文献导读（三）：
机器学习与疫情空间传播模式分析

导读人：陈露，东南大学

点评人：叶信岳，新泽西理工学院

北京时间：2020年5月30日上午8点半–9点

U.S. Time: 8:30-9:00 PM, May 29, 2020 (Eastern Time)

议题 Topics

1. 研究背景
2. 数据
3. 方法
4. 结果解读
5. 总结与讨论

文献一：空间数据研究——研究背景

Pare, Philip E.; Liu, Ji; Beck, Carolyn L.; Kirwan, Barrett E. and Basar, Tamer. "Analysis, Estimation, and Validation of Discrete-Time Epidemic Processes." IEEE TRANSACTIONS ON CONTROL SYSTEMS TECHNOLOGY, 2020, 28(1), pp. 79-93.

SIS模型
Susceptible Infected Susceptible Model

传染病SIS模型假设传染源头的感染个体（记作I，Infected），通过一定的概率β将传染病传给易感个体（记作S，Susceptible），同时感染个体以一定的概率γ恢复为易感状态，另一方面，易感人群一旦被感染，就又成为新的感染源，其感染机制可以由下式描述：

$$\begin{cases} S(i) + I(j) \xrightarrow{\beta} I(i) + I(j) \\ I(i) \xrightarrow{\gamma} S(i) \end{cases}$$

文献一：空间数据研究——研究数据

- 第一个数据集是Snow编制的霍乱数据集。Snow绘制了1854年伦敦Soho区霍乱死亡人数的地图，以说明霍乱是通过几个特殊的水井，而不是像当时人们认为的那样通过空气传播的。斯诺的这项开创性工作引领了现代流行病学领域

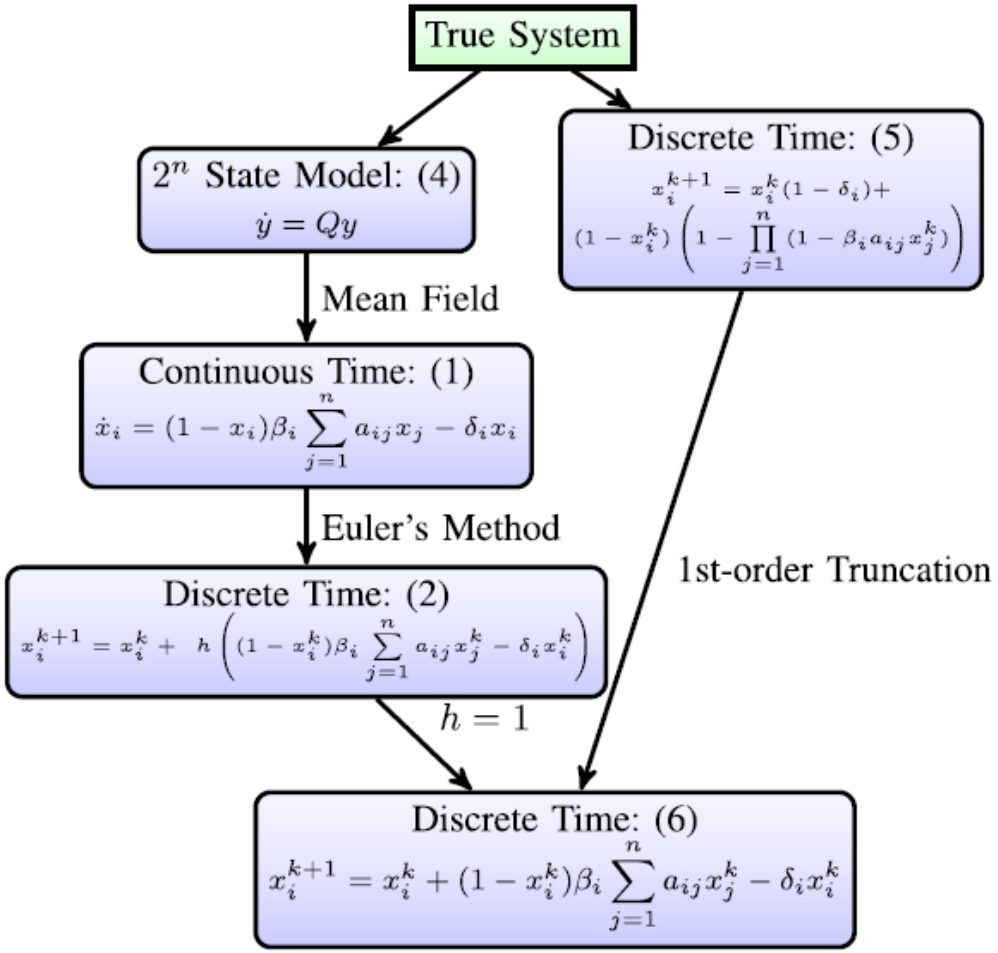
- 这里使用的第二个数据集记录了2008年至2013年美国农业部(USDA)向农场或者农民支付的所有由USDA资助的补贴项目的所有款项。主要用来模拟了农业技术的采用和传播是否遵循传染病模型

文献一：空间数据研究——研究方法

这篇文章重点研究一个非线性网络相关的常微分方程模型

本文的创新性在于，前人未对这些模型的数据的传播参数进行过估计。许多使用这些模型的病毒传播论文声称使用真实的数据来测试他们的模型，但是没有对重要的网络依赖的SIS传播模型进行真正的验证

据我们所知，还没有人使用一组实际的扩展数据来验证依赖于网络的SIS模型



文献一：空间数据研究——结果解读

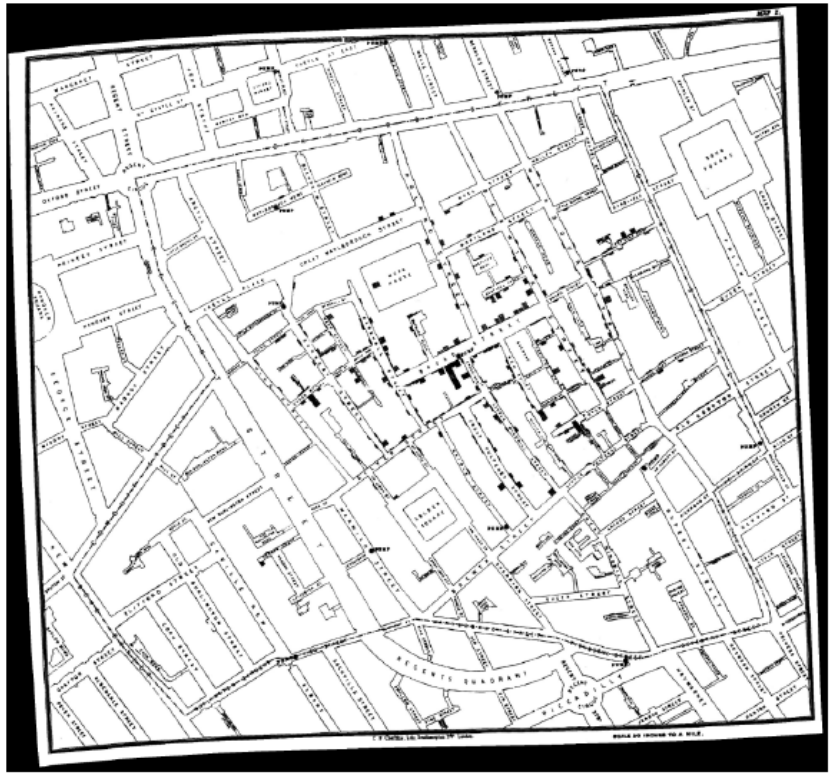


Fig. 5. This is the map of cholera spread in London in 1854 compiled by Snow [22].

斯诺在该地区的地图上描绘了1854年伦敦苏豪区每户人家死于霍乱的人数。在图5中，显示了地图，其中每个小矩形对应于该地址的一个死亡。斯诺绘制这幅地图是为了向官员们说明，霍乱的传播是通过来自宽阔街道的水泵的受感染的水，而不是像当时人们普遍认为的那样通过空气传播

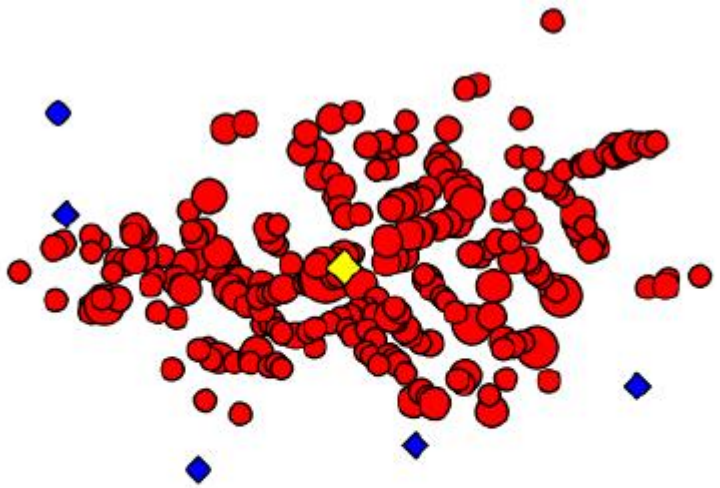


Fig. 6. Digitization of Fig. 5. The uncontaminated water pumps are depicted by blue diamonds, the contaminated water pump by the yellow diamond, and household deaths by red dots with the diameters scaled by the number of deaths, respectively.

未受污染的水泵用蓝色钻石表示，受污染的水泵用黄色钻石表示，家庭死亡人数用红点表示，红点的直径与死亡人数成正比

文献一：空间数据研究——结果解读

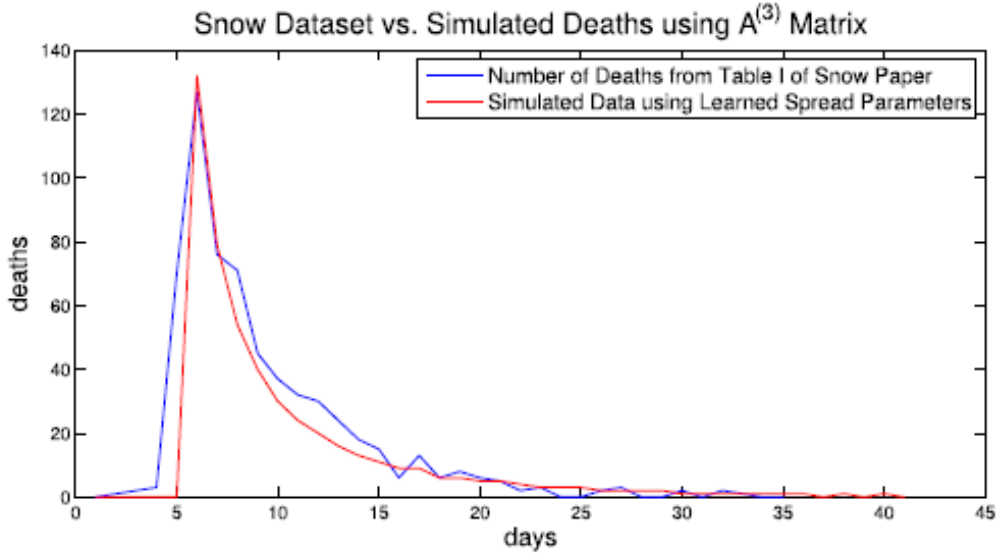


Fig. 14. Comparison of Figs. 7 and 11. Note that there is a difference in the magnitude, but the general shapes are very similar. The Euclidean distance between the two plots is 75.16, and the infinity norm is 70. One of the reasons for this discrepancy is due to the fact that we used the spatial data set in Figs. 5 and 6, which had only 489 documented deaths, while the cumulative data from [22, Table I], shown in Fig. 7 and the blue in this plot, have a total of 616 deaths. The difference of 127 has caused the discrepancy.

如图14所示拟合结果非常好。这篇文章的基于空间离散的SIS模型很好地从John Snow的1854年的数据集中捕捉了霍乱流行的模拟行为

表现最好的事实支持斯诺的假设，即宽街泵是霍乱爆发的源头，霍乱不通过空气传播，这是今天已知的事实

文献一：空间数据研究——总结与讨论

本文主要创新点

- 创新了基于空间网络的一种新的SIS流行病传播模型
- 利用John Snow的研究对文章进行了拓展

本文的可重复与可复制性

- 如果有详细的基于时间序列的COVID-19的时空数据，可以复制这一结果

文献二：零售商店如何复工复产？——研究背景

Avi Goldfarb, Catherine Tucker. Which Retail Outlets Generate the Most Physical Interactions? NBER Working Paper 27042, http://www.nber.org/papers/w27042.

现在是时候考虑首先开放经济的哪些领域了。哪些企业应该继续营业，哪些企业应该关闭?很少有数据能够帮助决策者做出这些决定。

这一决定将部分取决于在就业、工资和其他行业的外部性方面开放这些企业的经济效益。更重要的是，这将取决于这些营业场所是否存在风险，即开放这些业务会不会将导致疾病的指数级增长。

这篇文献主要回答：是小型零售商还是拥有多家店面的零售品牌风险更大？

文献二：零售商店如何复工复产？——研究数据

- We employ February 2020 data from SafeGraph, which tracks over 40 million mobile devices in the U.S.
- 使用SafeGraph提供的数据来研究2020年2月冠状病毒的传播情况。这些数据建立在一个包含445464.5亿台设备的面板上，这些设备收集匿名位置数据。这代表了美国大约10%的设备。这些设备的每一个用户都允许自己的位置被各种移动应用程序跟踪。SafeGraph将这些设备的位置与美国境内品牌实体零售店的各种位置进行匹配，其主要业务集中于向公司和分析师提供零售流量数据。
- 这些零售营销数据包括访问次数和独立访问者、到零售商的平均距离、在零售店花费的平均时间以及当月的外国访问者数量。

文献二：零售商店如何复工复产？——研究方法

- 描述性统计
- 线性回归

$$VisitationPattern = \beta NumberofOutlets + \alpha_k + \gamma_z + \epsilon_z$$

文献二：零售商店如何复工复产？——研究方法

Table 1: Summary Statistics

	Mean	Median	Std Dev	Min	Max
Distance (km)	19.6	8.57	124.9	0.0020	6561.0
Average Time Spent	31.2	20	41.3	4	1429.5
Visitor Count	176.0	74	371.3	5	39996
Visit Count	282.8	125	637.9	5	72219
Visitor to Visit Ratio	0.66	0.67	0.20	0.016	1
Canadian Visitors	0.0033	0	0.057	0	1
Mexican Vistors	0.0061	0	0.078	0	1
> 5000 Outlets	0.098	0	0.30	0	1
1000-4999 Outlets	0.11	0	0.31	0	1
100-999 Outlets	0.12	0	0.32	0	1
10-99 Outlets	0.055	0	0.23	0	1
2-9 Outlets	0.100	0	0.30	0	1
Single outlet	0.52	1	0.50	0	1
Outlets	1221.4	1	2902.4	1	15718
Food Store	0.14	0	0.35	0	1
Gas Station	0.11	0	0.32	0	1
Drug Store	0.033	0	0.18	0	1

在我们的数据中，到零售店的平均距离是19.6公里。但是中位数是8.6公里。或者平均每个人在商店里花30分钟，中间值是20分钟。在SafeGraph的数据中，平均每家商店每月有176名单独的访客，283名访客来自设备。“访问者访问比率”衡量的是同一访问者多次访问的程度。

文献二：零售商店如何复工复产？——结果解读

Table 2: Summary Statistics By Different Store Types

	Auto	Clothing	Electronics	Food	Furniture	Gas	General	Health	Home	Leisure	Misc
Visit Count	147.4	165.2	160.8	334.8	137.0	330.7	967.4	225.6	196.8	279.5	249.6
Visitor Count	81.44	116.4	95.59	197.9	86.94	221.5	591.2	148.5	114.7	177.5	157.4
Visitor to Visit Ratio	0.604	0.768	0.632	0.654	0.660	0.689	0.664	0.706	0.588	0.638	0.645
Distance (km)	17.36	35.26	14.54	17.93	14.62	25.37	13.34	11.92	13.24	27.80	17.98
Average Time Spent	34.15	31.43	41.94	25.42	36.67	14.87	22.32	30.77	40.07	40.93	36.14
Single outlet	0.544	0.502	0.686	0.528	0.588	0.229	0.154	0.519	0.627	0.740	0.675
Total	148620	89986	25842	155189	51069	124300	67277	98730	87394	98047	155089
Observations	1101543										

表2按数据中不同三位数NAICS代码的分类对这些汇总统计信息进行了分类。很明显，不同类型的商店有很大的差异，服装商店吸引人们从最远的距离。一般商店(Walmart和Target等品牌)吸引了最多的访客。

家庭商店拥有最多的回头客。而服装店的回头客最少。只有一家直营店的商店主要集中在休闲类，涵盖音乐。书籍,体育。和爱好。在我们的数据中，没有哪个品牌的家具店拥有超过5000家门店。

文献二：零售商店如何复工复产？——结果解读

Table 3: Stand-Alone Retailers and Large Footprint Retailers Attract the Smallest Number of Visitors

	All			Non-Essential Businesses		
	(1)	(2)	(3)	(4)	(5)	(6)
	Visitor Count	Visit Count	Visitor to Visit Ratio	Visitor Count	Visit Count	Visitor to Visit Ratio
> 5000 Outlets	22.14***	-32.35***	0.0845***	-230.9***	-504.2***	0.109***
	(1.752)	(3.129)	(0.000755)	(4.368)	(8.074)	(0.00116)
1000-4999 Outlets	262.4***	426.2***	0.0836***	339.9***	567.1***	0.0962***
	(2.161)	(3.727)	(0.000719)	(2.939)	(5.149)	(0.000878)
100-999 Outlets	221.5***	306.2***	0.104***	147.8***	170.6***	0.130***
	(1.918)	(2.965)	(0.000803)	(2.154)	(3.134)	(0.000906)
10-99 Outlets	96.18***	130.3***	0.0718***	71.10***	82.74***	0.0832***
	(1.786)	(2.781)	(0.000943)	(2.189)	(3.312)	(0.00123)
2-9 Outlets	20.60***	26.47***	0.0262***	21.37***	26.88***	0.0298***
	(0.728)	(1.255)	(0.000674)	(0.878)	(1.508)	(0.000783)
Zipcode Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
NAIC Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1102358	1102358	1102358	783889	783889	783889
R-Squared	0.225	0.195	0.193	0.259	0.235	0.217

Dependent variable is the number of visitors to the store in Columns (1) and (4). Dependent variable is the number of visits to the store in Columns (2) and (5). Dependent variable is the ratio of visitors to visits in Columns (3) and (6) * p <0.10, ** p <0.05, *** p <0.01. Data for February 2020. Robust standard errors clustered at the zipcode level. * p <0.10, ** p <0.05, *** p <0.01.

表3考察了品牌的实体店足迹如何影响访问模式。更多的访客和更多的访客可能意味着更大的传染风险。列(1)-(3)查看整个数据集，列(4)-(6)不包括加油站、食品店和药店。这种排斥的动机是，大多数居家订单不包括药店、加油站和食品店，因为它们是必要的服务。

文献二：零售商店如何复工复产？——结果解读

Table 4: Mid-Tier Brands Attract People from the Greatest Distance

	All	Non-Essential Businesses	All	Non-Essential Businesses
	(1)	(2)	(3)	(4)
	Distance (km)	Distance (km)	Average Time Spent	Average Time Spent
> 5000 Outlets	-6.374***	-8.332***	-19.57***	-24.14***
	(0.389)	(0.534)	(0.132)	(0.209)
1000-4999 Outlets	-4.971***	-3.999***	-17.11***	-18.35***
	(0.328)	(0.388)	(0.123)	(0.156)
100-999 Outlets	2.238***	-1.454***	-15.00***	-15.03***
	(0.411)	(0.410)	(0.120)	(0.148)
10-99 Outlets	2.668**	5.129***	-11.31***	-10.80***
	(0.836)	(1.162)	(0.159)	(0.216)
2-9 Outlets	0.638	1.319**	-4.881***	-5.207***
	(0.376)	(0.498)	(0.143)	(0.165)
Zipcode Fixed Effects	Yes	Yes	Yes	Yes
NAIC Fixed Effects	Yes	Yes	Yes	Yes
Observations	1098838	780960	1102358	783889
R-Squared	0.413	0.455	0.120	0.106

Dependent variable is median distance traveled in kilometers to the store in Columns (1)-(2). Dependent variable is the median length of time spent in the store in Columns (3)-(4). * p <0.10, ** p <0.05, *** p <0.01. Data for February 2020. Robust standard errors clustered at the zipcode level.

表4以回归形式给出了旅行距离和花费时间的描述性统计。在第(1)和(3)栏中，我们给出了完整样本的结果。在第(2)和(4)栏中，我们只显示非必要业务的结果结果与图3相似。人们去商店的路程是非线性的。一般来说，如果一个品牌足够大，足够普遍，并且有数百家商店，那么人们就不太可能去很远的地方。当一个品牌拥有5000家门店时，它可能会尽量减少潜在客户的出行时间。

文献二：零售商店如何复工复产？——结果解读

Table 5: Stand-Alone Stores And Large Footprint Brands Appear to Attract the Fewest Canadian Visitors

	All	Non-Essential Businesses	All	Non-Essential Businesses
	(1)	(2)	(3)	(4)
	Canadian Visitors	Canadian Visitors	Mexican Vistors	Mexican Vistors
> 5000 Outlets	-0.00328***	-0.00237***	-0.000517*	-0.000786
	(0.000383)	(0.000511)	(0.000206)	(0.000406)
1000-4999 Outlets	-0.00191***	0.00142***	-0.000133	-0.0000793
	(0.000375)	(0.000409)	(0.000138)	(0.000187)
100-999 Outlets	0.00678***	0.000224	0.00104**	0.000676
	(0.000709)	(0.000287)	(0.000352)	(0.000425)
10-99 Outlets	0.000989**	0.000598	0.000418	0.000687
	(0.000361)	(0.000445)	(0.000307)	(0.000487)
2-9 Outlets	0.000131	0.000143	0.0000666	0.000112
	(0.000216)	(0.000214)	(0.000186)	(0.000247)
Zipcode Fixed Effects	Yes	Yes	Yes	Yes
NAIC Fixed Effects	Yes	Yes	Yes	Yes
Observations	257964	187958	239285	167130
R-Squared	0.0626	0.0504	0.0878	0.107

Dependent variable is whether the store received visitors from Canada. * p <0.10, ** p <0.05, *** p <0.01. Sample restricted to states that border Canada. Data for February 2020. Robust standard errors clustered at the zipcode level.

表5显示了访问商店的外国访问者如何随商店的足迹而变化。一般来说。加拿大游客更有可能去中等规模的商店，而不太可能去比较普遍或只有一个店面的商店。一种简单的解释是，这些访问反映了加拿大人越过边境，试图在低于加拿大价格的商店购物，而在加拿大不存在的以价格为导向的商店往往有这些中间阶层的足迹。墨西哥游客也有类似的情况，但我们的许多估计并不重要。

文献二：零售商店如何复工复产？——结论

- 小型企业，特别是只有一个零售店面的商店首先开业。这些商店吸引的访客越来越少，回头客越来越多
- 与那些品牌相对普遍的商店相比，他们的顾客确实走得更远，在店里花的时间也更长

文献二：零售商店如何复工复产？——总结与讨论

本文主要创新点

- 数据创新

本文的可重复与可复制性：对数据复制要求较高

文献三：机器学习方法的应用——研究背景

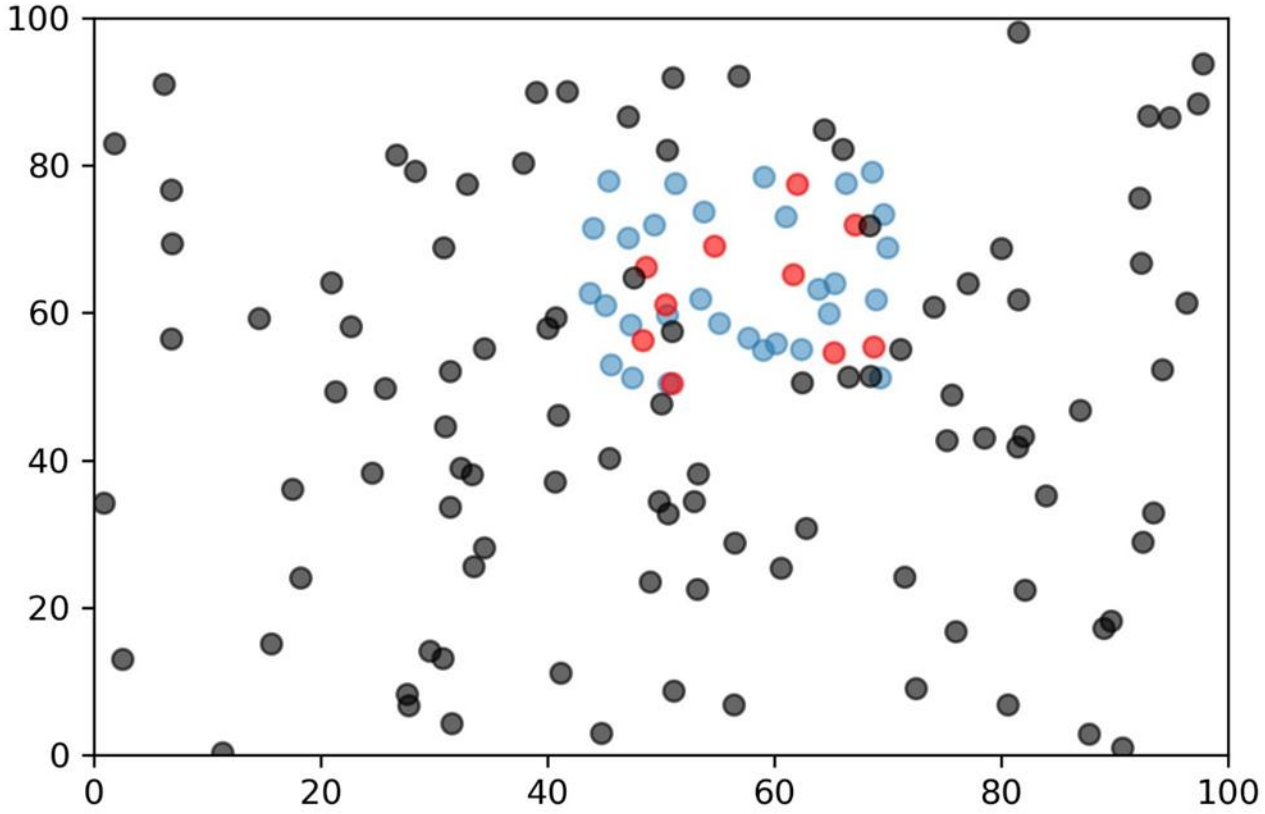
Billings, S. B., and E. B. Johnson. Agglomeration within an Urban Area[J]. Journal of Urban Economics, 2016, (91): 13-25.

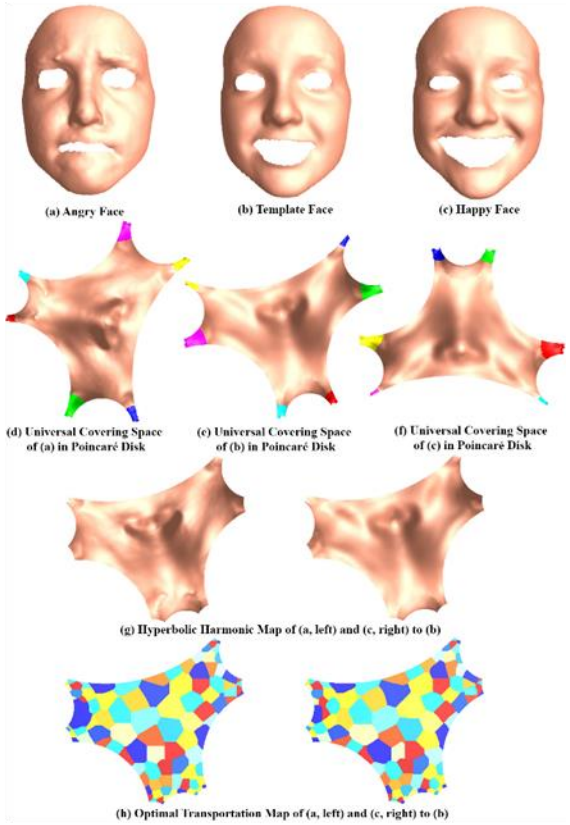
Previous work:

- Wasserstein距离是度量两个概率分布之间的距离度量函数，在数学领域的概率论和数理统计中被广泛应用，近年来由于其优异的数学特性，在计算机领域得到了进一步拓展，被应用于人脸识别、图片分析等人工智能的机器学习训练领域

- 在经济学领域，Wasserstein距离被应用到产业空间分布相似性的衡量上，并且通过蒙特卡洛模拟确定相似度的显著概率

文献三：机器学习方法的应用——方法





文献三：机器学习方法的应用——方法

Step1: Calculate the Wasserstein Distance of spatial distribution between two industries

$$W_{j,k}=\left(\inf_{\gamma\in\Gamma(f_j,f_k)}\int_{\mathbb{R}^2\times\mathbb{R}^2}d(x,y)^2\,d\gamma(x,y)\right)^{1/2}\qquad(1)$$

在实际计算过程中，两个分布的样本规模对计算复杂度的影响较大，采用线性规划求解的结果存在精度不高与计算复杂度过高的双重问题。为了在提高求解精度的情况下缩短求解的时间，Cuturi（2013）通过证明提出了矩阵扩张算法中的Sinkhorn（1964）不动点迭代和增加熵正则化约束项以实现加速求解，这一方法在Carlsson et al.（2018）的应用研究中得到了改进，并由Luise et al（2018）进一步证明了使用Sinkhorn算法后的Wasserstein距离结果的稳健性。增加熵约束项后的Wasserstein距离为：

$$W_p(\mu(x),\nu(y)):=\left(\inf_{\gamma(x,y)\in\Gamma(\mu,\nu)}\int_{\mathbb{R}^n\times\mathbb{R}^n}d(x,y)^p\,\gamma(x,y)dxdy\text{-}\alpha\Omega(\gamma)\right)^{1/p}\qquad(2)$$

进一步定义核密度（Kernel）函数：$\kappa(x,y)=e^{-td(x,y)^p}$（3）

对于所有 t > 0的情况下，该式是正定的，将式（3）代入式（2）中，可得：

$$W_p(\mu(x),\nu(y))_\kappa:=\left(\inf_{\gamma(x,y)\in\Gamma(\mu,\nu)}\int_{\mathbb{R}^n\times\mathbb{R}^n}\gamma(x,y)[\alpha\ln(\gamma(x,y))-\ln(\kappa(x,y))/t]dxdy\right)^{1/p}\qquad(4)$$

式（4）即为基于Sinkhorn算法的熵正则化Wasserstein距离计算公式，在实际求解过程中一般令$t=1/\alpha$，在本文的计算中取α的经验值为1。

文献三：机器学习方法的应用——方法

Step2: 构建反事实样本 construct our counterfactual of randomly located (pseudo) industries

原假设（**Null hypothesis**）：

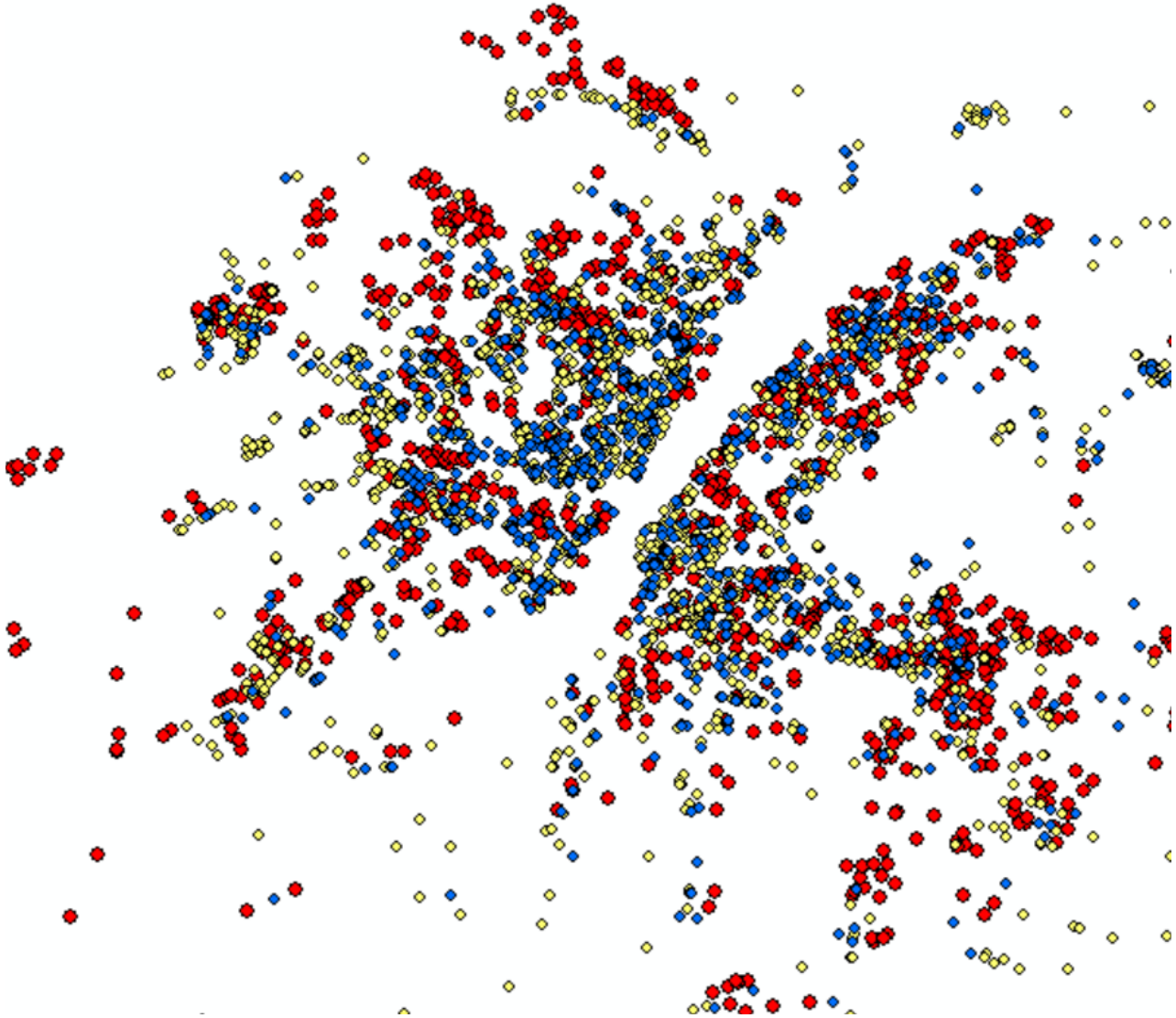
在基于分布j空间分布既定的情况下，分布j和分布k之间不存在空间分布相似性。
no spatial similarity between industry j and industry k conditional on the spatial density of industry j $\hat{k}$ $\hat{k}$

即在 $\hat{k}$ 上随机选取对于分布j而言的模拟分布区位集 ，分布区位集 必须满足两个条件，第一个条件是分布 的样本数量与分布k的数量应当相同；第二个条件是选取随机分布区位集策略是从空间上所有已知的所有潜在所在点的集合中进行抽取。这一抽取将重复1000次。 $W_{j,\hat{k}}$

$W_{j,\hat{k}}$

将分布j与实际分布k的Wasserstein的距离$W_{j,k}$与这1000次反事实结果进行比较，将$W_{j,k}$小于的次数记为n，进而定义产业共聚指数$Coagg_{j,k}=n/1000$，例如，如果1000次模拟中，有950次$W_{j,k}$小于 ，则$Coagg_{j,k}$为0.95，则可以理解为，在95%的置信度下可以拒绝原假设，接受备择假设，即基于分布j空间分布既定的情况下，分布j和分布k之间存在空间分布相似性，这种空间相似性是由分布j向分布k的方向共聚产生的。

文献三：机器学习方法的应用——方法



病例发病的空间分布
公共场所的空间分布

文献三：机器学习方法的应用——初步结果

1	超级市场	0.001	0.144677	0.001	0.141773	0	0.142069	0.002	0.142723
1	外国餐厅	1	0.133138	1	0.128979	1	0.129269	1	0.130446
2	特色商业	0.973	0.137957	0.953	0.134608	0.966	0.134786	0.976	0.135574
2	影剧院	1	0.131681	1	0.127262	1	0.127518	1	0.128867
3	学校	0	0.144361	0.056	0.139606	0.015	0.140044	0	0.141929
3	培训机构	1	0.138398	1	0.134487	1	0.134788	1	0.136055
4	诊所	0	0.145289	0	0.142634	0	0.142854	0	0.143375
4	综合医院	1	0.134942	1	0.131121	1	0.131364	1	0.132505
4	专科医院	1	0.134179	1	0.130345	1	0.130588	1	0.131602
6	便民商店	0	0.147165	0	0.144202	0	0.144494	0	0.145309
6	餐饮相关	1	0.136331	1	0.132892	1	0.133152	1	0.134057
7	家居建材	0	0.149961	0	0.14756	0	0.147827	0	0.148354
7	专卖店	1	0.139202	1	0.136046	1	0.13628	1	0.13703

参考文献

1. Pare, Philip E.; Liu, Ji; Beck, Carolyn L.; Kirwan, Barrett E. and Basar, Tamer. "Analysis, Estimation, and Validation of Discrete-Time Epidemic Processes." IEEE TRANSACTIONS ON CONTROL SYSTEMS TECHNOLOGY, 2020, 28(1), pp. 79-93.

2. Avi Goldfarb, Catherine Tucker. Which Retail Outlets Generate the Most Physical Interactions? NBER Working Paper 27042, http://www.nber.org/papers/w27042.

3. Billings, S. B., and E. B. Johnson. Agglomeration within an Urban Area[J]. Journal of Urban Economics, 2016, (91): 13-25.