

МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования

**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»**

Кафедра теории функций и стохастического анализа

**ПРОСТРАНСТВЕННОЕ МОДЕЛИРОВАНИЕ В
РЕГИОНАЛЬНЫХ ИССЛЕДОВАНИЯХ НА ПРИМЕРЕ
АНАЛИЗА УРОВНЯ БЕЗРАБОТИЦЫ В ВОСТОЧНОЙ И
ЗАПАДНОЙ ЧАСТИ РОССИИ**

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

Студентки 4 курса 451 группы
направления 38.03.05 — Бизнес-информатика

механико-математического факультета
Семеновой Веры Михайловны

Научный руководитель
доцент, к. э. н.

А. Р. Файзлиев

Заведующий кафедрой
д. ф.-м. н., доцент

С. П. Сидоров

Саратов 2022

ВВЕДЕНИЕ

Актуальность темы. Пространственный анализ является актуальным научным направлением и имеет большое значение для социального и экономического развития регионов Российской Федерации, а также зарубежных стран.

Целью данной работы является разработка программного кода для выявления роли пространственной автокорреляции в моделировании на примере уровня безработицы.

Для достижения цели были поставлены следующие задачи:

- Изучить факторы, влияющие на уровень безработицы.
- Собрать данные по выбранным факторам.
- Построить корреляционную матрицу для данных статистических факторов.
- Выбрать зависимые переменные.
- Выбрать матрицу пространственных весов.
- Провести тест Морана на наличие автокорреляции.
- Построить диаграмму рассеивания.
- Построить пространственную модель.
- Истолковать результаты.

Структура и содержание бакалаврской работы. Работа состоит из введения, четырех разделов, заключения, списка использованных источников, содержащего 20 наименований и двух приложений.

В **первом** разделе аргументируется выбор языка программирования, а также описываются используемые библиотеки.

Во **втором** разделе приводится описание данных, используемых для исследования, их отбор и картографическая визуализация.

Третий раздел посвящен таким аспектам пространственной эконометрики, как пространственная матрица весов, тест Морана, диаграмма рассеивания и пространственные модели.

В **четвертом** разделе представлены результаты построения пространственных моделей, а также их интерпритация.

Основное содержание работы

Выбор языка программирования. Для написания кода был выбран язык программирования R.

R — язык, позволяющий пользователю использовать операторы циклов, чтобы последовательно анализировать несколько наборов данных. Также язык R предоставляет возможность объединить в отдельную программу различные статистические функции для проведения более сложного анализа.

Достоинства среды R:

- является бесплатной и кроссплатформенной;
- имеет богатый арсенал статистических методов;
- обладает качественной векторной графикой;
- насчитывает более 7000 проверенных пакетов;
- является гибкой в использовании:
 - позволяет создавать и редактировать скрипты и пакеты,
 - взаимодействует с другими языками, например, с C, Java и Python,
 - может работать с форматами данных для SAS, SPSS и STATA;
- объединяет активное сообщество пользователей и разработчиков;
- поддерживает регулярные обновления и тех. Поддержку;
- обрабатывает большое количество данных (до 1 Гб);
- дает возможность использовать ее как программу (на основе готовых пакетов);
- позволяет создавать таблицы произвольной структуры;
- параллельные вычисления (пакеты `foreach` и `doParallel`)

В таблице 1 представлены все используемые в коде библиотеки.

Таблица 1 – Используемые библиотеки

Название библиотеки	Описание
spdep	Набор функций для создания объектов матрицы пространственных весов из многоугольников, набор тестов для пространственной автокорреляции
maptools	Инструменты для обработки пространственных объектов

Таблица 1 – Используемые библиотеки

Название библиотеки	Описание
leaflet	Позволяет легко создавать карты
RColorBrewer	Предоставляет цветовые схемы для карт (и другой графики)

Название библиотеки	Описание
gstat	Пространственное и пространственно-временное геостатистическое моделирование, прогнозирование и моделирование
RANN	Пакет R, обеспечивающий быстрый поиск ближайших соседей
rgdal	Привязки для библиотеки абстракции геопространственных данных
varhandle	Функции для более эффективной работы с переменными
geosphere	Вычисляет расстояния и связанные с ними меры для углообразующих местоположений
igraph	Сетевой анализ и визуализация
dils	Объединение сетей с несколькими отношениями в единую взвешенную сеть
splm	Эконометрические модели для пространственных панельных данных
plm	Линейные модели для панельных данных
rlang	Набор инструментов для работы с базовыми типами, основными функциями R, такими как система условий, и основными функциями Tidyverse, такими как tidy evaluation.
readxl	Чтение файлов Excel

Название библиотеки	Описание
reshape2	Пакет для изменения формы данных

Данные. Для оценки моделей использовались данные различных экономических и демографических показателей о 78 регионах России за 2016 - 2019 год. Данные о регионах обеспечены Федеральной службой государственной статистики. Из рассмотрения исключены Ненецкий автономный округ, Ямало-Ненецкий автономный округ, Ханты-Мансийский автономный округ - Югра, Чукотский автономный округ и Республика Северная Осетия-Алания из-за неполноты данных по ним.

Перечень объясняющих переменных представлен в таблице 3.

Таблица 3 – Объясняющие переменные

Переменные	Описание
Высшее образование	Доля занятых с высшим образованием, %
Городское население	Доля городского населения в общей численности населения, %
Миграция	Коэффициент миграционного прироста, чел. на 10 тыс. чел. населения
Доступность жилья	Отношение средних цен на вторичном рынке жилья (руб./м ²) на среднедушевые денежные доходы населения (руб./месяц)
Валовой региональный продукт	ВРП на душу населения (руб)
Плотность	Плотность населения, чел./м ²
Зарботная плата	Среднемесячная начисленная заработная плата работников организаций (руб.)

Корреляционная матрица. Корреляционная матрица показывает зависимость величин друг от друга. Элементами корреляционной матрицы являются коэффициенты корреляции величин.

В данной работе матрица строилась с помощью функции `cor()`.

```
df=ru.poly@data[,c("urbpopulation2019","higheduc2019","vrp2019","migrpr2019",
"houseaff2019","density2019","unemp2019")]
cor(df)
```

На вход функции подается фрейм данных `df`, содержащий выбранные объясняющие переменные, далее функция вычисляет коэффициенты корреляции между данными в фрейме и на выходе получается таблица 4, представленная ниже.

Таблица 4 – Корреляционная матрица для данных за 2019 год

	Городское население	Высшее образование	Валовый региональный продукт	Миграционный прирост	Доступность жилья	Плотность населения	Уровень безработицы
Городское население	1.000	-0.061	0.503	-0.083	-0.234	0.351	-0.589
Высшее образование	-0.0619	1.000	0.135	0.341	-0.101	0.452	0.068
Валовый региональный продукт	0.503	0.135	1.000	-0.016	-0.173	0.294	-0.306
Миграционный прирост	-0.083	0.341	-0.016	1.000	0.002	0.112	-0.067
Доступность жилья	-0.234	-0.101	-0.173	0.002	1.000	0.249	0.291
Плотность населения	0.351	0.452	0.294	0.112	0.249	1.000	-0.194
Уровень безработицы	-0.589	0.068	-0.306	-0.067	0.291	-0.194	1.000

Глядя на таблицу ??, можно заметить, что некоторые объясняющие переменные имеют корреляционную зависимость. Например, процент городского населения коррелирует с плотностью, из-за чего придется исключить один из показателей из модели.

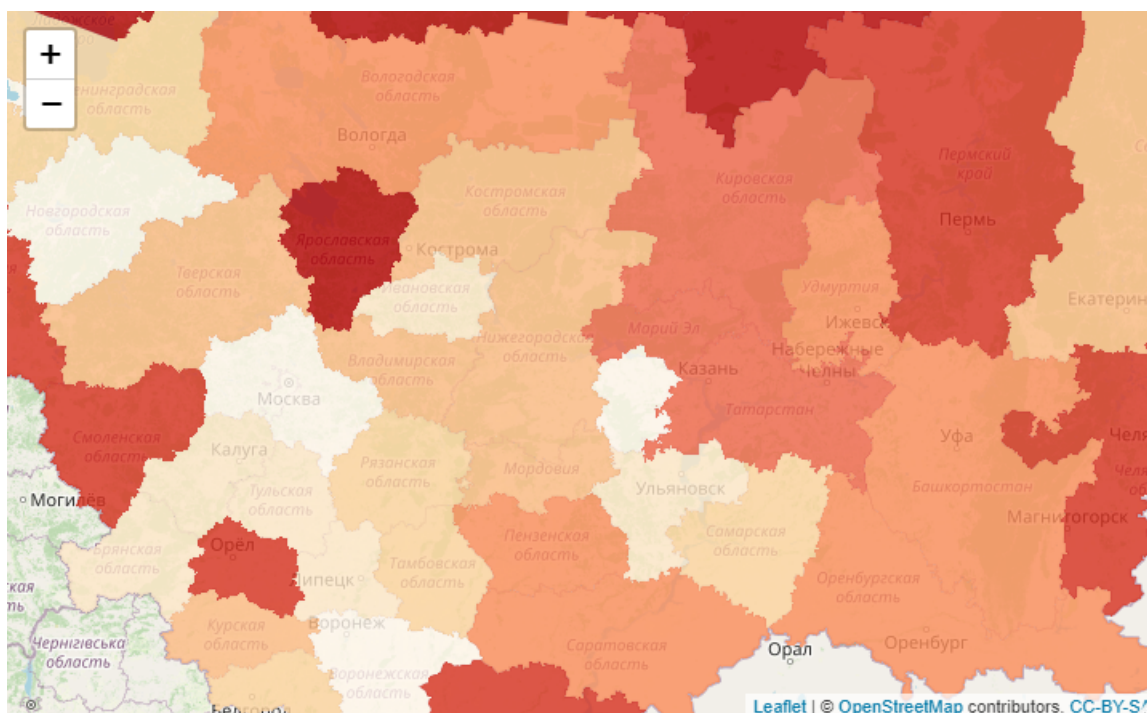
Исключая таким образом коррелирующие переменные, получаем список объясняющих переменных, не коррелирующих между собой, которые будут включены в модель:

Таблица 5 – Объясняющие переменные

Переменные	Описание
Высшее образование	Доля занятых с высшим образованием, %
Городское население	Доля городского населения в общей численности населения, %
Миграция	Коэффициент миграционного прироста, чел. на 10 тыс. чел. населения
Доступность жилья	Отношение средних цен на вторичном рынке жилья (руб./м ²) на среднедушевые денежные доходы населения (руб./месяц)

Картографическая визуализация. Для построения пространственно-эконометрической модели помимо статистической информации, необходимо

также иметь географическую информацию о регионах, для включения в модель пространства. Помимо использования географической информации в построении пространственной матрицы весов для включения ее в модель, географическая информация также используется в картографической визуализации данных.



С помощью цвета выделены регионы с низким(светлые цвета) и высоким(темные цвета) уровнем безработицы.

В работе была построена нормированная матрица весов W , так как данная матрица позволяет отразить тот факт, что одна и та же территориальная единица может оказывать неодинаковое влияние на соседние единицы. Список соседей по смежности можно получить с помощью функции `poly2nb()`.

умолчанию функция находит соседей по правилу ферзя.

Для объектов типа `nb` в пакете `spdep` определена своя функция `plot()`, которая позволяет визуализировать граф соседства. Функция требует на вход координаты точек, в случае площадных единиц для этого используют центры площадей, которые можно получить функцией `coordinates()`.

```
list.queen<-poly2nb(ru.poly, queen=TRUE)
W<-nb2listw(list.queen, style="W", zero.policy=TRUE)

coords<-coordinates(ru.poly)
plot(W,coords)
```

тест Морана. Оценка пространственных взаимодействий экономических систем, учет пространственной компоненты являются необходимым элементом современных экономических исследований. Результативность исследований определяется используемым инструментарием.

Для выявления пространственных взаимодействий широко используется Коэффициент Морана, выявляющий наличие пространственной автокорреляции между регионами.

При выполнении теста на наличие пространственной автокорреляции, были получены значения, представленные в таблице 6.

Таблица 6 – Значения коэффициента Морана.

Год	Moran I statistic	Expectation	Variance	Moran I statistic standard deviate	p-value
2016	0.474	-0.012	0.005	6.747	7.536e-12
2017	0.434	-0.012	0.005	5.973	1.161e-09
2018	0.465	-0.012	0.005	6.4854	4.426e-11
2019	0.471	-0.012	0.005	6.669	1.285e-11

Диаграмма рассеивания. Диаграммы рассеивания для уровня безработицы за 2019 год представлены для западной и восточной части России представлены на рисунках 2 и 3.

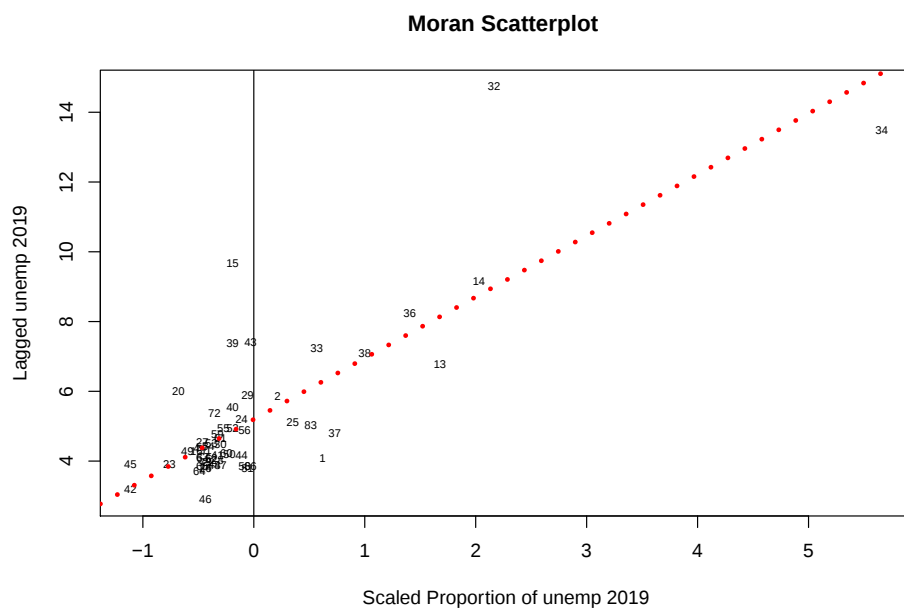


Рисунок 2 – Диаграмма рассеивания для уровня безработицы в западной части России.

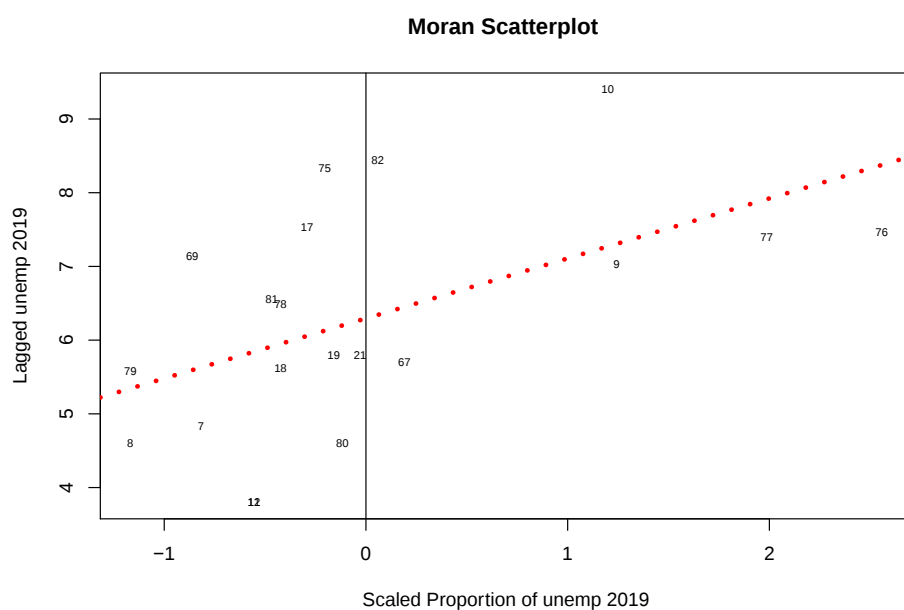


Рисунок 3 – Диаграмма рассеивания для уровня безработицы в восточной части России.

Результаты и интерпретация оценки моделей. Из-за разных расстояний между регионами и их размеров, данные разделили на восточную часть и западную. В восточную часть вошли Дальневосточный и Сибирский федеральные округа, в западную - все остальные. На первом этапе была построена модель линейной зависимости, которая не учитывает пространство.

В таблице 7 показаны результаты линейной зависимости за 2019 год.

Таблица 7 – Результаты модели линейной зависимости.

Западная часть				
Коэффициенты	Оценка	Стандартная ошибка	Статистика Стьюдента	Значимость
Константа	14.691	3.474	4.228	9.16e-05 ***
Миграционный прирост	-0.014	0.009	-1.570	0.122
Доля городского населения	-0.174	0.033	-5.358	1.78e-06 ***
Доступность жилья	2.266	1.368	1.656	0.104
Восточная часть				
Константа	6.992	3.822	1.829	0.086
Миграционный прирост	-0.012	0.012	-1.004	0.330
Доля городского населения	-0.082	0.033	-2.473	0.025 *
Доступность жилья	3.319	1.010	3.285	0.005 **

Был проведен тест Морана на наличие автокорреляции в остатках и построена диаграмма остатков. Результаты теста Морана и диаграмма показаны в таблице 8.

Таблица 8 – Значения коэффициента Морана.

Год	Moran I statistic	Expectation	Variance
Западная часть	0.016	-0.053	0.026
Восточная часть	0.434328850	-0.012987013	0.005607459

В последующие модели было добавлено пространство, учитывая географическое близость и местоположение регионов. Для определения соседних регионов была взята весовая матрица смежности регионов. Предполагалось, что регионы являются соседними, если они имеют общую границу. Веса распределялись одинаково для каждого соседнего региона, сумма весов соседних регионов равнялась единице.

Для западной и восточной части России были построены следующие модели: модель савторегрессионным пространственным лагом (SAR), модель с пространственной ошибкой (SEM) и модель Дарбина (SDM). Для каждой модели был проведен тест Морана и построена диаграмма остатков.

Построив пространственные модели для восточной и западной части России, получили следующие результаты:

Результаты оценивания моделей по данным 2016-2019 годов не сильно различаются.

В моделях, основанных на данных западной части России коэффициенты при объясняющих переменных значимо отличны от нуля. Значимое влияние на уровень безработицы оказали доля городского населения и доступность жилья. Лучшей моделью оказалась модель SDM, так как она имеет наибольший показатель логарифма правдоподобия.

Что касается моделей восточной части России, то как показал Тест Морана, из-за больших расстояний между регионами, пространственная зависимость отсутствует, в связи с чем использование пространственных моделей нецелесообразно.

Модель SDM для западной части России представлена в таблице 9.

Таблица 9 – Результаты модели Дарбина для западной части России.

Коэффициенты	Модель SDM			
	Оценка	Стандартная ошибка	Z- оценка	Значимость
2019				
Константа	14.622	5.638	2.594	0.0029567
Миграционный прирост	-0.007	0.008	-0.886	0.376

Таблица 9 – Результаты модели Дарбина для западной части России.

Коэффициенты	Модель SDM			
Доля городского населения	-0.075	0.032	2.310	0.021
Доступность жилья	2.446	1.087	2.250	0.024
lag.Миграционный прирост	-0.003	0.011	-0.319	0.750
lag.Доля городского населения	-0.079	0.049	-1.626	0.104
lag.Доступность жилья	-3.212	2.183	-1.471	0.141
Rho	0.65859			
LR test value	16.035			
Log likelihood	-133.0759			

Основные результаты

- Был выбран язык программирования для написания кода
- Изучены факторы, влияющие на уровень безработицы
- Собраны статистические данные по выбранным факторам
- Был написан программный код для пространственного анализа на примере уровня безработицы в восточной и западной части России. Для проведения исследования были отобраны переменные на основе корреляционной матрицы, была построена матрица пространственных весов, проведен тест Морана на пространственную автокорреляцию и построена диаграмма рассеивания, а также построены пространственно-эконометрические модели.
- Дана оценка результатов исследования.