

第十五届数学中国数学建模网络挑战赛

地址：数学中国数学建模网络挑战赛组委会
电话：0471-4969085

邮编：010021

网址：www.tzmcm.cn
Email: service@tzmcm.cn

第十五届“SPSSPRO 杯”数学中国

数学建模网络挑战赛 承 诺 书

我们仔细阅读了第十五届“SPSSPRO 杯”数学中国数学建模网络挑战赛的竞赛规则。

我们完全明白，在竞赛开始后参赛队员不能以任何方式（包括电话、电子邮件、网上咨询等）与队外的任何人（包括指导教师）研究、讨论与赛题有关的问题。

我们知道，抄袭别人的成果是违反竞赛规则的，如果引用别人的成果或其他公开的资料（包括网上查到的资料），必须按照规定的参考文献的表述方式在正文引用处和参考文献中明确列出。

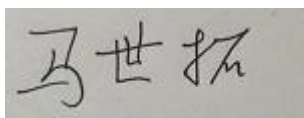
我们郑重承诺，严格遵守竞赛规则，以保证竞赛的公正、公平性。如有违反竞赛规则的行为，我们接受相应处理结果。

我们允许数学中国网站(www.madio.net)公布论文，以供网友之间学习交流，数学中国网站以非商业目的的论文交流不需要提前取得我们的同意。

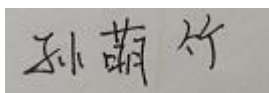
我们的参赛队号为：1019

参赛队员（签名）：

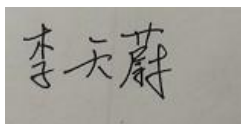
队员 1:



队员 2:



队员 3:



参赛队教练员（签名）：

参赛队伍组别（例如本科组）：本科组

第十五届数学中国数学建模网络挑战赛

地址：数学中国数学建模网络挑战赛组委会
电话：0471-4969085

邮编：010021

网址：www.tzmcm.cn
Email: service@tzmcm.cn

第十五届“SPSSPRO 杯”数学中国

数学建模网络挑战赛 编号专用页

参赛队伍的参赛队号：1019

竞赛统一编号（由竞赛组委会送至评委团前编号）：

竞赛评阅编号（由竞赛评委团评阅前进行编号）：

第十五届数学中国数学建模网络挑战赛

地址：数学中国数学建模网络挑战赛组委会
电话：0471-4969085

邮编：010021

网址：www.tzmcm.cn
Email: service@tzmcm.cn

2022 年第十五届“SPSSPRO 杯”数学中国 数学建模网络挑战赛第一阶段论文

题 目 风险视域下基于改进聚类法的 COVID-19 监测点分
布研究

关 键 词 复杂网络, KMeans, 风险视域, 粒子群算法, 神经网络

摘 要:

在新冠病毒持续爆发而全面检测困难的当下,从污水中进行采样成为了一种高效快捷的方法。但以美国本土作为研究对象,其污水采样点分布的合理性和效率仍然存在一些不合理。基于此,针对提出的问题,给出了统计意义上的一系列方案,并进行了较为全面的挖掘与思考。

问题一要求对监测点分布进行评价并再新增 10 个点,我们通过爬虫解析获取每个站点的经纬度,并从分布整体角度出发探究分布合理性。我们基于 **KMeans 聚类法**,得到当其聚类数为 5 的时候效果最好并基于地理信息学中的一些评价原则做出评价;另外我们构建了**复杂网络**模型,基于图论原理和拓扑性质对网络做出评价。我们认为,监测点分布存在整体上分散局域上集中的特性,即监测点以小集团的形式分布,并且承载服务量过高,不利于工作开展。而对于新增点的选择,我们基于 **K-Means**,以 10 个点的坐标为变量引入进行聚类,再以 **SSE** 为目标函数使用**粒子群算法**优化。同时我们还对某地点传播新冠的规模和被监测的注意力进行定义并引入 **SSE** 中,形成**风险视域**下的改进聚类法优化模型。经求解,我们提出的改进模型能有效解决风险问题,使结果更加合理。

问题二要求对新冠规模进行预测,我们同样从官方获取了这些站点 3 个月的历史数据,考虑到数据特征与体量我们使用 **BP 神经网络**构建分类模型确定每个地点的危险等级。最终分类准确率达到 0.954, F1 分数为 0.96,是相当好的水准。我们基于 **BP 神经网络**模型预测一日后、一周后、半月后和一月后的结果进行了初步分析,并撰写了建议信附后。

模型结合了多种方法的优点,具有可靠性,并且结果具有良好的可解释性。从常用的衡量指标来看,模型的表现无疑是非常优秀的,能够高效率地进行分布评价、预测以及背后原因探究,并生成最优策略,在实际应用中有一定价值。

所属类别	2021 年“SPSSPRO”杯全国大学生数学建模竞赛	参赛编号
本科组		1019

风险视域下基于改进聚类法的 COVID-19 监测点分布研究

摘 要

在新冠病毒持续爆发而全面检测困难的当下，从污水中进行采样成为了一种高效快捷的方法。但以美国本土作为研究对象，其污水采样点分布的合理性和效率仍然存在一些不合理。基于此，针对提出的问题，给出了统计意义上的一系列方案，并进行了较为全面的挖掘与思考。

问题一要求对监测点分布进行评价并再新增 10 个点，我们通过爬虫解析获取每个站点的经纬度，并从分布整体角度出发探究分布合理性。我们基于 **KMeans 聚类法**，得到当其聚类数为 5 的时候效果最好并基于地理信息学中的一些评价原则做出评价；另外我们构建了**复杂网络**模型，基于图论原理和拓扑性质对网络做出评价。我们认为，监测点分布存在整体上分散局域上集中的特性，即监测点以小集团的形式分布，并且承载服务量过高，不利于工作开展。而对于新增点的选择，我们基于 **K-Means**，以 10 个点的坐标为变量引入进行聚类，再以 **SSE** 为目标函数使用**粒子群算法**优化。同时我们还对某地点传播新冠的规模和被监测的注意力进行定义并引入 **SSE** 中，形成**风险视域**下的改进聚类法优化模型。经求解，我们提出的改进模型能有效解决风险问题，使结果更加合理。

问题二要求对新冠规模进行预测，我们同样从官方获取了这些站点 3 个月的历史数据，考虑到数据特征与体量我们使用 **BP 神经网络**构建分类模型确定每个地点的危险等级。最终分类准确率达到 0.954，F1 分数为 0.96，是相当好的水准。我们基于 **BP 神经网络**模型预测一日后、一周后、半月后和一月后的结果进行了初步分析，并撰写了建议信附后。

模型结合了多种方法的优点，具有可靠性，并且结果具有良好的可解释性。从常用的衡量指标来看，模型的表现无疑是非常优秀的，能够高效率地进行分布评价、预测以及背后原因探究，并生成最优策略，在实际应用中有一定价值。

关键词：复杂网络，**KMeans**，风险视域，粒子群算法，神经网络

一、 问题重述

1.1 问题背景

冠状病毒大流行给世界各地的生活和经济带来了混乱。考虑到大多数人不会接受检测，世界各地的十多个研究小组已经开始分析新型冠状病毒的废水，作为估计社区感染总数的一种方法。到目前为止，研究人员已经在荷兰、美国和瑞典的污水中发现了该病毒的痕迹。分析废水——通过排水系统进入处理设施的废水——是研究人员追踪通过尿液或粪便排出的传染病的一种方法，如 SARS-CoV-2[8]。

1.2 问题提出

我们需要解决的问题如下：

问题一：根据附注网站上的数据，判断美国污水监测采样点的分布的合理性，建立数学模型，增加十个采样点，选择最合理的位置设置采样点。

问题二：分析数据，对于可能出现的大规模疫情进行预警，并给当地的政府写一封信，说明对疫情的预判，对情况严重性的估计，并给出一定的防控建议。

二、 问题分析

2.1 问题一的分析

在问题一中，我们首先研究美国污水检测采样点分布是否合理。我们应该确定的是，此处的评价不能仅仅对于单个点进行，而应该对整个分布进行评价。评价的方法有两类。其一是图论计算指标，其二是聚类。我们通过计算可得将点分成 5 类。其次再研究如何增加 10 个采样点，使其位置最为合理。在前述分析中我们将点分成 5 类，同类的点可以用一个**聚类簇**完全覆盖。因此我们可以得到 5 个**聚类簇中心**的位置，再将五个**簇心**两两距离的平方加和得到对应的 SSE。再添入 10 个随机的点，重新将所有点分为 5 类，得到 5 个类后重新计算**簇心**对应的 SSE，最合适的 10 个点位置即对应最小 SSE 的取值。而在后续的分析中，我们对该模型进行了改善。我们发现风险传播和盲区分级在建立模型中是不可或缺的。我们将风险传播和盲区分级通过加权引入到 SSE 中，重新构造了目标函数，同理求最小。该模型我们将其称为“风险视域下的 K-Means 聚类优化模型”。

2.2 问题二的分析

在问题二中，我们通过爬取得到污水中检测的历史数据。由于数据量颇大，近 19 万。而这海量数据给了我们一个启发，通过神经网络对模型进行构建，从其他属性与先前的数据中推算下一日污水中病毒的含量，进而对疫情进行预判。通过结果我们发现最终的契合度达到了 95%，事实证明，这种方法是可行的。

建议信见后文。

三、 模型假设

针对这些问题，我们的模型假设如下：

- 1.观察到设点位置和非设点位置经纬度悬殊较大，我们假设类别的失衡不会造成严重影响，或者说造成的影响在我们模型的误差范围之内。
- 2.假设在官网抽取的数据能代表数据的一般情况。

四、 符号说明

符号变量的具体释义以正文解释为准，这里列举一些重要变量：

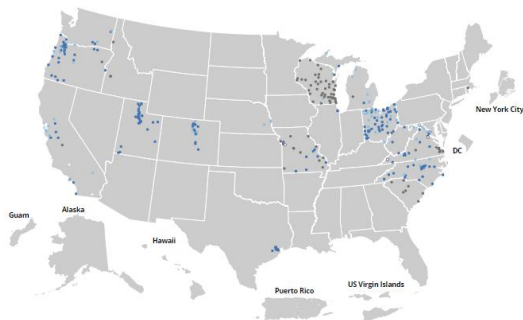
符号	说明
n	数据点数量
d_{ij}	ij 之间的距离
c	15 日百分比变化
P	传播风险
M	监测盲区风险
SSE	距离平方和
ρ	密度衡量
R	分布指数衡量
A	研究区域面积
W, b	权重与偏置项
f	激活函数

五、模型的建立与求解

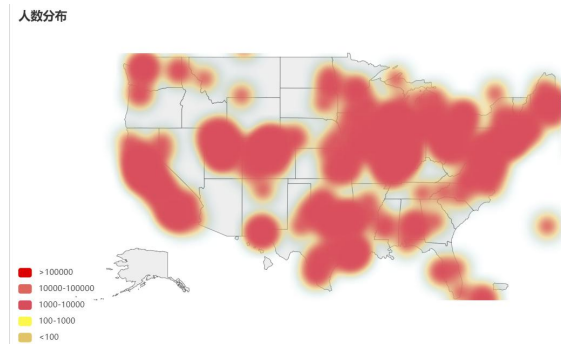
5.1 问题一模型的建立与求解

5.1.1 数据的探索与描述

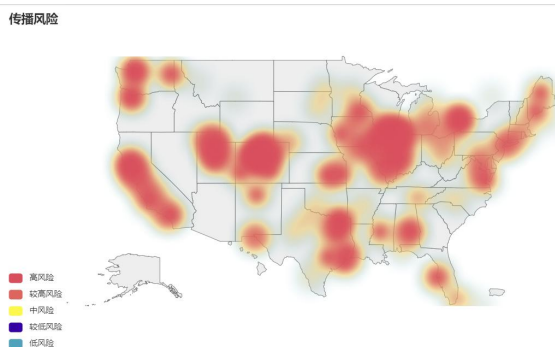
数据来源于美国 CDC 官网[1]，由于其每天都在发生变化，这里使用的是 2022 年 4 月 7 日北京时间 20:00 整获取的数据。该数据给出了 701 个废水监测站点的分布图像，数据缺失率为百分比变化缺失 42.6%，绝对浓度缺失 33.1%，可以看到其分布如图 1（1）所示：



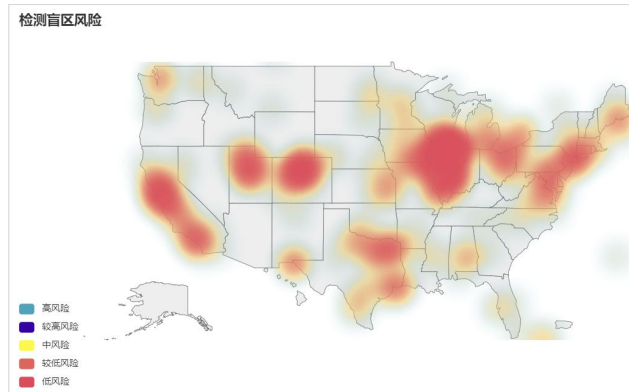
(1). 散点图



(2). 人数分布



(3). 传播风险



(4). 检测盲区风险

图 1. 散点分布图&人数分布图&风险图&注意力盲区图

我们通过百度地图 api 解析获得了这些站点的经纬度数据，若服务同一个城市则经纬度按照城市的经纬度计算，同时服务不同城市则按照这些城市地理位置的中心计算。平均每个城市的服务人数为 160415.01，从图 1（2）当中可以很明显看出，红色区域明显面积很大，每个站点承载量是比较高的，同时分析其统计分布，它是一个不均匀的偏峰分布，说明站点承载量也不够均匀。

与此同时，我们对监测点的风险有如下两重定义：

第一是传播风险，我们以 15 日平均变化率定义危险程度，按照李克特五级量表原则定为 5 个等级，分别为低风险、较低风险、中风险、较高风险和高风险五个等级，这 5 个等级的阈值界定如表 1 所示。

第二是注意盲区，我们以与监测点的距离衡量某一点被监测到的可能性（注意力）。同样的，我们也定义了密切观察区、高注意力、中等注意力、低注意力和检测

盲区五个等级，阈值界定也在表 1 中列出。

表 1. 两种风险的阈值界定[2]

危险等级	传播风险	阈值界定	注意盲区风险	阈值界定
1	低风险	$c < 0\%$	密切观察区	$d < 20\text{km}$
2	较低风险	$0\% < c < 100\%$	高注意力	$20\text{km} < d < 50\text{km}$
3	中风险	$100\% < c < 1000\%$	中等注意力	$50\text{km} < d < 100\text{km}$
4	较高风险	$1000\% < c < 2000\%$	低注意力	$100\text{km} < d < 200\text{km}$
5	高风险	$c > 2000\%$	检测盲区	$d > 200\text{km}$

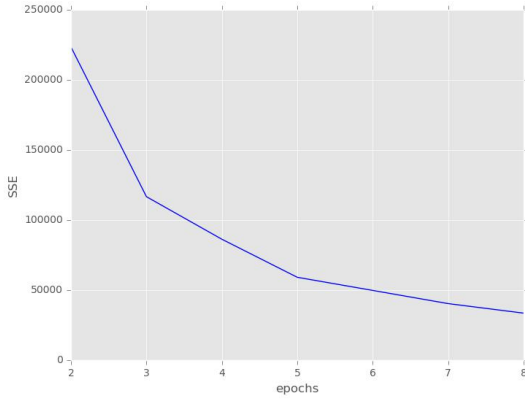
图 1（3）和图 1（4）分别描述了传播风险和注意盲区风险。可以看到，传播风险区高风险分布较为集中，尤其是美国西海岸，说明西海岸应该设置更多监测点然而并没有；而注意盲区则是空白区域面积，通过统计，盲区面积占到了美国本土面积的 67%，说明有相当一部分地区是处于传播的高危风险区域，却没有被得到应有的检测重视。

5.1.2 基于聚类方法的点分布评价

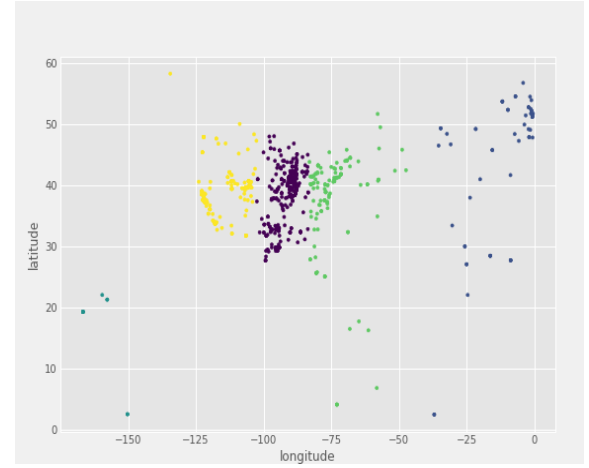
从单个点的角度考虑不够全面，没有考虑到整体情况。为了考察点的整体分布，下面我们结合聚类的数据来进行基本分析。

K-means 算法是最为经典的基于划分的聚类方法，是十大经典数据挖掘算法之一。简单的说 K-Means 就是在没有任何监督信号的情况下将数据分为 K 份的一种方法。聚类算法就是无监督学习中最常见的一种，给定一组数据，需要聚类算法去挖掘数据中的隐含信息[3]。它基于欧几里得距离，在数据集中根据一定策略选择 K 个点作为每个簇的初始中心，然后观察剩余的数据，将数据划分到距离这 K 个点最近的簇中，也就是说将数据划分成 K 个簇完成一次划分，但形成的新簇并不一定是最好的划分，因此生成的新簇中，重新计算每个簇的中心点，然后在重新进行划分，直到每次划分的结果保持不变。在实际应用中往往经过很多次迭代仍然达不到每次划分结果保持不变，甚至因为数据的关系，根本就达不到这个终止条件，实际应用中往往采用变通的方法设置一个最大迭代次数，当达到最大迭代次数时，终止计算。

我们以这些点的经纬度坐标、服务人口和 15 日浓度变化率为变量进行探索性 K-Means 聚类，聚类肘部图和最终结果如图 2 所示。从图 2（1）的肘部图中可以清晰看到 SSE 随着 k 值的不同而变化，这一结果表明聚类数为 5 时最合适。空间点的分布是相对较散的。图 2（2）的聚类图中横纵坐标分别为经纬度，可以看到 5 类点在空间中的集中性。



(1). 肘部图



(2). 聚类结果

图 2. 聚类肘部图和聚类结果描述

图 2(1) 中肘部图从 $K=5$ 开始斜率明显变缓, 说明聚类以 $k=5$ 为拐点。而图 2(2) 中聚类数量为 5 时聚类图主要按照纬度将其划分为 5 个区域。美国东北部面临五大湖, 水质资源较好, 同时与加拿大接壤, 所以重点监测; 中频带是横跨美国东中西步; 绿色区域主要集中在美国东部, 而低纬度地区则是以墨西哥湾、美国中部附近, 是人流量大的密集区域, 分布较为密集。但若从经度的角度看, 美国西海岸的重点城市例如芝加哥、西雅图等城市的采样则分布较为稀疏。

另外, 我们还通过密度分布法衡量点分布的分散性。一种方法是用每个点与其他点距离最小值的倒数均值, 计算方法为:

$$\rho = \frac{1}{n} \sum_{i=1}^n \frac{1}{d_{\min i}} \quad (1)$$

经求解, 这一结果为 158.7, 说明实际上有不少采样点分布非常紧密

另一种方法基于如下原理[4]:

$$R = \frac{\frac{1}{n} \sum_{i=1}^n d_{\min i}}{\frac{1}{2} \sqrt{\frac{A}{n}}} \quad (2)$$

经求解, 这一结果为 $2.737 > 1$, 说明从整体上看分布是较为均匀的, 似乎与上面的结果形成了冲突。但这两个结果实际上并不矛盾, 恰恰是分集中的基团相隔太远所以造成了这一假象。

5.1.3 基于复杂网络方法的点分布评价

另一种方法是基于点的空间距离构造复杂网络。由于数据点密集分布, 我们认为基于复杂网络理论去构建图论图可以从网络的拓扑性质入手进行分析, 从而得到更准确的结果。

我们以 200km 的盲区阈值为边界, 小于 200km 的边则添加到图中去, 形成的复杂网络关系图如图 3 所示:

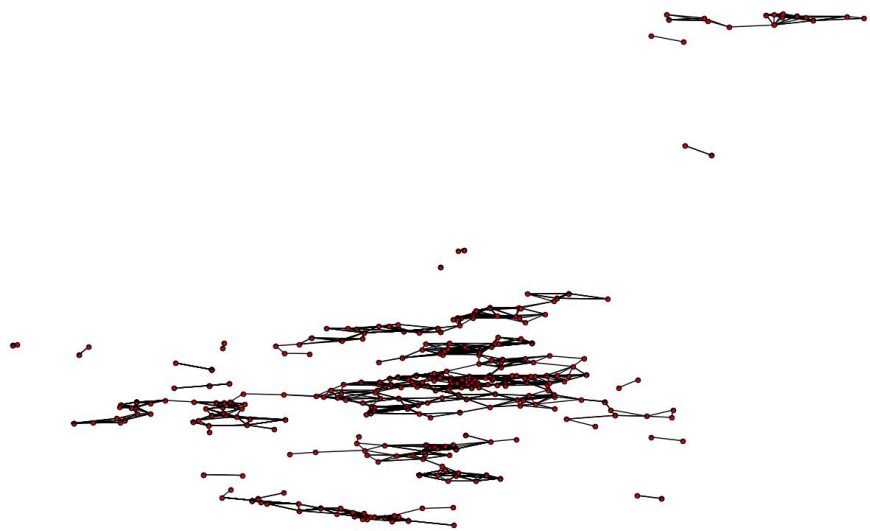


图 3. 复杂网络关系图

基于这一幅图，我们计算出了它的一些重要指标，并列举在表 2 中：

表 2. 复杂网络的几个指标值

编号	网络指标	数值
1	平均度数	5.37
2	平均聚类系数	0.429
3	网络传递性	0.393
4	连通子图规模	18
5	平均最短路	8.219
6	平均度中心性	0.0176
7	平均距离中心性	0.0528
8	平均点介数中心性	0.004

表 2 中的几个指标能够反映点的分布情况。聚类系数用于衡量节点聚集的程度。有证据表明，大多数现实世界的网络中，特别是在社交网络中，节点倾向于创建相对紧密联系的群体；这种可能性往往大于在两个节点之间随机建立关系的平均概率[4,5]。聚类系数仅为 0.429，几个中心性指标也都不高，说明它们从整体上来看是比较分散的，连通子图个数有足足 18 个。但平均度数反映一个监测点平均与多少监测点更近，指标显示平均一个节点有平均超过 5 个节点与其相邻，平均最短路也仅为 8.219km，从这个角度来看网络又是比较密集的。从这些结果当中可以看出，整个网络整体较为分散，但局部很集中，这是非常危险的一个信号。

5.1.4 基于聚类方法的最优点选址

5.1.2 中我们已经通过聚类的方法得到 5 类最优，那么下面我们基于这一原理进行目标函数的构造。

假设以目标 10 个点的经纬度坐标为变量（一共 20 维），添加至原数据以后再一

次进行 $K=5$ 的 K-Means 聚类，所求得的 SSE 即为优化目标。以数学形式表达，记五个聚类簇中心为 C_j ，那么 SSE 的值可以表示为：

$$SSE = \sum_{i=1}^{n+10} \sum_{j=1}^5 \|x_i - C_j\|_2^2 \quad (3)$$

由于函数维度过高且非常规，不适合用传统优化方法求解，这里我们采用粒子群算法进行求解。

粒子群算法是由 Eberhart 和 Kennedy 于 1995 年提出的一种全局搜索算法，是一种模拟自然界的生物活动以及群体智能的随机搜索算法[6]。粒子群优化算法要求每个粒子在寻优的过程中维护两个向量，速度向量 v 和位置向量 p ，粒子的速度决定了其运动的方向和速度，而位置则体现了粒子所代表的解在解空间中的位置，是评估该解质量的基础。算法同时还要求每个粒子各自维护一个自身的历史最优位置向量($pBest$)和群体还维护一个全局最优向量($gBest$)。算法的伪代码如下：

算法 1. 粒子群算法

```
init x=[xi1,xi2,xi3,...,xin]
cal fit for each xi
t=0
while t<MaxIteration:
    for xi in x:
        if fit is better than p_best:
            p_best=fit
        if fit is better than g_best:
            g_best=fit
        v[t+1]=w*v[t]+c1*(p_best-x[t])+c2*(g_best-x[t])
        x[t+1]=x[t]+v[t+1]
    t=t+1
print best_result
return x
```

经求解，得到的坐标点如表 3 和图 4 所示。粒子群算法的迭代优化曲线如图 5(1) 所示。可以看到最终得到的坐标点相对还是较为密集，说明这一方法还有待改进。

5.1.5 基于风险视域下改进聚类方法的最优点选址

5.1.4 中的模型没有考虑到目标点的风险。这里我们在 SSE 函数中引入两种风险函数对模型进行修正：

对传播风险的量化定义为：每个监测点到目标点的传播风险叠加。而单个监测点到目标点的传播风险我们用这样一个式子表示：

$$P_{ij} = ce^{-\lambda d_{ij}} \quad (4)$$

其中 λ 为参数，为了方便起见取 0.05。这里 c 代表该监测点的浓度， d 为目标点与监测点的距离，以此衡量风险传播水平

而盲区风险 M 则直接基于表 1 的界定给出其数值即可。

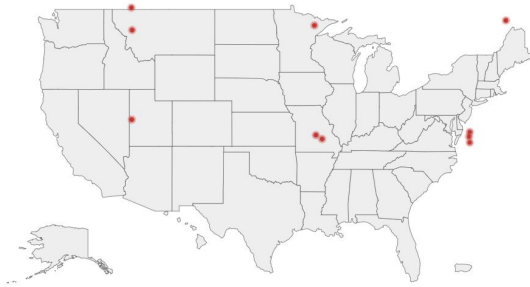
那么在引入两种风险以后，目标函数改进为：

$$FSSE = \sum_{i=1}^{n+10} \sum_{j=1}^5 M_i \left(\sum_{k \neq i} P_{ik} \right) \|x_i - C_j\|_2^2 \quad (5)$$

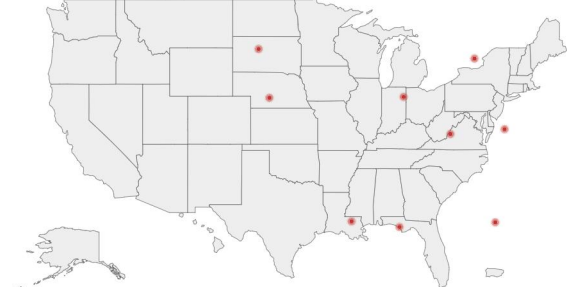
这一函数更复杂，我们同样用粒子群算法求解，并将求解结果绘制在表 3 和图 4 中，迭代曲线如图 5（2）所示。

表 3. 10 个点的坐标

编号	SSE 经度	SSE 纬度	FSSE 经度	FSSE 纬度
1	-74.22	38.26	-101.18	44.91
2	-92.40	47.61	-77.22	44.12
3	-91.50	37.65	-99.98	40.86
4	-74.31	37.86	-74.91	30.53
5	-113.76	39.35	-85.54	30.14
6	-113.81	49.16	-90.86	30.61
7	-74.23	37.34	-73.88	38.26
8	-92.22	37.99	-79.89	37.86
9	-70.00	48.04	-95.63	20.01
10	-113.72	47.20	-85.09	40.94

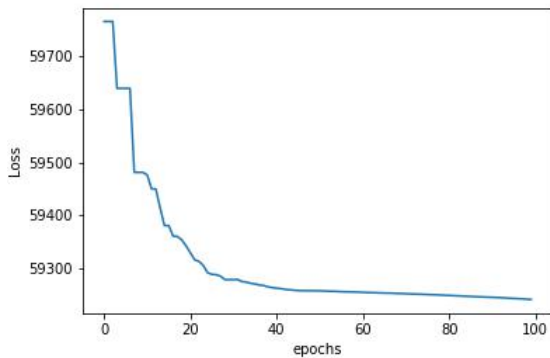


(1). SSE 模型的散点

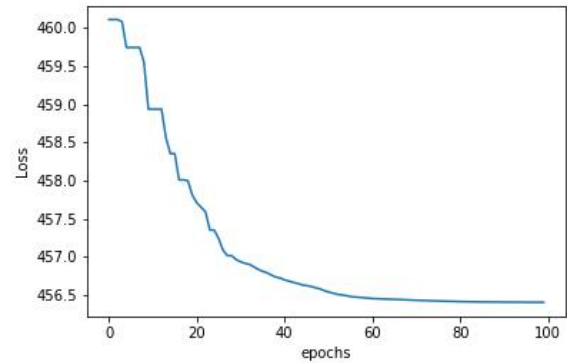


(2). 风险视域下改进 SSE 模型的散点

图 4. 两种模型的结果散点图



(1). SSE 模型的粒子群优化曲线



(2). 风险视域下改进 SSE 模型的粒子群优化曲线

图 5.粒子群算法的两条迭代曲线

可以看到，引入风险视域以后检测盲区和高危区域得到了更多的注意力，初步达成了我们的目的。

5.2 问题二模型的建立与求解

5.2.1 神经网络模型预测新冠发生

为了更充分地利用数据，我们获取了历史这些监测点的 **ptc** 百分比变化和绝对浓度变化的数据，平均时间跨度大约 3 个月。

我们以地点经纬度、时间、服务人口、服务区域面积（同样通过解析得到）等作为属性，并分别以前一日统计的 **ptc** 百分比变化和绝对浓度作为属性，对应的，则以当日 **ptc** 百分比变化与绝对浓度按照阈值划分危险等级做分类。由于数据体量达到 19 万余条数据，适合使用神经网络处理，故这里使用神经网络。

BP 神经网络是一种多层前馈神经网络，特点是：信号向前传播，误差反向传播。通俗理解就是，BP 神经网络通过层与层向前传播，得到最终实际输出后，与期望输出做对比，通过“梯度下降”策略，逐层调节权重和阈值，最终得到与期望输出在误差允许范围内的神经网络模型[7]。图 6 为一个 BP 神经网络的模型图，若隐藏层和输出层的激活函数分别为 f_1 和 f_2 ，权重矩阵分别为 W_1 和 W_2 ，则神经网络的模型可以表示为下式：

$$\hat{y} = f_2(W_2 \cdot f_1(W_1 X + b_1) + b_2) \quad (6)$$

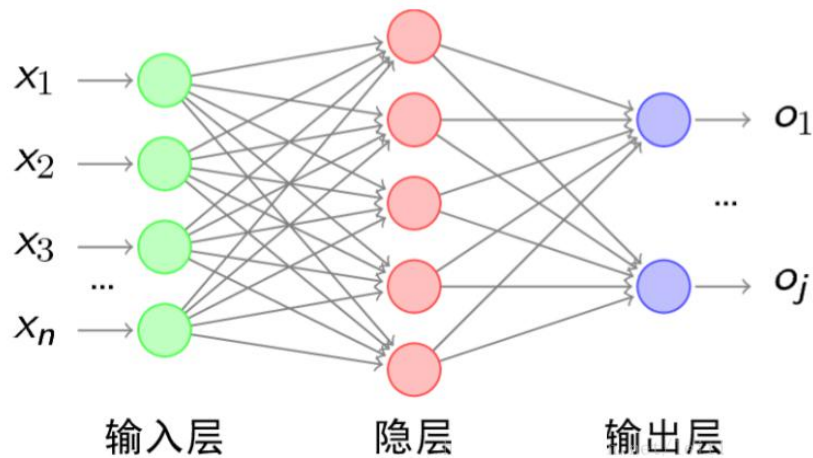
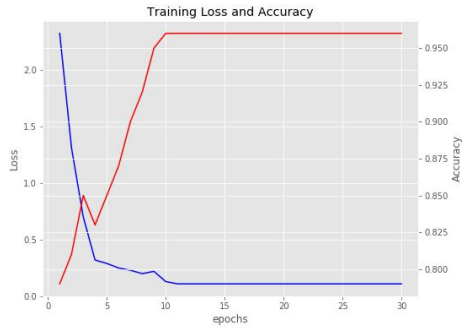


图 6. BP 神经网络的原理图

由于我们所进行的是分类神经网络，故我们以交叉熵损失作为训练损失，以分类准确率为目标，设置两层隐藏层分别包含 50、100 个隐藏单元，在这样的设置下我们将两个模型分别训练 30 轮，得到的曲线如图 7 所示：



(1). 使用变化百分比的模型



(2). 使用绝对浓度的模型

图 7. 神经网络的迭代曲线

从图 7 中我们很明显可以看到，基于绝对浓度去衡量危险的神经网络比 **ptc** 变化更能准确反映危险等级。因为 **ptc** 变化是不规律的，在数据点当中突变是一种很正常的现象，神经网络对突变点异常点的处理极不敏感，容易造成梯度无法收敛甚至梯度逸散的情况。故选用绝对浓度衡量危险等级的模型进行预测，在测试集上得到效果如表 4 所示：

表 4. 神经网络的效果评价指标

类别	precision	recall	F1-score	support
0	1.00	0.98	0.99	42076
1	0.71	0.77	0.74	2771
2	0.75	0.81	0.78	2856
3	0.72	0.71	0.72	1273
4	0.97	0.96	0.96	4556
accuracy			0.95	53532
Weighted-avg	0.96	0.95	0.96	53532

这里，F1 分数和准确率都是衡量模型准确性的评价指标，都是越接近 1 越高。可以看到，分类的准确率达到 0.96，是相当高的水平。因此我们根据美国经纬坐标范围和空间插值法分别生成次日、一周后、半月后和 1 月后的数据，并将这些数据绘制如图 8 所示：

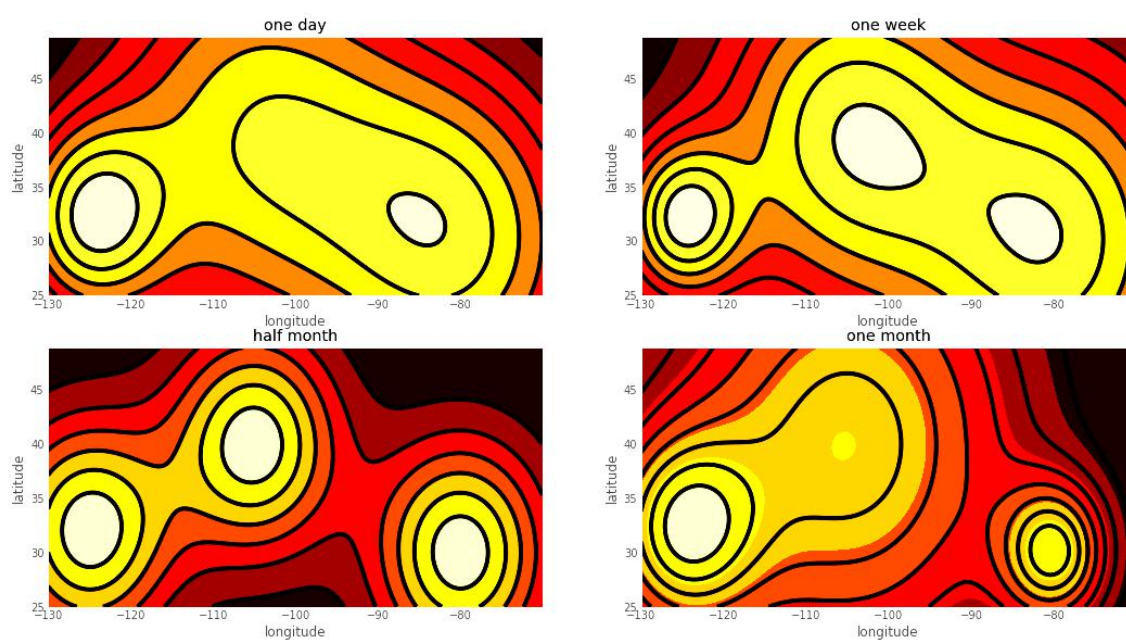


图 8. 风险的四个时间节点预测

图 8 中颜色越亮则表明风险越高。从图 8 中可以看出起始阶段美国东海岸与东北部、中部风险更高，另外是西海岸。但随着时间的迁移危险逐渐从东向西部转移，这将对美国西海岸西雅图、芝加哥等要地造成重大损失。

建议信

尊敬的政府领导：

您好！

我们对疫情的评判如下：在不对其进行人工干预的情况下，病毒的扩散会愈发迅速，并在一个月后呈现出大规模流行的局势。在我们所建立的模型中，不难发现病毒在开始时大多分布在沿海区域，而随着时间不断推进，病毒不断向内陆地区蔓延。这是一种必然的趋势，而在疫情防控中，我们所要采取的措施必须要抑制病毒蔓延并从根源上解决病毒传染问题。

对此，我有以下建议望您采纳：

1、建立应急指挥协调系统

建立并健全高效快捷的应急指挥协同系统，完善疫情防控应急预案，针对制定集中疫情暴发、病例散发等不同疫情形势的特定方案。实时分析判断疫情局势，严格执行 24 小时轮换工作制度，完善信息报告工作，提高工作效率。

2、健全疫情防控机制

教育并引导民众养成良好习惯，做好个人防护。加强对火车站、机场、地铁、公交车等交通站点及工具的管理检查，做好消毒通风工作。增强生物安全意识，探索构建生物安全等级标准体系。进一步加强区级疾病预防控制机构和社区卫生服务机构建设，全面提升基层防控和服务能力。针对疫情源头，制定详细的应对方案并严格执行。

3、完善疫情防控物资供应机制

设立区域应急物资储备中心，整合全州应急仓库，统筹物资储备管理。开发应急储备物资进出系统，建立统一的应急物资仓储信息平台。准确分类物资并按区块放置，使物资在紧急情况下得到最高效的利用。

4、完善专业化队伍培养机制

完善医疗队伍、社区人员队伍、保供队伍、志愿者队伍建设，建立应急专业队伍。设立卫生应急和疫情防控培训基地，建立常态化培训制度。政府加大资金支持，用于专业人才的知识培训，提高其应对能力。不断强化智库建设，为政府疫情防控和决策提供专业支撑，提高公共卫生事件的处理能力，巩固防疫基础力量。

5、完善疫情监测预警机制

健全公共卫生监测预警体系，加强基层医疗机构基础设施建设，提高医院、社区卫生服务中心发热筛查哨点预警能力，提升基层医疗卫生机构的传染病疫情初筛初诊能力，强化疾病预防控制业务现代化信息系统建设，搭建疫苗接种、核酸检测、流调追溯等系统性全民健康平台建设。

6、健全疫情医疗救治机制

完善分层级、分区域应急医疗救治体系建设，建立健全应急医疗救治机制，做到“早发现、早隔离、早诊断、早治疗、早报告”。建议建立新冠肺炎医院，切实保障人民群众身体健康和生命安全。

7、完善疫情应急信息化建设机制

加强公共卫生应急信息化建设，加快“互联网+”疫情医务建设，夯实疫情防控信息化技术基础。在疫情防控中，应广泛采用智能设备并利用信息技术平台，减少“人与人”直接接触，开展信息填报、数据采集、统计分析、会议研讨等工作，最大限度切断疫情扩散传播链。

8、完善疫情防控信息传播机制

进一步完善重大疫情信息公开制度，一旦疫情发生，政府相关部门要及时、准确、公开、透明地发布疫情，回应社会关切。健全舆情引导机制，以正确的信息引导社会。

六、 模型的评价、改进与推广

6.1 模型的优点

该模型从数据驱动的角度出发，通过基于聚类 and 复杂网络理论对分布进行评价，再基于聚类和风险传播构建优化模型，保证了结果的可靠性，随后我们根据一些模型评价指标评估，认为模型在本数据集上取得了较好的效果。我们认为，我们在这一系列问题中提出的模型有以下优点：

- 1.使用多种不同视角的模型进行评估，评估结果相较而言更为客观公正。
- 2.模型对整个分布进行评价，更为宏观整体。
- 3.模型逐步迭代更新，层层深入，不断改进，更具备递进性、创新性和说服力。
- 4.数据来源准确，分析方法合理，能够得出有一定参考价值的结论。

6.2 模型的缺点

尽管模型整体表现良好，但我们认为还存在一些不足之处，例如：

- 1.对于引入风险视域的改进 SSE 优化，模型形式相对更为复杂，求解时间较长。能否在编程过程中达成程序的并行化计算还有待探究。
- 2.对于传播风险和检测盲区风险的量化定义和风险传播机制建模还比较简单直接，需要考虑更有连续性的表达形式。

6.3 模型的改进

该模型除了可以使用神经网络作为分类器以外，其他的机器学习算法比如 SVM 也可以用于尝试。另外，对于改进的 SSE 模型优化除了传统粒子群算法还可以使用高度并行化的改进群智能算法进行求解。聚类也可以替换为 DBSCAN 聚类等。

6.4 模型的推广

该问题的求解方法不仅可以用于废水中新冠检测，更适用于一般空间中点分布的评价问题，以及对中-大规模的时空序列处理方案是能够迁移到一般时空序列处理的方法中的。除了进行 COVID-19 病毒的浓度检测，还可以适用于普通污染物的浓度检测、风险预测、风险传播等领域中去。

七、 参考文献

- [1] <https://data.cdc.gov/browse?limitTo=datasets>.
- [2] 张新秀,程耀东,张志华.基于 GIS 的城区急救中心空间分布合理性评价研究[J].应用科技,2014,41(01):43-47..
- [3] 王千,王成,冯振元,叶金凤.K-means 聚类算法研究综述[J].电子设计工程,2012,20(07):21-24.DOI:10.14022/j.cnki.dzsjgc.2012.07.034.Bicici Ufuk Can,Akarun Lale. Conditional information gain networks as sparse mixture of experts[J]. Pattern Recognition,2021,120:
- [4] D. J. Watts and Steven Strogatz (June 1998). "Collective dynamics of 'small-world' networks". Nature. 393 (6684): 440 - 442. Bibcode:1998Natur.393..440W. doi:10.1038/30918. PMID 9623998.)
- [5] Wu Cong,Li Hongxin,Ren Jiajia. Research on hierarchical clustering method based on partially-ordered Hasse graph[J]. Future Generation Computer Systems,2021(prepublish):
- [6] 陈贵敏,贾建援,韩琪.粒子群优化算法的惯性权值递减策略研究[J].西安交通大学学报,2006(01):53-56+61.杜丽英.基于数据挖掘的决策树算法分析[J].吉林建筑工程学院学报,2014,31(05):48-50.
- [7] 刘天舒. BP 神经网络的改进研究及应用[D].东北农业大学,2011.Chen T , Tong H , Benesty M . xgboost: Extreme Gradient Boosting[J]. 2016.
- [8] Mallapaty S . How sewage could reveal true scale of coronavirus outbreak[J]. Nature, 2020, 580(7802).

附录

环境: OS: Windows 10; CPU: Intel i7; GPU: NVIDIA GEFORCE 1650

Language:

Python 3.8.2 Jupyter notebook

文件列表:

图片:

代码: 复杂网络.ipynb, 神经网络.ipynb, 热力图.py, 数据处理.py

配置要求: 有 sklearn, 安装 tensorflow 2.2.0 即可

热力图.py 代码如下:

```
# 绘制瑞士地图

from pyecharts import options as opts
from pyecharts.charts import Geo
from pyecharts.datasets import register_url
from pyecharts.globals import GeoType
import pandas as pd

try:
    register_url("https://echarts-maps.github.io/echarts-countries-js/")
except Exception:
    import ssl
    ssl._create_default_https_context = ssl._create_unverified_context
    register_url("https://echarts-maps.github.io/echarts-countries-js/")

data=pd.read_csv("newdata.csv")
data=data[data['15-Day Percent Change ']!=-1].reset_index(drop=True)
geo = (
    Geo()
    .add_schema(maptype="美国")
    .set_global_opts(title_opts=opts.TitleOpts(title="美国"))
)
data_pair=[]
for i in range(data.shape[0]):
    geo.add_coordinate(data['County'][i], data['longitude'][i],data['latitude'][i])
    data_pair.append((data['County'][i],data['Population Served'][i]))
# 将数据添加到地图上
geo.add('', data_pair, type_=GeoType.HEATMAP, symbol_size=5)
# 设置样式
geo.set_series_opts(label_opts=opts.LabelOpts(is_show=False))
# 自定义分段 color 可以用取色器取色
pieces = [
    {'max': 100, 'label': '<100', 'color': '#E2C568'},
    {'min': 100, 'max': 1000, 'label': '100-1000', 'color': '#FCF84D'},
```

#1019

```
        {'min': 1000, 'max': 10000, 'label': '1000-10000', 'color': '#D94E5D'},
        {'min': 10000, 'max': 100000, 'label': '10000-100000', 'color': '#DD675E'},
        {'min': 100000, 'label': '>100000', 'color': '#DD0200'} # 有下限无上限
    ]
    # is_pieewise 是否自定义分段， 变为true 才能生效
    geo.set_global_opts(
        title_opts=opts.TitleOpts(title="人数分布"),
        visualmap_opts=opts.VisualMapOpts(is_pieewise=True, pieces=pieces)
    )
    geo.render("geo_chart_countries_js2.html")
```