广州市疾控中心空气质量健康风险指数的构建建模方案

1 建模目标

针对广州疾控中心“研究大气污染对人群健康的影响，建立时空动态评估模型”的需求，进行“空气质量健康风险指数”的构建。

建模流程示意图如下：

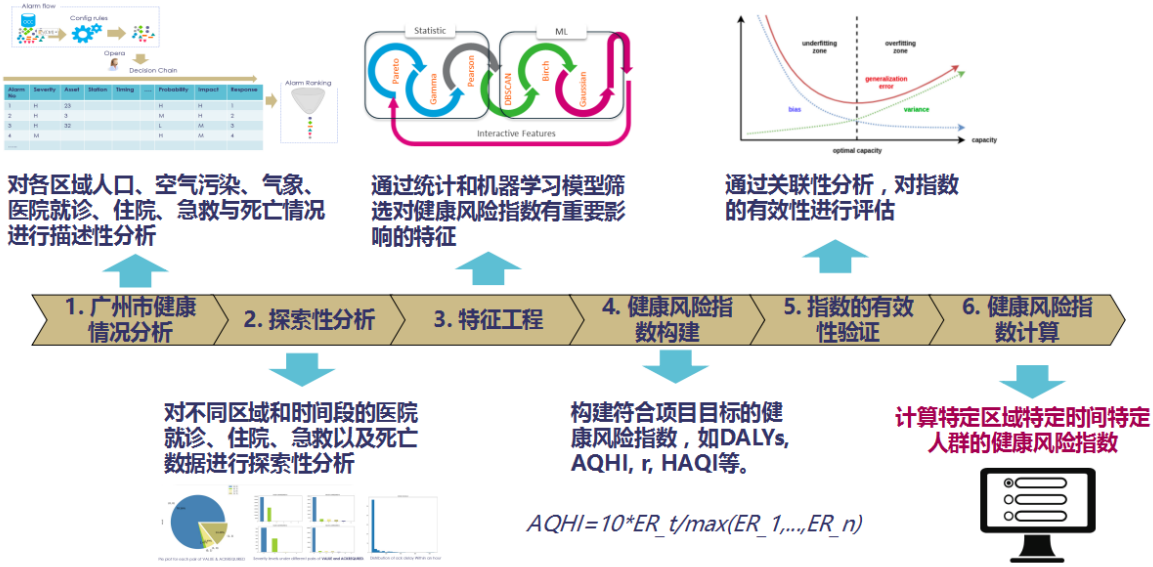


图1-建模流程示意图

2 指数应用

- 疾控中心的学术目标：探索环境因素与人体健康之间的关系，研究如何构建可反映这种关系的指数体系。
- 疾控中心内部工作指导：通过构建环境健康风险指数，为指导疾控中心体系内部的环境监控和公共卫生的疾病预防提供依据。
- 为政府部门提供评估环境健康风险评估标准；向公众发布健康风险指数，使他们能够采取适当的行动，以降低环境改变带来的健康风险。

3 背景介绍

环境空气污染是全球公共卫生面临的最大环境风险。据2017年对全球195个国家或领地疾病负担研究的系统分析，环境空气污染每年导致410万人过早死亡。据世界卫生组织2016年的报告，世界上91%的人口都居住在超过世界卫生组织建议的空气质量环境下，这种情况在中低经济收入国家更为严重。

同时，环境空气污染也会带来极大的经济损失。根据世界银行2016年的报告，室外空气污染给中国带来了10%的国民生产总值(GDP)的损失。政府部门及时地公布可反映环境变化带来的风险情况，可以指导人们进行适当的行为反应，以减轻户外空气污染对人体带来的短期健康风险。

4 数值公式的发展

总体架构

我们将用浓度反应函数来构建空气质量健康指数。每日死亡率、每日急救率、每日门诊率及每日住院率时间序列研究均可被用作系数的来源，以表示环境污染物混合物对增加健康风险的来源。

数据来源

环保监测点数据、气象数据、人口数据、急救数据、死因数据、医院门诊数据和医院住院数据均来自于广州市疾病预防控制中心。关于环保监测点数据，我们获得了广州市51个站点的一氧化碳(CO)，二氧化氮(NO_2)，臭氧(O_3)，二氧化硫(SO_2),和空气动力学直径小于10微米和2.5微米的微粒物质(PM,分别为PM_{10}和$PM_{2.5}$)浓度数据。一氧化碳(CO)，二氧化氮(NO_2)，二氧化硫(SO_2),$PM_{2.5}$和PM_{10}均每天测量一次（用什么方法？是否是每天测量一次？)，其中O_3可提供日最大8小时均值和O_3日最大1小时均值（如何计算的？如何选择？）。关于气象数据，我们获得了广州市的平均气压、最高气压、最低气压、平均温度、最高温度、最低温度、平均相对湿度、最大湿度、最小湿度、降水量、日平均风速和日照小时数。这些值均为每天检测一次。关于人口数据，我们有广州市11个区170个街道或镇的辖区内常住居民数（什么年份）。关于死因数据，我们有每日死者的性别、年龄、生前常住地址、根本死因ICD编码。可以考虑将死因数据的分析仅限于内因死亡（国际疾病分类[ICD]-9代码小于800和ICD-10代码A00-R00）。关于急救数据，我们有每日被急救者的性别、年龄、初步诊断和呼救地址。我们可以将死亡人数及急救人数按年龄段分成幼年、青壮年和老年三类进行考虑；也可以将呼吸道疾病类的死者和被急救者以及循环系统类疾病的死者和被急救者进行分类考虑；在数据支持的情况下，住院和就诊数据也可以进行类似分析。

除了环保监测点数据以外，我们也从广州市疾病预防控制中心获得了$PM_{2.5}$成分监测数据，包括硫酸盐(SO42-)、硝酸盐(NO3-)、氯离子(Cl-)、氟离子(F-)、铵盐(NH4+)、锑(Sb)、铝(Al)、砷(As)、铍(Be)、镉(Cd)、铬(Cr)、汞(Hg)、铅(Pb)、锰(Mn)、镍(Ni)、硒(Se)、铊(Tl)、钡(Ba)、钴(Co)、铜(Cu)、铁(Fe)、钼(Mo)、银(Ag)、钍(Th)、铀(U)、钒(V)、锌(Zn)、铋(Bi)、锶(Sr)、锡(Sn)、锂(Li)、萘、苊烯、芴、苊、菲、蒽、荧蒽、芘、屈、苯并[a]蒽、苯并[b]荧蒽、苯并[k]荧蒽、苯并[a]芘、二苯并[a,h]蒽、苯并[g,h,i]苝和茚并[1,2,3-cd]芘。每月选择连续七天从越秀区、番禺区和从化区进行采样，将$PM_{2.5}$的成分分离出来并进行记录。

数据治理

目前针对我们已有的数据，我们可以将常住地址找到死因数据和急救数据的死者和被急救者在地图上的经纬度。通过克里金插值法，我们可以将51个站点各种污染物的浓度扩展到每个街道。通过经纬度，我们将某人与他/她所在街道的各污染物浓度匹配起来。对于死因数据，根据ICD-10分为22类，以便于筛选希望进一步研究的疾病种类。对于年龄组，我们可以按照不同的年龄段分类进行探索，目前是每五岁分一类。在论文The construction and validity analysis of AQHI based on mortality risk: A case study in Guangzhou, China里人群年龄被划分为<5岁，5-65岁和大于等于65岁三组，我们接下来可以参考。

我们将每个区域的所有站点的污染物按天取平均，然后通过日期将分年龄组、分性别、分疾病类型的死因数据、急救数据医院门诊数据、医院住院数据与环保监测点数据、气象数据、人口数据进行合并。

可视化分析

- 空间维度

1. 使用克里金插值法（Kriging），以51个站点数据为基础，绘制每天/每种污染物在整个广州市的污染物热力分布图:

$$\hat{z_0} = \sum_{i=0}^{N} \lambda_i z_i$$

其中，$\hat{z_0}$是点(x_0, y_0)处的污染物估计值，即$\hat{z_0} = z(x_0, y_0)$，λ_i是污染物权重系数。

克里金插值的假设条件为，空间属性z是均一的。对于空间任意一点$(x,\ y)$，都有同样的期望E与方差σ^2：

$$E(z(x,y)) = E(z) = c, Var(z(x,y)) = \sigma^2 \tag{2}$$

换一种说法：任意一点处的污染物数值$z(x,\ y)$，都由区域平均值c和该点的随机偏差$R(x,\ y)$组成，即:

$$z(x,y) = E(z(x,y)) + R(x,\ y) = c + R(x,\ y) \tag{3}$$

$$Var(\sum_{i=0}^{N}\lambda_i z_i - z_0) = Var(\sum_{i=1}^{N}\lambda_i z_i) - 2Cov(\sum_{i=1}^{N}\lambda_i z_i, z_0) + Cov(z_0, z_0) \tag{4}$$

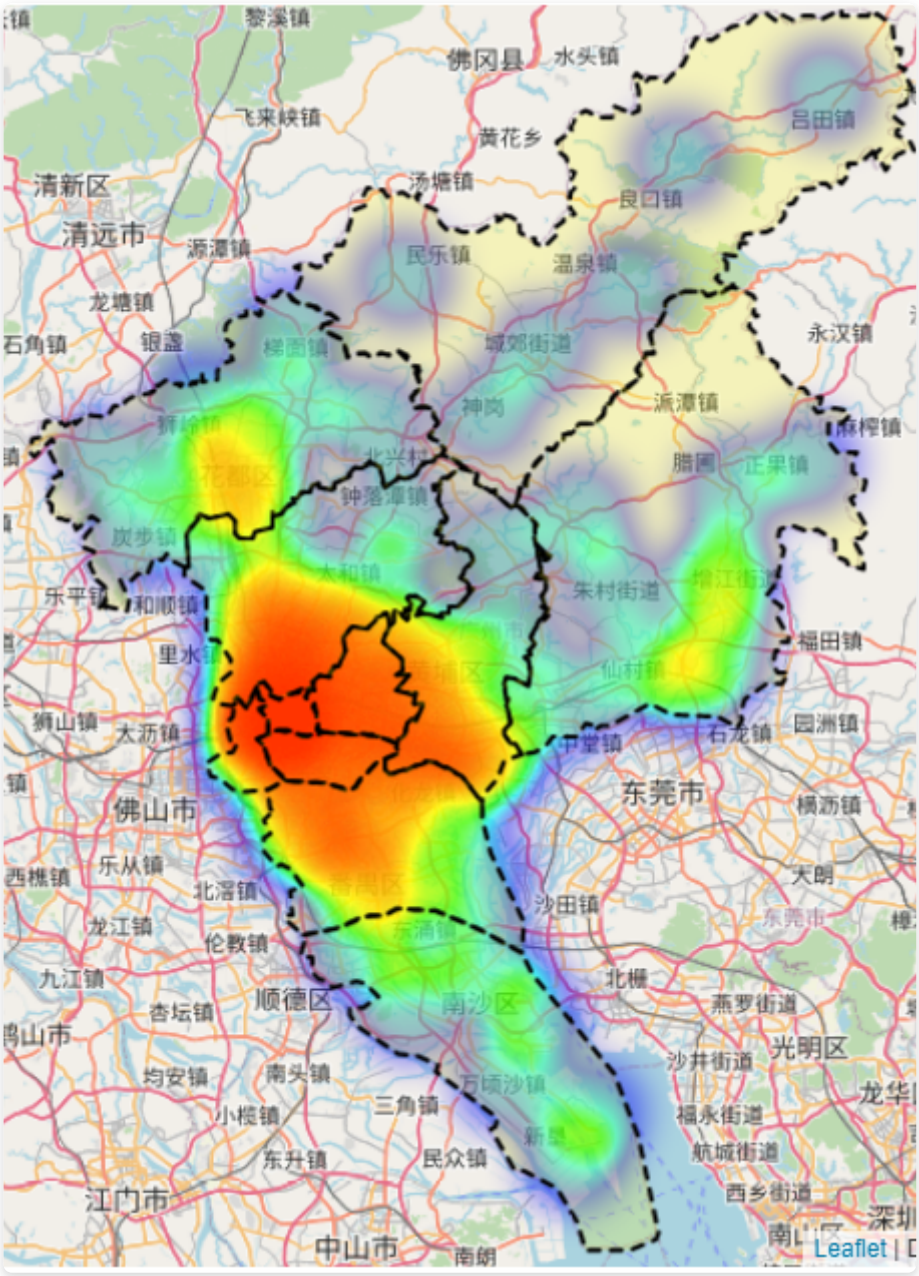
2. 根据(1)所得的全市污染物热力图，通过各街道经纬度坐标，得到每个街道中每种污染物的估测数据，并绘制全市污染物热力分布图。

   目前现有的数据，无论是医疗数据（急诊/住院/死亡），还是人口数据，最小颗粒度都是以街道为单位；因此，我们需要将空气污染数据也处理到以街道为单位，这样几种数据的最小颗粒度就可以保持一致，便于之后的建模，将各个街道的数据单独计算，既提高了数据的准确性，也便于最后高风险区域的获得，更利于模型的训练和优化。

	经度	纬度	街道名	PM2.5
0	113.273899	23.132627	广州市越秀区北京街道	37.828592
1	113.274916	23.143590	广州市越秀区洪桥街道	41.182862
2	113.261748	23.139342	广州市越秀区六榕街道	41.246485
3	113.266805	23.151288	广州市越秀区流花街道	42.210719
4	113.265087	23.127736	广州市越秀区光塔街道	38.278554
...	...	...	...	...
165	113.526936	23.473641	广州市从化区太平镇	34.194059
166	113.706103	23.618497	广州市从化区温泉镇	34.400906
167	113.767268	23.765372	广州市从化区良口镇	30.578995
168	113.951436	23.842807	广州市从化区吕田镇	34.427051
169	113.422839	23.629415	广州市从化区鳌头镇	34.418614

上图为2018-2-4各街道PM2.5污染值列表

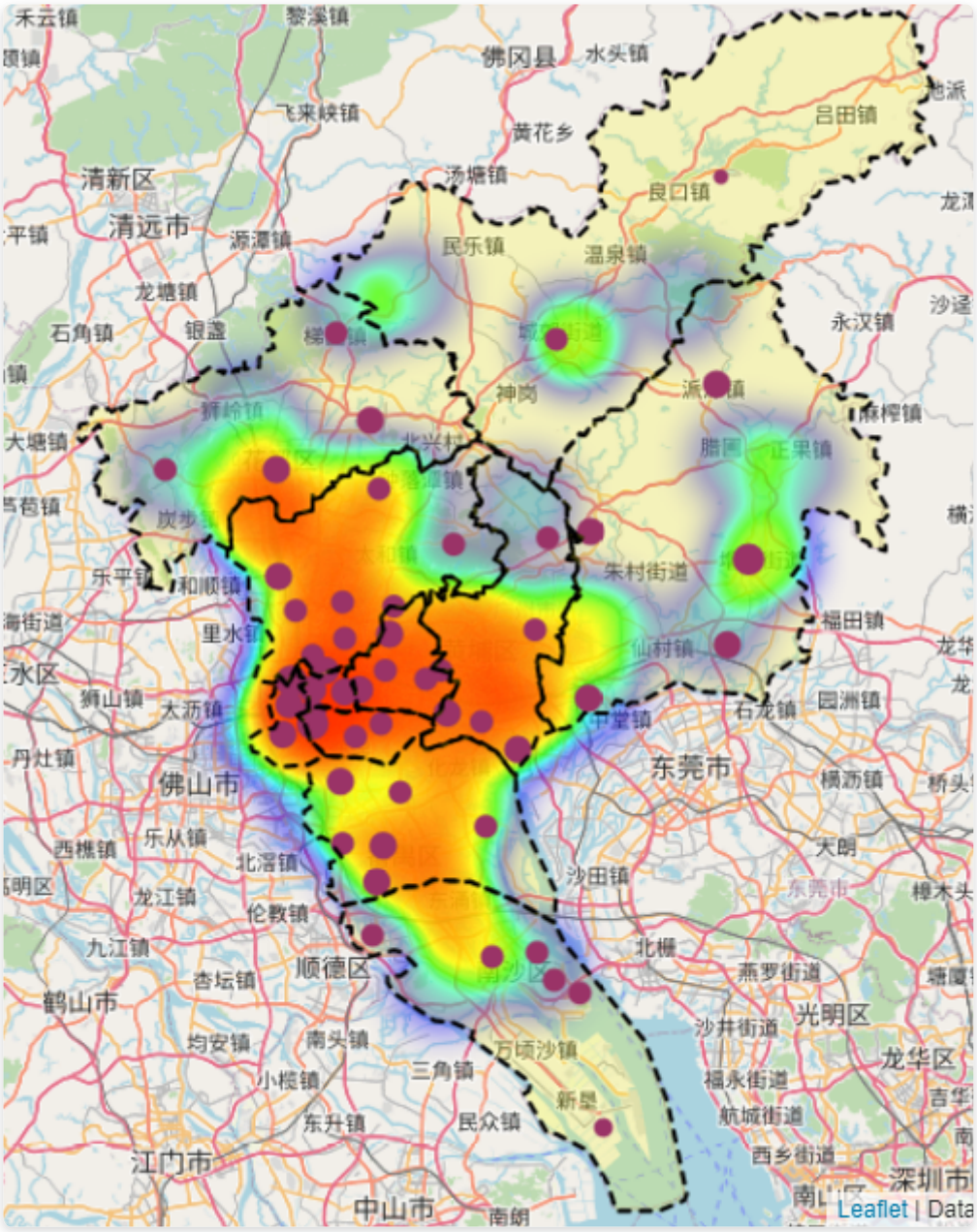
全市的某种污染物的热力图绘制如下：



PM2.5热力图（2018-02-04）

上图为以街道基本单元绘制的$PM_{2.5}$污染物热力图，红色区域为空气污染较严重的区域，绿色区域为空气质量比较好的区域。

3. 将三个月的急救死亡数据（包含总人数与按疾病种类/年龄分类的人数）分别绘制全市分布情况热力图，并增加各站点污染物的中位数/峰值一并制图，直观反应两者的关联性。



呼吸系统急诊患者分布

上图为三个月呼吸系统急诊患者分布图，红色区域表示急诊人数密集的区域；紫色圆点代表各空气污染物监测站点，半径长度反映了$PM_{2.5}$检测值的浓度，站点的圆圈越大表示此站点的$PM_{2.5}$浓度值越大。

- 时间维度

  空气污染与健康风险的相关关系不仅在空间上有所体现，也往往体现于两者的时间序列存在一定的相依性。对此，可以使用格兰杰因果检验对空气污染物浓度和急救/死亡人数的时间序列在时间相依性上进行大致了解。格兰杰因果检验可用于检验一组时间序列x是否为另一组时间序列y的原因，但是需要注意的是，这里的因果并非我们通常理解的因与果的关系，而是说x的前期变化能有效地解释y的变化，所以称其为“格兰杰原因”。而进行格兰杰因果检验的前提是两个时间序列需要满足平稳性或存在协整关系，接下来具体介绍。

  1. 平稳性检验

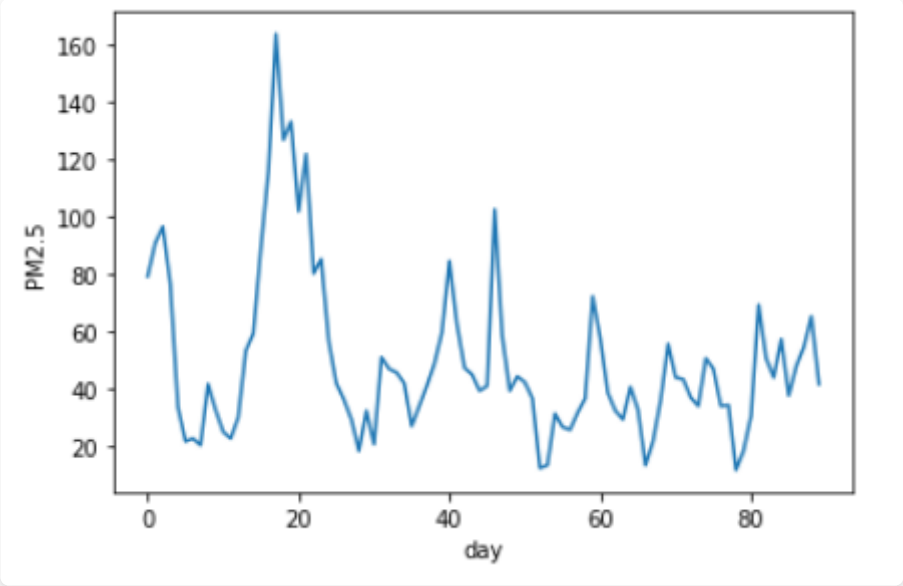
在数学中，平稳随机过程（Stationary random process）是在固定时间和位置的概率分布与所有时间和位置的概率分布相同的随机过程，即随机过程的统计特性不随时间的推移而变化。这样，数学期望，方差等参数也不随时间和位置变化。即

$$E(y_t) = E(y_{t+m}) \tag{5}$$
$$cov(y_t, y_{t+k}) = cov(y_{t+k}, y_{t+k+m})$$

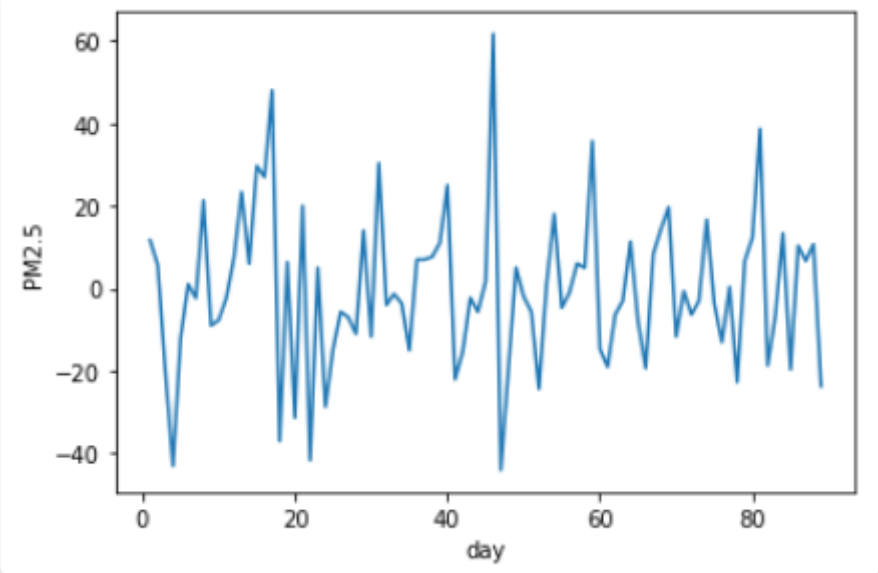
简单来说，就是时间序列数据的基本特性能在包括未来阶段的一个长时期里维持不变。

对于时间序列数据进行建模分析往往要求序列是平稳的，如果数据无法达到平稳性要求，可以通过一阶差分和滤波器过滤方法处理数据，使其平稳之后，再进行建模分析。而对于格兰杰因果检验，如果两个不平稳的序列存在协整关系，也可以不进行平稳化而直接检验。

在此，我们使用ADF检验（Augmented Dickey-Fuller Test）和PP检验（Phillips and Perron Test），验证污染物和医疗时序数据的平稳性。验证后发现，污染物和医疗数据的平稳性不符合置信要求，因此通过一阶差分（计算两天之间的污染物变化量）来进行数据处理，得到的结果符合要求，并将结果制图。



三个月PM2.5变化图



三个月PM2.5每日变化量（第二天与第一天差值）的变化图

分析上面两张图，也可以直观看出，每日$PM_{2.5}$的变化量在0附近震荡，从长期看符合数据平稳性的要求。

2. 协整检验

非平稳序列很可能出现伪回归，此时格兰杰因果检验的结果就不再具有借鉴性。协整检验的意义就是检验两个序列的回归方程所描述的因果关系是否是伪回归，即检验两个序列之间是否存在稳定的关系。所以，非平稳序列的因果关系检验就是协整检验。由平稳性检验我们可以看到，污染物和死因数据的时序原始数据都不是平稳的，为了避免差分可能带来的信息损失，如果两个序列存在协整关系，我们依然可以使用原始数据进行分析。

以呼吸系统疾病为例，在检验中，我们发现$PM_{2.5}$浓度与呼吸道死亡人数之间存在协整关系，这表明$PM_{2.5}$浓度和呼吸系统死亡人数之间会存在一个长期稳定的关系。

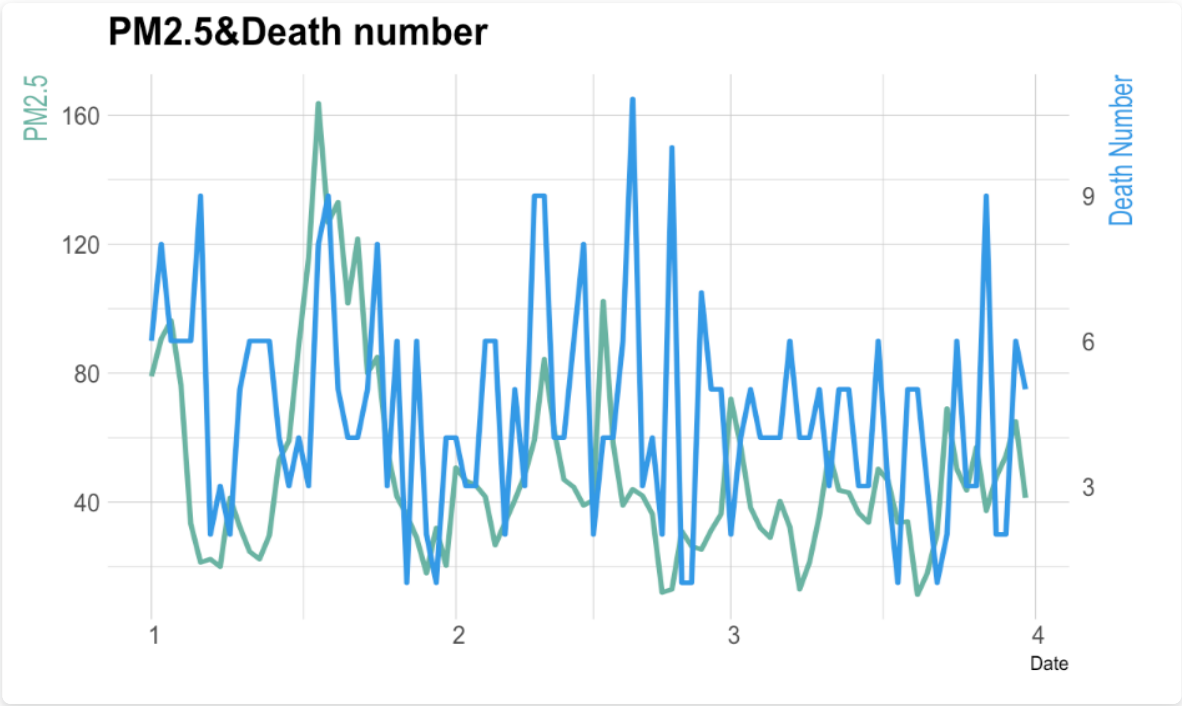
3. 格兰杰因果检验

接下来，对呼吸系统死亡人数时间序列和$PM_{2.5}$时间序列进行格兰杰因果检验，判断两者是否存在时间上前后的因果关系。分别以呼吸系统死亡人数或$PM_{2.5}$为响应变量进行格兰杰因果检验。

检验结果表示，当原假设为“呼吸系统死亡人数不是引起$PM_{2.5}$变化的格兰杰原因”,P值为0.7809,接受原假设,认为呼吸系统死亡人数不是引起$PM_{2.5}$变化的格兰杰原因; 而当 原假设为”$PM_{2.5}$不是引起别呼吸系统死亡人数变化的格兰杰原因”时,P值为0.00242<0.05，我们拒绝原假设。

因此，可以认为: $PM_{2.5}$是引起呼吸系统死亡人数变化的格兰杰原因。 可以理解为,

- $PM_{2.5}$有助于预测呼吸系统死亡人数
- 呼吸系统死亡人数不应当有助于预测$PM_{2.5}$。



PM2.5浓度与呼吸系统疾病死亡人数变化图

4. 伪回归检验

通过Durbin-Wartson方法，检验上述回归模型是否是伪回归，伪回归指两个没有因果关系的时间序列之间，基于一些其他的外在因素，推断出因果关系。检验统计量DW = 2.1045，查验临界值DW分布表后发现，DW统计量在接收域内，不存在明显的自相关，不认为这里是伪回归。由此，可以验证$PM_{2.5}$和呼吸系统死亡数之间存在协整关系，他们之间是长期稳定的。

长期均衡式为:

呼吸道系统死亡人数= 3.801943 + 0.015526pm2.5 + ϵ_t

AQHI的构建

下面用死亡人数作为例子来构建AQHI，同样可以考虑急救人数、就诊人数或者入院人数来构建AQHI。或者可以从疾病的不同阶段或严重程度采用加权的方法得出最终的AQHI。参照论文[^4]和论文[^6]建立AQHI的过程如下。

- 时间序列分析

以死亡人数为例，用广义线性模型建立空气污染物与死亡人数的相关性。我们选择用$PM_{2.5}$[^4]有三个原因：1. $PM_{2.5}$可以反应广州市68.84%的PM_{10}情况；2. $PM_{2.5}$和PM_{10}具有高相关性；3. 研究表明小颗粒物对人体的伤害更大。因此这篇论文在构建AQHI的时候包含了$PM_{2.5}$而未包含PM_{10}。

假设每日死亡人数服从泊松分布， 我们用泊松链接的广义加性模型来检验空气污染物与死亡人数之间的关系。论文The construction and validity analysis of AQHI based on mortality risk: A case study in Guangzhou, China里使用每年7个自由度的平滑函数的日历时间来控制死亡率长期趋势和季节性的波动，对于当日温度使用6个自由度，相对湿度使用3个自由度。每一周的某一天也作为虚拟变量被包含在模型中。

单污染物模型：

$$log[E(Y_t)] = \alpha + \beta_i X_i + s(t, df = 7/year) + s(temp, df = 6) + s(humidity, df = 3) + DOW$$

多污染物模型：

$$log[E(Y_t)] = \alpha + \sum \beta_i X_i + s(t, df = 7/year) + s(temp, df = 6) + s(humidity, df = 3) + DOW$$

这里$E(Y_t)$是第t天预期死亡人数，α为模型截距，X_i为污染物i的浓度，β_i为X_i的回归系数。t可以调整长期趋势和季节性的时间。$s()$表示一个基于惩罚平滑样条的平滑，它捕捉了时间趋势协变量和天气参数与每日死亡率的非线性关系。df是自由度。每日平均气温和相对湿度在所有模型中被用来控制混淆。

由于这些变量存在高度相关性，我们每次就建模单个因素与日死亡率的相关性来评估每种污染物造成的死亡风险。0-3日的单日延迟被用来决定那一天的暴露与死亡率有最大的相关性。关联性最强的延迟时间被用来做后续的分析。为了保证污染物与死亡率的相关性是稳健的，当某种污染物被引入多污染物模型不失去它的统计显著性时，它就被包含在AQHI的构建中。在构建AQHI的过程中，使用单污染物模型的参数来构建。

- 过量风险

  过量风险是指每一种空气污染物每增加$10\mu g/m^3$时日死亡率的增长率及其95%的置信区间。过量风险的计算公式如下：

  $$ER_{it} = 100 \times [\exp(\beta X_{it} - 1)]$$

  这里ER_{it}代表与第t天第i种污染物相关联的死亡率变化，β_i代表单污染物时序模型里污染物i的回归参数，X_{it}是污染物i第t天第i种污染物的浓度。

- AQHI的计算

  根据每个区域选择每日死亡率百分比。然后对所有可用区域每天的这一百分比进行平均

  $$Mortality\ weighted\ excess\ deaths(\%) = \sum_{j=1,\ldots,n} [(m_j / \sum_{j=1,\ldots,n} m_j) \sum_{i=1,\ldots,p} 100(e^{\beta_i x_{ijt}} - 1)]$$

  这里 β_i 是一个泊松模型的回归参数将第i个空气污染物变量与死亡率联系起来， x_{ijk} 是第i个污染物在第j个城市第t小时日过量死亡率对应的浓度值（区域为1至n，污染物为1至p），m_j为第j个区域的日平均死亡人数。

得到第1至q天内的最大值

$$c = \max_{t=1,\ldots,q} \{ \sum_{j=1,\ldots,n} [(m_j / \sum_{j=1,\ldots,n} m_j) \sum_{i=1,\ldots,p} 100(e^{\beta_i x_{ijt}-1})]\}$$

要建立一个从0至10的简单数值刻度，我们可以用最大值来进行比例缩放。将所有的过量风险加和乘以10再除以c。

$$AQHI = (10/c) \sum_{i=1,\ldots,p} 100(e^{\beta_i x_i} - 1)$$

这里β_i为同一个泊松模型将第i个空气污染物与死亡率连接的回归系数，x_i是第i种污染物的浓度，c是比例因子。

AQHI的计算过程如下：

$$AQHI_t = 10 \times daily\ total\ ER_t / \max(daily\ total\ ER_1, daily\ total\ ER_2, \ldots, daily\ total\ ER_n)$$

这里每日总过量风险ER_t代表所有污染物在第t天的过量风险之和。

- 将AQHI分为5个组，包括"低健康风险"、“中健康风险”、“对脆弱人群不健康”、“高风险”和“严重高风险”。我们可以根据WHO-空气质量标准(AQG)设定高风险截断值，并据此将AQHI划分为好几个同样的长度段。
    1. WHO-AQG规定SO_2的24小时均值截断值为$20\mu g/m^3$，$PM_{2.5}$的24小时均值截断值为$25\mu g/m^3$，O_3的8小时均值截断值为$100\mu g/m^3$。WHO-空气质量标准无NO_2的标准，可以借鉴香港的指标$129.8\mu g/m^3$。我们可以根据这个方法找到其他的污染物指标标准。我们可以用这些值来计算加和的ER并进一步用它来计算相关的AQHI，代表“高风险”的下限作为高风险与中风险的截断值，记录为$AQHI_{high}$。
    2. 然后我们将$AQHI_{high}$的50%作为中风险与低风险的截断值$AQHI_{moderate}$。最后将$AQHI_{high}$的1.5倍作为高风险与严重风险的截断值$AQHI_{serious}$。
    3. 脆弱人群（<5岁或者>65岁的人群）更容易受空气污染对死亡率的影响。我们可以计算value1=幼年的过量风险中位数（<5岁）/总人群的过量风险中位数，value2=老人的过量风险中位数（65岁以上）/总人群的过量风险中位数，我们计算得到regulate factor=max(value1, value2)。中度健康风险与脆弱人群健康风险的截断值$AQHI_{vulnerable} = \frac{AQHI_{high}}{regulate\ factor}$。
- AQHI有效性的验证

    用分裂样本方法来评估AQHI作为风险预测的有效性。先用奇数年来计算回归系数建立AQHI，然后再将AQHI作为自变量在偶数年上与死亡率做回归。

    比较AQHI和API[^3]以及AQI[^4]与健康事件的相关性，结果证明AQHI与健康事件的相关性比较高。

可改进的地方

在论文[^5]中，对中国的272个城市的权重平均计算每日健康风险。我们可以依照这个方法对广州市的11个区的权重计算每日健康风险。

参照论文[^5]的格式，我们也可以根据性别、特定的疾病和年龄组来计算AQHI中各污染物的系数关系。论文[^5]单独计算了由冠心病（CHD）、中风（stroke）和慢性阻塞性肺病(COPD)引起的死亡人数构建的AQHI。

5. 结果

1. 计算分析污染物之间的相关性
2. 根据单一污染物模型的回归系数，估计在每种污染物的平均浓度下死亡率增加的百分。可以尝试不同的时滞性下不同的自由度的不同时间的自然样条函数，看看是否有差别。

3. 对比温暖的季节和寒冷的季节的污染物浓度。
4. 在这个时滞性和平滑度下，计算死亡率的增长与每种污染物浓度的浓度值。
5. 检验每个指标与每个区域的污染物浓度来污染复合物。
6. AQHI的验证，看看AQHI在偶数年份与死亡率的相关性。计算包含了$PM_{2.5}$或者PM_{10}的AQHI与污染物的相关性。或者也可以将寒冷季节或者温暖季节分开计算AQHI。

6. 表现形式

主要观众是对空气污染有很大风险的人们，特别是那些和健康状态（已有呼吸道疾病或者心肺疾病的人），年龄（老年人和小孩），以及暴露率较高的人（室外工作者和那些参加室外运动的人）。

表达基本组成部分：

1. 数值测量（比如0-10+）
2. 颜色测量（可以利用颜色的深浅来反映风险的高低，比如颜色浅的代表AQHI低并且健康风险低，颜色深的代表AQHI高并且健康风险高）
3. 分类标签（低风险，中风险，高风险，非常高风险）
4. 健康信息（比如说对不同的人群提供不同的外出活动参考建议）
5. “你知道吗”部分（给大家普及一些相关的空气污染或者室外活动的常识）

7. 已发表论文的局限性

1. AQHI在郊区的测量和对人们的活动指导不一定是准确的，因为大部分数据是从城市中心区域搜集的。
2. 只给人们提供信息不一定能指导人们改变行为。但提供更加准确的信息是帮助做出有效决定的先决条件。
3. 根据AQHI提供的建议比较少。
4. 由于数据的限制，AQHI的建立都是基于死因数据的，没有基于急救数据、就诊数据和住院数据的。由于短时的空气污染致死率应该较低，加入急救数据、就诊数据和住院数据的数据应该对AQHI的建立有所帮助。
5. 在对区域污染物浓度进行研究时，很多论文选用站点平均数，这个不一定是最好的选择。可以尝试使用第75个百分位或者最大值来建立模型，看看效果是否更好。

参考文献

1. World Health Organization. 2009. Global health risks: mortality and burden of disease attributable to selected major risks.
2. Xiao Lin,Yu Liao,Yuantao Hao. The burden associated with ambient PM 2.5 and meteorological factors in Guangzhou, China, 2012–2016: A generalized additive modeling of temporal years of life lost[J]. Chemosphere,2018. 1-12.
3. Chen R, Wang X, Meng X, et al., 2013. Communicating air pollution-related health risks to the public: An application of the air quality health index in shanghai, China. Environment international, 51: 168–173.
4. Li X, Xiao J, Lin H, et al., 2017. The construction and validity analysis of aqhi based on mortality risk: A case study in guangzhou, China. Environmental Pollution, 220: 487–494.
5. Dux, Chen R,Mengx, et al., 2020. The establishment of national air quality health index in China. Environment International, 138: 105594.
6. Stieb D M, Burnett R T, Smith-Doiron M, et al., 2008. A new multipollutant, no-threshold air quality health index based on short-term associations observed in daily time-series analyses. Journal of the Air & Waste Management Association, 58(3): 435–450.