影响纽约州各郡县居民健康水平的因素探究

4 组　刘子鸣　王逸轩　韩子钊　吴典

Abstract

本文中，我们着力于分析过去十年来影响纽约州各郡县公民健康水平的因素。首先通过熵权法整合社区健康的各个指标，确定健康指标，以刻画各郡县健康问题的严重程度。接着在餐饮卫生检查、人口统计与环境辐射的数据中，先验地提取出可能对健康指标有影响的因素，综合利用多层感知机、熵权法及 Spearman 值判别法，筛选并整合出对健康指标有显著影响的餐饮与人口指标。然后援引多元线性回归法，用这两者拟合健康指标的分布。最终得出结论，餐饮食品是影响纽约州各郡县公民健康水平的主要因素，且具有相当强的关联性。

1. 引言

近年来，在美国联邦政府的政策指导下，纽约州卫生署响应号召，制定了针对美国人健康状况的政策和目标，来解决健康问题、促进健康生活。我们针对数据集中提供的数据，进行定量分析，探究具体哪些方面对于纽约州各郡县公民的健康水平有着较为显著的影响。

2. 整体思路

对于我们提出的问题，需要通过以下几个步骤来给出答案。首先，如何定义和衡量公民的健康水平？我们选择餐饮卫生检查所覆盖的 57 个具有统计意义的郡县（余下的郡县的餐饮卫生检查数据太少，在之后的计算中没有显著的统计意义），分析它们的健康指标。由于健康指标由不同的特征反映，不同的特征又可分类为正面、负面及中性评价，我们利用熵权法确定系数，整合得到每个郡县健康指标的整体评价。

其次，如何筛选得到对于健康指标有显著影响的因素？我们针对余下 4 组关于热线电话、餐饮卫生、人口特征、环境辐射的数据，先验地提出一些可能会显著影响健康指标的因素，其中包括热线电话的投诉类型数量，每种投诉类型的解决速度；餐饮行业的每种类型餐馆分布数量，每种食品服务类型违规数量；人口方面的平均家庭收入、平均收入、平均社保、平均退休金、平均辅助社保、平均公众援助金、食品卷福利、家庭平均人数、私立医疗保险比例、公立医疗保险比例；环境辐射的不同类型同位素数量等等。

经过初步的数据处理，我们发现热线电话覆盖的郡县数太少，只覆盖了 57 个中的 7 个，因此我们舍弃了这些特征。另外环境辐射

的 Spearman 系数太低，与健康指标的关系不显著，故也舍去。

对于餐饮卫生来说，各个因素很可能会有关联比如食品服务类型违规的数量与餐馆类别。我们利用多层感知机处理餐饮卫生的多个特征，利用健康指标来进行有监督的机器学习，得到整体的餐饮指标。

对于人口特征来说，各个因素基本上互相无关，于是我们再次利用熵权法来确定各因素对于人口特征的权重，得到整体的人口指标。

最后，如何刻画不同因素对于健康的影响占比？我们利用多元线性回归，得到健康指标与餐饮指标以及人口指标的关系。

3. 主要模型和方法

3.1. 熵权法

在实际问题中，衡量一个郡县的健康指标以及人口生活指标（包括收入、保险、退休金等）时，我们需要联合考虑多个特征来进行评判。我们对于这种多个特征的情况，找到已经获得的多组数据，采用熵权法决定各个特征的权值，最后求和给出指标.

设郡县依次为 C_1，C_2，...，C_m. 设健康指标（或人口生活指标）由 k 个不同的特征而决定，特征的数据构成的集合为 X_1，X_2，...，X_k. 其中 X_i=$\{x_{i1}$，x_{i2}，…，$x_{im}\}$.

下面我们构造图表：

特征/郡县	C_1	C_2	...	C_m
X_1	x_{11}	x_{12}	...	x_{1m}
X_2	x_{21}	x_{22}	...	x_{2m}
...	...	...	...	...
X_k	x_{k1}	x_{k2}	...	X_{km}

接下来利用熵权法确定指标权重

$$W_1，W_2，...，W_k.$$

第一步 数据标准化:

设各特征标准化处理后取值构成集合为 Y_1，Y_2，...，Y_k，其中

$$y_{ij} = \frac{x_{ij} - \min_l(x_{il})}{\max_l(x_{il}) - \min_l(x_{il})}$$

第二步 求各个指标的信息熵:

根据信息熵的定义，Y_i 对应的信息熵

$$E_i = -\frac{1}{\ln m}\sum_{j=1}^{m} f_{ij}\ln f_{ij}$$

其中 $f_{ij} = \frac{y_{ij}}{\sum_{l=1}^{m} y_{il}}$.

特殊情况的声明：若 $\max_l(x_{il}) - \min_l(x_{il})$=0, 则定义 E_i=0；若 f_{ij}=0，则定义 $f_{ij}\ln f_{ij}$=0.

第三步 确定各指标权重:

通过信息熵 E_1，E_2，...，E_k 计算各指标权重 W_1，W_2，...，W_k，其中

$$W_i = 1 - E_i.$$

说明：上式中，$0 \le W_i \le 1$，这是由 $x\ln x$ 的凸性以及 Jensen 不等式所保证的。另外，我们这里与一般的熵权法稍有不同的是没有标准化权重 W_i，这是由于接下来我们给出的指标有正有负，只需要一个相对值.

3.2. 多层感知机 (MLP)

随着数据科学的发展，人们越来越认识到传统统计方法的局限性——比如模型的表示能力弱，需要比较强的先验假设等等。基

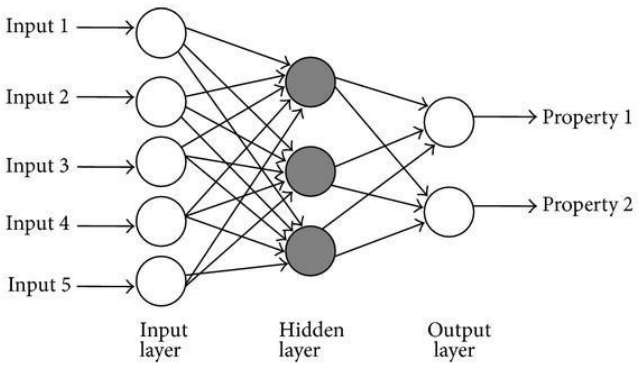

图 1. 多层感知机

于神经网络的机器学习技术，有利于补充统计学习的短板。本模型中用到的多层感知机 (MLP)，是一种表示能力极强的前馈神经网络。数学上已经可以证明，一个具有足够多神经元的单层感知机，可以刻画任何连续函数；而具有足够多神经元的双层感知机，足以刻画任何函数。多层感知机与传统统计的层次分析法很相近，但却不需要任何先验知识。事实上，利用反向传播算法，网络内的参数可以自动调节，从而摆脱对先验知识的依赖。多层感知机由多个单层感知机构成，单层感知机的输入 X_{in} 是 m 维实向量，输出 X_{out} 是 n 维实向量, 它进行的操作是:

$$X_{out} = \sigma(A_{m\times n}X_{in} + b_{n\times 1}) \tag{1}$$

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{2}$$

其中 $\sigma(x)$ 是 “激活” 函数 (Sigmoid 函数)，这是个非线性变换，可以增加模型的表示能力。将多个单层感知机串联起来，就组成了多层感知机。

在本模型中的餐饮指标部分，我们采用多层感知机来建模。我们统计出每种类型餐馆分布数量，每种食品服务类型违规数量，把这些数据输入多层感知机，输出就代表一个郡县公民整体的餐饮指标。

为了设计合理的网络结构，防止过拟合，我们借助贝叶斯信息准则 (BIC, Bayesian Information Criterion), 公式为 $BIC = ln(n)k - 2ln(\hat{L})$。其中 n 为样本数，k 为模型参数个数 (也就是网络中权重的个数)，$ln(\hat{L})$ 是将模型参数最优化后模型的拟合程度 (liklihood function)，即一种网络结构所能达到的最佳拟合效果，可用方均误差 (MSE) 度量。我们通过改变网络结构寻找最小的 BIC 值，来确定最为合理的网络结构。特别地，考虑贝叶斯信息准则应用在多层感知机时，该准则会强烈倾向结构更简单的结构。我们提出一种修正的 BIC，来减弱这种效应。改进的 BIC 为 $ADJBIC = \frac{k}{n^2}ln(n) - 2ln(\hat{L})$。我们后面将通过数值实验验证，通过这个准则挑选出的模型效果十分良好；以及通过增加随机无关数据来进行扰动，我们模型预测的结果几乎不变，验证了模型的稳定性。

3.3. Spearman 系数

Spearman 系数也称为等级相关系数，其特点是在计算每个样本在两个变量上的等级时，不仅要区别二者的高低差异，而且要计算二者差异的确切数值。Spearman 系数定义如下:

$$r_s = 1 - \frac{6}{n(n^2-1)}\sum_{i=1}^{n} D_i^2 \tag{3}$$

其中 n 是样本数, D_i 为同一个样本 i 在两个变量上的等级差。根据公式定义，r_s 的范围为 $[-1, 1]$。如果 r_s 的绝对值接近零，我们就可以以一定的置信程度说明两个变量之间没有相关性。对于连续型变量，排序后使

用 Spearman 系数，比直接使用 Pearson 系数，效果更佳稳定和鲁棒。

3.4. 多元线性回归

我们最终将利用对于健康指标有着较为显著影响的餐饮指标 F 与人口指标 D, 来对健康指标 H 进行多元线性回归，即

$$H = a_F F + a_D D + b \tag{4}$$

用最小二乘法确定比例系数 a_F, a_D 与残差 b. 通过比例系数的相对大小，我们可以判定 F, D 两个影响因素在 H 中的占比.

4. 具体分析

4.1. 健康指标的分析

为了探究哪些因素对于公民的健康有较为显著的影响，我们首先需要刻画每个郡县公民的整体健康指标。我们由人工将社区健康的 307 个因素分为正面、负面以及中性评价，分别见 health_indicator_names1.txt，health_indicator_names2.txt，health_indicator_names3.txt。通过熵权法的计算，我们得到这 307 个因素的权重，见 health_problem_weight.txt。
最后，我们根据权重和标准化的特征给出每个郡县的健康指标 $H = h_i$:

$$h_i = \sum_j \sigma_j w_j y_{ij} \tag{5}$$

其中特征 j 与健康程度负相关时，$\sigma_j = 1$；特征 j 与健康程度正相关时，$\sigma_j = -1$; 特征 j 与健康程度的关系呈中性时，$\sigma_j = 0$.
由于熵权法的特点，这里得到的健康指标 H 反映了某个郡县与其他郡县相比，健康程度的改进空间，其中 h_i 越高代表郡县 i 的健康问题越严重。

4.2. 餐饮类服务影响的分析

采用固定变量的思想，我们控制住其它变量，仅仅考虑餐饮对健康指标的影响。

对每一个郡县而言，我们发现食品服务类型有 4 类，问题类型有 89 类，如果将两者直积得到 $89 * 4 = 356$, 会得到一个过于稀疏的表格，这在统计上就失去了意义。基于这个考虑，我们将这两个不同类型的特征直接拼接起来成 $89 + 4 = 93$ 维向量。多层感知机具有强大能力，使得我们不必要对不同种类的数据进行进一步的归一化。因此，每个郡县对应多层感知机一个训练样本 (总共 $n = 57$ 个样本)，输入是一个 93 维的向量，输出是上小节计算出的健康指标，是一个标量。下面我们将通过尝试不同层数和结构的多层感知机，并结合 ADJBIC, 选择一个最优的结构。

我们总共训练了深度为 3、4、5、6 的四个多层感知机，学习率 $lr = 0.0002$, 训练轮数为 2000 轮。网络结构的设计遵循着相邻层数等比的原则 (除了最后一层 1 个神经元、和倒数第二层 2 个神经元)，这是希望每层网络不至于损失过多信息。计算各自的参数个数 k、方均误差 mse 和 $ADJBIC$，结果如表格所示。(注意有 $-ln(\hat{L}) = ln(MSE)$)

中间层宽度	MSE	k	ADJBIC
$46 \to 20 \to 10 \to 3 \to 2$	0.17	5499	2.37
$31 \to 10 \to 3 \to 2$	0.19	3278	0.2
$19 \to 4 \to 2$	0.1709	1879	-1.5
$15 \to 2$	0.3204	1506	-0.6

注意到 $ADJBIC$ 这个指标数值越小，效果

越好，因此 5 层的感知机具有最好的效果。最后我们利用这个感知机通过 93 个因素，来预测某个郡县公民的健康指数。

神经网络的一个主要问题是：训练结果可能高度敏感于数据。为了验证我们模型的稳定性，我们随机加入五个高斯噪声数据，得到 62 个郡县作为训练数据，训练一个新的模型。新模型与老模型在原本的 57 个郡县公民的健康指数的预测上，非常一致。这体现了我们模型的鲁棒性。图中显示了我们多层感知机对健康指数的预测能力。

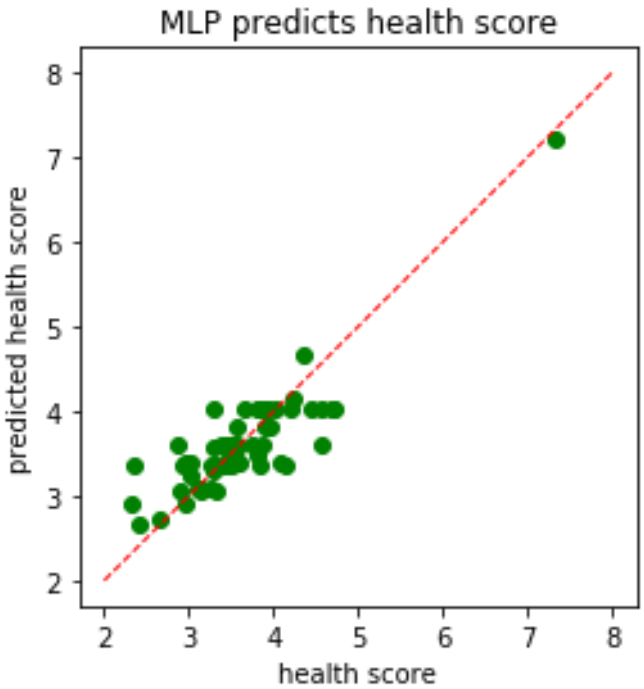

图 2. 多层感知机对健康指数的预测效果

4.3. 人口生活指标的分析

对于我们列出考虑的平均家庭收入、平均收入、平均社保、平均退休金、平均辅助社保、平均公众援助金、食品卷福利、家庭平均人数、私立医疗保险比例、公立医疗保险比例这十个可能的影响因素，我们援引分析健康指标时候的想法，利用熵权法确定各个指标的权重，刻画出人口生活指标 D.

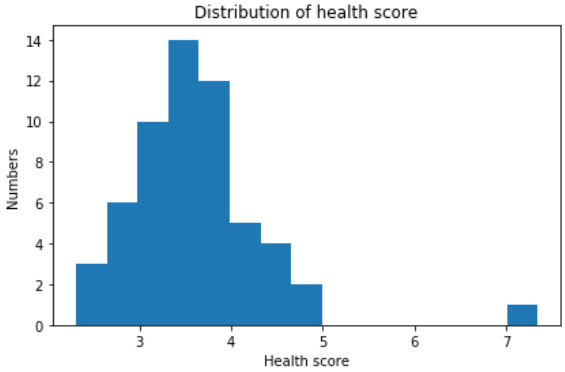

图 3. 不同郡县公民健康指数的分布

4.4. 环境辐射的分析

环境辐射是指某地区环境介质中（如土壤、水源）以及生物（如水源中的鱼类）体内的辐射和放射性物质的水平。对环境辐射的监测在放射性设施附近十分重要，例如电厂附近需要监测电磁辐射、核设施附近需要监测放射性物质的水平，过量的辐射会对附近人员的身体健康造成短期或慢性的危害。

为了分析纽约州环境辐射对公民健康的影响，我们将纽约州各个辐射监测点的监测数据按照郡县划分，并对一个郡县内各个监测点监测某种辐射的测量值进行平均，用来衡量一个郡县某种辐射的辐射水平。我们利用 Spearman 相关系数来检验各种辐射的辐射水平与郡县公民健康指数的相关性。具体地，对于每一种辐射，我们将涉及到的郡县分别按辐射水平和健康指数进行排序，对两个序列计算其 Spearman 相关系数。

经过对辐射数据的处理，我们选取了同时在 5 个以上郡县有非零测值的辐射类别，分析它们与郡县公民健康指数的相关性，结

果如下表所示：

同位素名称	郡县数	Spearman 值	临界值
GROSS BETA	13	0.368	0.484
POTASSIUM-40	8	0.000	0.643
RADIUM-224	7	-0.071	0.714
BERYLLIUM-7	6	-0.600	0.829

可以看出，这几种主要辐射的测量值与郡县公民健康指数的 Spearman 相关系数均未达到相关显著（置信概率 90%）临界值，因此可以得到结论，环境辐射与郡县公民健康没有显著的相关性。事实上，自然或常规生产中产生的辐射对环境和公民健康的影响是很小的，除非发生意外事故，公民们不必过于担心环境辐射的危害。

4.5. 311 热线的分析

311 热线贯穿了纽约市民们的生活，这可以从数据的大量性和抱怨类型的多样性看出。抱怨类型包括卫生、啮齿类动物、食物中毒、餐馆的手续合法性和就餐环境、室内空气、吸烟、石棉等等。这里列出的八类是主要的抱怨类型。57 个郡县和 8 种主要的抱怨类型，可以仿照餐饮类服务影响的分析，利用多层感知机来进行预测。但是表格数据，出现了一些问题。311 热线的电话信息中，有城市信息，和我们的郡县信息不太相符，最后只有 5 个郡县的数据行非零；后来我们抛弃了城市信息，利用 python 的地图包 geopy 通过经纬度去查询郡县信息，也只有 7 个郡县的数据行非零。这个特征说明热线电话的信息并没有收集完整，或者在地理位置的分划上出现了模糊的问题。我们可以更进一步地收集更完整的数据，但鉴于时间的限制，我们不得不放弃对这个因素的分析。

4.6. 最终影响占比的确定

最后我们使用餐饮指标 F 与人口指标 D, 来对健康指标 H 进行多元线性回归，即

$$H = a_F F + a_D D + b \tag{6}$$

由最小二乘法得

变量	a_F	a_D	b
预测值	1.05	-0.584	0.0744
p 值	1.6×10^{-16}	0.2	0.8

可以看出，a_D 与 b 的 p 值均大于阈值 0.05，可以从模型中排除它们的影响。接下来我们只用 F 对 H 进行线性回归，即

$$H = a_{F2} F \tag{7}$$

由最小二乘法得 $a_{F2} = 1.04$，p 值为 1.8×10^{-18}，相关系数 R 为 0.87。因此，餐饮指标 F 能够解释影响健康指标 H 的绝大部分因素。

5. 模型的稳定性分析

为了验证我们模型的稳定性，我们随机加入 5 个高斯噪声数据，得到 62 个郡县作为训练数据，重新训练模型，发现之前 57 个郡县的健康指标 H、餐饮指标 F 与人口指标 D 的变化不超过 5%。接着我们随机加入 5 个餐饮因素，得到 98 个餐饮因素作为训练数据，重新训练 MLP，发现餐饮指标 F 的变化仍然不超过 5%。这体现了我们的模型对于无关数据有很好的稳定性。

6. 模型评价

6.1. 模型的优点

1. 本模型整合了多种数据处理方法，从经典统计学中的熵权法、Spearman 系数判别法、多元线性回归到机器学习中的多层感知机，将不同的分析方法有机地结合起来，达到了更好的效果。
2. 同时，本模型避免了粒子群算法、深度学习等“黑箱”算法，具有良好的可解释性。
3. 模型着眼于全局，立足于细处，多角度、全方面对健康水平因素进行探究。

6.2. 模型的缺点

1. 模型里做出了许多简化的假设，无法精确地考察到每个因素的影响。比如在先验地提出潜在影响因素时，不可避免。
2. 模型为了尽可能考虑更多的对健康指标的影响因素，舍弃了探究各因素随着时间的变化情况，即我们只考虑了过去十年的整体情况，没有考察具体的变化情况。

7. 总结与展望

从我们的模型得到的数据试验结果中，我们可以推断出，食品餐饮行业是过去近十年中影响纽约各郡县公民的健康指标的主要显著因素，其中该影响是每种类型餐馆的分布以及食品服务类型违规数量的分布的综合作用。因此我们应该建议州政府倡导各郡县大力推广餐饮行业的多样性，并对食品安全质量进行严格的监管，只有公民的食品得到了保障，各个郡县的总体健康水平才能有所提升。另外，特别针对那些相对来说食品卫生与服务质量较差的郡县，政府应该加大力度保障公民的食品安全与质量。