社会科学中的空间分析：概念、技术和应用实例[[1]](#footnote-2)

**摘要：**在社会科学的研究中，空间分析占有非常重要的地位。对这一重要概念的忽略，有可能遮蔽我们对于现实世界的更深入了解。但是，遗憾的是，在国内社会科学的相关研究领域中，空间分析的概念一直是缺失的。本文试图简单地梳理空间分析的一些核心概念和分析技术，并通过一个社会组织的实例，演示如何在社会学的分析模型中纳入空间这一概念，以期对学界有所裨益。

**关键词：** 空间分析 空间回归 社会组织

**Spatial Analysis in Social Science: Concept, Technique and Application**

**Abstract**

Spatial analysis is a very important issue in social science research, such as residential segregation, crime, suicide, and urban sub-culture, etc. But unfortunately, this issue is absent in Chinese sociology studies. This paper aims to introduce some key concepts and techniques of spatial analysis. Using a dataset of social organization from Shanghai, this paper gives an example to show how to use spatial analysis in sociology.

**Keywords**: Spatial Analysis, Spatial Regression, Social Organization

**引言**

在西方社会科学研究领域中，空间概念和技术一直得到广泛的应用，包括流行病学、公共卫生与健康、人口学、社会学、政治学、公共管理等学科。在社会科学的发展史上，空间分析占有非常重要的地位。1920年代芝加哥学派的兴起，很大程度上源于对于城市空间居住形态及其机制的阐释，随后发展出来的对于城市居住模式、居住隔离、犯罪与自杀问题、城市亚文化研究等等议题，空间分析的概念一直受到重视。[[2]](#footnote-3)[[3]](#footnote-4)[[4]](#footnote-5)[[5]](#footnote-6)[[6]](#footnote-7)[[7]](#footnote-8)

自创立之初，社会学就格外关注社会底层群体，如城市中少数群体的居住隔离与聚居；城市贫困与贫民区；邻里、社区及地方社会结构在城市发展中的作用；城市中越轨行为的空间分布模式；城市内的居住迁移模式；城市公共服务设施的可达性等等，都是城市社会学研究经久不衰的经典题目，而这些研究领域，都与空间分析概念紧密结合在一起。

在新近的一篇文章中，罗根（John Logan）呼吁社会科学研究更多地关注空间概念。他认为，在当代社会科学的发展过程中，随着大规模包含地理信息的调查数据日益增多，空间分析正在获得强劲的发展，空间分析的可用性也在不断增长，空间分析在社会科学中有更大的研究需求。同时，随着计算机技术的发展，以及统计程序在集群抽样、空间相关性、多层建模和空间回归方面的进展，这一研究领域应该获得越来越多的注意。[[8]](#footnote-9)

在真正的社会科学研究中，很多现实现象具有明显的空间聚类特征，即相似的事情在地理空间上也是彼此接近的。如自杀率具有明显的空间聚集效应，即使控制了很多可能的变量，周边地区实际上还是会影响观测值，这种现象称为“空间滞后”（spatial lag），如果忽略这一空间效应，那么模型的估计结果将会存在很大偏误。[[9]](#footnote-10) 在社会科学的研究工作中，通过空间数据的可视化，可以使学者们很容易发现许多有意思的现象。在西方城市研究的空间分析中，很多都依赖地理信息系统（Geographical Information System，GIS）技术。这一分析技术在空间分析方面具有诸多优势：首先，很多零散的数据一旦做成地图，很多空间特征就会一目了然，从而可以发现一些传统研究方法容易忽略的特征。如在大都市地区不同社会群体在居住和工作在空间分布上的不同形态，可能存在明显的空间不平等现象。其次，快速发展的空间统计方法，如空间相关指数（Moran's I），空间回归分析模型（Spatial Regression Model）等，可以修正传统数据模型中可能存在的空间自相关问题，从而可以让学者对不同变量之间的关系进行更精确的分析。

近年来，社区效应，或者称为邻里效应（Neighborhood Effect），也得到越来越广泛的关注。[[10]](#footnote-11)研究表明，社区环境对个体的影响，并不仅仅是通过使用多层次模型（Multilevel Model）就可以解决，也就是说，不仅仅是个体所在的社区会影响到个体的某些行为特征，他也会受到周围社区的影响，这也是“空间滞后”的影响效果。[[11]](#footnote-12)

遗憾的是，相对于西方社会科学界日益广泛应用的空间概念和空间分析技术，国内社会科学界对空间概念的实证分析是非常缺乏的。究其原因，一方面是因为带有空间属性数据的缺乏，另一方面是因为具体分析技术（包括软件应用）方面的缺乏。[[12]](#footnote-13)本文试图用简单易懂的语言梳理空间分析的一些核心概念和技术，并通过一个现实的社会组织的分析示例，展示如何在社会学中引入空间这一分析概念，以期望对学界有所裨益。

# 一、空间分析模型

在本节，使用一个假设的例子（上海市某中心区的5个街道），简单介绍空间分析中的一些重要概念和计算方法，包括空间相关性的概念、空间相邻性、空间权重、空间滞后量、空间相关系数、空间回归模型。试图通过简单介绍空间分析的基本概念和方法，促进社会学领域学者们对于这一技术的了解，对以后相关议题的研究有所裨益。

**空间相关性**

对于空间相关性的研究，最早起源于对伦敦地区霍乱的分析。19世纪，随着伦敦地区的快速城市化，大量人口涌入伦敦地区。但是，由于公共卫生设施的不足，伦敦受到霍乱的严重冲击。在当时，人们一致认为霍乱是由空气传播的。John Snow对这个观点颇为怀疑, 通过分析1854年英国Broad大街的霍乱爆发地点，他发现，死亡发生地都在Broad街道中部一处水源(水井)周围，有明显的地理规律，而市内其它水源周围极少发现死者。进一步调查，他发现这些死者都饮用过这里的井水。据此，他最终确定了霍乱的源头及传播机制。[[13]](#footnote-14)这一研究开启了学者对于空间相关分析的关注。



图1：1854年伦敦霍乱地图

来源: <http://en.wikipedia.org/wiki/Spatial_analysis#mediaviewer/File:Snow-cholera-map.jpg>.

从技术层面讲，通常情况下，在我们使用最小二乘法进行回归分析的时候，假设我们关心的变量Y有n个观测值y1、y2、y3、……、yn，一般我们假设这些观测值之间是相互独立并且服从同一分布的，均值为μ，方差为σ2（iid假设）。

但是，如果观测值之间存在空间相关，通过经典检验的结果将是有偏的。如果yi之间存在空间相关，如观测值yi与yj在空间上隔得越近相似性越大，那么其方差将会大于σ2；这种相关导致均值的真实标准误近似于：



对于存在空间相关性的数据，如果忽略这种相关性，会使得基于独立同分布假设（iid）的统计检验出现严重的问题，将会导致对数据真实方差（或标准误）的低估：σ<σy；根据公式Z=b/σ，如果σ被低估，那么统计值Z值就会被高估；如高估了统计值，会使得研究者更容易拒绝原假设，而这种拒绝是错误的。[[14]](#footnote-15)而且，更为严重的是，这种作用并不会随着样本量的增大而消失。[[15]](#footnote-16)

**空间相邻性**

假设上海某区的5个街道，分别用1-5来表示（图2）。要标示这5个区域之间的空间相邻性，可以有多种方法，如地理上是否相连、行政中心是否在某个距离之内，是否有地铁直达，是否能步行10分钟到达，等等。这种界定，更多地取决于理论假设，如我们甚至可以根据“街道领导人是否高中同学、是否曾经是同事、是否参加过同一个党校学习”等来界定两个街道是否属于“相邻”状态。

|  |  |
| --- | --- |
| 图2: 上海某中心区5街道地图 | 图3: 上海某中心区5街道的空间相邻图 |

一般来说，我们通常使用空间上是否接壤来界定空间相邻性。最常用的测量空间地理相邻有3种不同的方法：是否有共同的边界（Rook），共同顶点（Bishop），同时包括边界和顶点（Queen）。使用地理空间是否接壤（Queen）的标准来界定这5个街道在地理位置上是否相邻，可以发现，街道1和街道2、4相邻；街道2和街道1、3、4相邻；街道3与街道2、4、5相邻；街道4和其他所有4个街道都相邻；街道5与街道3、4相邻（图3）。可以把上图转化为下表（一个类似于社会网络的表格）：

表1：空间相邻

|  |  |
| --- | --- |
| 街道 | 相邻街道 |
| 1 | 2,4 |
| 2 | 1,3,4 |
| 3 | 2,4,5 |
| 4 | 1,2,3,5 |
| 5 | 3,4 |

再进一步，我们可以把上表用一个5\*5的矩阵C来表示，Cij表示空间区域i与空间区域j是否相邻，Cij =1表示空间区域i与空间区域j处于相邻状态，Cij =0则表示空间区域i与空间区域j不相邻。这一矩阵形式被称为空间连接矩阵C。

表2：空间相邻矩阵

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 | 5 |
| 1 | 0 | 1 | 0 | 1 | 0 |
| 2 | 1 | 0 | 1 | 1 | 0 |
| 3 | 0 | 1 | 0 | 1 | 1 |
| 4 | 1 | 1 | 1 | 0 | 1 |
| 5 | 0 | 0 | 1 | 1 | 0 |

**空间权重**

在连接矩阵C的基础上，将矩阵中每一个数字除以每行的总连接数Ci\*，使得每行加总为1，这一矩阵称为行标准化的空间权重矩阵W。

表3：行标准化的空间权重矩阵

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 | 5 |  | 行百分比 |
| 1 | 0 | 0.5 | 0 | 0.5 | 0 |  | 1 |
| 2 | 0.33 | 0 | 0.33 | 0.33 | 0 |  | 1 |
| 3 | 0 | 0.33 | 0 | 0.33 | 0.33 |  | 1 |
| 4 | 0.25 | 0.25 | 0.25 | 0 | 0.25 |  | 1 |
| 5 | 0 | 0 | 0.5 | 0.5 | 0 |  | 1 |

**空间滞后量**

给定一个特定的观测值Y，如街道的人口数、GDP、社会组织数量等任意一个我们感兴趣的指标，将街道i周围相邻单位j的观测值乘以相对应的空间权重Wij后相加，得到街道i所有相邻观测值的平均值，Yis=ΣWij\*Yi，称为空间单位（街道）i的空间滞后变量，也可以记做Ys=WY。

以街道1为例，它与街道2、4相邻，街道2、4的Y值（如人口）分别为25、20，因此街道1的空间滞后为25\*0.5+20\*0.5=22.5，这表示街道1周围街道的平均人口为22.5。

表4：空间滞后量

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 | 5 | Y的空间滞后 |
| 1 | 0 | 0.5 | 0 | 0.5 | 0 | 22.5 |
| 2 | 0.33 | 0 | 0.33 | 0.33 | 0 | 26.4 |
| 3 | 0 | 0.33 | 0 | 0.33 | 0.33 | 16.5 |
| 4 | 0.25 | 0.25 | 0.25 | 0 | 0.25 | 22.5 |
| 5 | 0 | 0 | 0.5 | 0.5 | 0 | 15.0 |
| Y值 | 50 | 25 | 10 | 20 | 5 |  |

**空间相关系数**

以观测值Y为横坐标，以Y的空间滞后量Ys为纵坐标，计算Y与Ys之间的相关系数，就是空间相关系数，也称作Moran’s I系数：[[16]](#footnote-17)



其中wij是行标准化的空间权重矩阵，y是我们所关注的变量。

Moran’s I是一种基于全局分析的空间自相关（Global Spatial Autocorrelation）的测量指标，表示该区域空间聚集或离散的总体情况。Moran’s I 指标的值介于-1到1之间，当指标相似的地区存在空间聚集时，Moran’s I为正（正相关），表示观测值高的地区其周围的观测值也高，观测值低的地区其周围的观测值也低；当指标相反的地区存在空间聚集时，Moran’s I为负（负自相关），表示观测值高的地区，其周围的观测值偏低；而当不相关时，为-1/(n-1)（当n足够大时，其值接近于0）。[[17]](#footnote-18)

通过下图可以看出，我们上面例子中的变量Y与其滞后变量之间存在比较明显的正相关关系。某个地区单位（街道）的Y（人口数）越大，其周围街道的人口数也越大。如街道1的人口数为50，周围街道的平均人口数为22.5；而街道5的人口数为5，周围街道的平均人口数为15。

|  |  |
| --- | --- |
|  |  |

图4: Moran’s I及显著性检验

如同常规的相关系数一样，Moran’s I系数在统计上是否显著，也是可以进行统计检验的。对于Moran’s I系数的检验，一般采取两种方法。一种是假设数据分布是正态的，[[18]](#footnote-19)一种是假设数据分布是随机的，使用蒙特卡洛模拟（Monte Carlo Simulation）进行检验。[[19]](#footnote-20)通常情况下这两种方法会得到相似的结果。使用第二种方法，对上述5个街道的Y进行空间相关性检验，得到Moran’s I系数为2.37，p值为0.0089。即这5个街道的Y存在显著的空间聚集现象，Y值比较大的街道，其周围街道的平均Y值也比较大，反之亦然。

**空间回归模型**

如前所示，如果观测值存在显著的空间相关，那么就不应忽略这种会影响统计结果的因素，而应将之纳入统计分析模型。在过去几十年中，统计学家已经发展出来了比较成熟的空间回归模型，来处理这个问题。在当今的社会科学界，有2种空间回归模型应用最广：空间滞后模型与空间误差模型。

传统的最小二乘法回归（OLS）可以如下表达：

Y=Xβ+ε………………………………………………………………………………（1）

如果我们将观测变量Y的空间滞后变量WY认为是Y的一种本质特征，而不是一种统计上的干扰，即某一特定观测单位i周围空间单位的平均Y会对地区i的Y产生影响，如GDP高的地区，其周围地区的平均GDP也会比较高，那么，可以把这种空间滞后变量的影响效果独立出来，并纳入回归模型进行分析。可以将上式中的误差项分解为两项：因变量的空间滞后项和自变量的误差，即：ε=ρWY+μ

代入式（1），就得到空间滞后回归模型（spatial lag model）：

Y=Xβ+ ρWY+μ ………………………………………………………………………………（2）

ρ表示因变量Y的空间滞后项WY对因变量的影响效果。

或者，如果我们把空间相关看做一种统计干扰，认为模型中的误差项虽然在空间上是相关的，但并非由因变量的空间滞后导致，那么，可以将误差项拆分为两个部分：包含空间因素的误差项与空间不相关的误差项，即：ε=λWξ+μ

代入式（1），就得到空间误差回归模型（spatial error model）：

Y=Xβ+ λWξ+μ ………………………………………………………………………………（3）

λ表示相邻观测值的空间误差项Wξ的相关程度，如果观测值i和观测值j之间的误差项在空间上不存在相关关系，则λ=0，上式可简化为传统的OLS模型。

# 二、一个例子：上海市社会组织的空间分析

在本节中，作者使用上海市2010年社会组织的数据，进行一个实际的空间回归分析，对空间分析的应用进行一个演示。

**数据说明**

上海市社会组织的数据，来自于上海社团管理部门网站（“上海社会组织”网站[[20]](#footnote-21)）所公示的上海所有在册社会组织信息。通过“网络爬虫”，我们获得了上海所有在册社会组织的如下变量：组织名称、组织注册代码、注册时间、证书有效时间、组织类型、注册地、主管单位、法人代表、地址、邮编、电话、网址、主要业务内容以及奖惩情况等变量信息。网络爬取时间为2014年1月1日。

国内社会组织分为三种类型，包括“民办非企业组织”、“社会团体”和“基金会”。民办非企业组织，是指企业事业单位、社会团体和其他社会力量以及公民个人，利用非国有资产举办的从事非营利性社会服务活动的社会组织，如各类民办学校、医院、文艺团体、科研院所、体育场馆、职业培训中心、福利院、人才交流中心等。[[21]](#footnote-22)社会团体是指我国公民行使结社权利自愿组成，为实现会员的共同意愿，按照其章程开展活动的非营利性社会组织，包括各类使用学会、协会、研究会、促进会、联谊会、联合会、商会等。[[22]](#footnote-23)基金会，是指利用[自然人](http://baike.baidu.com/view/15620.htm)、[法人](http://baike.baidu.com/view/10873.htm)或者其他组织捐赠的财产，以从事[公益事业](http://baike.baidu.com/view/360377.htm)为目的，按照规定成立的[非营利性](http://baike.baidu.com/view/1867460.htm)法人。[[23]](#footnote-24)

在本文中，仅仅分析2010年及之前成立的社会组织。[[24]](#footnote-25)从2010年数据来看，上海市这三类组织的数目分别为5762、3399、113。根据上海市社会团体管理局发布的数据，2011年1月，这三个数字分别为6190，3568，116，与我们的数据库结果基本一致。[[25]](#footnote-26)在本文的实际分析中，仅保留了“民办非企业组织”，作为分析的示例。[[26]](#footnote-27)

本文使用的软件为R，下文所有结果均由R软件计算得到。[[27]](#footnote-28)

**自变量**

用于分析的自变量数据，主要来自于《中国2010年人口普查分乡、镇、街道资料》中的上海部分数据，包括总人口、14岁以下人口、60岁以上人口、户籍人口等几个核心指标。

还有一部分地理信息的自变量，通过R中的GIS包生成，包括街道/乡镇面积、是否属于乡镇（虚拟变量，乡镇=1。在上海，乡镇具有独立的财政，而街道没有），是否属于中心城区（虚拟变量，中心城区=1）。

**社会组织的空间分布**

在进行模型分析之前，首先进行分析数据的前期准备工作，包括：（1）读取上海地图数据，一般是GIS软件通用的shape格式（polygon）；（2）读取社会组织数据，通常是外部文件，包括stata格式、文本格式、excel格式等等，然后转化为shape格式的点图（point）格式；（3）读取人口数据，通常为excel格式；（4）将人口数据与地图数据合并、生成我们需要的一系列测量指标（如街道/乡镇面积、是否属于乡镇、是否属于中心城区），等。

在数据准备完成之后，可以先进行数据的地图展示，比如社会组织的空间分布、历史发展、频数分布等。

|  |  |
| --- | --- |
|  |  |

图5: 上海社会组织分布图

从上图可以看出，上海社会组织的分布，存在明显的空间聚集，具体表现为：（1）在中心城区，存在密集的空间分布形态；（2）同样，在各个区县的中心地区，社会组织的分布更密集一些，呈现明显的中心化趋势。

从成立年份可以看出，民办非企业组织在2000年之后开始迅猛发展，且一直保持比较高的发展速度，每年都有500个左右新的民办非企业组织成立。

下图是民办非企业组织的分布密度图。从图中可以看出，虽然这个变量的分布不是完全正态分布，但比较接近正态分布，因此本文使用线性模型（Linear Model）进行估计。[[28]](#footnote-29)

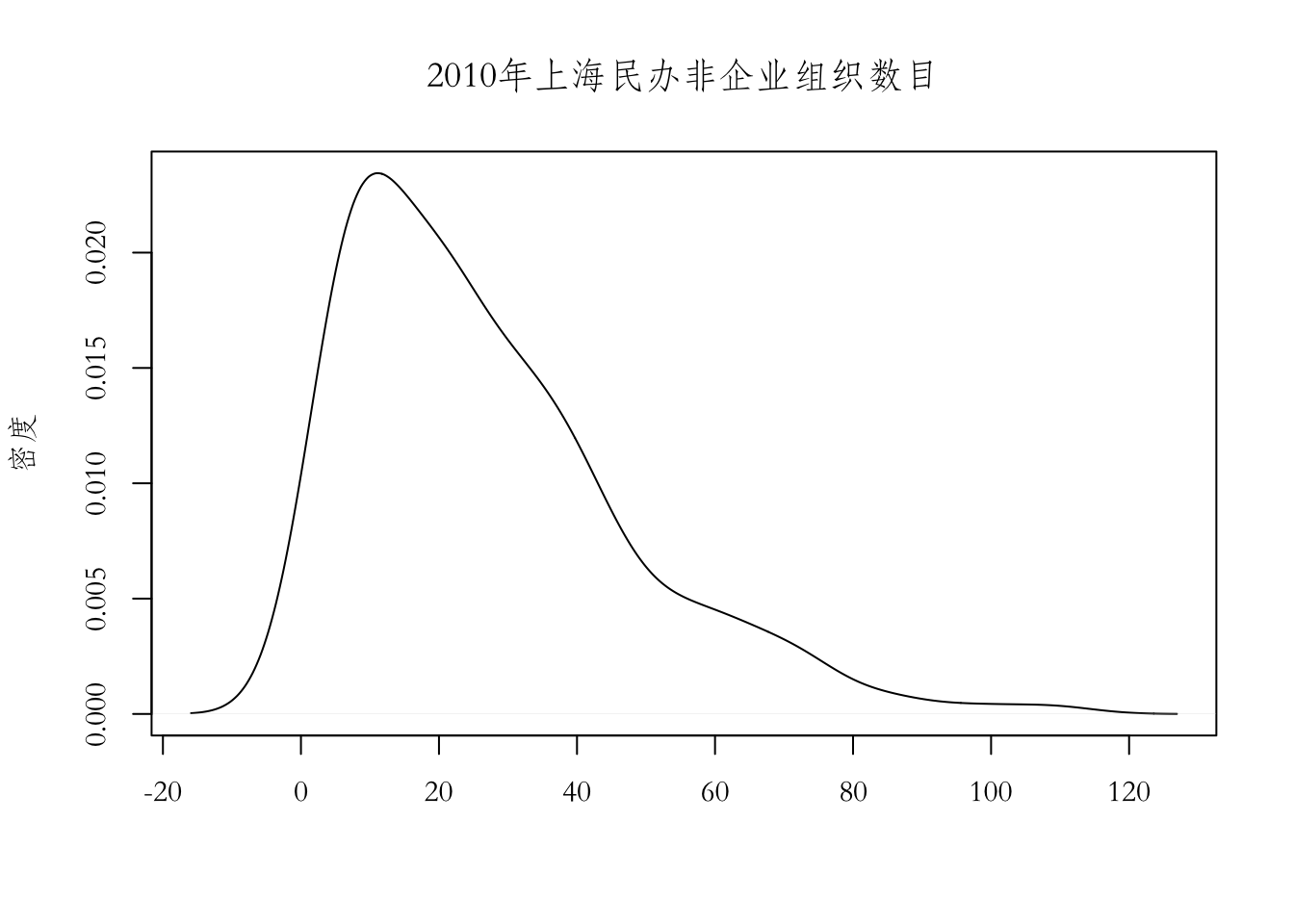


图6: 上海民办非企业组织的分布密度

**空间相邻与空间权重**

空间权重矩阵是进行空间回归分析的必要前提。引入不同的空间权重矩阵，可能会导致不同的模型结果（当然，对于不同空间权重的界定，更过需要来源于理论上对于观测变量的理解。）本文使用地理空间是否接壤（Queen）来界定上海市街道的空间相邻。利用空间相邻矩阵C，生成用于后文分析的空间权重矩阵W，空间权重矩阵采用行标准化的方法。下图展示了本文使用数据的空间相邻图：

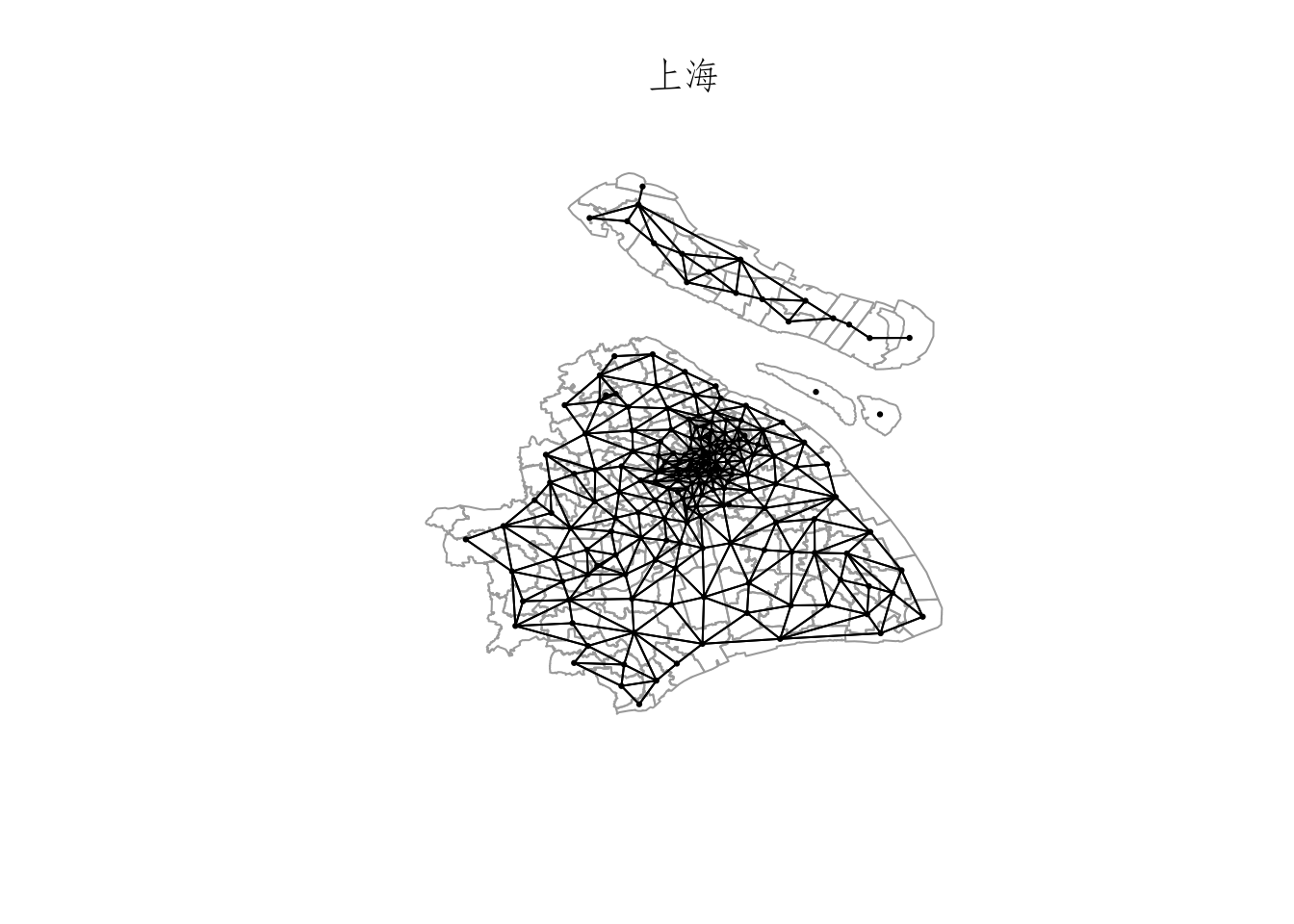


图8: 上海街道的空间相邻图

**Moran’s I 检验**

我们关注的观测变量，是否存在空间相关性，可以使用Moran’s I系数进行检验。如果检验不显著，说明并不存在明显的空间相关，那可以使用传统的OLS进行估计；如果检验显著，则认为存在明显的空间相关关系，也就是说，某个街道的社会组织数目，受到周围街道社会组织数目的影响，在这种情况下，忽略空间影响的效果，OLS估计的结果将是有严重问题的，需要使用空间回归模型来进行估计。

通过Moran’s I检验，“民办非企业组织”的Moran’s I统计值为7.76，在0.0001的水平上显著，也就是说，“2010年上海市民办非企业组织”这一观测值在空间上存在统计显著的空间聚集效应，民办非企业组织数目多的街道，其周围街道的民非组织平均数也高。

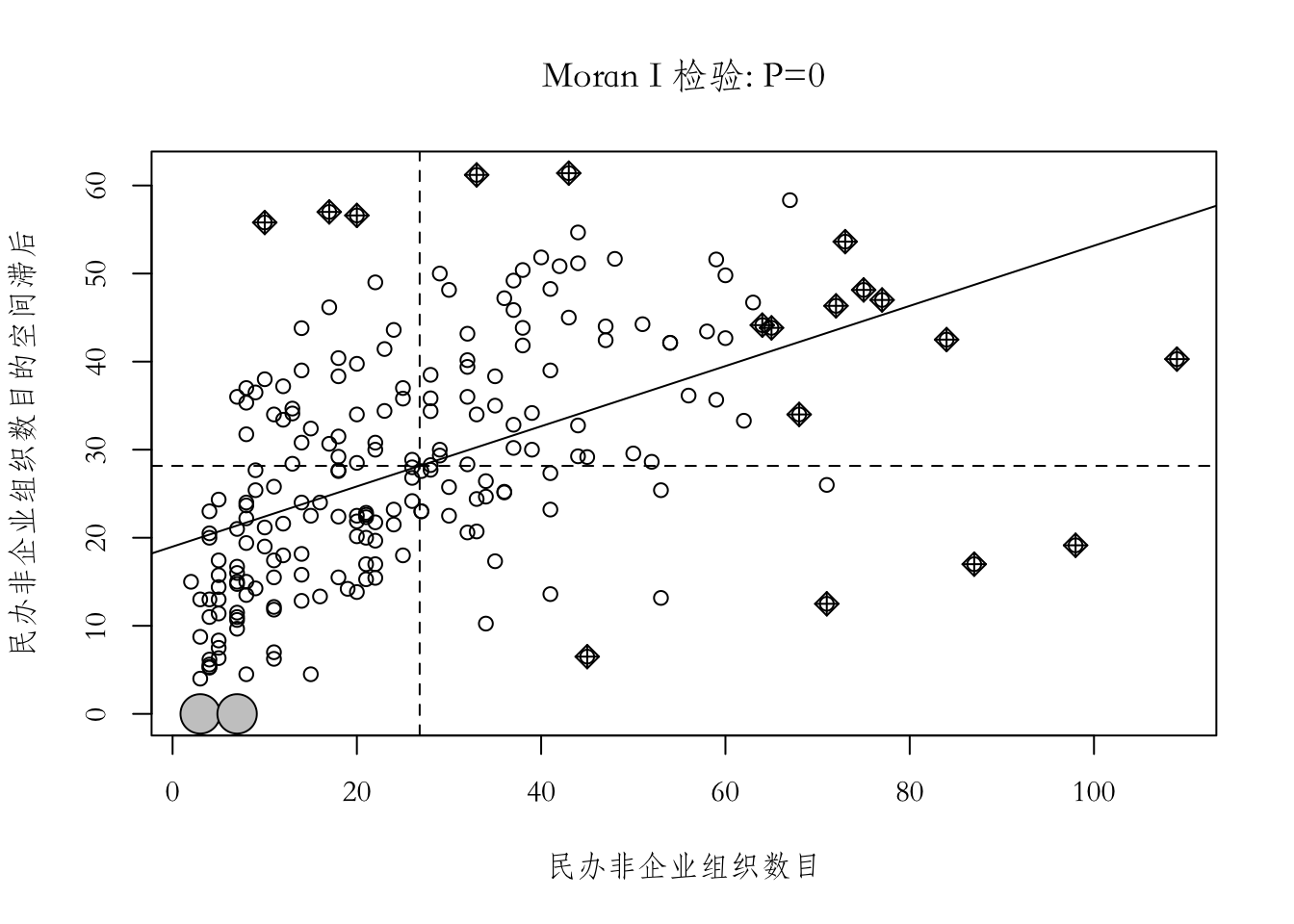


图9: 上海民办非企业组织的空间相关性检验

**空间回归模型结果**

从上图的Moran’s I检验可以看出，“民办非企业组织”存在明显的空间滞后效应，社会组织多的街道/乡镇，其周围街道/乡镇的社会组织数目也比较高。因此，在这种情况下，对于社会组织的分析，应用将空间滞后变量纳入模型估计当中，需要使用空间回归模型而非OLS模型来进行估计。

对于空间滞后模型或者空间误差模型的选择，更多取决于理论的假设。如果我们把空间相关看做一种统计干扰，那么选择空间误差模型；如果我们将观测变量的空间滞后变量认为是影响观测变量的一种本质特征，即某一特定观测单位周围空间单位的数值会与周围互相影响，则选择空间滞后模型。一般来说，社会科学家更多倾向于选择空间滞后模型，即把观测变量的空间滞后项作为影响观测变量的因素纳入模型。本文仅考虑空间滞后的回归分析。

事实上，对于空间回归模型的选择，可以进行一个统计检验，看具体是存在空间滞后效应还是空间误差效应。LMerr与RLMerr用来测量是否存在空间误差效应；LMlag和RLMlag用来测量是否存在空间滞后效应。前者显著，则认为存在空间误差效应；后者显著，则认为存在空间滞后效应。如果LMerr和LMlag的检验都是显著的，那需要进一步考察RLMerr和RLMlag的检验结果。[[29]](#footnote-30)本文也进行了检验，结果显示，模型存在明显的空间滞后效应，而非空间误差效应。

对于空间滞后效应回归模型的估计，有两种不同的估计方式：一种是使用最大似然法（Maximum Likelihood， ML）；另外一种是二阶段估计方法（two stage OLS），通过把解释变量的空间滞后项作为因变量的工具变量（IV），在同步联立方程组中使用二阶段最小二乘法来拟合模型。[[30]](#footnote-31)前者在R中使用“spdep”包的“lagsarlm”命令，后者使用“stsls”命令，本文使用第二种方法进行估计。

在空间回归模型Y=Xβ+ ρWY+μ中，系数ρ表示Y的空间滞后项的影响效果，系数ρ显著则表示存在空间滞后效应；系数ρ不显著，则表示不存在空间滞后效应，即空间滞后项对Y没有显著影响。下表分别汇报了最小二阶段回归模型（OLS）、使用二阶段估计方法的空间滞后回归模型（two stages Spatial Lag Model）的结果。[[31]](#footnote-32)

表5：民办非企业组织的回归结果

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | OLS | | |  | Spatial Lag Model | | |
|  | b | SE | Pr(>|t|) |  | b | SE | Pr(>|t|) |
| ρ |  |  |  |  | 0.4722 | 0.1464 | 0.0013 |
| (Intercept) | 14.4800 | 3.4637 | 0.0000 |  | 0.1161 | 5.5563 | 0.9833 |
| 人口 | -1.6870 | 0.8075 | 0.0379 |  | -1.6387 | 0.7744 | 0.0344 |
| 14以下人口 | 23.3240 | 6.9178 | 0.0009 |  | 21.1487 | 6.6678 | 0.0015 |
| 60岁以上人口 | -6.7240 | 6.9279 | 0.3329 |  | -8.4289 | 6.6643 | 0.2059 |
| 户籍人口 | 3.5690 | 1.4792 | 0.0167 |  | 3.9843 | 1.4243 | 0.0052 |
| 面积 | -409.5460 | 537.7115 | 0.4471 |  | 56.6320 | 535.5031 | 0.9158 |
| 乡镇=1 | -9.4830 | 3.2974 | 0.0045 |  | -5.0925 | 3.4426 | 0.1391 |
| 中心城区=1 | 15.2260 | 3.3682 | 0.0000 |  | 11.8124 | 3.3989 | 0.0005 |

从模型结果可以看出，“民办非企业组织”的二阶段估计方法的空间滞后回归模型（Spatial Lag）的结果显示，系数ρ是非常显著的。这说明，街道i周围街道的社会组织数目，会对街道i的社会组织数产生显著的影响；而且，这一系数的符号为正，说明街道的社会组织数目与周围的社会组织数目存在正向的相关关系，即周围的社会组织数目越多，街道本身的社会组织数目也会越多。

在这种情况下，如果仅仅使用OLS方法进行分析，会忽略非常重要的空间聚集效应，在这种情况下，OLS的估计是有问题。对于这一点，从模型结果的常数项可以看出，在OLS模型中，常数项是非常显著的，而在控制空间滞后效应滞后，常数项则变得不显著。

从对现实世界的实质理解上，这表明，社会组织的发展存在明显的空间聚集，这也许与地方政府之间的组织模仿行为有关。[[32]](#footnote-33)[[33]](#footnote-34)[[34]](#footnote-35)在经验模型中，如果忽略这点，对于经验世界的理解是有问题的。

从民办非企业组织的回归结果可以看出，在加入空间之后效应之后，原先在OLS模型中显著的变量如“是否乡镇”变得不显著，说明这一变量的作用与空间效应是重合的，在控制空间效应之后，这一变量对于民非组织的影响效果消失了。在这种情况下，如果忽略空间滞后效果，仅仅使用OLS模型进行估计，得到的结果是有问题的。其余几个在空间回归模型中仍然显著的变量，在控制空间滞后效应滞后，其系数明显减少，如总人口的影响效果（绝对值）从1.7减少为1.6，14岁以下人口的效应从23.3减少为21.1，中心城区的效果（绝对值）从9.5减少为5.1。

从其他人口变量来看，一个有意思的发现是“民办非企业组织”对14岁以下人口比较敏感，而对60岁以上人口不敏感。即60岁以上人口的数量并不会影响到“民办非企业组织”的数目，而14岁以下人口数量越多，“民办非企业组织”的数量也会越多。如果我们进一步考察针对幼儿和年轻人的社会组织分布，以及针对老年人的社会组织分布，会发现这一现象更加明显：针对幼儿和年轻人的社会组织空间分布，与14岁以下人口的空间分布高度吻合；而针对老年人的社会组织分布，与老年人口的空间分布密度，在空间上基本不相关。当然，对于社会组织分布形态的更详尽的分析，超出了本文的容量，作者拟在后续的研究中进一步探讨。

**三、小结**

一直以来，在社会科学的很多重要议题中，如居住隔离、犯罪与自杀问题、城市亚文化研究等等，空间分析的概念和技术都扮演着非常重要的角色。但是，遗憾的是，在国内社会科学尤其是社会学的研究领域中，空间分析的概念一直付诸阙如。而对这一重要概念的忽略，有可能遮蔽我们对于现实世界的更深入了解。因此，本文试图简单地梳理空间分析的一些核心概念和分析技术，并通过一个社会组织的实例，演示如何在社会学的分析模型中纳入空间这一概念，以期对学界有所裨益。

当然，我们需要看到，将空间分析引入中国的社会科学研究，仍然面临一些问题。首先，从理论层面来看，国内对于城市社会学的关注，具有非常明显的人文主义色彩，对于比较量化的实证分析则比较缺乏，尤其是对于西方城市研究中已经非常成熟的社会区因子分析的研究范式，在国内社会学的城市研究中尚未得到足够重视，这种情况，在某种程度上可能会阻碍我们对于当今中国快速城市化背景下的城市研究。

其次，从数据方面来看，在国内，获取含有GIS信息的社会科学技术比较困难，在西方，这部分工作多由政府来完成，如含有GIS信息的普查数据的公开发布，而这一点，在国内仍然具有一定的困难，尤其是对于社会学家而言。

再次，从技术层面来看，对于空间分析需要的技术方法、软件操作等等，都对社会科学研究人员提出了新的要求，如ArcGIS软件，R软件等。这些知识的掌握，需要一个学科的人才培养、课程设置等方面都做出一系列的调整。

最后，空间概念如何纳入现有研究议题中，仍然是一个需要不断摸索、讨论的过程。西方的城市空间研究，已经形成了定性与定量相结合的发展模式。国内城市社会学的研究，即使引入量化研究的范式，也应该从理论层面讨论中国城市发展的形成机制与内在逻辑，只有这样，我们才会形成真正有学术生命力的城市研究。

任重而道远。

附录：R软件的程序代码

# 导入上海地图 #

library(maptools)

shanghai<- readShapeSpatial("/town")

shanghai.t<- readShapeSpatial("/county")

# 导入社会组织数据：STATA格式#

library(foreign)

t0<-read.dta("/NGO-SH-2013.dta")

t0<-subset(t0,t0$fyear<=2010)

levels(t0$orgtype)<-c("Social Group","NGO","Foundation")

#将STATA数据转化为地图 #

t1<-SpatialPointsDataFrame(data.frame(as.numeric(t0$lon),as.numeric(t0$lat)),t0)

# 计算每个街道中的社会组织数目 #

k=over(t1,shanghai)

t1@data=data.frame(t1@data,k$CODE)

names(t1)[names(t1)=="k.CODE"]="CODE"

t1@data$n=1

number=tapply(t1@data$n,data.frame(t1@data$CODE,t1@data$orgtype),length)

number[is.na(number)]=0

dimnames(number)[2][[1]]<-c("y0","y1","y2")

number<-data.frame(number)

number$code=as.numeric(row.names(number))

#画地图 #

plot(shanghai, border = "grey60",lwd=.5)

plot(t1, add = TRUE, pch = 19, cex = 0.1,alpha=.5,col="tomato1")

plot(shanghai.t,add=T, border="royalblue")

title(main="2010年上海社会组织分布",font.main= 4, col.main= "blue")

kk<-table(t0$fyear,t0$orgtype)

kk<-data.frame(kk)

names(kk)<-c("year","type","freq")

kk<-kk[as.numeric(as.character(kk$year))>=1990,]

kk<-subset(kk,kk$type=="NGO")

library(ggplot2)

mytheme<-theme\_bw()+

theme(panel.border=element\_blank(),

axis.line=element\_line(colour="black"))

ggplot(kk,aes(x=year,y=freq,shape=type,group=type))+

geom\_line()+

geom\_point(size=4,colour="blue")+

mytheme+

ggtitle("Social Organization")

plot(density(number$y1), main="2010年上海民办非企业组织",xlab="", ylab="密度")

# 读取自变量数据：Excel格式 #

popu<-read.csv("/population2010.csv")

popu[,c(2:11)]<-popu[,c(2:11)]/10000

shanghai=shanghai[!is.na(shanghai$CODE),]

shanghai@data["code"]=as.numeric(as.character(shanghai@data[,"CODE"]))

shanghai@data = merge(shanghai@data,popu,by.x="code",by.y="code",all.x=TRUE,sort=F)

shanghai@data = merge(shanghai@data,number,by.x="code",by.y="code",all.x=TRUE,sort=F)

# 生成新的自变量 #

area=sapply(slot(shanghai,"polygons"),slot,"area")

shanghai@data=data.frame(shanghai@data,area=area)

shanghai@data$density=(shanghai@data$population/shanghai@data$area)/100

shanghai@data$male.r=shanghai@data$male/shanghai@data$female

shanghai@data$hh.r=shanghai@data$hhnum/shanghai@data$population

shanghai@data$popu1.r=shanghai@data$popu1/shanghai@data$population

shanghai@data$popu3.r=shanghai@data$popu3/shanghai@data$population

shanghai@data$migrant.r=(shanghai@data$population-shanghai@data$local)/shanghai@data$local

shanghai@data$rural=substr(as.character(shanghai@data$CODE),7,7)

shanghai@data$rural[shanghai@data$rural=="5"]="0"

shanghai@data$rural[shanghai@data$rural=="2"]="1"

shanghai@data$center<-0

shanghai@data$center[round(shanghai@data$code/1000000)<310112]<-1

# 生成空间邻近矩阵数据和空间权重矩阵数据 #

library(spdep)

shanghai\_nb <- poly2nb(shanghai)

shanghai\_wt <- nb2listw(shanghai\_nb, zero.policy=T)

plot(shanghai, border = "grey60")

plot(shanghai\_nb, coordinates(shanghai), add = TRUE, pch = 19, cex = 0.3);title(main="上海")

# 空间相关性检验：Global Moran's I #

moran.test(shanghai$y1,listw=shanghai\_wt,zero.policy=T)

moran.plot(shanghai$y1,listw=shanghai\_wt,labels=F,zero.policy=T,main=paste("Moran I 检验: P=",round(moran.test(shanghai$y1,listw=shanghai\_wt,zero.policy=T)$p.value,4),sep=""),xlab="民办非企业组织数目",ylab="民办非企业组织数目的空间滞后")

# 空间回归模型 #

y1.lm<-lm(y1~population+popu1+popu3+local+area+rural+center,data=shanghai)

y1.lag2<-stsls(y1~population+popu1+popu3+local+area+rural+center,data=shanghai,listw=shanghai\_wt,zero.policy=T)

1. 作者简介：孙秀林（1978-），上海大学社会学院副研究员，主要研究方向为组织社会学、空间分析。

   2 基金项目：本文是教育部哲学社会科学研究重大课题攻关项目“新时期加强社会组织建设研究”（项目编号：11JZD027）的阶段性成果。感谢陈伟、梁海祥、张璨、邹韵、肖亦宗、施润华等多位同学在数据收集、数据清理过程中的努力工作。感谢上海大学“公共性论坛”中各位同仁的意见。文责自负。 [↑](#footnote-ref-2)
2. F.W. Boal, “Ethnic Residential Segregation”, In D.T. Herbert and R. J. Johnston(ed), *Spatial Processes and Form*, John Wiley & Sons, 1976. [↑](#footnote-ref-3)
3. R. A. Murdie, “Spatial Form in the Residential Mosaic”, In D.T. Herbert and R. J. Johnston(ed), *Spatial Processes and Form*, John Wiley & Sons, 1976. [↑](#footnote-ref-4)
4. John Logan, R.D. Alba, and W. Zhang, “Immigrant Enclaves and Ethnic Communities in New York and Los Angeles”, In *American Sociological Review*, vol.67 (2002), pp.299-322. [↑](#footnote-ref-5)
5. John Logan and Limei Li, “The Impact of Housing Tenure on Residential Segregation in Beijing, China”, in Thomas Maloutas and Kuniko Fujita(ed),  *Residential Segregation Around the World: Why Context Matters*, London: Ashgate Publishing, 2012, pp.69-88. [↑](#footnote-ref-6)
6. Haiwang Zhou and Ying Bo, *The Patterns, Causes, and Policies of the Separation of Registered and Actual Residences in Shanghai*, Shanghai: Shanghai Social Science Academy, 2012. [↑](#footnote-ref-7)
7. [美]罗根、张玮玮，徐宏伟，《空间概念及其思考方法在近期人口学研究中的应用》，载梁在主编《人口学》，第九章，北京：中国人民大学出版社2012年。 [↑](#footnote-ref-8)
8. John Logan, “Making a Place for Space: Spatial Thinking in Social Science,” in *Annual Review of Sociology*, vol.38 (2012), pp.507-524. [↑](#footnote-ref-9)
9. R.D. Baller, L. Anselin , S.F. Messner, G. Deane, and D. F. Hawkins, “Structural covariates of US county homicide rates: incorporating spatial effects,” in *Criminology*, vol. 39 (2001), pp.561–590. [↑](#footnote-ref-10)
10. Robert Sampson, *Great American City*, Chicago and London: The University of Chicago Press, 2012. [↑](#footnote-ref-11)
11. Basile Chaix, Juan Merlo, S V Subramanian, John Lynch, Pierre Chauvin, “Comparison of a spatial perspective with the multilevel analytical approach in neighborhood studies: the case of mental and behavioral disorders due to psychoactive substance use in Malmo, Sweden, 2001”, in *American Journal of Epidemiology*, vol.162, no.2, (2005), pp.171-182. [↑](#footnote-ref-12)
12. 孙秀林：《城市研究中的空间分析》，《新视野》2015年第1期。 [↑](#footnote-ref-13)
13. <http://zh.wikipedia.org/zh-cn/%E7%BA%A6%E7%BF%B0%C2%B7%E6%96%AF%E8%AF%BA_(%E5%8C%BB%E5%AD%A6%E5%AE%B6> [↑](#footnote-ref-14)
14. [美]沃德、格里蒂奇：《空间回归模型》，宋曦译，上海：格致出版社2012年版。 [↑](#footnote-ref-15)
15. Noel Cressie, *Statistics for Spatial Data (Rev.ed.),* New York: Wiley, 1993. [↑](#footnote-ref-16)
16. P. A. P. Moran, 1950. “Notes on Continuous Stochastic Phenomena”, in *Biometrika, vol. 1* (1950), pp.17–23. [↑](#footnote-ref-17)
17. M. F. Goodchild, *Spatial Autocorrelation,* Catmog 47, Geo Books, Norwich, 1986. [↑](#footnote-ref-18)
18. A.D. Cliff and J. K. Ord. “Evaluating the Percentage Points of a Spatial Autocorrelation Coefficient”, in *Geographical Analysis*, vol. 4(1971), pp.51-62. [↑](#footnote-ref-19)
19. Roger S Bivand, Edzer J. Pebesma and Virgilio Gomez-Rubio, *Applied Spatial Data Analysis with R,*  Springer US, 2008. [↑](#footnote-ref-20)
20. http://stj.sh.gov.cn/Index.aspx [↑](#footnote-ref-21)
21. http://www.shstj.gov.cn/Info.aspx?ReportId=7cdb4ece-9de2-4bae-90f3-1787f2b82ce8 [↑](#footnote-ref-22)
22. http://www.shstj.gov.cn/Info.aspx?ReportId=7a2464d1-e3a6-457f-8535-6ce7e3e6fdc0 [↑](#footnote-ref-23)
23. http://baike.baidu.com/view/21375.htm?fr=aladdin [↑](#footnote-ref-24)
24. 之所以选取2010年作为分析，是因为我们手头的自变量主要来自于2010年六普数据。 [↑](#footnote-ref-25)
25. http://www.shstj.gov.cn/Info.aspx?ReportId=e0a28450-8abd-4bd8-a24c-85f59dab3b71 [↑](#footnote-ref-26)
26. 本文所有代码都可以在附录中查到。 如果读者对本文的数据、模型、代码有兴趣，欢迎给作者来信索取。 [↑](#footnote-ref-27)
27. R Core Team, *R: A language and environment for statistical computing,* R Foundation for Statistical Computing, Vienna, Austria, 2014, URL <http://www.R-project.org/>. [↑](#footnote-ref-28)
28. 根据观测变量的分布，空间回归模型已经发展出了适用于二项分布、poisson分布等的不同模型。这些模型相对更加复杂，本文不进行介绍。 [↑](#footnote-ref-29)
29. Lucas Anselin, I. Syabri, and Y. Kho, “GeoDa: An Introduction to Spatial Data Analysis”, in *Geographical Analysis*, vol.38(2006),pp.5-22. [↑](#footnote-ref-30)
30. Lucas Anselin, *Spatial Econometrics: Methods and Models,*  Kluwer, Dordrecht, 1988. [↑](#footnote-ref-31)
31. 为了检验模型的稳健性（Robustness），本文尝试使用了不同的变量界定方式，如取百分比方式、取log方式等，模型结果都是一致的。 [↑](#footnote-ref-32)
32. Paul J. DiMaggio and Walter W. Powell, “The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields”, In *American Sociological Review*, vol. 2(1983), pp.147-160. [↑](#footnote-ref-33)
33. Xueguang Zhou, “Occupational Power, State Capacities and the Diffusion of Licensing in the American States: 1890-1950”, in .*American Sociological Review,* vol. 58(1993), pp.536-552. [↑](#footnote-ref-34)
34. Dirk M. Zorn, “Here a Chief, There a Chief: The Rise of the CFO in the American Firm”, in .*American Sociological Review,* vol. 69(2004), pp.345-364. [↑](#footnote-ref-35)