阿片类处方药在空间和时间上的相关性

S610 项目

David Terner*

Jiacheng Zhong†

Fall 2019

GitHub Repository : https://github.com/EphaneZhong/Research_Proposal

目录

*Department of Economics, Indiana University Bloomington

†Department of Operations and Decision Technologies, Indiana University Bloomington

1 研究建议书概要

1.1 数据描述

1. Medicare D 类阿片类药物处方数据: 2013-2017 年每年州、县和邮政编码一级的阿片类药物处方率的年度数据。这些数据是去标识化的。阿片类药物处方率是利用医疗服务提供者开具的 Medicare D 类索赔数据得出的。

2. Medicaid 阿片类药物处方数据: 与上述医疗补助数据类似，Medicaid 数据是关于 2013 年至 2017 年期间去掉身份标识的 Medicaid 阿片类药物报销数据。不过，这些数据只是州一级的数据。阿片类药物处方率是利用医疗服务提供者开出的处方药的 Medicaid 数据推算出来的，并由各州向 Medicare 和 Medicaid 服务中心（CMS）报告。(县和邮政编码一级的更精细的数据或许可以用于该项目。)

3. 美国的收入和失业数据：2013 年至 2017 年，每个州的每个县都有家庭收入中位数和失业率。

4. 美国的人口数据：2013 年至 2017 年，一个州的每个县的人口都有县和邮编一级的数据。

5. “zipcode” R-包: 该包有助于识别任意两个县之间的距离。[1]

1.2 方法

1.2.1 Moran's I 指数

Moran's I 指数被定义为

$$I = \frac{N}{W}\frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2} \tag{1}$$

这里 x_i 是目标 i 的位置变量, N 是穿越指标 i 和 j 的独特的位置数目, w_{ij} 是从位置 i 到 j 的测量 (注 $w_{ii} = 0$), 并且 $W = \sum_i \sum_j w_{ij}$.[2] (1) 实际上是一个关于距离相关性的统计指标。对统计指标 I 最好的检验就是:

$$z_I = \frac{I - \mathbb{E}[I]}{\sqrt{V[I]}} \tag{2}$$

其中

$$\mathbb{E}[I] = -\frac{1}{N-1} \quad \text{并且} V(I) = \underbrace{\frac{NS_4 - S_3 S_5}{(N-1)(N-2)(N-3)W^2}}_{=\mathbb{E}[I^2]} - (\mathbb{E}[I])^2$$

1. 我们用的是 `housingData` library, 其中包含县域中心坐标 (alliteration not intended).
2. 参考 Moran 1950。

另外

$$S_1 = \frac{1}{2}\sum_i\sum_j (w_{ij}+w_{ji})^2 \qquad S_4 = (N^2-3N+3)S_1 + NS_2 + 3W^2$$

$$S_2 = \sum_i \Big(\sum_j w_{ij} + \sum_j w_{ji}\Big)^2 \qquad S_5 = (N^2-N)S_1 = 2NS_2 + 6W^2$$

$$S_3 = \frac{\frac{1}{N}\sum_i (x_i-\bar{x})^4}{\Big(\frac{1}{N}\sum_i (x_i-\bar{x})^2\Big)^2}$$

1.2.2 第 1 种距离测量方式

显然，I 对距离度量的选择和相应的权重系数 w_{ij} 很敏感。如果我们把距离简单地看成是相邻性，那么距离矩阵就会变得相当稀疏，其中如果位置 i 和 j 是相邻的，则 w_{ij} 为一，否则为零。然而，考虑到我们数据的定义方式与建模偏好的合理性，我们把距离作为一个连续变量。取 i 和 j 的经纬度坐标，我们将计算出大圆距离 d_{ij}，然后进行反演。[3]因此，我们定义 w_{ij} 为

$$w_{ij} = \begin{cases} \frac{1}{d_{ij}} & i \neq j \\ 0 & \text{o.w.} \end{cases} \tag{3}$$

1.2.3 第 2 种距离测量方式

按照Oden 1995的例子，我们将通过人口权重调整距离权重 w_{ij}。特别是，

$$w_{ij}^P = \begin{cases} e^{-4\left(\frac{d_{ij}}{k_i}\right)^2} & i \neq j \\ 0 & \text{o.w.} \end{cases} \tag{4}$$

其中 d_{ij} 的定义同前，$k_i = d_{im_i}$ 和 $m_i = \max\{j : u_{j(i)} \leq \lambda\}$。$u_{(j)i}$ 是地理单位 i 和其所有邻居 j 的总人口。λ 实际上是一个空间聚类的度量：λ 的值越大/小，人口聚类的权重就越大/小。[4]

1.2.4 全局 vs 局部

无论 (3) 或 (4), I 是一个衡量密切相关变量的全局指标。然而，如果对比 $d_{ij} = 50$ 英里和 $d_{ij} = 500$ 英里，那么单位 i 和 j 之间的相关性的方向和强度会有所不同。因此我们变化了 d_{ij} 的计算方式

$$\hat{d_{ij}} = \begin{cases} d_{ij} & \text{if } d_{ij} \leq \bar{d} \\ 0 & \text{o.w.} \end{cases} \tag{5}$$

这里 $\bar{d}$ 是用户指定的最大距离，在本项目中，$\bar{d}_k$ 是距离分布的第 k 个十分位数（跨越所有地点 i 和 j）。

3. 根据老师的要求，我们可以直接实现大圆距离函数，或者从头开始手工编码一些大地测量距离函数；这里我们选择手工编码距离。

4. 在我们实现(3)的过程中，我们最终改变了 d_{ij}/k_i 的指数化方式：参见(12)。我们做了这个改变(3)，因为如果不改变，大多数观测值的权重基本为零。

1.3 研究建议

鉴于我们在县级的 Medicaid 和 Medicare 附表 D 中关于阿片类药物索赔的数据，我们将：

1. 构建通用函数来计算 (1), (7), (3), (4), 和 (5) 基于纬度/经度坐标和一些我们感兴趣的变量。
2. 对于我们样本中的每一年，我们将：
   (a) 确定 Moran's I 指数及以下每个的检验统计/p 值：
      i. 距离
      ii. 距离的十分位数 $\bar{d}_k$
   (b) 画出结果
3. 确定阿片类药物处方和失业率的 Moran's I 指数。给定一种阿片类药物，确定具有相同 Moran's I 指数的区域。给定一个时间段内的 Moran's I 指数值，确定两类地区是否重叠。最有趣的是，失业率和阿片类药物处方如何随着时间的推移而相互影响（Hollingsworth, Ruhm, and Simon 2017）
4. 分析并且总结我们的研究发现

2 其他方法

2.1 新的测量

经过初步提议，我们决定加入两个新的空间自相关度量：Geary‘s C 和 Getis Ord'sG 统计量。

2.1.1 Geary's C 指数

Geary's C 指数被定义为

$$C = \frac{(N-1)\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - x_j)^2}{2\Big(\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\Big)\sum_{i=1}^{n}(x_i - \bar{x})} \tag{6}$$

而(1)的交叉乘积项 (分子中) 是基于偏离平均数的，Geary'sC 的交叉乘积项涉及每个位置对的观测值之间的差异。在设计上，$C \in [0, 2]$ 与 I 的解释基本相反：$C \to 0$ 表示完美的正自相关，$C \to 2$ 表示完美的负自相关，$C = 1$ 表示随机聚类。C 的显著性检验涉及: E[C] 和 Var[C] 以及 Z 值为:

$$z_C = \frac{C - \mathbb{E}[C]}{\sqrt{\mathbb{V}\text{ar}[C]}} \tag{7}$$

这里

$$\mathbb{E}[C] = -\frac{1}{N-1} \quad \text{and } \mathbb{V}\text{ar}[C] = \frac{(2S_1 + S_2)(N-1) - 4S_0^2}{2(n-1)S_0^2}$$

并且

$$S_0 = \sum_{i}^{N}\sum_{j}^{N} w_{ij} \quad S_1 = \frac{\sum_{i}^{N}\sum_{j}^{N}(w_{ij} + w_{ji})^2}{2} \quad \text{and } S_2 = \sum_{i}^{N}(w_{i\cdot} + w_{\cdot i})^2$$

2.1.2 Getis Ord's G 指数

Getis Ord's G 指数不同于 (1)和(6)，因为 G 用于" 热点" 分析。虽然 I 和 C 提供了一个聚类的度量，但都不能描述聚类存在的种类。例如，考虑两个区域 A 和 B，每个区域有 n 的子区域。如果所有 A_n(子)地区都是高收入地区，而所有 B_n 地区都是低收入地区，那么 $I(A) \approx I(B) \approx 1$ 和 $C(A) \approx C(B) \approx 0$。相比之下，$G(A) \neq G(B)$ 因为 G 是为了区分不同程度的空间集中值的区域和值的差异水平。G 的大/小表示高/低值聚集在一起。当 G 以其一般形式出现时，如(8)中所定义的那样，它反应了 G 与其预期值之间的差异。

$$G = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}x_i x_j}{\sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j} \; \forall j \neq i \tag{8}$$

G 的 z-score 的计算方法是:

$$z_G = \frac{G - \mathbb{E}[G]}{\sqrt{V[G]}}$$

其中

$$\mathbb{E}[G] = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}}{n(n-1)} \; \forall j \neq i \quad \text{and } V[G] = \mathbb{E}[G^2] - \mathbb{E}[G]^2$$

并且:

$$\mathbb{E}[G^2] = \frac{A+B}{C}$$

$$A = D_0\Big(\sum_{i=1}^{n} x_i^2\Big)^2 + D_1\sum_{i=1}^{n} x_i^4 + D_2\Big(\sum_{i=1}^{n} x_i\Big)^2\sum_{i=1}^{n} x_i^2$$

$$B = D_3\sum_{i=1}^{n} x_i\sum_{i=1}^{n} x_i^3 + D_4\Big(\sum_{i=1}^{n} x_i\Big)^4$$

$$C = \Big[\Big(\sum_{i=1}^{n} x_i\Big)^2 - \sum_{i=1}^{n} x_i^2\Big]^2 n(n-1)(n-2)(n-3)$$

$$D_0 = (n^2 - 3n + 3)S_1 - nS_2 + 3W^2$$

$$D_1 = -\left[(n^2 - n)S_1 - 2nS_2 + 6W^2\right]$$

$$D_2 = -\left[2nS_1 - (n+3)S_2 + 6W^2\right]$$

$$D_3 = 4(n-1)S_1 - 2(n+1)S_2 + 8W^2$$

$$D_4 = S_1 - S_2 + W^2$$

$$W = \sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} w_{ij}$$

$$S_1 = \frac{1}{2}\sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n}(w_{ij} + w_{ji})^2$$

$$S_2 = \sum_{i=1}^{n}\Big(\sum_{\substack{j=1 \\ j\neq i}}^{n} w_{ij} + \sum_{j=1}^{n} w_{ij}\Big)^2$$

2.2 局部化

上一节，全局 vs 局部 讨论了我们如何采用全局性的测量方法 I，C，和 G，并通过设定地理区域间的最大距离上限来改变观测数量。例如，如果我们设置 $\bar{d} = 500$ 英里，那么我们从权重矩阵 W（根据 $\bar{d} = 500$ 进行调整后）中计算出的 I 的估计值将表明相距不超过 500 英里的所有地理对 (i, j) 之间的空间自相关程度。此外，如果 $\bar{d}$ 大得多/小得多，那么我们相应的 I 就会反映出一个较大/小集合中较大/小 (范围) 的地理对的空间自相关程度。然而，无论 I 被如何地参数化，I，C 和 G 都只能产生一个全局化的度量。

出于这个原因，我们开发并实现了三种自相关度量的局部化版本；也就是说，如果一个地理样本有 n 个观测值，那么以下任何一个局部化的自相关函数都会产生 $n \times 1$ 个估计值。$\bar{d}$ 的选择仍然很重要，但现

在扮演了不同的角色。例如，如果我们要估计位置 i 的局部 Moran's I 指数，就像在(9)中一样，我们仍然需要确定与 i 进行比较的邻居（以及邻居的数量）。因为每一个局部化的测量都只在所有 n 个地理区域中执行一年，所以 $\bar{d}$ 必须被设置为对于所有位置 i，i 都有至少两个邻居。[5]

2.2.1 局部 Moran' s I

将位置 i 的局部 Moran 定义为:

$$I_i = \frac{x_i - \bar{x}}{S_i^2} \sum_{\substack{j=1 \\ j \neq i}}^{n} w_{ij}(x_j - \bar{x}) \quad \text{given } S_i^2 = \frac{1}{n-1} \sum_{\substack{j=1 \\ j \neq i}}^{n} w_{ij} - \bar{x}^2 \tag{9}$$

与 I 不同的是，I_i 可以在 ± 1 之外取值，而 I_i 的值越高/低，对应的空间自相关越正/负。在零假设下 (如随机化)，I_i 的期望值和方差由以下公式给出:

$$\mathbb{E}[I_i] = \frac{1}{n-1} \sum_{\substack{j=1 \\ i \neq j}}^{n} w_{ij} \quad \text{and } V[I_i] = \frac{n - b_2}{n-1} \sum_{\substack{j=1 \\ j \neq i}}^{n} w_{ij}^2 - \frac{(2b_2 - n)}{(n-1)(n-2)} \sum_{\substack{k=1 \\ k \neq i}}^{n} \sum_{\substack{k=1 \\ k \neq i}}^{n} w_{ik} w_{ik} - \Big(\mathbb{E}[I_i]\Big)^2$$

其中 $b_2 = n \sum_{\substack{i=1 \\ i \neq j}}^{n} (x_i - \bar{x})^4 \times \Big(\sum_{\substack{i=1 \\ i \neq j}}^{n} (x_i - \bar{x})^2 \Big)^{-2}$.

2.2.2 局部 Geary's C

Geary's C 指数的局部版本，表示为 C_i，定义为:[6]

$$C_i = \frac{1}{2} \sum_{j=1}^{n} w_{ij}(z_i - z_j)^2 \tag{10}$$

方程10以标准化的 z-score 的形式呈现。对于这个特殊的函数，我们对数据进行标准的正态变换。该公式与Anselin 1995的不同之处在于将检验统计量调整了 2 倍；在零假设下，$\mathbb{E}[C_i] = 1$。这样一来，C 和 C_i 之间的解释相似度就强多了。

2.2.3 Local Getis-Ord G

局部化的 Getis-Ord ‘s G_i 与全球化的 G 相比，更适合” 热点” 分析。

$$G_i = \frac{\sum_{j=1}^{n} w_{ij} x_j - \bar{x} \sum_{j=1}^{n} w_{ij}}{S \sqrt{\frac{1}{n-1} \Big(n \sum_{j=1}^{n} w_{ij}^2 - \Big(\sum_{j=1}^{n} w_{ij} \Big)^2 \Big)}} \tag{11}$$

已知

$$S = \sqrt{\frac{1}{n} \sum_{j=1}^{n} x_j^2 - (\bar{x})^2}$$

等式11是一个标准化的 z-score。Getis and Ord 2010断言在特定条件下 (例如 $n \to \infty$)，这个版本的 G_i 的分布使得 $G_i \sim \mathcal{N}(0,1)$。出于这个原因，我们计算了每个 G_i 的单尾 p. 检验来进行假设检验。[7]。

5. 我们的代码可以被修改为以这样的方式设置 $\bar{d}$ 以及 $m \geq 2$ 邻居。如果 $m < 2$，那么在局部测量计算中使用的地理区域总数将少于 3 个，估计的均值/方差将无法定义。

6. 请参考:
https://www.biomedware.com/files/documentation/spacestat/Statistics/Gearys_C/Geary_s_C_statistic.htm.

7. 考虑到冷-热点的解释取决于 G_i 的符号，双尾检验将不太有用

3 结果

3.1 数据工作

关于数据的第一个判断是决定地理单位。由于数据的局限性–例如，如果我们的研究在邮编或人口普查跟踪层面上进行，会更有价值–我们选择美国的县（特别是它们的 FIPS 代码）作为我们的地理标识符。这种数据局限性被证明是一种变相的祝福。给定 n 个独特的地理区域，计算距离矩阵 $D_{n\times n}$ 需要 $\frac{n(n-1)}{2}$ 的计算，所以是 $\mathcal{O}(n^2)$。相对于在 数据描述部分的内容，我们做了一些关键的数据修改。特别是，我们

1. 对每个县的坐标**使用了** `housingData`。
   美国各县中心点的地理坐标中位数存在于 `geoCounty` 数据框架内中的新的库中。存在于 `geoCounty` 中的县由美国人口普查 FIPS 代码标记，并有其他标识符，使以后的地图构建/可视化更加容易。

2. **将我们的注意力限制在 OPR 和失业率上**。
   我们分析中的主要变量是 Medicare D 类阿片类药物预付率（以下简称 OPR）。OPR 被定义为阿片类药物索赔数除以总体索赔数并乘以 100。就我们的目的而言，在县一级列出了每年的 $t \in \{2013 \ldots 2017\}$。人口数据来自美国农业部（USDA）.[8]。最后，我们将劳动力数据的空间分析范围缩小到只包括我们样本中每一年县级的失业率（以下简称 UR）；数据来自美国农业部。[9]

我们遇到的另一个挑战是将我们的 OPR 数据与失业数据合并，以便保留尽可能多的观察值。首先看一下我们的 OPR 数据，在 3,027 个独特的 FIPS 县中，有 3,005 个县在我们样本的所有五年中都存在。然而，当我们合并失业数据时，在所有年份中只剩下 1,163 个县（在减少的 3,024 个县中）。

Year	OPR	OPR+UR
2013	3,018	3,015
2014	3,018	3,015
2015	3,017	3,014
2016	3,017	1,189
2017	3,017	3,011

表 1: FIPS counts

表格1总结了每个数据集每年的观测数（如仅 OPR 和 OPR+UR），而图1和2 描述了每个数据集的共享观测的数目。

8. 见下文下载：
https://www.ers.usda.gov/webdocs/DataFiles/48747/PopulationEstimates.xls?v=2561.3
9. See: https://www.ers.usda.gov/data-products/county-level-data-sets/download-data/.

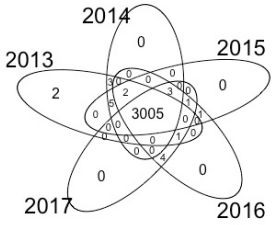

图 1: OPR data only

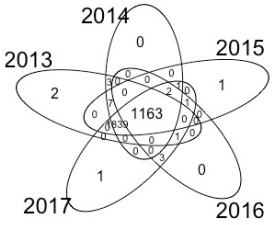

图 2: OPR +UR data

我们的分析将基于两个主要数据集：`OPR_test_data` 有 3,005 个一致的县在所有 5 年内的 OPR 率（3,005 次 5=15,025 美元的观测值）；以及 `test_data_unemployment_new`，这是 OPR+UR 数据的不平衡面板。[10] `OPR_test_data` 用于测试我们的时变函数，而 `test_data_unemployment_new` 则用于将某一年（如 2013 年）的 OPR 和 UR 数据进行关联。

3.2 图

3.2.1 `OPR_test_data` 的工作

图 3 是利用 `OPR_test_data` 的 3,005 个县创建的，描述了各县之间的距离分布。.[11]。县与县之间最小（最大）的距离是 4.82（2832.58）英里，在弗吉尼亚州阿灵顿县和哥伦比亚特区（缅因州华盛顿县和加州圣马特奥县）之间。第 25、中位和第 75 距离的百分位数分别为 471.40、756.90 和 1115.60 英里。在我们的工作中，距离矩阵 D 往往是计算量最大的任务；例如，为 `OPR_test_data` 的 3,005 个县创建距离矩阵大约需要 93 秒。出于这个原因，我们为所有包含相同地理区域的分析/绘图只计算一个距离矩阵。

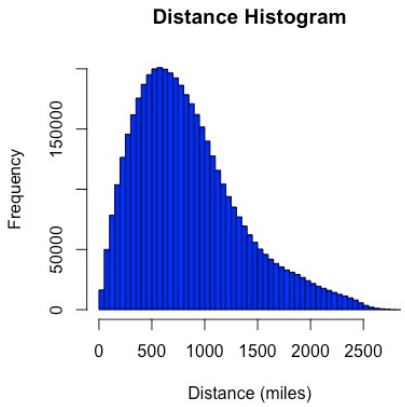

图 3

下一张图，图 4，描述了 Moran's I 在不同的 $\bar{d}$ 上的变化，以及仅对 OPR 数据的时间变化。我们根据一

10. 这个数据集也有每年的县人口。
11. 图 3 描绘了 D_{ij} $i \neq j$ 使得 $D_{ij} \neq 0$. 的分布

个由指数序列定义的选择一个 $\bar{d}$ 的序列。[12] 图4包括一个" 放大" 的面板，作为对较小的 $\bar{d}$ 的 $I(\bar{d})$ 可视化更好一点。图 4 和表格 2 指出了随时间变化的如下: (1) 期望值中的 I 下降 (更多的空间随机性)；(2)I 变得不那么分散；(3)I 在所有年份都在 24.7 英里左右达到最大值。

Year	Max	Mean	Sd
2013	0.431	0.139	0.0943
2014	0.361	0.116	0.0753
2015	0.355	0.107	0.0689
2016	0.373	0.103	0.0691
2017	0.345	0.0945	0.0640

表 2: Moran I (OPR) Summary

图 5 显示了如果我们仅使用 2013 年的数据，调整人口权重后，Moran' s I （OPR）是如何变化的。[13]我们使用与图4相同的距离带宽序列。我们对人口参数 λ 的选择遵循了 2013 年县级人口数据的第 10/50/90 百分位数。与非人口调整后的 $I(d)$'s 相似，$I(d;\lambda)$ 最大为 $\bar{d} \approx 25$ 英里，跨越所有 λ。此外，I 似乎 (几乎) 是随着 λ 单调增加的。$\lambda_{0.5}$ 和 $\lambda_{0.9}$ 产生的 I's 基本上无法区分。

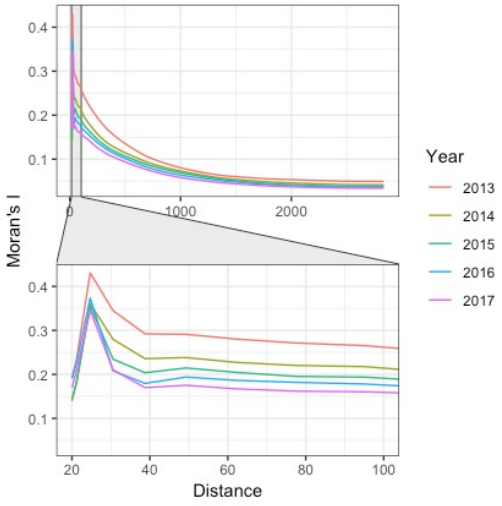

图 4: Moran I varying $\bar{d}$ over time

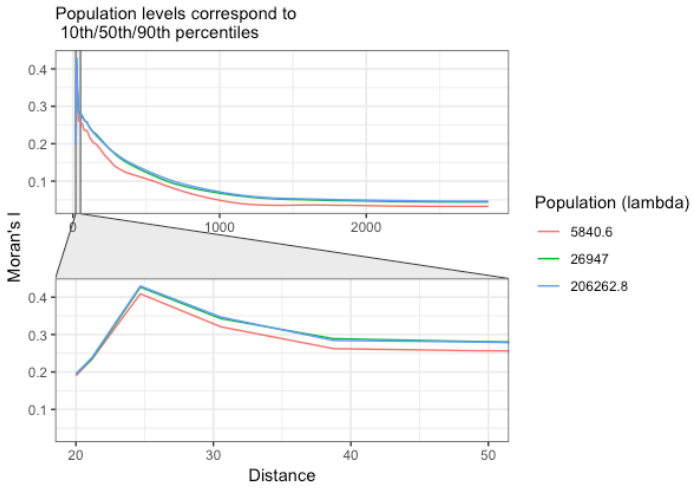

图 5: Moran I varying $\bar{d}$ w/ λ.

局部 Moran 'sI$_i$—就像我们空间自相关统计的几乎所有本地版本一样—需要相当长的时间来计算。我们对 2013 年 OPR 观测数据的计算需要 38 分钟的运行时间。我们选择了 $\bar{d} = 120$ 英里。通俗地说，这意味着对于每个县 i，我们计算了给定 120 英里以外的 i 和 $j \neq i$ 县的(9)。

12. 我们将最小距离设置为 20 英里，最大为 $\bar{d} = 2832.58$（即 D 的最大距离），以及指数调整参数 $\theta = 2$。我们使用这种指数" 网格" 来计算较小的 $\bar{d}$ 的更多 I 实例，因为相对于较大的 $\bar{d}$ 而言，I 很可能在小的 $\Delta\bar{d}$ 中变化更大。

13. 图5 所使用的数据来源于 `test_data_unemployment_new`.

County	State	OPR	Local Moran I
Clay County	Georgia	8.1	1,296
Shelby County	Alabama	9.56	1,234
Hardin County	Ohio	4.02	484
Sebastian County	Arkansas	6.25	451
Franklin County	Arkansas	5.66	259
⋮	⋮	⋮	⋮
Highland County	Virginia	6.57	-245
Perry County	Ohio	5.23	-273
Okmulgee County	Oklahoma	6.66	-312
Osceola County	Michigan	8.36	-840
Christian County	Missouri	9.18	-937

表 3: Extreme Local Moran's

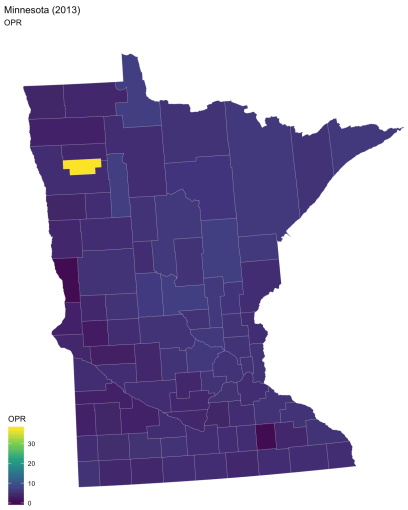

图 6: OPR data only

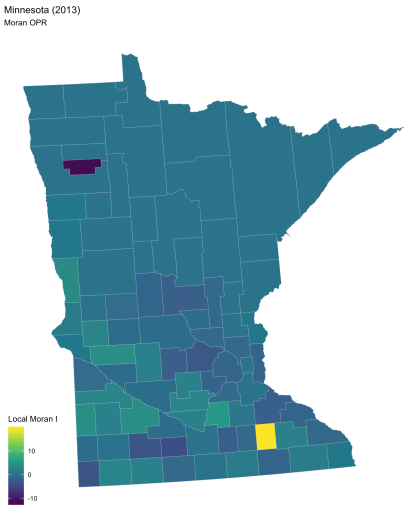

图 7: Local Moran OPR data

将表2和图8的结果耦合在一起，I_i 的离群值的存在使得我们的结果可视化变得非常困难。图8通过只对当 $|I_i| < 10$ 时的 i 进行颜色编码来部分弥补这个问题，然而，即使这样的修改也不足以有意义地区分 I_i 的局部差异。因此，图6 和图7 被包括在内。请注意，图 6中亮黄色的县对应的是红湖县。2013 年，红湖县的 OPR 为 37.61。由于红湖县的大多数邻居的 OPR 要低得多，所以红湖县的 I_i 得分非常低（全州最低），约为-12；图7表示这种差距，因为红湖县被填充了一个非常深的色调，以表示强烈的负空间分配。相反，图中7 中亮黄色的县是道奇县。道奇县的 OPR≈ 0,，由于道奇县的大部分邻居的 OPR 也很低，所以道奇县的 I_i 得分非常正，在 19 左右。

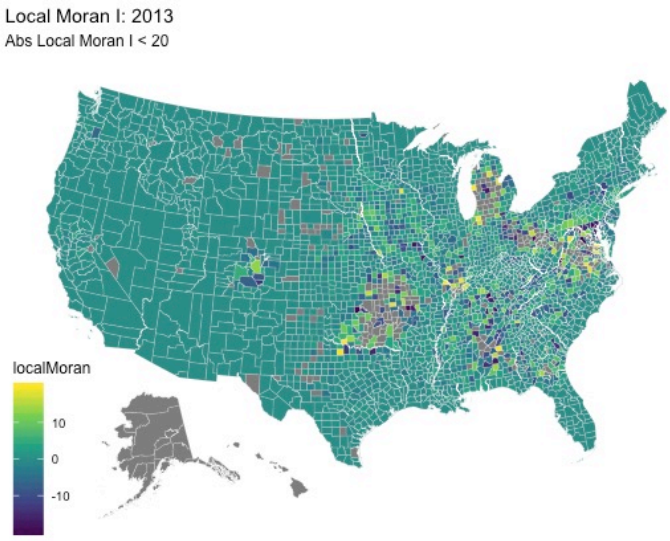

图 8: I_i OPR Scores for US

3.2.2 `test_data_unemployment_new` 工作

- **全局 Geary's C**

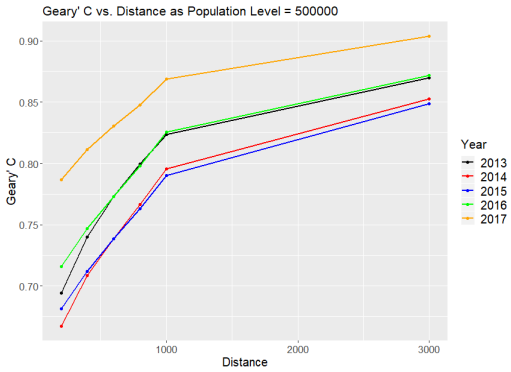

图 9: Geary's C for Opioid Prescription Rate

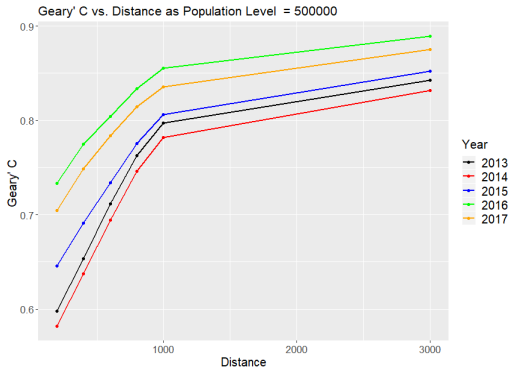

图 10: Geary's C for Unemployment Rate

上面两张图显示了全局 Geary's C 和 $\bar{d}$ 之间的关系，给定 $\lambda = 500,000$ 的阿片类药物处方率和失业率。平均而言，最强的空间正相关性发生在 2017 年的阿片类处方率和 2016 年的失业率。在每一年中，Geary's C 都随 $\bar{d}$ 单调递增。回想一下，$C \to 0$ 表示越来越强的正向空间自相关，而 $C \to 1$ 表示没有空间自相关。由于 $I \propto 1 - C$，这就解释了当涉及到 OPR 数据时，图 4 和图9之间的视觉对比。奇怪的是，当4 表明 OPR 随着时间的推移在空间上变得不那么集中（几乎是以一种统一的方式），9则没有这样明确的模式。

- **热点探测?**

我们决定测试 G_i，而不是一般的 G 统计，因为 G_i 是用于" 热点" 分析的。在我们的 OPR 背景下，我们的 G_i 估计值可以用来确定那些与它们的近邻一起具有高 OPR 流行率特征的县。简而言之，这些县可以

被认为是 OPR ” 热点”。同样，G_i 也可以用来检测” 冷点”，或者说，这些县与它们的近邻一起 OPR 率的流行率较低。县 i 的邻居被定义为不超过 $\bar{d} = 100$ 英里的地方。

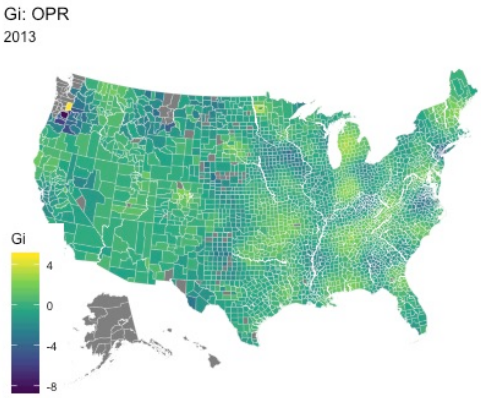

图 11: G_i OPR

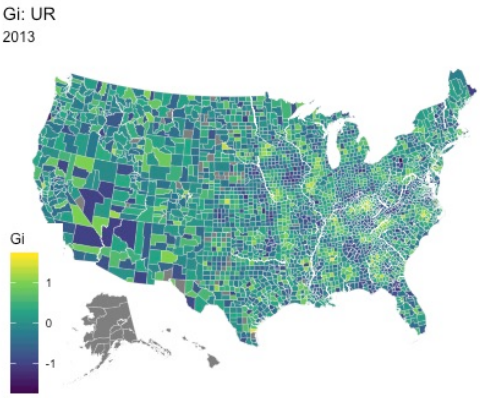

图 12: G_i UR

图11 和图 12 分别描述了 2013 年 OPR 和失业率 (如 UR) 的 G_i。回忆一下，在这种情况下，浅色对应” 更热” 的点，而深色表示” 更冷” 的点。图 11和图12 之间一个显著的比较是视觉平滑度的差异；我们的意思是，OPR G_i 的颜色梯度相对于 UR G_i 来说不那么” 毛躁”。如果读者可以用地形类比，OPR 的极端 (即非常热或非常冷) 斑点往往在区域层面上聚集在一起，因此高原和洼地有一个逐渐上升/下降的过程。与此相反，UR G_i 分数可能被概念化为特别陡峭的山脉和夹杂在不平坦地形中的急剧下滑的峡谷。

- **随时间变化?**

可以说，我们的工作可以回答的最政策导向的研究问题是各县是否随着时间的推移而升温，因为它与 OPR 的变化有关.[14]利用这些结果，我们进行了两个背离信封的分析。

1. **变化为当场得分分布**。我们将选择的 G_i 限制为那些对 $\alpha = 10\%$ 稳健的 p 值，并比较 2013 年和 2014 年之间 G_i 分数的分布。请注意，不同时期的（重要的 G_i 得分）县的构成会发生变化。然而，由于我们主要是想看看分布是如何变化的，所以我们结果（大部分）不受影响。从图13来看，当 $G_i > 0$ 时，2013 年的 G_i's 相对于 2014 年的分布来说密度更大。事实上，2014 年约有 5% 的县存在” 热点”（即 $G_i > 0$），而 2013 年的相应数字接近 7%；因此，与 2013 年相比，2014 年的热点数量（至少相对而言）有所下降。

2. **平均热点得分变化**. 2013 年至 2014 年，平均来看，OPR 热点是否升温? 或者说，县域平均得分有什么变化? 为了回答这两个问题，我们将表4改变。[15]我们将县级 G_i 分解为四个的类别：2013 年至 2014 年间，县域 i 是升温还是降温，一开始是热还是冷? 表4的结果表明美国的” 热门”OPR 县确实

14. 为了回答这个问题，我们使用 2013 年和 2014 年的 OPR 数据计算 G_i 分数。

15. 以后的工作需要在县的比较方面提高。例如，我们是否应该只比较那些在 2013 年、2014 年或两者都有显著 G_i 得分的县? 截至目前，我们无视了显著性。

从 2013 年相对于 2014 年热了起来，但总的来说，大多数县平均降温了。有趣的是，一旦我们对县的数量进行调整，[16] 升温的冷县基本上抵消了降温的热县。此外，由于降温的冷县数量和平均热量变化的幅度都超过了升温的热县，这就导致了我们的净降温。

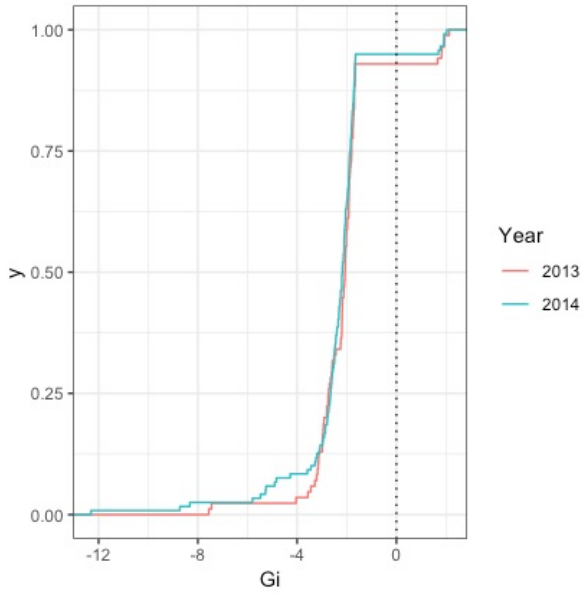

图 13: Empirical CDF of G_i (OPR) Scores

Direction	Count	Mean Change	Share	Adjusted Mean
Cold getting Colder	511	-0.514	0.182	-0.0937
Cold getting Hotter	994	0.647	0.355	0.230
Hot getting Colder	927	-0.699	0.331	-0.231
Hot getting Hotter	370	0.292	0.132	0.0386
Net Change				-0.057

表 4: Change in G_i

最后，图14给出了 2013-2014 年之间 G_i OPR 分数变化的汇总级别视图。请注意，虽然配色方案发生了变化，但明亮的颜色仍然对应着更高的水平，在这个例子中，更高的变化水平。与之前两者一样，灰色阴影的县在 2013 年或 2014 年都没有 G_i 分数。

16. 这个项目未来将根据人口和/或人口密度而不是县数进行调整。

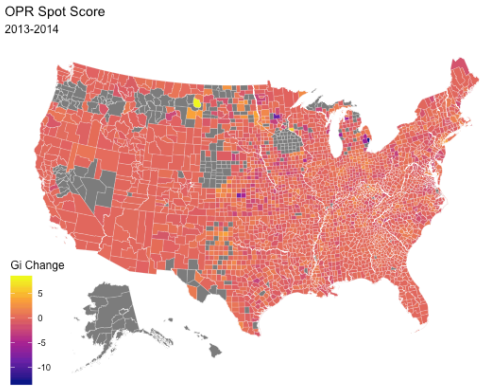

图 14: ΔG_i OPR Map

再次，相对较大的 ΔG_i 的分散度阻碍了视觉上的区分。可以说，密歇根州有很多暗点（降温），而东北部各州的光照比较均匀（升温）；事实上，表6列出了 2013 年至 2014 年的平均 ΔG_i，并证实密歇根州的平均 ΔG_i 最低，而东北部各州则倾向于越来越热。

4 附录

4.1 代码的开发

表 5: Function Summaries

Function	Arguments	Output
`gcd.hf`	4, $n \times 1$ vectors with longitude and longitude coordinates.	Haversine great circle distance in miles.
`distance.matrix`	A $n \times 3$ dataframe; the first column is a list of geographic identifiers (e.g. place name) and the last two columns list longitude/latitude coordinates per each identifier.	$n \times n$ symmetric distance matrix D; d_{ij} is the haversine distance between locations i and $j, d_{ii} = 0$.
`weight_distance_matrix`	Distance matrix D, distance bandwidth parameter $d_{max} > 0$, $n \times 1$ population vector, population clustering parameter λ, and options to population-weight distances.	Weight matrix $W_{n \times n}$. See additional details.
`moranI`	Spatially lagged variable $y_{n \times 1}$, weight matrix $W_{n \times n}$, weight scaling option (default is false), and p.value options.	Moran's index I (as in (1)) and (one or two-sided) p.value. If weight scaling option is set to `TRUE` (default is `FALSE`), then W is normalized such that $\sum_i W_{ij} = 1$ (which lowers the magnitude of I).
`moran_time_dist`	Spatially lagged and time-varying $y_{(n \times t) \times 1}$, a column vector of time variables $((n \times t) \times 1)$, distance matrix $D_{n \times n}$, a sequence of distance bandwidths $d_{max}(d \times 1)$, and a year sequence $(t \times 1)$.	Moran's index I and p.value (default set to two.sided) for each distance bandwidth and year in the year sequence. Note that weight-scaling is defaulted to `TRUE`. Output is a $(d \times t) \times 4$ dataframe. See additional details.

Continued on next page

表 5– continued from previous page

Function	Arguments	Output
`MoranI_pop`	Spatially lagged variable $y_{n\times 1}$, population vector $p_{n\times 1}$, distance matrix $D_{n\times n}$, a sequence of distance bandwidths $(d\times 1)$, and sequence of population clustering-parameters $\lambda_{l\times 1}$.	Moran's index I using population adjusted weights for each distance bandwidth parameter and population clustering-parameter. Note that the weight scaling option and the p.value are defaulted to `FALSE` and two.sided, respectively. Output is $(d \times l) \times 4$ dataframe.
`LocalMoran`	Spatially lagged variable $y_{n\times 1}$, distance matrix $D_{n\times n}$, distance bandwidth $d_{max} > 0$, weight scaling option (default is still `FALSE`), and p.value options.	Local Moran's I_i (as in (9)) for each n locations and corresponding p.value (default here is set to two-sided). Output is formatted as a list of two lists: the first list contains I_i and the second list contains associated p.values.
`Getis_Ord`	Spatially lagged variable $y_{n\times 1}$, weight matrix $W_{n\times n}$, and p.values (default is set to one.sided).	G statistic (from (8)), Spot type which is merely $G - \mathbb{E}[G]$ (if this value is positive, return Hot, o.w. Cold), and one.sided p.value. Note that the output is formatted as a list containing the aforementioned items. See additional details.
`Getis_Ord_local_z`	Spatially lagged variable $y_{n\times 1}$, distance matrix $D_{n\times n}$, and distance bandwidth $d_{max} > 0$.	G_i (as in (11)), one sided p.value, and spot status ($G_i < 0$ returns Cold, o.w. Hot). Unlike, `Getis_Ord`, this output is formatted as a dataframe.
`find_global_Geary_C`	Year from test data and distance bandwidth $d_{max} > 0$.	C statistic as in (6). See additional details.
`local_GC`	Spatially lagged variable $y_{n\times 1}$, distance matrix $D_{n\times n}$, and distance bandwidth $d_{max} > 0$.	$n \times 2$ dataframe with local Geary's c_i as in (10) and corresponding two tailed p.value.

Continued on next page

表 5– continued from previous page

Function	Arguments	Output
`grid_spacing`	Start/end points a/b, number of grid points n, and exponential tuning parameter θ (if $\theta = 1$, then output is a sequence of linearly spaced points).	$x_{n\times 1}$ vector such that: $x_i = a + (b-a)\left(\frac{i-1}{n-1}\right)^{\theta}$

4.2 更多函数的细节

1. `weight_distance_matrix`.
   权重矩阵 W 有两种计算方式。

   (a) **群众没有被人口调整**. 比如带宽 $d_{max} > 0$ 则

$$w_{ij} = \begin{cases} \frac{1}{d_{ij}} & \text{if } i \neq j \\ 0 & \text{if } i = j \text{ or } d_{max} < d_{ij}. \end{cases} \tag{12}$$

   (b) **权重被人口调整**

$$w_{ij} = \begin{cases} \frac{\alpha_{ik}}{d_{ij}} & \text{if } i \neq j \\ 0 & \text{if } i = j \text{ or } d_{max} < d_{ij}. \end{cases} \tag{13}$$

   参数 α_{ik} 和指数 k 需要一些解释。让 $\mathcal{S}_i$ 表示位置 i 与所有位置 (包括它自己) 的距离集，并按递增顺序排序。$\mathcal{S}_i$ 不是跟踪这个排序后的距离列表中的每个元素，而是简单地跟踪排序后列表中的元素指数 (从 D 的第 i 行开始)。一个典型的 S_i, k 元素将对应于 s 离 i 最近的位置。例如 $S_i(1) = 1$，因为 i 是离 i 最近的位置的索引。让 $RP_i(k)$ 表示 i 的最近邻居的区域人口。人口聚类参数 $lambda > 0$ 是 $RP_i(k)$ 的上界，现在让 $\mu_i > 0$ 表示 i 的人口。指数 k 和相应的 α_{ik} 通过以下算法确定:

Algorithm 1 Determining α_{ik}

if $\mu_i \geq \lambda$ **then return** $\alpha_{ik} = d_{ik}$
$RP_i = \mu_i$
for ($s \in S_i$ $s = 2, \ldots, n$) {
  $k = \mathcal{S}_i(s)$
  $RP_i = \mu_k + RP_i$
  if $RP_i \geq \lambda$ **then return** $\alpha_{ik} = d_{ik}$
}

2. `moran_time_dist`.
   需要注意的是，这个函数的存在主要是为了方便计算不同 $d_{\max}$ 和不同时间段的 I，并使作图更加方便。

3. 函数 `find_global_Geary_C` 和 `plot_OPR` 是在考虑到特定测试数据的情况下实现的。原则上，这些函数可以重新设计，以适应更广泛的数据集。

4. `Getis_Ord`
   我们在 处理 G 时遇到的一个问题是确定 $V(G)$。特别是每次我们试图计算 $V(G)$ 时，都会遇到 $\mathbb{E}[G^2] < \mathbb{E}[G]^2$，这样我们最终得到的是负方差。这是很糟糕的。最有可能的是，一些占位变量 (如 A、B 和/或 C) 在一开始就被错误地编码了。

本项目的所有相关文件/输出可以在以下 GitHub 仓库中找到: https://github.com/EphaneZhong/Research_Proposal

Testing Resources

- 我们的实际数据和清理后的数据见 `data_folder.zip`.
- `test_file.R` 包含我们函数/代码的通用实现。
  `find_global_Geary_C and a testthat for it.R` 包括优秀的测试材料。

4.3 Assorted Results

State	Mean Change
New York	0.4166
North Dakota	0.4144
Montana	0.347
Rhode Island	0.3428
Delaware	0.2556
New Jersey	0.2414
Massachusetts	0.228
Georgia	0.1724
Kentucky	0.1707
Nevada	0.1345
Tennessee	0.1051
Texas	0.0957
Alabama	0.0914
South Carolina	0.0695
New Hampshire	0.0401
Ohio	0.0301
Florida	0.0164
Arkansas	0.0152
Illinois	0.0141
Vermont	0.0138
Louisiana	0.0101
Wyoming	-0.0156
Utah	-0.0204
Arizona	-0.0264
Washington	-0.0328
Oklahoma	-0.0357
Oregon	-0.0395
Colorado	-0.0482
New Mexico	-0.0706
California	-0.0709
West Virginia	-0.0712
Mississippi	-0.0812
Indiana	-0.1195
Idaho	-0.1314
Missouri	-0.1573
South Dakota	-0.1629
North Carolina	-0.1721
Minnesota	-0.1782
Wisconsin	-0.1786
Maryland	-0.2411
Kansas	-0.2956

References

Anselin, Luc. 1995. "Local indicators of spatial association—LISA." *Geographical analysis* 27 (2): 93–115.

Getis, Arthur, and J Keith Ord. 2010. "The analysis of spatial association by use of distance statistics." In *Perspectives on Spatial Data Analysis,* 127–145. Springer.

Hollingsworth, Alex, Christopher J Ruhm, and Kosali Simon. 2017. "Macroeconomic conditions and opioid abuse." *Journal of health economics* 56:222–233.

Moran, Patrick AP. 1950. "Notes on continuous stochastic phenomena." *Biometrika* 37 (1/2): 17–23.

Oden, Neal. 1995. "Adjusting Moran's I for population density." *Statistics in Medicine* 14 (1): 17–26.