第三十章 空间统计方法及案例应用

在前面的章节中，我们详细介绍了各种非空间统计模型的应用。然而，在实际的公共卫生和医学研究中，因变量（结局事件）经常是具有空间属性的分类变量。例如，在传染病的流行病学研究中，某地区的疾病发病率或某人群的健康结局可能受到地理位置的影响；在环境健康研究中，空气污染物的空间分布可能影响不同地区居民的健康状况。此时，传统的回归模型无法有效地捕捉空间依赖性和空间异质性，而空间统计回归模型则能够较好地解决这些问题。

本章将通过一个实际的公共卫生案例，演示如何使用 R 语言中的相关软件包（如 spdep、sf、spatialreg 等）进行空间自回归模型（SAR）、空间滞后模型（SLM）和空间误差模型（SEM）的构建与分析。之后，将对空间统计模型的相关理论及 R 软件包进行具体解释。通过本章内容的学习，将能够掌握如何在公共卫生和医学研究中有效应用空间统计回归模型，以揭示空间因素对健康结局的影响。

一、 研究案例简介

本章研究案例来自文献：“Jing F, Li Z, Qiao S, Ning H, Lessani MN, Li X. From neighborhood contexts to human behaviors: Cellphone-based place visitation data contribute to estimating neighborhood-level depression prevalence in the United States. Cities. 2024 May 1;148:104905.”，以下关于案例的相关说明均源自该文献。

现有关于健康行为与心理健康关系的实证研究主要基于调查。调查数据可以提供关于每个研究参与者的健康和特征的具体信息。这使得能够深入探讨可能影响健康的个体变量，如遗传、行为和个人经历。个体层面的研究结果可以为个性化的健康干预和建议提供信息。然而，这些发现可能不容易推广到更广泛的人群或社区。同时，尽管使用传统调查可以收集关于行为和健康结果的具体信息，但这种方法受限于样本量小、成本高、参与者招募时间长以及潜在的回忆偏差和社会期望偏差。有必要开发一种新方法，以在社区层面收集大样本并进行分析。

为了解决这一研究空白，目前的研究旨在利用美国全国范围内的城市大数据，考察健康行为对社区层面精神疾病患病率的解释力。我们的研究旨在调查通过手

机基于位置访问数据测量的社区层面健康行为在多大程度上可以解释美国全国范围内社区层面的精神疾病患病率，以抑郁症为例。

本研究选取文献中美国南卡罗来纳州的数据作为案例数据源，通过 https://github.com/FengruiJing/SouthCarolinaSHP/tree/main 可下载。

二、 案例研究设计

本案例采用横截面研究设计。这项研究的结局变量是南卡罗来纳州的社区抑郁症患病率，定义为 18 岁或以上调查对象中，报告曾从医生、护士或其他合格医务人员处确诊抑郁症的比例；自变量为南卡罗来纳州的社区 3 种积极和消极健康行为平均次数（健身房访问、自然公园访问、酒吧访问），定义为基于 2019 年 GPS 轨迹数据的社区居民年均健康行为次数，即该社区居民在该年度访问该类设施点（健身房、自然公园、酒吧）的平均次数。

（一）研究对象的选择

我们选择 2019 年作为研究年份，是因为在新冠疫情爆发前，人类流动模式相对稳定。我们选择人口普查区块组（Census Tracts）作为研究的分析单位，因为较低层级的人口普查街区组（CBG）数据在健康结果数据方面不可用。此外，SafeGraph 提供的一些健康行为的访问数据在 CBG 层级过于稀疏，无法具有代表性。我们利用南卡罗来纳州 1103 个人口普查区块组的数据，其余区块由于这些数据源中数据不足而被排除。通过结合多种数据源，我们能够估算整个南卡罗来纳州的抑郁症患病率。

（二）研究变量的定义

研究中使用的主要结果是抑郁症的患病率，其操作化定义为 18 岁及以上的调查受访者中，报告曾被医生、护士或其他合格医疗从业者诊断为抑郁症的比例。

输入变量包括社区背景变量、兴趣点（POI）环境特征、健康行为变量。

社区背景变量包括美国农业部（USDA 2020）定义的城乡分类、人口密度、女性比例、65 岁及以上的个人比例、黑人比例、没有健康保险的个人比例、租房比例、社区劣势程度以及美国环境保护署（EPA 2019）定义的步行性。社区劣势是一个包含五个因素的指数，即平民失业率、女性主导的家庭比例、25 岁及以上

的高中辍学率、年收入低于 15,000 美元的家庭比例以及接受公共援助的家庭比例。

第二类变量与构建和自然环境的健康相关兴趣点（POI）特征有关，包括健身中心、自然公园、提供酒精饮料的饮酒场所、药房、医生诊所、综合医院、专科医院和宗教场所。这些 POI 根据 NAICS 分类为“713940（健身和休闲运动中心）”、“712190（自然公园和其他类似机构）”、“722410（饮酒场所（酒精饮料））”、“456110（药房和药品零售商）”、“621111（医生诊所（精神卫生专家除外））”、“622110（综合医疗和外科医院）”、“622310（专科（精神病和药物滥用除外）医院）”、“813110（宗教组织）”。

第三类变量与基于 POI 访问的健康行为信息有关，分为两类，即积极和消极健康行为（PNHB）以及健康服务利用行为（HSUB）。在本研究中，PNHB 包括两种与心理健康相关的积极行为，即健身和自然公园访问，以及一种与心理健康相关的消极行为，即访问提供酒精饮料的饮酒场所。HSUB 定义为包含五种行为，即访问药房、医生诊所、综合医院、专科医院和宗教场所。随后，这些从 SafeGraph 手机数据中获得的个体层面健康行为（PNHB 和 HSUB）被汇总为社区层面数据。

三、 数据结构和数据预处理

（一）数据结构

本研究依赖于三个关键数据来源，即 SafeGraph 的访问和兴趣点（POI）数据、人口普查数据以及健康数据。

SafeGraph 是美国领先的手机数据提供商，提供关于 POI 和访问模式的全面数据，包括根据北美行业分类系统（NAICS）标准确定的大约 700 万个 POI（U.S. Census Bureau 2022）。SafeGraph 主要提供 NAICS 分类下的 POI 的访问信息。POI 访问数据来自 SafeGraph 的 2019 年每周 Patterns 数据集，数据层级为人口普查街区组（CBG），计算每周每个 CBG 对每个 POI 的访问次数。将数据聚合到 2019 年的人口普查区块组（Census Tract）层级后，计算每个区块组对每种 POI 类型的人均访问次数。例如，假设 T1 区块组有 1000 人，其中 100 人在 2019 年访问医生诊所（被视为一种 POI 类型）。在这种情况下，2019

年 T1 区块组内医生诊所的人均访问次数为 0.1，通过将 100 除以 1000 得出。值得注意的是，T1 区块组的居民可以选择访问其本区块组内或外的医生诊所。

从美国疾病控制与预防中心（CDC）的 PLACES 项目（PLACES 2023）收集了 2019 年人口普查区块组层级的精神疾病结果数据。此外，依赖美国社区调查（ACS）2015-2019 年的数据来获取人口普查数据。我们特意选择了 ACS 五年估计数据，而不是一年或三年估计数据，以包括人口较少（少于 20,000）的区块组的信息（ACS n.d.）。

（二）空间数据提取流程

根据案例中研究对象的选择和研究变量的定义制定空间数据提取流程，主要包括四个步骤。

1. 提取研究所用变量

首先，从地理空间数据集中提取案例所需空间矢量文件。从 https://github.com/FengruiJing/SouthCarolinaSHP/blob/main/Reduced_SouthCarolina.zip?raw=true 提取南卡罗来纳州的空间矢量文件，以及从 https://github.com/FengruiJing/SouthCarolinaSHP/blob/main/DepressionData.csv?raw=true 提取南卡罗来纳州的 DepressionCSV 数据集，存放于研究者指定的文件夹 C:/Rbook/Chapter30/Data 中。如不能顺利通过程序读取，可通过以上链接下载。程序见 R_prog_30.1。

R_prog_30.1 从数据平台中获取原始数据集

```
library(httr)
library(utils)
shapefile_url <-
"https://github.com/FengruiJing/SouthCarolinaSHP/blob/main/Reduced_SouthCarolina.zip?raw=true"
shapefile_dest <- "C:/Rbook/Chapter30/Data/Reduced_SouthCarolina.zip"
csv_url <-
"https://github.com/FengruiJing/SouthCarolinaSHP/blob/main/DepressionData.csv?raw=true"
csv_dest <- " C:/Rbook/Chapter30/Data/DepressionData.csv"
## 增加超时时间并下载文件
options(timeout = 1000) # 设置超时时间为 1000 秒
## 下载 Shapefile 文件
download.file(shapefile_url, shapefile_dest, mode = "wb")
## 下载 CSV 文件
download.file(csv_url, csv_dest, mode = "wb")
## 设置解压缩目标路径
unzip_dir <- "C:/AP/Teaching/SpatialRegression/R_BOOK/"
## 解压缩 Shapefile 文件
unzip(shapefile_dest, exdir = unzip_dir)
```

程序说明：

(1) library(httr) 和 library(utils)，加载 httr 和 utils 包，用于下载文件和解压缩文件；

(2) 设置 Shapefile 和 CSV 文件的 URL 和目标文件路径，定义下载文件的 URL 和保存路径；

(3) download.file()，下载文件，使用 download.file 函数从指定 URL 下载文件到目标路径；

(4) 设置解压缩目标路径，定义解压缩文件的目标路径。

(5) unzip()，解压缩文件，使用 unzip 函数将下载的 ZIP 文件解压到目标路径。

2. 合并原始数据集

将横截面数据集和空间矢量文件进行横向合并，合并后的数据集 SouthCarolina 存放于指定的文件夹 C:/Rbook/Chapter30/Data。

R_prog_30.2 合并原始数据集

```
library(sf)
library(dplyr)
library(readr)
# 读取 Shapefile 数据
SC_sf <- st_read("C:/Rbook/Chapter30/Data/Reduced_SouthCarolina.shp")
# 设置坐标参考系（CRS）
sf::st_crs(SC_sf) <- 4326
# 输入 CSV 文件
depression_data <- read_csv("C:/Rbook/Chapter30/Data/DepressionData.csv")
# 确保 GEOID 的类型一致
SC_sf <- SC_sf %>% mutate(GEOID = as.character(GEOID))
depression_data <- depression_data %>% mutate(GEOID = as.character(GEOID))
# 基于 GEOID 字段进行连接
SC_sf <- SC_sf %>% left_join(depression_data, by = "GEOID")
# 保存连接后的 Shapefile 文件
st_write(SC_sf, "C:/Rbook/Chapter30/Data/SouthCarolina.shp")
```

程序说明：

(1) library(sf)，加载 sf 包，用于处理空间数据；

(2) library(dplyr)，加载 dplyr 包，用于数据操作；

(3) library(readr)，加载 readr 包，用于读取 CSV 文件；

(4) st_read()，读取 Shapefile 数据，将数据读取到对象 SC_sf 中；

(5) sf::st_crs()，设置坐标参考系（CRS）为 EPSG:4326；

(6) read_csv()，输入 CSV 文件，将数据读取到对象 depression_data 中；

(7) mutate()，确保 GEOID 字段的类型一致，将 GEOID 字段转换为字符类型；

(8) left_join()，基于 GEOID 字段进行连接，将 depression_data 数据合并到 SC_sf 中；

(9) st_write()，保存连接后的 Shapefile 文件，将连接后的 SC_sf 数据保存为新的 Shapefile 文件。

程序 R_prog_30.2 运行后，得到 Final 数据集，包含所有需要的原始变量。

3. 对原始变量进行预处理

根据案例研究设计，对结局变量和输入变量进行预处理。处理数据中的异常值，如缺失值检查、数据实例、数据字段检查、数据变量名称检查等。本案例中使用的变量清单见表 30-1。对原始变量预处理的程序见 R_prog_30.3。

表 5-1 案例中使用的变量清单

变量名	描述	变量说明
Population density	人口密度	每单位面积内的居民人数，反映该区域的居住密度。
females	女性比例	区域内女性居民占总人口的比例。
aged65andover	65 岁及以上人口比例	区域内年龄在 65 岁及以上的居民占总人口的比例。
noHealthInsurance	无健康保险比例	区域内没有健康保险的居民占总人口的比例。
Walkability	步行性	衡量区域内步行便利程度的指标，包括步行道、步行设施等。
HouseRenting	租房比例	区域内租房居民占总人口的比例。
FitnessPOI	健身场所兴趣点数量	区域内健身场所的数量，如健身房、体育中心等。
DrinkingPOI	饮酒场所兴趣点数量	区域内饮酒场所的数量，如酒吧、夜总会等。
PharmaciesPOI	药房兴趣点数量	区域内药房的数量。
physiciansPOI	医生诊所兴趣点数量	区域内医生诊所的数量。
GeneralhospitalsPOI	综合医院兴趣点数量	区域内综合医院的数量。
SecialtyhospitalsPOI	专科医院兴趣点数量	区域内专科医院的数量（不包括精神病和药物滥用医院）。
ReligiousPlacesPOI	宗教场所兴趣点数量	区域内宗教场所的数量，如教堂、寺庙等。
Fitnessvisits	健身场所访问次数	区域内居民访问健身场所的次数。
Drinkingvisits	饮酒场所访问次数	区域内居民访问饮酒场所的次数。
Pharmaciesvisits	药房访问次数	区域内居民访问药房的次数。
physiciansvisits	医生诊所访问次数	区域内居民访问医生诊所的次数。
Generalhospitalsvisits	综合医院访问次数	区域内居民访问综合医院的次数。
Secialtyhospitalsvisits	专科医院访问次数	区域内居民访问专科医院的次数（不包括精神病和药物滥用医院）。
ReligiousPlacesvisits	宗教场所访问次数	区域内居民访问宗教场所的次数。

R_prog_30.3 原始变量的预处理

```
# 加载必要的包
library(sf)
library(dplyr)
SC_sf <- st_read("C:/Rbook/Chapter30/Data/SouthCarolina.shp")
# 修改字段名称
names(SC_sf)[names(SC_sf) == "Population"] <- "population_density"
names(SC_sf)[names(SC_sf) == "Populati_1"] <- "females"
names(SC_sf)[names(SC_sf) == "Populati_2"] <- "age65andover"
names(SC_sf)[names(SC_sf) == "NoHealthIn"] <- "noHealthInsurance"
names(SC_sf)[names(SC_sf) == "Neighbor_D"] <- "neighborhooddisadvantage"
names(SC_sf)[names(SC_sf) == "Walkabilit"] <- "Walkability"
names(SC_sf)[names(SC_sf) == "X_HouseRen"] <- "HouseRenting"
names(SC_sf)[names(SC_sf) == "FitnessPO"] <- "FitnessPOI"
names(SC_sf)[names(SC_sf) == "DrinkingPl"] <- "DrinkingPOI"
names(SC_sf)[names(SC_sf) == "Pharmacies"] <- "PharmaciesPOI"
names(SC_sf)[names(SC_sf) == "physicians"] <- "physiciansPOI"
names(SC_sf)[names(SC_sf) == "Generalhos"] <- "GeneralhospitalsPOI"
names(SC_sf)[names(SC_sf) == "Secialtyho"] <- "SecialtyhospitalsPOI"
names(SC_sf)[names(SC_sf) == "ReligiousP"] <- "ReligiousPlacesPOI"
names(SC_sf)[names(SC_sf) == "Fitness"] <- "Fitnessvisits"
names(SC_sf)[names(SC_sf) == "Drinking_1"] <- "Drinkingvisits"
names(SC_sf)[names(SC_sf) == "Pharmaci_1"] <- "Pharmaciesvisits"
names(SC_sf)[names(SC_sf) == "physicia_1"] <- "physiciansvisits"
names(SC_sf)[names(SC_sf) == "Generalh_1"] <- "Generalhospitalsvisits"
names(SC_sf)[names(SC_sf) == "Secialty_1"] <- "Secialtyhospitalsvisits"
names(SC_sf)[names(SC_sf) == "Religiou_1"] <- "ReligiousPlacesvisits"
# Check data
head(SC_sf);
str(SC_sf);
names(SC_sf)
st_write(SC_sf, "C:/AP/Teaching/SpatialRegression/R_BOOK/SouthCarolina1.shp")
```

程序说明：

(1) library(sf)，加载 sf 包，处理空间数据；

(2) library(dplyr)，加载 dplyr 包，便于数据操作；

(3) st_read()，读取 Shapefile 数据，将数据读取到对象 SC_sf 中；

(4) names()，修改字段名称，将 Shapefile 中原字段名称更改为新的描述性名称；

(5) head() 和 str()，检查数据的前几行和数据结构，names() 检查所有字段名称；

(6) st_write()，将修改字段名称后的 SC_sf 数据保存为新的 Shapefile 文件。

四、 统计方法

在模型选择之前，我们使用两个指标检查了变量和误差项的空间依赖性：空间误差和空间滞后。空间误差意味着不同空间单元的误差项是相关的，空间滞后则意味着地点 M 的因变量 V 受地点 M 和 N 的自变量的影响。为了评估回归的空间依赖性，我们使用普通最小二乘（OLS）回归，将抑郁症患病率作为因变量，所有相关输入变量作为自变量，并引入权重矩阵。残差的 Moran's I 结果显著（$p<0.001$），表明残差存在强烈的空间自相关，可以使用空间滞后回归（SLR）和空间误差回归（SER）。为了比较这两种空间模型的拟合度，我们使用了一个常见的非空间模型——OLS 线性回归（LR），评估了健康相关行为的纳入如何影响模型拟合度。

在模型性能评估中，我们采用了 Akaike 信息准则（AIC）和贝叶斯信息准则（BIC）作为评估指标。因此，我们报告了所有三个模型的 AIC 和 BIC 值。AIC/BIC 值更低的模型被认为具有更好的拟合度。文献统计分析采用 R 软件。

五、 R 程序和结果解读

按照文献中的分析思路和策略，采用 R 软件进行空间自相关分析，在此基础上运算空间误差模型和空间滞后模型，并比较模型效果。R 程序如下：

R_prog_30.4 结局变量的空间可视化：

```
# 加载必要的包
library(sf)
library(dplyr)
library(classInt)
SC_sf <- st_read("C:/Rbook/Chapter30/Data/SouthCarolina1.shp")
# 生成四分位数分组间隔
breaks <- classIntervals(SC_sf$Depression, n = 4, style = "quantile")$brks

# 使用生成的分组间隔绘制地图
plot(SC_sf['Depression'], graticule = sf::st_crs(4326),
        main = 'Depression prevalence', breaks = breaks, axes = TRUE)
```

程序说明：

(1) library(sf)，加载 sf 包，处理空间数据。

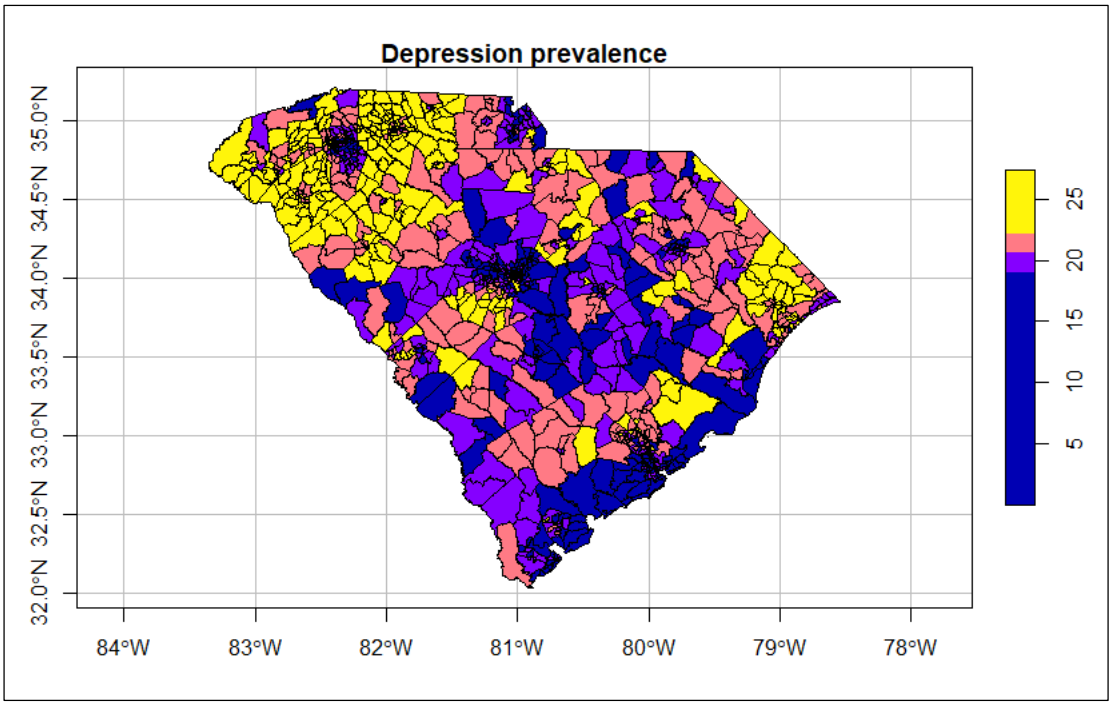
(2) library(dplyr)，加载 dplyr 包，便于数据操作。

(3) library(classInt)，加载 classInt 包，用于生成分组间隔。

(4) 计算 Depression 变量的四分位数分组间隔，使用 classIntervals 函数生成四分位数分组间隔，并提取分组边界：

(5) 使用生成的分组间隔绘制抑郁症患病率地图，调用 plot 函数，绘制 SC_sf 对象中的 Depression 变量地图，并设置经纬网、标题、分组间隔和坐标轴：

结果见 R_output_30.4 研究对象的空间可视化。



接下来进行 OLS 线性回归分析，程序见 R_prog_30.5：

R_prog_30.5 OLS 线性回归分析及残差空间自相关检验

```
# Create a simple linear regression
olsRslt <- lm(Depression ~ population_density + females + age65andover +
noHealthInsurance + neighborhooddisadvantage + Walkability + HouseRenting +
                        FitnessPOI + DrinkingPOI + PharmaciesPOI + physiciansPOI +
GeneralhospitalsPOI + SecialtyhospitalsPOI + ReligiousPlacesPOI +
                        Fitnessvisits + Drinkingvisits + Pharmaciesvisits + physiciansvisits +
Generalhospitalsvisits + Secialtyhospitalsvisits + ReligiousPlacesvisits,
                    data = SC_sf)
summary(olsRslt)


# Derive the residuals from the regression. Need to handle those missed values
lmResiduals <- rep(0, length(SC_sf$Depression))
resIndex <- olsRslt$residuals %>% names() %>% as.integer();
lmResiduals[resIndex] <- olsRslt$residuals

# Test if there is spatial autocorrelation in the regression residuals (errors).
SCNbList %>%
   spdep::moran.test(lmResiduals, ., zero.policy = TRUE)
```

程序说明：

(1) library(sf)，加载 sf 包，处理空间数据；

(2) library(spdep)，加载 spdep 包，进行空间依赖性分析；

(3) st_read()，读取 Shapefile 数据；

st_read("C:/AP/Teaching/SpatialRegression/R_BOOK/SouthCarolina.shp")，将数据读取到对象 SC_sf 中；

(4) sf::st_crs()，设置坐标参考系，sf::st_crs(SC_sf) <- 4326；

(5) lm()，建立 OLS 线性回归模型，Depression ~ Population_density + %females + age65andover + noHealthInsurance + neighborhooddisadvantage + Walkability + HouseRenting + FitnessPOI + DrinkingPOI + PharmaciesPOI + physiciansPOI + GeneralhospitalsPOI + SecialtyhospitalsPOI + ReligiousPlacesPOI + Fitnessvisits + Drinkingvisits + Pharmaciesvisits + physiciansvisits + Generalhospitalsvisits + Secialtyhospitalsvisits + ReligiousPlacesvisits，data = SC_sf；

(6) summary()，查看模型参数估计的结果；

(7) residuals，提取回归模型的残差，处理缺失值；

(8) spdep::moran.test()，检验回归残差的空间自相关性。

R_output_30.5　OLS 线性回归分析及残差空间自相关检验结果

```
## R_output_30.5 OLS 线性回归及残差检验
Call:
lm(formula = Depression ~ population_density + females + age65andover +
    noHealthInsurance + neighborhooddisadvantage + Walkability +
    HouseRenting + FitnessPOI + DrinkingPOI + PharmaciesPOI +
    physiciansPOI + GeneralhospitalsPOI + SecialtyhospitalsPOI +
    ReligiousPlacesPOI + Fitnessvisits + Drinkingvisits + Pharmaciesvisits +
    physiciansvisits + Generalhospitalsvisits + Secialtyhospitalsvisits +
    ReligiousPlacesvisits, data = SC_sf)

Residuals:
Min          1Q  Median      3Q     Max
-7.6007 -1.5150 -0.0599  1.4306  8.0578
                            Estimate   Std. Error  t value  Pr(>|t|)
(Intercept)                 2.89E+00   3.73E-01      7.756  2.02E-14  ***
population_density          8.42E-05   7.60E-05      1.107   0.26852
females                     2.90E-01   1.05E-02     27.782   < 2e-16  ***
age65andover               -4.29E-02   1.04E-02     -4.125  3.99E-05  ***
noHealthInsurance           1.27E-01   1.43E-02      8.843   < 2e-16  ***
neighborhooddisadvantage   -7.96E-01   1.69E-01     -4.723  2.63E-06  ***
Walkability                -2.68E-01   3.62E-02     -7.398  2.78E-13  ***
HouseRenting                4.74E-02   6.05E-03      7.826  1.19E-14  ***
FitnessPOI                 -1.46E-01   4.72E-02     -3.101   0.00198  **
DrinkingPOI                 5.50E-02   4.83E-02      1.139   0.25493
PharmaciesPOI              -5.07E-02   5.10E-02     -0.996   0.31961
physiciansPOI              -7.68E-03   4.65E-03     -1.652   0.09892  .
GeneralhospitalsPOI         4.32E-02   1.15E-01      0.375    0.7074
SecialtyhospitalsPOI       -1.99E-01   1.34E-01     -1.485   0.13778
ReligiousPlacesPOI          4.34E-02   1.57E-02      2.756   0.00595  **
Fitnessvisits              -2.22E-02   1.23E-02     -1.808   0.07082  .
Drinkingvisits             -2.19E-03   3.02E-02     -0.073    0.9422
Pharmaciesvisits            1.20E-01   1.69E-02      7.108  2.14E-12  ***
physiciansvisits            1.87E-01   3.62E-02      5.163  2.89E-07  ***
Generalhospitalsvisits      1.89E-02   6.66E-03      2.837   0.00465  **
Secialtyhospitalsvisits    -2.15E-01   2.24E-01     -0.959   0.3378
ReligiousPlacesvisits      -9.62E-03   1.31E-02     -0.735   0.46244
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 2.234 on 1081 degrees of freedom
Multiple R-squared:  0.7038,    Adjusted R-squared:  0.698
F-statistic: 122.3 on 21 and 1081 DF,  p-value: < 2.2e-16
```

续 R_output_30.5

```
        Moran I test under randomisation

data:  lmResiduals
weights: .

Moran I statistic standard deviate = 23.19, p-value < 2.2e-16
alternative hypothesis: greater
sample estimates:
Moran I statistic       Expectation          Variance
     0.4138085635     -0.0009074410     0.0003198103
```

莫兰检验结果显示，线性回归的残差具有显著的正空间自相关。莫兰指数为 0.414，p 值小于 0.01。因此，我们可以运行空间误差模型以考虑这种自相关。R 代码见 R_prog_30.6 如下：

R_prog_30.6 空间误差模型回归分析

```
library(spatialreg)
# use spdep package to run the spatial error model
# Use spatialreg::errorsarlm to run the same model

#error start
serrRslt <- spatialreg::errorsarlm(Depression ~ population_density + females + age65andover +
noHealthInsurance + neighborhooddisadvantage + Walkability + HouseRenting +
                                     FitnessPOI + DrinkingPOI + PharmaciesPOI + physiciansPOI +
GeneralhospitalsPOI + SecialtyhospitalsPOI + ReligiousPlacesPOI +
                                     Fitnessvisits + Drinkingvisits + Pharmaciesvisits +
physiciansvisits + Generalhospitalsvisits + Secialtyhospitalsvisits + ReligiousPlacesvisits,
                                   data = SC_sf,
                                   listw = SCNbList,
                                   zero.policy = TRUE,
                                   na.action = na.omit);
summary(serrRslt)



# Derive the residuals from the regression. Need to handle those missed values
seResiduals <- rep(0, length(SC_sf$Depression))
resIndex <- serrRslt$residuals %>% names() %>% as.integer();
seResiduals[resIndex] <- serrRslt$residuals

# Test if there is spatial autocorrelation in the regression residuals (errors).
SCNbList %>%
   spdep::moran.test(seResiduals, ., zero.policy = TRUE)
```

程序说明：

(1) library(spatialreg)，加载 spatialreg 包，用于运行空间误差模型；

(2) spatialreg::errorsarlm()，使用 spatialreg 包中的 errorsarlm 函数运行空间误差模型；

(3) summary()，查看模型参数估计的结果；

(4) 提取回归模型的残差，处理缺失值；

(5) spdep::moran.test()，检验回归残差的空间自相关性；

R_output_30.6 空间误差模型回归分析参数估计结果。

Call:spatialreg::errorsarlm(formula = Depression ~ population_density + females + age65andover + noHealthInsurance + neighborhooddisadvantage + Walkability + HouseRenting + FitnessPOI + DrinkingPOI + PharmaciesPOI + physiciansPOI + GeneralhospitalsPOI + SecialtyhospitalsPOI + ReligiousPlacesPOI + Fitnessvisits + Drinkingvisits + Pharmaciesvisits + physiciansvisits + Generalhospitalsvisits + Secialtyhospitalsvisits + ReligiousPlacesvisits, data = SC_sf, listw = SCNbList, na.action = na.omit, zero.policy = TRUE)

Residuals:

Min	1Q	Median	3Q	Max
-6.693911	-1.080604	-0.015916	1.026230	7.825888

Type: error

Coefficients: (asymptotic standard errors)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.76E+00	3.38E-01	8.1677	2.22E-16
population_density	1.59E-04	6.72E-05	2.3653	0.018016
females	2.77E-01	8.39E-03	33.0395	< 2.2e-16
age65andover	-2.21E-02	9.09E-03	-2.436	0.014853
noHealthInsurance	1.02E-01	1.25E-02	8.1973	2.22E-16
neighborhooddisadvantage	-4.73E-01	1.45E-01	-3.2695	0.001078
Walkability	-1.01E-01	3.54E-02	-2.8467	0.004418
HouseRenting	4.38E-02	4.97E-03	8.8147	< 2.2e-16
FitnessPOI	-9.12E-02	3.65E-02	-2.4999	0.012423
DrinkingPOI	-5.55E-03	3.82E-02	-0.1453	0.884444
PharmaciesPOI	-7.77E-02	3.79E-02	-2.0518	0.040187
physiciansPOI	-3.68E-03	3.52E-03	-1.0452	0.295948
GeneralhospitalsPOI	4.54E-02	8.46E-02	0.5363	0.591739
SecialtyhospitalsPOI	-1.25E-01	9.81E-02	-1.2782	0.201187
ReligiousPlacesPOI	2.91E-02	1.23E-02	2.3631	0.018121
Fitnessvisits	-1.66E-03	1.29E-02	-0.1286	0.8977
Drinkingvisits	-1.16E-02	3.14E-02	-0.3689	0.712204
Pharmaciesvisits	1.10E-01	2.04E-02	5.3884	7.11E-08
physiciansvisits	4.86E-02	3.72E-02	1.3079	0.190917
Generalhospitalsvisits	1.02E-02	7.71E-03	1.3246	0.185289
Secialtyhospitalsvisits	-3.18E-01	2.20E-01	-1.4455	0.148327
ReligiousPlacesvisits	1.71E-02	1.19E-02	1.4465	0.148046

续 R_output_30.6.

```
Lambda: 0.7205, LR test value: 423.11, p-value: < 2.22e-16
Asymptotic standard error: 0.026235
      z-value: 27.463, p-value: < 2.22e-16
Wald statistic: 754.23, p-value: < 2.22e-16

Log likelihood: -2229.024 for error model
ML residual variance (sigma squared): 2.9589, (sigma: 1.7201)
Number of observations: 1103
Number of parameters estimated: 24
AIC: 4506, (AIC for lm: 4927.2)

       Moran I test under randomisation

data:    seResiduals
weights: .

Moran I statistic standard deviate = -3.2757, p-value = 0.9995
alternative hypothesis: greater
sample estimates:
Moran I statistic          Expectation             Variance
 -0.0594750006        -0.0009074410         0.0003196707
```

从 AIC 来看，空间误差模型比线性模型表现更好（较低的 AIC 基本上意味着更好的拟合）。Lambda 值为 0.721，也具有统计显著性，表明误差项具有空间自回归特性。虽然模型系数变化不大，但它们的大多数绝对 Z 值更高，这意味着估计更稳健，方差更低。

现在很明显，残差中不再存在明显的空间自相关，因为莫兰指数现在接近于零，这意味着我们不能拒绝莫兰指数为零的假设。

最后，实现空间滞后模型，程序见 R_prog_30.7。

R_prog_30.7 空间误差模型及结果

```
# use spdep package to run the spatial lag model
# spatialreg::lagsarlm

slmRslt <- spatialreg::lagsarlm(Mental_hea ~ New_rural + Population +
                                    Populati_1 + Populati_2 +
                                    NoHealthIn + X_Black + Neighbor_D +
                                    X_Immigran + X_HouseRen,
                                data = SC_sf,
                                listw = SCNbList,
                                zero.policy = TRUE,
                                na.action = na.omit);

summary(slmRslt)
```

程序说明：

(1) library(spatialreg)，加载 spatialreg 包，用于运行空间滞后模型；

(2) spatialreg::lagsarlm()，使用 spatialreg 包中的 lagsarlm 函数运行空间滞后模型；

(3) summary()，查看模型参数估计的结果。

R_output_30.7 空间滞后回归模型的结果

R_output_30.7

Call:spatialreg::lagsarlm(formula = Depression ~ population_density +
females + age65andover + noHealthInsurance + neighborhooddisadvantage +
Walkability + HouseRenting + FitnessPOI + DrinkingPOI + PharmaciesPOI +
physiciansPOI + GeneralhospitalsPOI + SecialtyhospitalsPOI +
ReligiousPlacesPOI + Fitnessvisits + Drinkingvisits + Pharmaciesvisits +
physiciansvisits + Generalhospitalsvisits + Secialtyhospitalsvisits +
ReligiousPlacesvisits, data = SC_sf, listw = SCNbList, na.action = na.omit, zero.policy = TRUE)

Residuals:

Min	1Q	Median	3Q	Max
-5.54249	-1.26923	-0.01433	1.16914	8.35261

Type: lag

Coefficients: (asymptotic standard errors)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-5.6172	5.49E-01	-10.2353	< 2.2e-16
population_density	5.11E-05	6.58E-05	0.7768	0.43728
females	2.87E-01	9.06E-03	31.6765	< 2.2e-16
age65andover	-2.63E-02	9.01E-03	-2.9126	0.003584
noHealthInsurance	1.06E-01	1.25E-02	8.5025	< 2.2e-16
neighborhooddisadvantage	-7.89E-01	1.46E-01	-5.4118	6.24E-08
Walkability	-2.09E-01	3.14E-02	-6.6499	2.93E-11
HouseRenting	4.45E-02	5.24E-03	8.4915	< 2.2e-16
FitnessPOI	-1.17E-01	4.08E-02	-2.8569	0.004278
DrinkingPOI	1.62E-02	4.18E-02	0.3866	0.69908
PharmaciesPOI	-8.90E-02	4.42E-02	-2.0169	0.043709
physiciansPOI	-7.18E-03	4.02E-03	-1.7848	0.0743
GeneralhospitalsPOI	3.17E-02	9.96E-02	0.3182	0.750334
SecialtyhospitalsPOI	-1.26E-01	1.16E-01	-1.0876	0.276758
ReligiousPlacesPOI	4.55E-02	1.36E-02	3.3389	0.000841
Fitnessvisits	-1.07E-02	1.06E-02	-1.0061	0.314359
Drinkingvisits	2.15E-02	2.62E-02	0.8218	0.411163
Pharmaciesvisits	7.71E-02	1.49E-02	5.1913	2.09E-07
physiciansvisits	1.00E-01	3.15E-02	3.1772	0.001487
Generalhospitalsvisits	1.49E-02	5.77E-03	2.5867	0.009689
Secialtyhospitalsvisits	-6.70E-02	1.94E-01	-0.345	0.730113
ReligiousPlacesvisits	-2.24E-03	1.13E-02	-0.1974	0.843521

续 R_output_30.7

```
Rho: 0.44209, LR test value: 255.94, p-value: < 2.22e-16
Asymptotic standard error: 0.025136
    z-value: 17.588, p-value: < 2.22e-16
Wald statistic: 309.35, p-value: < 2.22e-16

Log likelihood: -2312.61 for lag model
ML residual variance (sigma squared): 3.7347, (sigma: 1.9325)
Number of observations: 1103
Number of parameters estimated: 24
AIC: 4673.2, (AIC for lm: 4927.2)
LM test for residual autocorrelation
test value: 78.638, p-value: < 2.22e-16



        Moran I test under randomisation

data:  slResiduals
weights: .

Moran I statistic standard deviate = 6.9415, p-value = 1.939e-12
alternative hypothesis: greater
sample estimates:
Moran I statistic       Expectation          Variance
   0.1232061082      -0.0009074410       0.0003196887
```

同样，空间滞后模型比线性模型表现更好，尽管不如空间误差模型。空间滞后项 ρ（Rho）也具有统计显著性。空间滞后模型表现不佳从莫兰检验中也可以看出。尽管残差的莫兰指数从 0.414 降到了 0.123，但下降幅度小于空间误差模型。

以上是加入所有变量（社区背景变量+兴趣点（POI）环境特征+健康行为变量）的 OLS 线性回归模型、空间误差模型、空间滞后模分析结果。案例中还建立了基于社区背景变量、兴趣点（POI）环境特征、健康行为变量 9 种不同配组的三类模型结果，篇幅限制，此次省略。

六、 案例研究结论

研究结果表明，OLS 线性回归模型的残差具有空间自相关，因此需要采取空间统计模型控制空间自相关的影响，减少模型结果的偏差。空间误差模型的表现结果好于空间误差模型，因此选取空间误差模型作为空间统计模型作为本研究的回归统计方法。

通过加入或不加入基于手机的地点访问数据变量，比较模型拟合度结果。研究发现，基于手机的地点访问数据可以显著提高社区层面心理疾病解释模型的拟合度，并提供社区健康行为与社区层面心理疾病之间关联的宝贵见解。具体分析显示，PNHB（积极和消极健康行为）行为的模型拟合度高于 HSUB（健康服务利用行为）行为，这些健康行为在预测社区层面自我报告的心理健康状态方面比抑郁症更有效。健身访问、饮酒场所（酒精饮料）访问、药房访问、综合医院访问和专科医院访问与社区层面的抑郁症显著相关。

七、 统计方法的进一步说明

空间自相关的处理

在空间数据中，通常会出现一些或所有结果测量值表现出空间自相关。当两个点的相对结果与它们之间的距离相关或两个多边形共享边界时，就会发生这种情况。分析空间数据时，检查自相关性非常重要。如果没有空间自相关的证据，则可以继续采用标准方法。但是，如果有空间自相关的证据，则标准非空间分析的一项基本假设可能被违反，结果可能无效。空间自相关衡量对象在与邻近对象或邻居比较时的相似或不同程度。

（一）空间误差模型

空间误差模型处理残差中的空间自相关。其思路是，这样的误差（回归的残差）是自相关的，即一个空间特征的误差可以建模为其邻居误差的加权平均。换句话说，这些误差具有空间自相关性。该模型可以表示为：

$$y = X\beta + u, u = \lambda ErrWu + \varepsilon \qquad 式(30-1)$$

其中：y 是一个(N×1)的观测向量，在每个 N 个位置上对响应变量的观测值；X 是一个(N×k)的协变量矩阵；β 是一个(k×1)的参数向量；u是一个(N×1)的空间自相关扰动向量；ε是一个 (N×1)的独立同分布扰动向量；λ_{Err} 是一个标量空间参数。

（二）空间滞后模型

空间滞后是一种变量，实质上是对一个位置的邻居值进行平均（每个邻居位置的值乘以空间权重，然后将这些乘积求和）。它可以用来将邻居的值与该位置本身的值进行比较。按照惯例，邻居中心的位置不包括在邻居的定义中，因此被设为零。根据这些空间滞后，我们可以使用空间滞后模型来解决因变量中的空间自相关问题。

$$y = \rho_{Lag} Wy + X\beta + \varepsilon \qquad 式(30-2)$$

其中：ρ_{Lag} 是一个标量空间参数，表示一个空间特征受到其邻居的影响程度。

这些空间回归模型帮助我们更准确地描述和预测带有空间依赖性的变量，克服简单线性回归模型在存在空间自相关时的局限性。

30.8 R 程序的进一步说明

在 30.7 中，我们已经介绍了空间自相关分析、空间误差模型以及空间滞后模型的理论知识，这一节我们进一步介绍实现这些回归模型的 R 语言程序。

30.8.1 空间自相关分析

包括全局空间自相关和局部空间自相关，这里主要介绍全局空间自相关分析。用到了 moran.test 函数：这是 spdep 包中的一个函数，用于执行莫兰检验（Moran's I test），以测试残差（errors）是否存在空间自相关。zero.policy = TRUE：这是 moran.test 函数的一个参数，指定如何处理零权重的情况。设置为 TRUE 表示在计算莫兰指数时忽略零权重。Moran's I test 的 R 语言语法如下：

```
##Moran's I test 语法：
moran.test(x, listw, randomisation = TRUE, zero.policy = NULL, alternative = "greater", spChk = NULL, resfun = weighted.residuals, na.action = na.fail, ...)
```

表 30-2 是 Moran's I test 函数的参数说明。

表 5-4 Moran's I test 函数参数说明

参数	描述	默认值
x	一个数值向量，通常是回归残差或其他需要检测空间自相关性的变量。	
listw	一个列表权重对象，定义了空间邻接结构，可以使用 nb2listw 函数生成。	
randomisation	逻辑值，表示是否使用随机化方法计算 p 值。	TRUE
zero.policy	逻辑值，表示如何处理零权重的情况。如果为 TRUE，则忽略零权重。	
alternative	字符串，指定备择假设，可以是 "greater"（默认值）、"less" 或 "two.sided"。	greater
spChk	逻辑值，表示是否检查空间对象的一致性。	
resfun	函数，用于计算残差。	
na.action	指定如何处理缺失值的函数。	

Moran's I test 详细使用方法见前述实例。

30.8.2 空间误差回归

在 R 语言中，可以用 spatialreg 包中的 errorsarlm 函数来拟合空间误差滞后模型，并使用 summary 函数来查看模型的摘要信息。errorsarlm 函数 的 R 语言语法如下：

```
errorsarlm(formula, data, listw, na.action, zero.policy=NULL, tol.solve=1e-10, returnHcov=FALSE, trs=NULL, interval=NULL, control=list())
```

表 30-3 中显示了 errorsarlm 函数的参数说明。

表 5-6　errorsarlm 函数的参数

参数	描述	默认值
formula	描述模型的公式，例如 y ~ x1 + x2 + ...。	
data	包含模型变量的数据框。	
listw	空间权重邻接列表对象，定义了空间邻接结构。	
na.action	处理缺失值的方法，常见的有 na.omit、na.exclude 等。	
zero.policy	逻辑值，指定如何处理零权重。TRUE 表示忽略零权重，FALSE 则会在零权重存在时报错。	
tol.solve	解方程时的容差	1e-10
returnHcov	逻辑值，指定是否返回异方差调整的协方差矩阵，默认值为 FALSE。	
trs	用于调试的工具，通常不需要修改。	
interval	用于线性搜索的参数，默认情况下不需要修改。	
control	一个列表，用于控制优化的参数设置。	

详细使用方法见前述实例。

30.8.3 空间滞后模型

在 R 语言中可以用 spatialreg 包中的 lagsarlm 函数来拟合空间滞后模型，并使用 summary 函数来查看模型的摘要信息。lagsarlm 函数的 R 语言语法如下:

```
lagsarlm(formula, data = list(), listw, na.action, Durbin, type, method, quiet, zero.policy=NULL, interval=NULL, tol.solve=1e-10, trs=NULL, control=list())
```

表 30-4 是关于 lagsarlm 函数的参数说明。

表 30-4　lagsarlm 函数的参数说明

参数	描述	默认值
formula	描述模型的公式，例如 y ~ x1 + x2 + ...。	
data	包含模型变量的数据框。	
listw	空间权重邻接列表对象，定义了空间邻接结构。	
na.action	处理缺失值的方法，常见的有 na.omit、na.exclude 等。	
Durbin	可选参数，指定是否包括滞后解释变量，或滞后解释变量的公式。	
type	指定类型，例如 "lag" 表示滞后模型。	
method	指定估计方法，可以是 "eigen", "LU", "Chebyshev" 等。	
quiet	逻辑值，指定是否在计算中保持安静。	
zero.policy	逻辑值，指定如何处理零权重。TRUE 表示忽略零权重，FALSE 则会在零权重存在时报错。	
interval	用于线性搜索的参数，默认情况下不需要修改。	
tol.solve	解方程时的容差，默认值为 1e-10。	1e-10
trs	用于调试的工具，通常不需要修改。	
control	一个列表，用于控制优化的参数设置。	

详细使用方法见前述实例。

参考文献：

1. Jing F, Li Z, Qiao S, Ning H, Lessani MN, Li X. From neighborhood contexts to human behaviors: Cellphone-based place visitation data contribute to estimating neighborhood-level depression prevalence in the United States. Cities. 2024 May 1;148:104905.

2. Bivand R, Piras G. Comparing implementations of estimation methods for spatial econometrics. Journal of Statistical Software. 2015 Feb 16;63:1-36.

扩展阅读与实践：

1. Grekousis, G., Feng, Z., Marakakis, I., Lu, Y., & Wang, R. (2022). Ranking the importance of demographic, socioeconomic, and underlying health factors on US COVID-19 deaths: A geographical random forest approach. Health & Place, 74, 102744.

[Geographical random forest]

2. Comber, A. J., Brunsdon, C., & Radburn, R. (2011). A spatial analysis of variations in health access: linking geography, socio-economic status and access perceptions. International journal of health geographics, 10, 1-11.

[Geographical Weighted Regression]

（敬峰瑞）