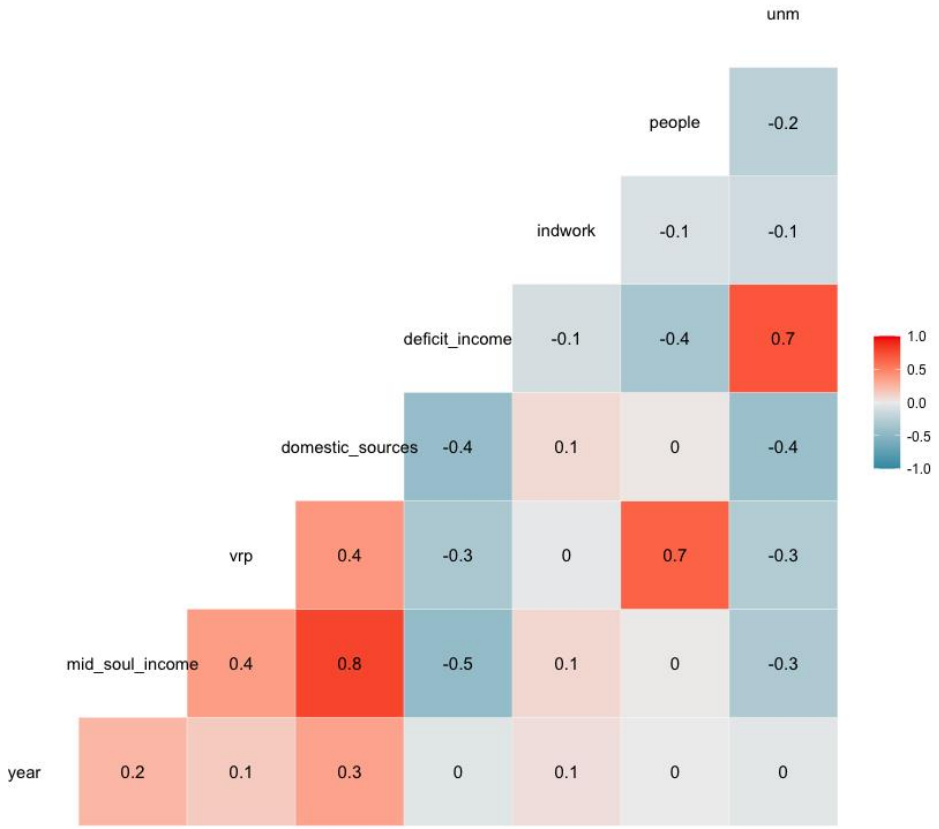

Рис. 5 Корреляционная матрица

Если значение корреляции между факторами равно или превышает по модулю 0.7, то исключим любой один из них.

Между такими факторами как Население и ВРП, Дефицит бюджета и Уровень безработицы значение корреляции по модулю = 0,7.

Исключим из нашей модели Население и Дефицит бюджета.

Проведем проверку на мультиколлинеарность.

Составим новую корреляционную матрицу Q.

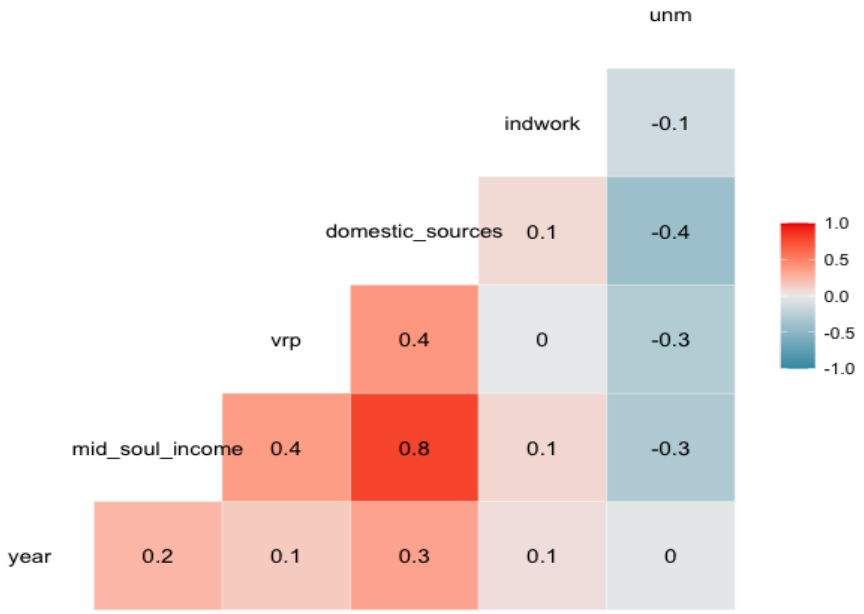

Рис. 6 Корреляционная матрица(обновленная)

Определитель матрицы Q = 0.2123487; не стремится к нулю, проверим дальше.

Вычислим Определитель матрицы (X^TX)= 0.5437161. Не стремится к нулю. Проведем проверку на наличие мультиколлинеарности с помощью VIF-критерия в следующей Таблице 1:

YEAR	VRP	DOMESTIC_SOURCES	INDWORK	UNM
1.151159	1.221576	1.515799	1.027153	1.285745

Таблица 1. VIF

Все значения статистик < 5, из этого следует что мультиколлинеарности нет. По результатам всех проверок мультиколлинеарности нет.

3.4. Построение моделей множественной линейной регрессии

Построив модель множественной линейной регрессии и получим следующие результаты (Таблица 2), также проверим гипотезу о значимости коэффициентов регрессии и значимости регрессии в целом на уровне значимости $\alpha = 0{,}05$:

МОДЕЛЬ 1.1

Коэффициенты	Оценка	P-value	Значимость
Свободный член	403,900	0.16776	Незначим
YEAR	-208.7	0.15152	Незначим
VRP	0.0007979	0.00685	Значим
DOMESTIC_SOURCES	1.24	< 2e-16	Значим
INDWORK	146.1	0.12186	Незначим
UNM	1370	0.13763	Незначим

Таблица 2. Коэффициенты модели 1.1

Из значения p-value: < 2.2e-16 (F-статистики) следует, что регрессия значима в целом.

Итоговая модель имеет следующий вид:

MID_SOUL_INCOME = 403,900 - 208.7YEAR + 0.0007979VRP + +1.24DOMESTIC_SOURCES + 146.1INDWORK+ 137UNM

В среднем, при увеличении на одну единицу значения любого фактора, за исключением переменной YEAR средние душевые денежные доходы будут увеличиваться.

Верификация модели:

1. $R^2 = 0.6308$
2. MAD = 4899.081
3. A = 15,23%

Для выполнения теоремы Гаусса-Маркова необходимо нормальное распределение остатков.

Проведем тест Шапиро-Уилка на нормальность распределения и построим гистрограмму остатков

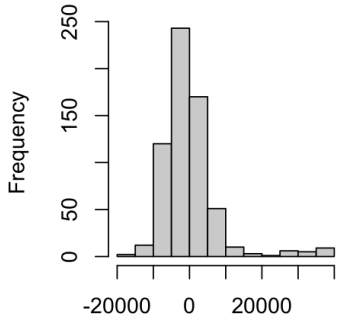

Рис. 7 Гистограмма остатков модели 1.1

Как уже можно заметить визуально на гистограмме есть правый «хвост», что демонстрирует не нормальное распределение остатков. P-value статистики Шапиро-Уилка : < 2.2e-16, из чего следует принять гипотезу о ненормальности распределения. Теорема Гаусса-Маркова не выполняется, из этого следует что оценки не являются лучшими в классе линейных несмещенных оценок.

Для обнаружения наличия автокорреляции проведем тест Дарбина-Уотсона, в тесте имеются следующие гипотезы:

H_0: между остатками нет корреляции,

H_A: остатки автокоррелированы.

Для определения нулевого математического ожидания воспользуемся критерием Стьюдента, предварительно выдвинув следующие гипотезы:

H_0: математическое ожидание $= 0$,

H_A: математическое ожидание $\neq 0$.

Для определения наличия гетероскедостичности проведем тест Гольдфреда-Квандта. Тест Гольдфреда-Квандта использует следующие гипотезы:

H_0: присутствует гомоскедастичность,

H_A: присутствует гетероскедастичность.

В Таблице 3 проведем тесты на гетероскедастичность, автокорреляцию и на нулевое математическое ожидание. Все тесты проведены на уровне значимости $\alpha = 0{,}05$

Тесты	Тест Дарбина-Уотсона	Тест на нулевое математическое ожидание	Тест Гольдфреда-Квандта
p-value	< 2.2e-16	0,98	4.34e-06
Выводы	Остатки автокоррелированы	Математическое ожидание $= 0$	Наличие гетероскедастичности

Таблица 3. Тесты модели 1.1

Исключив незначимые переменные, построим новую модель и проведем ее верификацию для оценки качества.

Обновленная модель имеет такой вид (Таблица 4), также проверим гипотезу о значимости коэффициентов регрессии и значимости регрессии в целом на уровне значимости $\alpha = 0{,}05$

МОДЕЛЬ 1.2

Коэффициенты	Оценка	P-value	Значимость
Свободный член	-400.4	0.690	Незначим
VRP	0.0007018	0.016	Значим
DOMESTIC_SOURCES	1.203	< 2e-16	Значим

Таблица 4. Коэффициенты модели 1.2

Из значения p-value: < 2.2e-16 (F-статистики) следует, что регрессия значима в целом.

Полученная модель:

MID_SOUL_INCOME= 400.4+0.0007018VRP+1.203DOMESTIC_SOURCES

В среднем, при увеличении на одну единицу значения любого фактора, средние душевые денежные доходы будут увеличиваться.

Верификация модели:

1. $R^2 = 0.6276$
2. MAD = 4868.624
3. A = 15,09%

Для выполнения теоремы Гаусса-Маркова необходимо нормальное распределение остатков.

Проведем тест Шапиро-Уилка на нормальность распределения и построим гистограмму остатков. Получим аналогичные результаты.

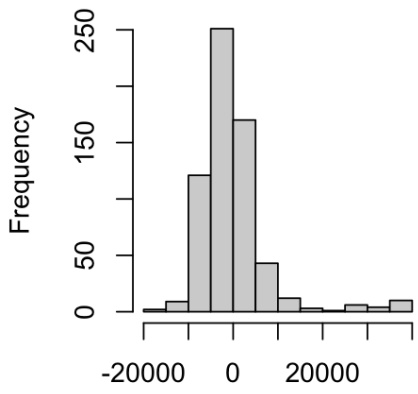

Рис. 8 Гистограмма остатков модели 1.2

P-value статистики Шапиро-Уилка: <2.2e-16, из чего следует принять гипотезу о ненормальности распределения. Теорема Гаусса-Маркова не выполняется, из этого следует что оценки не являются лучшими в классе линейных несмещенных оценок.

Дополнительно в Таблице 5 проведем тесты на гетероскедастичность, автокорреляцию и на нулевое математическое ожидание. Все тесты проведены на уровне значимости $\alpha = 0{,}05$

Тесты	Тест Дарбина-Уотсона	Тест на нулевое математическое ожидание	Тест Гольдфреда-Квандта
p-value	< 2.2e-16	0,96	5.875e-06
Выводы	Остатки автокоррелированы	Математическое ожидание = 0	Наличие гетероскедастичности

Таблица 5. Тесты модели 1.2

3.5. Визуализация данных

Визуализация данных была проведена посредством программных пакетов языка R.

Была построена серия карт, получившая схожую раскраску, и следовательно разбиение на кластеры.

Визуализируем по субъектам Российской Федерации среднедушевые денежные доходы населения за 2021 год:

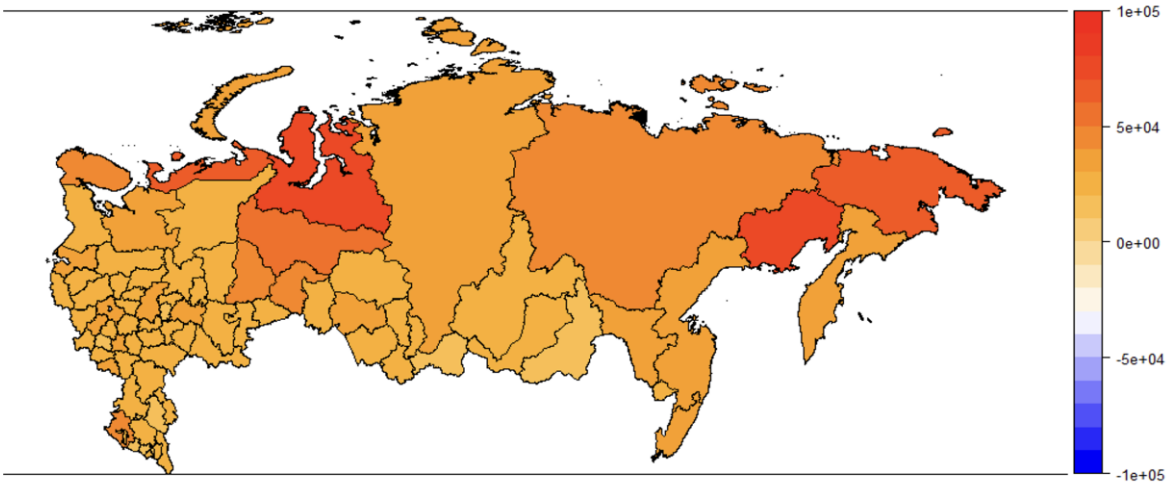

Рис. 9 Фактические данные MID_SOUL_INCOME за 2021

Визуально, можно заметить некий кластеры на дальнем востоке и районах крайнего севера, также между Центральным и Северо-Западным округом.

Проведем тест Морана для определения наличия пространственной автокорреляции и получим следующие результаты на данном уровне значимости:

Нулевая гипотеза (H_0): данные разбросаны случайным образом.

Альтернативная гипотеза (H_A): данные не рассредоточены случайным образом, т. е. они сгруппированы по заметным шаблонам.

Уровень значимости $\alpha = 0{,}05$

Матрица весов	P-value	Наличие пространственной автокорреляции(принимаемая гипотеза)	Значение статистики Морана
Нормированная W	<2e-16	Данные сгруппированы по кластерам	0.5370184
Бинарная Wbin	<2e-16	Данные сгруппированы по кластерам	0.4375301

Таблица 6. Индекс Морана

На данном уровне значимости гипотезу H_0 не имеем оснований опровергнуть для построенных матриц весов и значений среднедушевых денежных доходов.

3.6. Построение пространственной авторегрессионной модели с нормированной матрицей весов

В дальнеших построениях моделей будет использваться нормированная матрица весов W.

Построим пространственную авторегрессионную модель и получим следующие результаты (Таблица 7), также проверим гипотезу о значимости коэффициентов регрессии и значимости регрессии в целом на уровне значимости $\alpha = 0{,}05$:

МОДЕЛЬ 2.1

Коэффициенты	Оценка	P-value	Значимость

Параметр авторегрессии	0.99962	< 2.2e-16	Значим
YEAR	517.0	1.523e-06	Значим
VRP	0.0021452	< 2.2e-16	Значим
DOMESTIC_SOURCES	0.51714	< 2.2e-16	Значим
INDWORK	51.217	0.4447995	Незначим
UNM	-406.15	0.0002873	Значим

Таблица 7. Коэффициенты модели 2.1

Из значения p-value: < 2.2e-16 (F-статистики) следует, что регрессия значима в целом.

Полученная модель:

MID_SOUL_INCOME=517YEAR+0.0021452VRP+

+0.51714DOMESTIC_SOURCES+51.217INDWORK -406.15UNM + 0.99962W

В среднем, при увеличении на одну единицу значения любого фактора, за исключением переменной UNM, средние душевые денежные доходы будут увеличиваться, а с увеличением числа безработных следует то, что это способствует снижению средних душевых доходов.

Верификация модели:

1. $R^2 = 0.6552737$
2. MAD = 3249.096
3. A = 9,72%

Для выполнения теоремы Гаусса-Маркова необходимо нормальное распределение остатков. Проведем тест Шапиро-Уилка на нормальность распределения и построим гистограмму остатков. Получим аналогичные результаты.

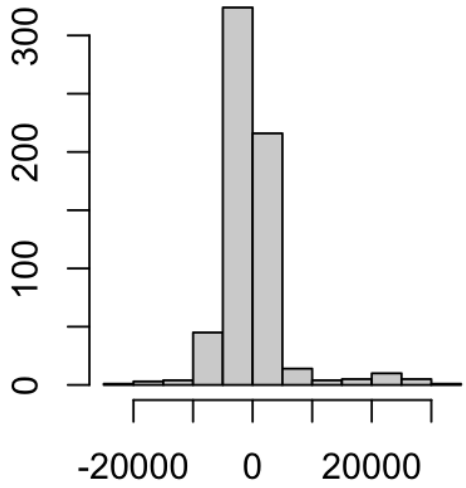

Рис. 10 Гистограмма остатков модели 2.1

P-value статистики Шапиро-Уилка : < 2.2e-16, из чего следует принять гипотезу о ненормальности распределения. Теорема Гаусса-Маркова не выполняется, из этого следует что оценки не являются лучшими в классе линейных несмещенных оценок.

Дополнительно в Таблице 8 проведем тесты на гетероскедастичность, автокорреляцию и на нулевое математическое ожидание. Все тесты проведены на уровне значимости $\alpha = 0{,}05$. Стоит отметить, что многие тесты на гетероскедастичность не точно дают оценку для пространственных моделей, и самым эффективных среди всех тестов является тест Глейзера [2]. Тест Глейзера использует следующие гипотезы:

H_0: присутствует гомоскедастичность,

H_A: присутствует гетероскедастичность.

Тесты	Тест Дарбина-Уотсона	Тест на нулевое математическое ожидание	Тест Глейзера
p-value	<2.2e-16	0.5808	2.2e-16
Выводы	Остатки автокоррелированы	Математическое ожидание = 0	Наличие гетероскедастичности

Таблица 8. Тесты модели 2.1

Исключив незначимые переменные, построим новую модель и проведем ее верификацию для оценки качества.

Построим обновленную модель (Таблица 9), также проверим гипотезу о значимости коэффициентов регрессии и значимости регрессии в целом на уровне значимости $\alpha = 0{,}05$.

МОДЕЛЬ 2.2

Коэффициенты	Оценка	P-value	Значимость
Параметр авторегрессии	0.99961	< 2.2e-16	Значим
YEAR	523.58	1.034e-06	Значим
VRP	0.0021538	< 2.2e-16	Значим
DOMESTIC_SOURCES	0.51431	< 2.2e-16	Значим
UNM	-414.62	0.0002003	Значим

Таблица 9. Коэффициенты модели 2.2

Из значения p-value: < 2.2e-16 (F-статистики) следует, что регрессия значима в целом.

Итоговая модель имеет следующий вид:

MID_SOUL_INCOME=523.58YEAR+0.0021538VRP+

+0.51431DOMESTIC_SOURCES – 414.62UNM +0.99961W

В среднем, при увеличении на одну единицу значения любого фактора, за исключением переменной UNM, средние душевые денежные доходы будут увеличиваться, а с увеличением числа безработных следует то, что это способствует снижению средних душевых доходов.

Верификация модели:

1. $R^2 = 0.6548562$
2. MAD = 3246.732
3. A = 9,7%

Для выполнения теоремы Гаусса-Маркова необходимо нормальное распределение остатков. Проведем тест Шапиро-Уилка на нормальность

распределения и построим гистограмму остатков. Получим аналогичные результаты.

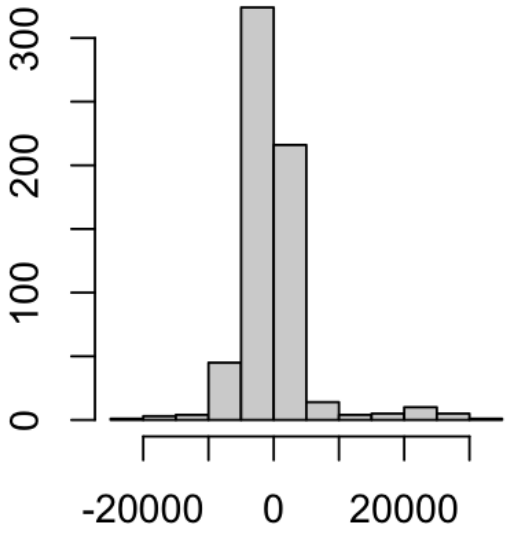

Рис. 11 Гистограмма остатков модели 2.2

P-value статистики Шапиро-Уилка: < 2.2e-16, из чего следует принять гипотезу о ненормальности распределения. Теорема Гаусса-Маркова не выполняется, из этого следует что оценки не являются лучшими в классе линейных несмещенных оценок.

Дополнительно в Таблице 10 проведем тесты на гетероскедастичность, автокорреляцию и на нулевое математическое ожидание. Все тесты проведены на уровне значимости $\alpha = 0{,}05$.

Тесты	Тест Дарбина-Уотсона	Тест на нулевое математическое ожидание	Тест Глейзера
p-value	< 2.2e-16	0.5843	< 2.2e-16
Выводы	Остатки автокоррелированы	Математическое ожидание $= 0$	Наличие гетероскедастичности

Таблица 10. Тесты модели 2.2

3.7. Визуализация модели 2.2

Визуализируем результаты за 2021 год полученной модели 2.2

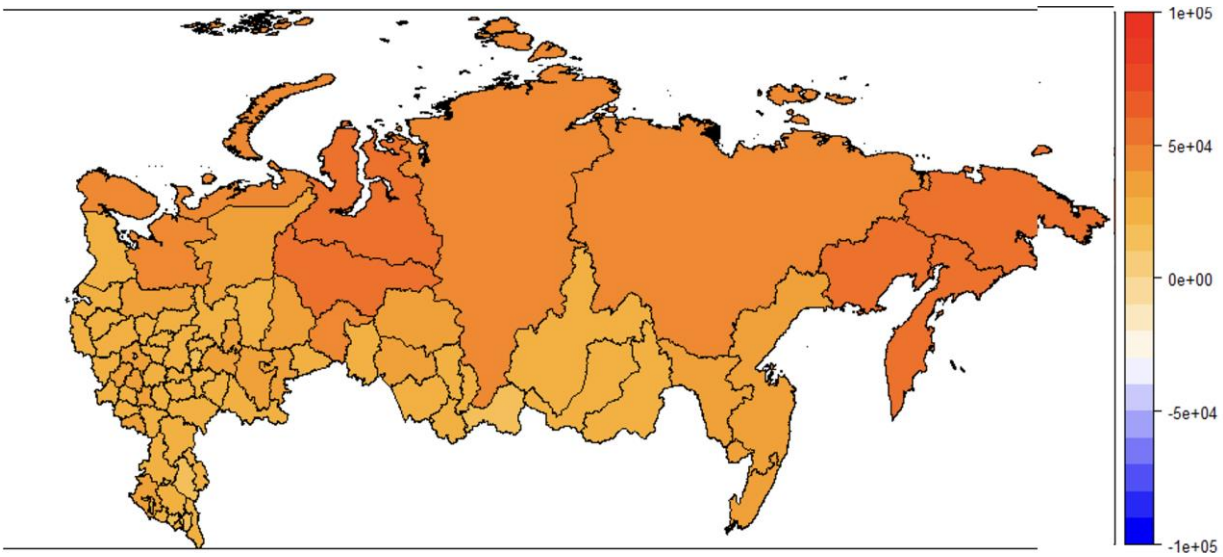

Рис. 12 Предсказанные значения модели 2.2 за 2021 год

Визуально заметно цветовое отличие от рис. 6, цвета карты у некоторых регионов стали бледнее, как например Ямало-Ненецкий авт. округ, Чукотский авт. округ. Наоборот, некоторые регионы стали на тон темнее, такие как, Камчатский край, Краснодарский край и другие.

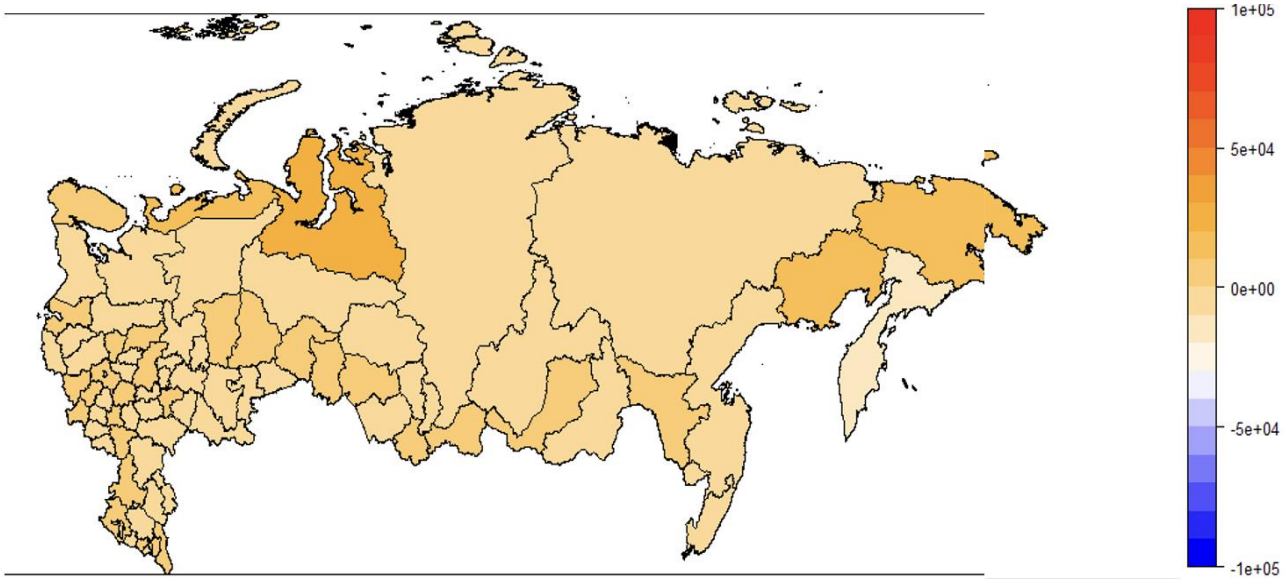

Рис. 13 Остатки модели 2.2 за 2021 год

3.8. Построение пространственной авторегрессионной модели с бинарной матрицей весов

В дальнеших построениях моделей будет использваться бинарная матрица весов Wbin.

Построим пространственную авторегрессионную модель с обновленными весами получим следующие результаты (Таблица 11), также проверим гипотезу о значимости коэффициентов регрессии и значимости регрессии в целом на уровне значимости $\alpha = 0{,}05$.

МОДЕЛЬ 3.1

Коэффициенты	Оценка	P-value	Значимость
Параметр авторегрессии	0.018661	< 2.2e-16	Значим
YEAR	3.0856	0.45314	Незначим
VRP	0.0019378	1.297e-11	Значим
DOMESTIC_SOURCES	0.87932	< 2.2e-16	Значим
INDWORK	50.205	0.51399	Незначим
UNM	-173.92	0.09657	Незначим

Таблица 11. Коэффициенты модели 3.1

Из значения p-value: < 2.2e-16 (F-статистики) следует, что регрессия значима в целом.

Итоговая модель имеет следующий вид:

MID_SOUL_INCOME=3.0856YEAR+0.0019378VRP+
+0.87932DOMESTIC_SOURCES+50.205INDWORK-
173.92UNM+0.018661Wbin

В среднем, при увеличении на одну единицу значения любого фактора, за исключением переменной UNM, средние душевые денежные доходы будут увеличиваться, а с увеличением числа безработных следует то, что это способствует снижению средних душевых доходов.

Верификация модели:

1. $R^2 = 0.6182284$
2. MAD = 3956.515
3. A = 12%

Для выполнения теоремы Гаусса-Маркова необходимо нормальное распределение остатков. Проведем тест Шапиро-Уилка на нормальность распределения и построим гистограмму остатков. Получим аналогичные результаты.

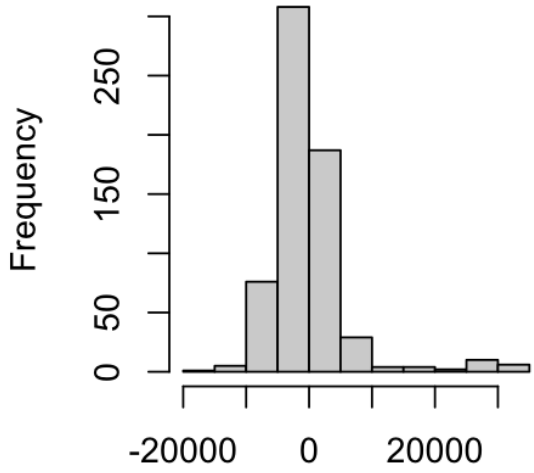

Рис. 14 Гистограмма остатков модели 3.1

P-value статистики Шапиро-Уилка: < 2.2e-16, из чего следует принять гипотезу о ненормальности распределения. Теорема Гаусса-Маркова не выполняется, из этого следует что оценки не являются лучшими в классе линейных несмещенных оценок.

линейных несмещенных оценок.

Дополнительно в Таблице 12 проведем тесты на гетероскедастичность, автокорреляцию и на нулевое математическое ожидание. Все тесты проведены на уровне значимости $\alpha = 0{,}05$.

Тесты	Тест Дарбина-Уотсона	Тест на нулевое математическое ожидание	Тест Глейзера
p-value	< 2.2e-16	0.5408	2.409e-08
Выводы	Остатки автокоррелированы	Математическое ожидание = 0	Наличие гетероскедастичности

Таблица 12. Тесты модели 3.1

Исключив незначимые переменные, построим новую модель (Таблица 13) и проведем ее верификацию для оценки качества, также проверим гипотезу о значимости коэффициентов регрессии и значимости регрессии в целом на уровне значимости $\alpha = 0{,}05$

МОДЕЛЬ 3.2

Коэффициенты	Оценка	P-value	Значимость
Свободный член	0.018496	< 2.22e-16	Значим
VRP	0.00136497	8.664e-07	Значим
DOMESTIC_SOURCES	1.16886761	< 2e-16	Значим

Таблица 13. Коэффициенты регрессии модели 3.2.

Из значения p-value: < 2.2e-16 (F-статистики) следует, что регрессия значима в целом.

Итоговая модель имеет следующий вид:

MID_SOUL_INCOME=0.00136497VRP+1.16886761DOMESTIC_SOURCES+
+0.018496Wbin

В среднем, при увеличении на одну единицу значения любого фактора, средние душевые денежные доходы будут увеличиваться.

Верификация модели:

1. $R^2 = 0.735895$
2. MAD = 4122.896
3. A = 12.53%

Для выполнения теоремы Гаусса-Маркова необходимо нормальное распределение остатков. Проведем тест Шапиро-Уилка на нормальность распределения и построим гистограмму остатков. Получим аналогичные результаты.

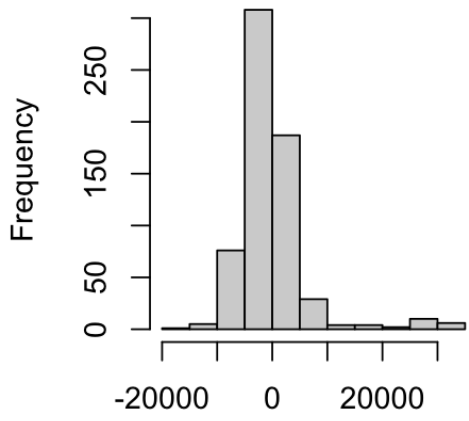

Рис. 15 Гистограмма остатков модели 3.2

P-value статистики Шапиро-Уилка: < 2.2e-16, из чего следует принять гипотезу о ненормальности распределения. Теорема Гаусса-Маркова не выполняется, из этого следует что оценки не являются лучшими в классе линейных несмещенных оценок.

линейных несмещенных оценок.

Дополнительно в Таблице 14 проведем тесты на гетероскедастичность, автокорреляцию и на нулевое математическое ожидание. Все тесты проведены на уровне значимости $\alpha = 0{,}05$.

Тесты	Тест Дарбина-Уотсона	Тест на нулевое математическое ожидание	Тест Глейзера
p-value	< 2.2e-16	0,192	< 2.2e-16
Выводы	Остатки автокоррелированы	Математическое ожидание $= 0$	Наличие гетероскедастичности

Таблица 14. Тесты модели 3.2

Выводы

Построим таблицу (Таблица 15) со значениями верификации каждой модели и введем столбец с информационным критерием Акаике:

Модель	R^2	MAD	A	AIC
Модель 1.1	0.6308	4899.081	15,23%	13130
Модель 1.2	0.6276	4868.624	15,09%	13129
Модель 2.1	0.6552737	3249.096	9,72%	12769
Модель 2.2	0.6548562	3246.732	9,7%	12767
Модель 3.1	0.6382284	3956.515	12%	12901
Модель 3.2	0.735895	4122.896	12.53%	12938

Таблица 15. Модели

Из таблицы видно, что пространственная авторегрессионая модель лучше описывает среднедушевые денежные доходы во всех способах оценки, чем множественная пространственная регрессия. Также важно заметить, что выбор построения весов играет важную роль в верификации модели. Среди двух пространственных авторегрессионных моделей лучшие результаты показывает модель с нормированной матрицей весов. Стоит отметить, что из-за возможных выбросов и корреляции остатков, остатки моделей не распределены нормально, из чего следует невыполнение теоремы Гаусса-Маркова, то есть, что оценки не являются лучшими в классе линейных несмещенных оценок. Тем не менее, модели показывают достаточно неплохие результаты. Средняя ошибка аппроксимации не выходит из интервала 9-16%, а значение R^2 колеблется на значении ~ 0,65.

По оценке R^2 лучшая модель является модель 3.2, по всем остальным оценкам лучшей оказалась модель 2.2, и которая является лучшей по информационному критерию Акаике.

Визуализируем фактические и спрогнозированные значения модели 2.2 в Приложении.

Заключение

В ходе исследования пространственной зависимости среднедушевых денежных доходов были выполнены все поставленные задачи.

Изучена необходимая теоретическая база для построения описанных моделей, способы построения весов, оценка параметров регрессии и способы оценки качества модели.

Были собраны и обработаны все необходимые данные для объясняющей и объясняемой переменных с официальных источников (Росстат), также были собраны географические данные административных регионов Российской Федерации, были исключены регионы, не имеющие территориальных границ с регионами и возможными выбросами.

Исследовано наличие пространственной автокорреляции между регионами по средствам индекса Морана. Построены 6 моделей (2 модели множественной регрессии и 4 модели пространственной авторегрессии) с двумя видами построения матрицы весов (бинарная и нормированная). Произведя оценку и верификацию моделей, а затем интерпретацию результатов было выяснено, что данные, имеющие пространственную привязку и зависимость лучше описывается пространственной авторегрессионной моделью, на примере данного исследования лучшей из построенных моделей оказалась модель 2.2:

MID_SOUL_INCOME=523.58YEAR+0.0021538VRP+
+0.51431DOMESTIC_SOURCES - 414.62UNM +0.99961W

Список литературы

3. Самсонов Т.Е. Визуализация и анализ географ. данных на языке R. Географический факультет МГУ, 2017. DOI: 10.5281/zenodo.901911.

4. Luc Anselin. Spatial Econometrics: Methods and Models, 1988. https://doi.org/10.1007/978-94-015-7799-1

5. Теория вероятностей и математическая статистика / В. М. Буре, Е. М. Парилина. — Санкт-Петербург: Лань, 2013. — 416 с. — ISBN 978-5-8114-1508

6. Методы прикладной статистики в R и Excel / В. М. Буре, Е. М. Парилина, А. А. Седаков. — 3-е изд., стер. — Санкт-Петербург: Лань, 2022. — 152 с. — ISBN 978-5-8114-2229-6.

7. Официальный сайт Федеральная служба государственной статистики. https://rosstat.gov.ru.

8. Демидова О. А. Пространственная авторегрессионная модель для двух групп взаимосвязанных регионов (на примере Восточной и Западной части России). Журнал «Прикладная эконометрика» №34(2) 2014 год.

9. Айвазян, С. А. Методы эконометрики; Московская школа экономики МГУ им. М.В. Ломоносова (МШЭ). — Москва: Магистр: ИНФРА-М, 2020. — 512 с. - ISBN 978-5-9776-0153-5.

10. Вакуленко Е. С. Эконометрический анализ факторов внутренней миграции в России. Журнал «Прикладная эконометрика» №25(1) 2012 год.

11. Ратникова Т.А. Лекции введение в эконометрических анализ панельных данных.

12. М.И. Баканов, М.В. Мельник, А.Д. Шеремет; под ред. М.И. Баканова. — 5-е изд., перераб. и доп. — М.: Финансы и статистика, 2007. — 536с.

13. Доугерти К. Д 71 Введение в эконометрику: Пер. с англ. — М.: ИНФРА-М, 1999. — XIV, 402 с.Абель Э., Бернанке Б. Макроэкономика

14. Tobler W., (1970) "A computer movie simulating urban growth in the Detroit region". Economic Geography, 46(Supplement): 234–240.

15. Григорьев А. А. Прстранственная автокорреляция образовательный достижений в Российской Федерации 2018. Т. №1. С. 164-173 DOI: 10.17323/1813-8918-2018-1-164-173

16. Hubert, L. J., R. G. Golledge, and C. M. Costanza (1981). Generalized © Procedures for Evaluating Spatial Autocorrelation. Geographical Analysis 13, 224-32. DOl: 10.1111/j.1538-4632. 1981.tb00731.

17. Официальный сайт базы данных глобальных административных районов. https://gadm.org/maps/RUS_1.html

18. Moran, PAP (1950), «Заметки о непрерывных стохастических явлениях», *Biometrika* , **37** , 17–33. DOI : 10.1093 / biomet / 37.1-2.17 JSTOR : 2332142

Приложение

График MID_SOUL_INCOME и предсказанные значения модели 2.2, (значения отсортированы):

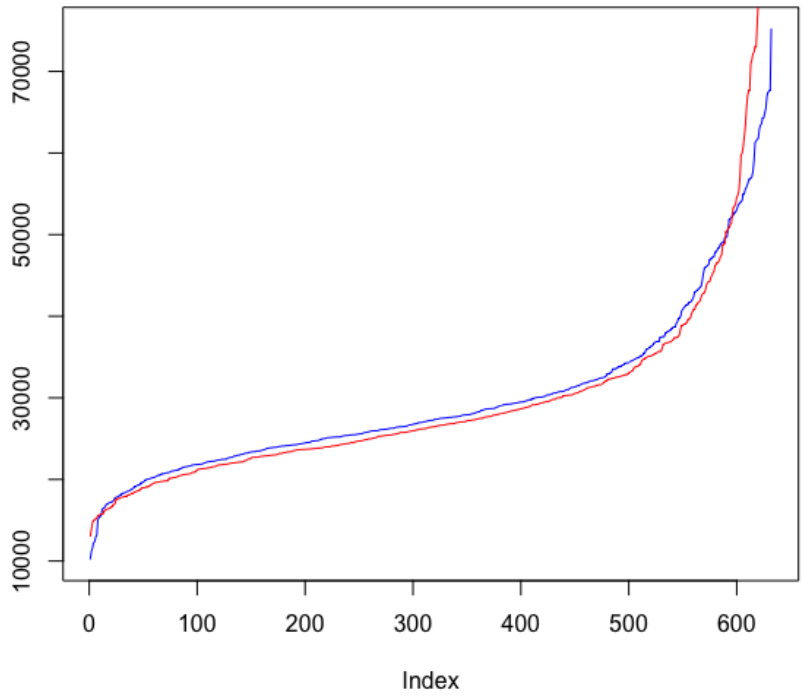

Рис. 16 MID_SOUL_INCOME (красный) и модель 2.2(синий)

График MID_SOUL_INCOME и предсказанные значения модели 2.2

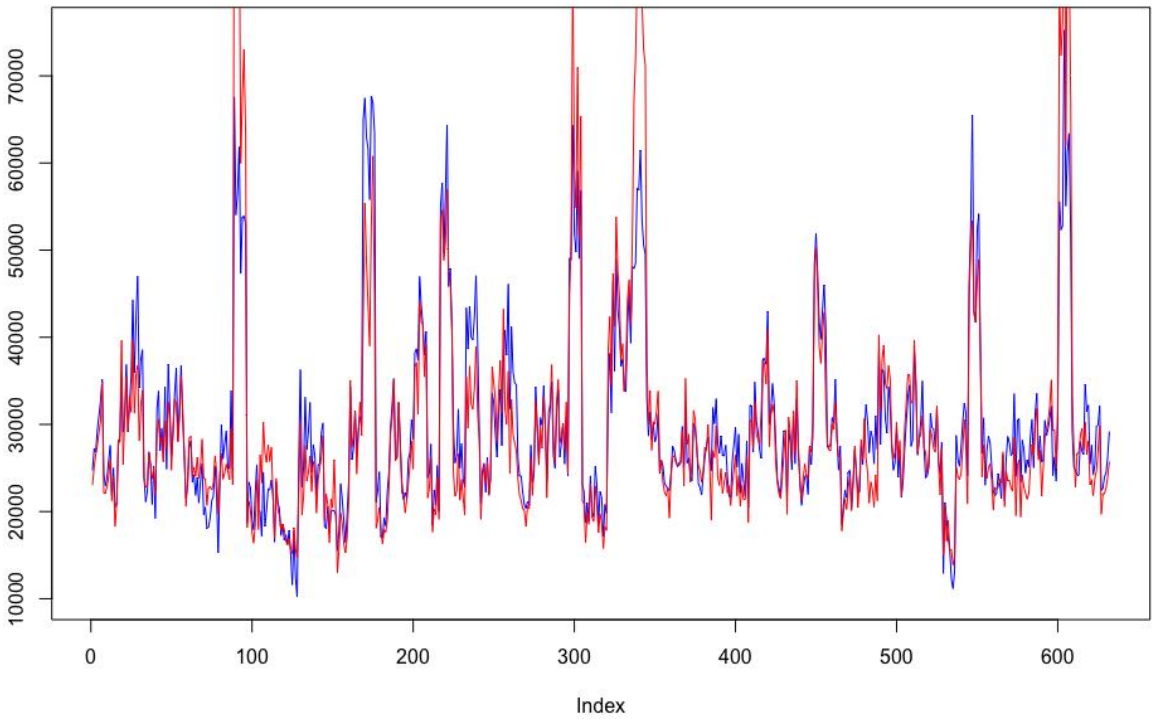

Рис. 17 MID_SOUL_INCOME (красный) и модель 2.2(синий)