

一种基于大数据的天气现象相关性分析方法

赵奕清

(北京外国语大学附属外国语学校, 北京 100089)

【摘要】本文使用了大数据、统计分析、机器学习等方法,研究雷电与降雨之间的相关性,论证了“雷声大雨点小”的正确性以及适用条件。

【关键词】气象; 大数据; 统计分析

本文的结构如下:

1. 采用了大数据方法。采用了全球 1.7 亿条美国国家环境信息中心气象数据,覆盖全球 26,584 个站点,时间跨度从 1929 至 2023 共 95 年;采用了中国科学院子午工程 961 万条武汉站高精度雷电数据,时间跨度从 2013 至 2023 共 11 年。

2. 采用了统计学方法,论证了“雷电-降水”相关性。从历史气象数据中挖掘隐藏的相关性规律,采用图形分析与皮尔逊相关系数,证明了雷电与降水之间存在弱线性相关性。

3. 采用了机器学习方法,构建了“雷电强度-降水量”线性回归模型,证明了:当雷声来源于云体间电场时,“雷声大雨点小”正确;当雷声来源于云体与地面间电场时错误,应为“雷声大雨点大”。

一、自然原理探究

(一) 打雷与闪电

雷电是夏天常见的一种大气现象。空气极不稳定时,容易发生强烈对流,从而形成积雨云^[5]。

积雨云中充满上下奔窜的水汽,就会产生静电:云的上端产生正电荷,云的下端产生负电荷,地面是正电荷。

若正负电荷间的电压差,大到可以冲破绝缘体空气,就会导致空气瞬间膨胀爆炸、发热发光。发光就是闪电,膨胀爆炸发出巨大声响就是打雷。

(二) 雷声大小与电压差

雷声与闪电的根源都是:“云体-云体”间或“云体-地面”间的电压差。这个电压差越大,绝缘击穿时爆发的雷声就越大。

相应的,电压差也可以分为两类:“云体-云体”电压差、“云体-地面”电压差。分别对应俗称的“云

闪”和“地闪”。

(三) 雷声大小与降水量

降水是由于空气降温促使空气中的水气凝结成云致雨所致。积云密布的时候有非常多的小云滴存在,它们比较小而悬于空中。雷声是一种外在因素,它产生声波震动,可能促使小云滴聚合为大雨滴。但降水量最终取决于单位质量空气降温的幅度与空气降温前含有的水气量。

二、“雷电-降水”相关性分析

(一) 气象数据处理

1. 气象数据来源。在“雷电-降水”相关性分析中,选用美国国家环境信息中心(National Centers for Environmental Information, NCEI)数据^[6]。美国国家环境信息中心的中国气象数据,是根据中美合作项目由中国气象局提供的;除了中国的气象数据,也包括世界上其他主要国家的气象数据。

2. 气象数据分析。气象大数据按照站点组织,每一个站点一个文件;按照年份组织文件夹,同一年份的气象数据文件放在同一个文件夹中。字段包括:站点编号、日期、纬度、经度、站点名称、降雨量、观测次数、气象事件编码。其中,气象事件编码为 6 位数字,每位取值 0 或 1,分别表示:是否有雾/下雨/下雪/冰雹/打雷/龙卷风。

3. 气象数据处理。气象数据读取包括了 5 个步骤:(1) 设置文件路径;(2) 遍历文件夹,循环读取文件路径;(3) 将文件路径组织成 DataFrame 格式,方便后续操作;(4) 分组读取文件。因为文件太多,共 570,497 个,一次性读取的程序运行时间很长。因此,采用分组读取的方式:每组读取

5 万个文件，分 12 组批次读取；（5）输出数据。在完成数据读取之后，进行气象信息的提取。由于数据文件共 57 万多个，数量比较大，一次性全部提取的程序运行时间太长。所以，将气象数据文件分 12 组、每组 5 万个分别提取信息。

基于 Knime 数据科学平台开发的气象信息提取程序包括 3 个步骤：

第 1 步：保留 FRSHTT 字段。原始文件中包括 28 个字段，本文仅使用 FRSHTT 字段。因此，该步骤从原始文件中过滤掉其他字段，仅保留 FRSHTT 字段。

第 2 步：信息提取。FRSHTT 是 6 位数字。从 FRSHTT 字段中提取第 2 位数字、倒数第 2 位数字，分别对应是否下雨、是否打雷，形成一个新的 2 位数字 RT。

第 3 步：分组统计 RT 的 4 种情况。即 RT 为 00、01、10、11 的次数，分别代表没打雷且没下雨、打雷但没下雨、下雨但没打雷、打雷且下雨的 4 种情况出现次数。

（二）气象信息合并

在 12 组文件都处理完成之后，将各组处理结果按列拼接到一起，形成汇总后信息，再按照按行求和，得到所有分组数据的 RT=00/01/10/11 出现的次数 SUM。这个数据合并的结果作为后续分析的基础。

三、图形分析法

（一）热力图分析

构建“雷电-降水”次数统计热力图。在热力图中，颜色越亮，出现的次数越多。从热力图可以看出，没下雨也没打雷的次数最多（正常天气情况），打雷了却没下雨的次数最少。通过热力图，可以快速得到一个数据分布的初步印象。

（二）柱状图分析

构建打雷与下雨的组合次数柱状图。分别计算：没打雷且没下雨的概率 0.77；没打雷但下雨了的概率为 0.23；打雷且下雨了的概率 0.77；打雷但没下雨的概率为 0.23。

（三）饼状图分析

构建“雷电-降水”次数统计饼状图。分别计算：没打雷且没下雨的概率 0.77；没打雷但下雨了的概率为 0.23；打雷且下雨了的概率 0.77；打雷但没

下雨的概率为 0.23。

四、相关系数法

皮尔逊相关系数的定义如下：

其中，表示变量 X 与 Y 的协方差，和分别表示变量 X 和 Y 的标准差，和分别表示变量 X 和 Y 的均值。

通过计算，得到变量 T（是否打雷）与变量 R（是否下雨）之间的皮尔逊相关系数 Pearson 为 0.275（弱线性相关），显著性水平 p 为 0（非常显著）。

五、“雷电强度-降水量”函数拟合

（一）雷电数据处理

1. 雷电数据来源。本文使用了中国科学院子午工程武汉九峰观测站 2013 至 2023 共 11 年、961 万条、6678 个文件 5.18GB 的雷电强度数据^[7]，降水量仍选用美国国家环境信息中心数据。子午工程是由中国科学院牵头的一个国家级大型空间环境地基监测系统。

2. 电场仪数据分析。（1）地面电场仪。在雷电观测中，使用地面电场仪观测“云体-地面”间电压，每分钟观测 1 次。在观测数据文件中，每行代表 1 次观测结果，每次观测得到 7 个数值：时间、6 个“云体-地面”电压水平分量^[8,9]。（2）大气电场仪数据。在雷电观测中，使用大气电场仪观测“云体-云体”间电压，每分钟观测 1 次。在观测数据中，也是每行代表 1 次观测，每次观测得到 2 个数值：时间、“云体-云体”电压^[8,9]。

3. 电场仪数据处理。图 1 给出了基于 Knime 工具的电场数据处理过程，包括 3 个主要步骤：（1）分别读取 3 类数据：降水量、“云体-云体”电压、“云体-地面”电压。（2）将这 3 类数据按照日期合并求和，得到每天的总量。（3）输出处理结果。

（二）线性函数拟合

1. 线性回归。令降水量表示为 R、“云体-云体”电压表示为、“云体-地面”电压表示为，则本文要寻找的线性关系为。本文使用历史数据，使用 python 中的线性回归模型进行训练，寻找参数和，从而确定降水量与雷电强度之间的函数关系。最终得出线性回归的参数 $\alpha = -0.00111724$ ； $\beta = 0.00010757$ 。因此，降水量 R 与“云体-云体”电压、“云体-地面”电压之间的函数关系为：。

2. 模型结果。上节我们得到了降水量 R 与“云体-云体”电压、“云体-地面”电压之间的函数

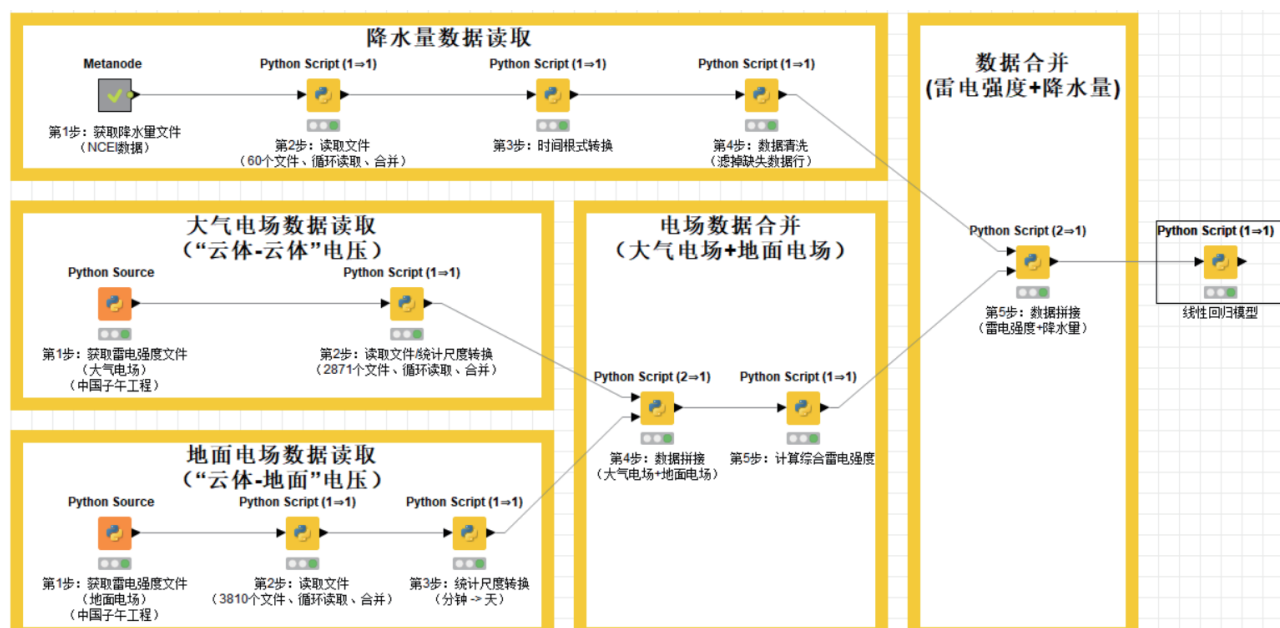


图1 基于 Knime 的雷电强度与降水量数据读取过程

关系为： ρ 。其中，“云体-云体”电压之前的系数为负数，说明“云体-云体”电压越大、降水量越小，证明了“雷声大雨点小”正确；“云体-地面”电压之前的系数为正数，说明“云体-地面”电压越大，降水量越大，证明了“雷声大雨点小”错误，应为“雷声大雨点大”。

3. 结果验证。上面为大数据给出的论证结果，我们再反推到自然原理中进一步印证：“云体-地面”电压大，说明空气上下对流强度大，单位质量空气降温的幅度大，则降水量大；而“云体-云体”电压大，说明空气水平对流强度大，单位质量空气降温的幅度小，则降水量小。

结束语：

本文使用大数据论证了“雷声大雨点小”的正确性，即雷电与降雨之间的相关性。首先收集并处理了美国国家环境信息中心 1.7 亿条气象数据（26,584 个站点 1929 至 2023 共 95 年）、中国科学院子午工程 961 万条（2013 至 2023 共 11 年）的雷电数据。

在此基础上使用统计学方法，证明了“雷电-降水”之间存在弱线性相关性；使用机器学习方法，证明了：当雷声来源于云体间电场时，“雷声大雨点小”正确；当雷声来源于云体与地面间电场时错误，应为“雷声大雨点大”。

从修辞学的角度，“雷声大雨点小”用来比喻做

起事来声势造得很大，实际行动却很少，这句俗语无可厚非。然而从科学的角度，“雷声大雨点小”缺乏严谨性。因此，这句俗语的正确性，取决于应用场景。

参考文献：

- [1] 维特根斯坦著，逻辑哲学论，First published in W. Ostwald's Annalen der Naturphilosophie, 1921 年。
- [2] 小宫山博仁著，王倩倩译，图解统计学 [M]. 北京时代华文书局，2021 年。
- [3] Luis G.Serrano 著，郭涛译，机器学习图解 [M]. 清华大学出版社，2023 年。
- [4] Knime 数据科学开发平台官网，<https://www.knime.com>，2023 年 9 月 4 日可用。
- [5] 郝秀书，张其林，袁铁等，雷电物理学 [M]. 科学出版社，2013 年。
- [6] 美国环境信息中心 (National Centers for Environmental Information, NCEI) 官网，<https://www.ncei.noaa.gov>，2023 年 9 月 4 日可用。
- [7] 中国子午工程数据中心，<https://data2.meridianproject.ac.cn>，2023 年 9 月 4 日可用。
- [8] 中国气象局，地面大气电场观测规范，中华人民共和国气象行业标准 QX/T 594-2020，2020 年 12 月 29 日。
- [9] 秦佳媚，钟鼎坤，冯学尚等，子午工程数据资源概述 [J]. 中国科学数据，2021,6(2)。