arXiv:2303.05484v1 [stat.AP] 9 Mar 2023

L让’S 谈谈天气：基于聚类的天气预报准确率研究 *

Jill Lundell
数据科学系 达纳-法伯癌症研究所
生物统计学系 哈佛大学陈曾熙公共卫生学院 马萨诸塞州波士顿 02215
jlundell@ds.dfci.harvard.edu

Brennan Bean
数学与统计学系 犹他州立大学 洛根，犹他州 84322
brennan.bean@usu.edu

Jürgen Symanzik数学与统计学系
犹他州立大学 洛根，犹他州 84322
Juergen.Symanzik@usu.edu

摘要

深入了解与美国天气预报准确率相关的特征，有助于气象学家做出更准确的预测，也有助于美国人更好地解读日常天气预报。本文考察美国各地的时空特征与预报准确度之间的关系。我们根据天气和地理特征，将美国划分为六个天气区域，并分析各天气区域内预报准确度的模式。随后，我们探讨这些天气区域内气候特征与预报准确度之间的关系。我们得出结论，预报误差的模式与各区域独有的气候密切相关。

***Key*词语** 气候 · 聚类 · Data Expo 2018 · 字形图 · 随机森林 · 可视化

从五大湖沿岸寒冷潮湿的冬季，到美国西南部炎热干燥的夏季，美国（U.S.）经历着范围广泛的气候极端现象。这些极端情况给天气预报带来了独特的挑战。要理解如此多样的地貌区域中的预报误差，同样困难重重，需要通过多维可视化来呈现空间、时间和气候测量值。深入了解美国各地预报误差的性质和模式，有助于气象学家努力改进天气预报。这也能帮助普通美国人判断，在重要活动当天，天气预报有多值得相信。

美国统计学会（ASA）统计计算与统计图形分会举办的 2018 年 Data Expo，为探索并比较美国各地的天气预报误差提供了机会。我们的分析聚焦于以下问题：

美国各区域的天气预报误差有何不同?

这一引发思考的问题促使我们进一步探讨以下问题：

- 美国的气象站是否会根据天气特征聚类成不同区域?
- 误差变量之间有何相关性？这些相关性是否因区域而异?
- 预报误差如何随地区和季节变化?

*引用： **JF Lundell、B Bean 和 J Symanzik**。《**来谈谈天气：一种基于聚类的天气预报准确性方法**》。

表1：分析中包含的天气变量列表。在分析之前，已剔除所有超出所示范围的观测值。

变量	Unit	范围
最低/最高气温	$^\circ F$	$[-37, 127]$
降水	in	$[0, 12.95]$
最低/最高露点	$^\circ F$	$[-50, 90]$
最低/最高湿度	%	$(0, 100]$
最低/最高海平面气压	inHg	$[28.2, 31.2]$
平均/最大风速	mph	$[0, 70]$
最小能见度	mi	$[0, 10]$
云量	okta	$\{0, 1, \cdots, 8\}$
距海岸距离	mi	$[0, 807]$
海拔	ft	$[3, 7422]$

- 哪些地方的预报准确度最高，哪些地方最低?
- 哪些变量对确定预报误差很重要?

初步分析结　　果发表于2018年联合统计会议论文集 [1]。

本文旨在回答这些问题。我们使用集成图形，全面呈现美国不同区域的天气预报误差。 [2]集成图形通过将数据的多种可视化结果与相邻文本联系起来，增强了传统分析。这种呈现方式能够比几幅彼此割裂的图形更有效地讲述连贯的数据故事。在第1节中，我们概述数据，并展示如何利用所提供的气候测量值、海拔和距海岸距离，将美国划分为六个界限分明的天气区域。这些聚类（即天气区域）构成了我们比较美国各地预报准确度的基础；第2节将通过一系列多维图和变量重要性分析对此进行阐述。在第3节中，我们介绍为拓展数据探索而创建的交互式应用程序。我们在第4节中得出结论：区分美国各天气区域的气候差异，也会形成因地区而异的模式，并导致预报准确度存在差异。本文末尾附有两个附录，分别说明数据清理方法，以及如何创建本文中使用的图元。

1 天气区域

数据包含2014年7月至2017年9月期间美国113个气象站的测量值和预报数据。这些数据可从我们的补充材料中获取，也可通过以下网址获取：

http://community.amstat.org/stat-computing/data-expo/data-expo-2018.

每个位置都记录了八种不同气象指标的每日测量值，包括温度、降水、露点、湿度、海平面气压、风速、云量和能见度。许多值得关注的天气事件也以文字形式记录下来，例如雷暴和雾。每项指标均记录了每日最低值、最高值和平均值。本文使用的天气特征列于表1。数据还补充了部分地理信息，并经过仔细检查和清理。有关数据清理、获取其他数据以及最终变量选择依据的详情，见附录A。

1.1 开发天气聚类

美国根据流域和气候等环境特征划分为不同区域 [3][4]。我们考察了现有的环境区域划分，但发现其中没有一种适用于本分析中的天气研究。我们根据表1中的指标，对气象站进行聚类，从而创建了自己的天气区域。因此，各聚类由每个气象站观测到的天气特征定义。我们利用这些聚类，确定天气预报误差模式与特定区域独特的气候测量值之间的关系。第1.2节将探讨现有天气区域及其与我们划分的天气区域之间的对应关系。我们对每个气象站的数据进行汇总，计算记录期内113个气象站的表1中各变量的均值和标准差。

我们使用欧几里得距离 [5] 和Ward最小方差聚类法 [6] 进行层次聚类，以识别各聚类。我们从空间分布上考察这些聚类，以评估聚类方法的表现并确定最终的聚类数量。我们希望确保气象站聚类的规模足够大，具有实用性。划分为五个聚类时，其中一个聚类涵盖了从美国中西部到美国东海岸的所有气象站。考虑到沿海气候与内陆气候的差异，我们认为这个聚类过大。划分为七个聚类时，其中一个聚类仅包含五个气象站，规模过小。因此，我们选择划分为六个聚类，将美国划分为不同的天气区域。
图1和图2展示了聚类分析的结果。图3展示了各天气区域特征的平行坐标图。我们计算了表1中每个变量的均值和标准差的Z分数，并将其绘制在平行坐标图上。六个天气区域彼此之间难以区分，因此我们创建了一个交互式应用，以便更清楚地查看每个聚类的特征。该应用将在第3节中讨论。

各天气聚类的名称和特征如下：

- **加利福尼亚—佛罗里达**（13个气象站）：温暖潮湿，露点和气压较高。几乎所有测量指标的变异性都较低。
- **美国东南部**（22个气象站）：温暖潮湿，降雨量大。降水的变异性较高，而温度的变异性较低。
- **美国东北部**（39个气象站）：寒冷、潮湿，能见度低。温度、露点和气压的变异性较高。
- **山间西部**（19 个气象站）：寒冷干燥，温度、风速和气压的变化幅度较大。降水量和露点的变化幅度较小。
- **美国中西部**（13 个气象站）：地处内陆，风速较高，温度、气压和风速的变化幅度较大。
- **美国西南部**（7 个气象站）：温暖、晴朗、干燥，温度或降水量变化很小。风速和湿度的变化幅度较大。

1.2 与现有气候区域的比较

其他研究已对美国的生态区和气候区域进行了划分。许多研究聚焦于美国较小的区域，但也有少数研究考察了整个美国。不同研究采用的聚类方法以及用于识别聚类的变量各不相同。环境合作委员会划定的北美生态区 [3] 以生态系统为依据划分区域。研究人员利用空气、水、土地和生物群（包括人类）来划分生态区。这些生态区呈现出明显的经度方向趋势，与我们的聚类所呈现的经度方向趋势十分吻合。研究人员并未采用统计聚类方法来确定聚类，而是通过仔细评估北美洲各地的生态属性来划分。

美国国家海洋和大气管理局（NOAA）制定了纳入季节性气温和降水信息的气候区域 [7]。这些区域与北美洲生态区域有很大差异，因为它们除了呈现纵向趋势外，还呈现横向趋势，并且受州界限制。NOAA 使用评估干旱和湿润期的谱曲线来定义这些区域 [8]。尽管区域边界由州界划定，NOAA 的区域与我们划分的一般天气区域大致相符。美国东部的南北分界与我们在该地区的聚类划分十分接近。我们的聚类中主要的东西向分界也与 NOAA 聚类的分界位置相近。

国际能源节约规范（IECC）对美国的气候聚类 [4] 以及 Hathaway 等人随后进行的重新分类 [9]，依据温度、露点、风速和辐射将美国划分为十四个区域。聚类方法包括 K均值聚类和蒙特卡洛筛选。两套区域划分都显示出美国东部明显的横向趋势。这些区域还清楚地将美国西海岸和美国西南部沙漠与美国西部其他地区区分开来。我们的聚类中也呈现出类似的趋势。我们的聚类中，美国东部的横向趋势没有那么明显，但这很可能是因为我们选择的天气聚类数量较少。加入距海岸距离、海拔和湿度等对横向趋势不敏感的其他变量，都会减弱我们聚类中的横向分离。

我们的天气区域与其他研究中所见区域的一个关键区别是，我们将佛罗里达州和太平洋沿岸合并为一个天气区域。这可能是因为我们在聚类分析计算中选择不考虑气象站的地理邻近性，而只考虑天气模式的相似性。佛罗里达州和太平洋沿岸的天气模式季节性都比美国其他地区弱。因此，这两个区域的许多气候变量的标准差都低于平均水平。这些较小的标准差使佛罗里达州与太平洋沿岸之间呈现出较高的接近程度，这或许可以解释为什么这两个

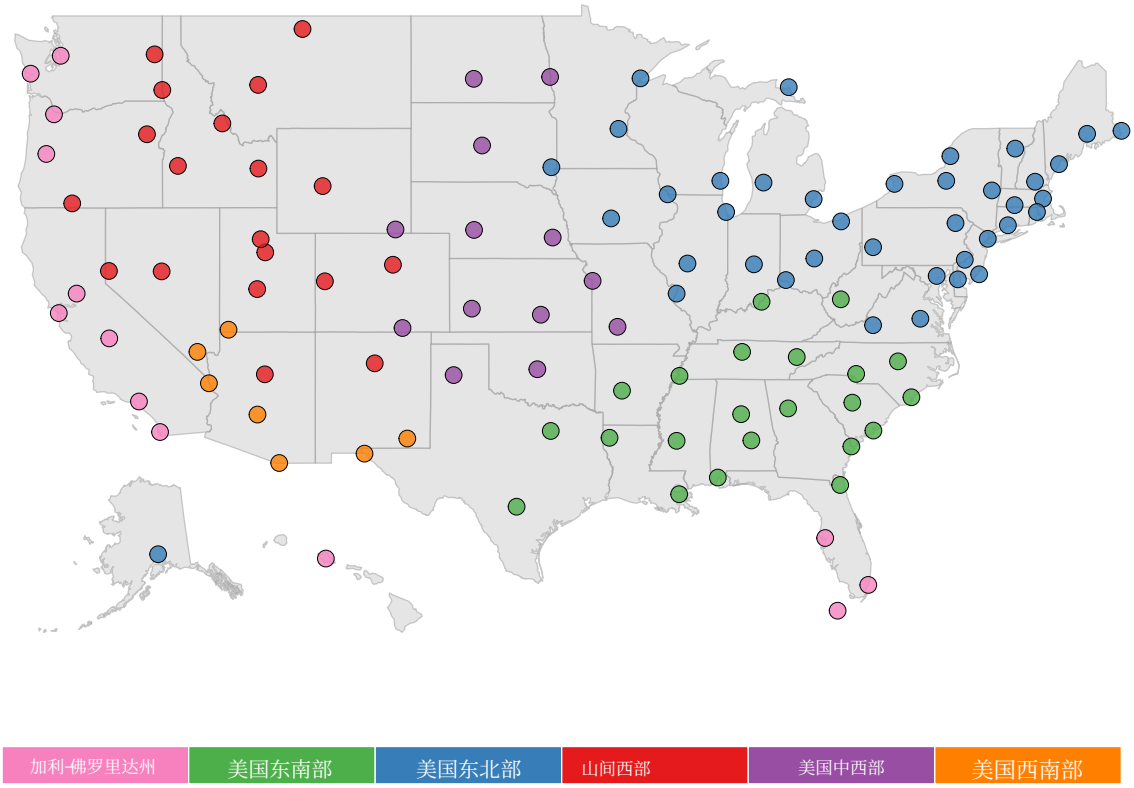


图1：六个天气区域的地图。底部的色带通过名称和颜色标示各个区域

地理区域在使用六个或更少聚类时会被归入同一个聚类。使用七个聚类时，佛罗里达州和太平洋沿岸的气象站会分属不同聚类，但太平洋沿岸有两个气象站例外，它们与佛罗里达州的气象站归为一类。文献中要么忽略夏威夷和阿拉斯加，要么将它们划入各自独立的区域。由于我们没有将空间邻近性作为聚类变量，并且将所有气象站都分配到六个天气聚类之一，夏威夷和阿拉斯加分别与加利福尼亚—佛罗里达和美国东北部归为一类。我们的聚类结果表明，天气模式通常具有较强的空间相关性，温带沿海区域则是一个显著例外。

2 预测误差探索

鉴于该国明显划分为不同的天气区域，我们希望确定这些区域之间的预测误差模式是否存在明显差异。分析仅针对最低气温、最高气温和降水概率的预报。最低气温和最高气温的预报误差按预报值与实测值之间的绝对差计算。降水的预报误差则使用布里尔技巧评分（BSS）衡量，这是一种广为人知的概率预报准确度衡量指标 [10]。对于某个特定气象站，其定义如下：

$$\mathrm{BSS} = 1 - \frac{\sum_{i=1}^{N}\sum_{j=0}^{M}\left(Y_{ij} - O_i\right)^2}{\sum_{i=1}^{N}\sum_{j=0}^{M}\left(P - O_i\right)^2} \tag{1}$$

其中

- $Y_{ij} \in [0,1]$ 是预报滞后为 j 时，对第 i 天降雨概率的预测；

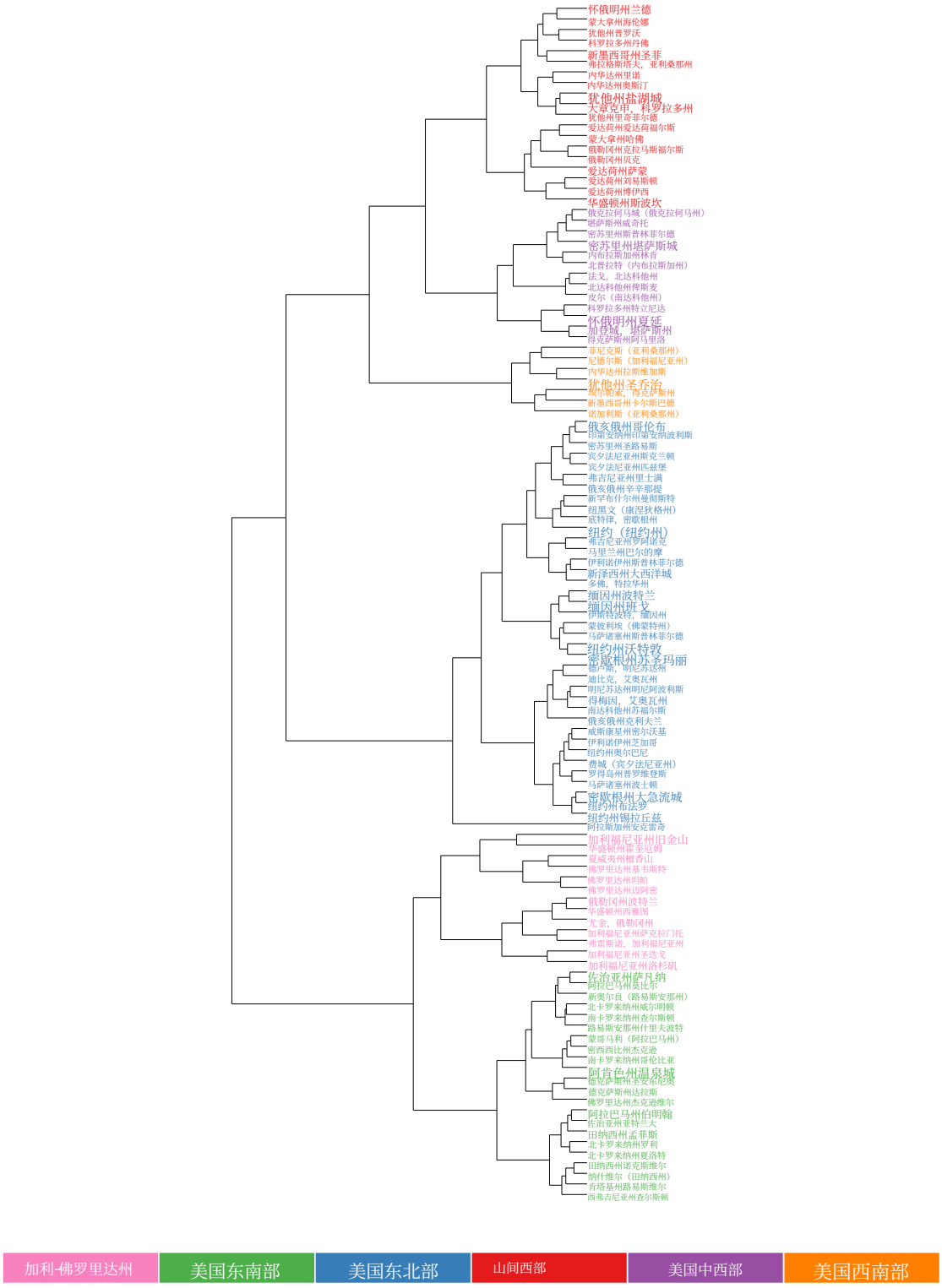


图2：图1中识别出的天气聚类树状图

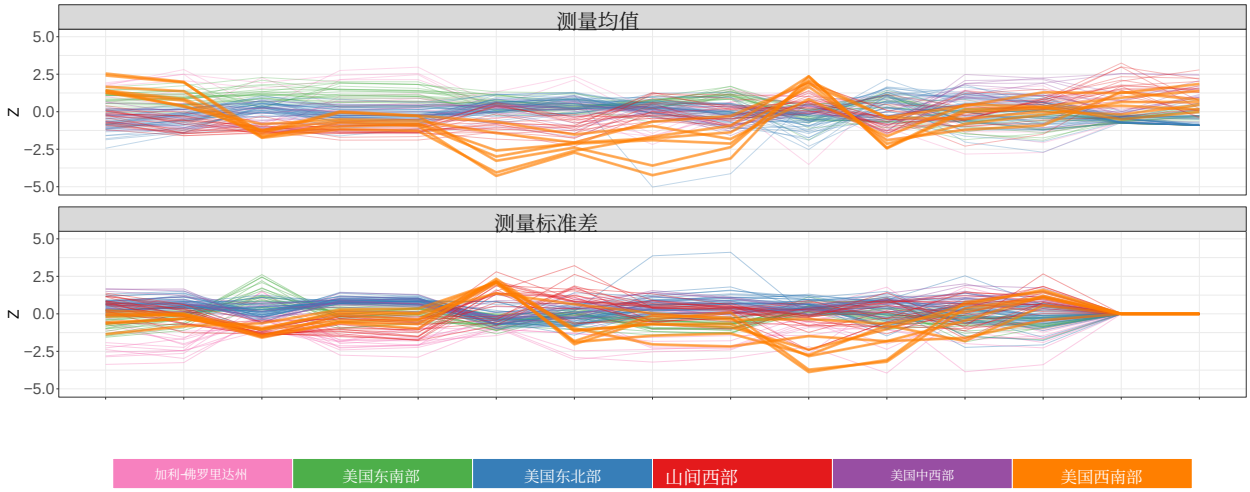


图3：表1所列天气变量均值和标准差的平行坐标图。图中每条线代表113个气象站中的一个。线条的颜色与气象站所属的天气区域相对应。另有交互式应用可供使用，以便更好地识别区域趋势。图中突出显示了西南地区，以强调其天气特征

- $O_i \in \{0, 1\}$ 是一个二元变量：如果当天有任何降水，则取值为1；否则取值为0。我们将降水事件定义为降水测量值大于零，或事件信息中包含“rain”或“snow”这两个词；N
- $P \in [0, 1]$ 是感兴趣时期内每日降水概率的平均值，定义为 $P = \frac{1}{N} \sum O_i$；$i=1$
- N 表示记录期间有降水记录的天数，$M \in \{0, \ldots, 5\}$ 表示预报滞后期数。

请注意，BSS $\in (-\infty,\ 1]$，其中 1 表示预报技巧完美，数值朝 $-\infty$ 变化则表示预报效果变差。我们选择使用 $1 - \text{BSS}$，以确保三个误差变量的方向一致。以下小节将探讨先前定义的天气区域之间及各区域内部的预报误差差异，这些区域在图1中以可视化方式呈现。每张图中的预报误差均按滞后期取平均，部分情况下还按月份取平均。以下小节中的可视化结果证实了我们的假设：不同天气区域的天气预报误差模式截然不同。

2.1 误差相关性

三种不同测量指标（即最低气温、最高气温和降水量）的预报误差彼此相关吗？这些关系在不同天气区域之间如何变化？我们通过图4中叠加在美国地图上的相关椭圆 [11] 来探究这些相关性。我们计算了每个聚类内各位置的每对测量指标之间的斯皮尔曼相关系数。相关系数的符号由椭圆的斜率表示，相关强度则由椭圆的宽度表示。

除了美国东北部的最低气温与另外两个变量之间的相关性外，所有误差变量之间的相关性均为正。最强的关系出现在美国中西部、南部和美国西南部，最弱的关系则出现在美国东北部。只有少数针对特定聚类的相关性达到显著水平。这可能是因为许多天气区域中的气象站数量较少。不过，113 个气象站的总体相关性均为正且显著。这表明，在某个预报变量上预测效果良好的区域，其他预报变量的预测效果通常也不错。最低气温与降水量预测之间的相关性最弱。尽管这三个天气预报变量之间存在关联，但这些关联并不特别强，而且各区域内的关联强度有所不同。通过这幅相关性椭圆地图得出的观察结果，展示了这种绘图方式如何便于对空间上的多维数据进行比较。有关相关性符号的计算和实现方法，请参见附录B。

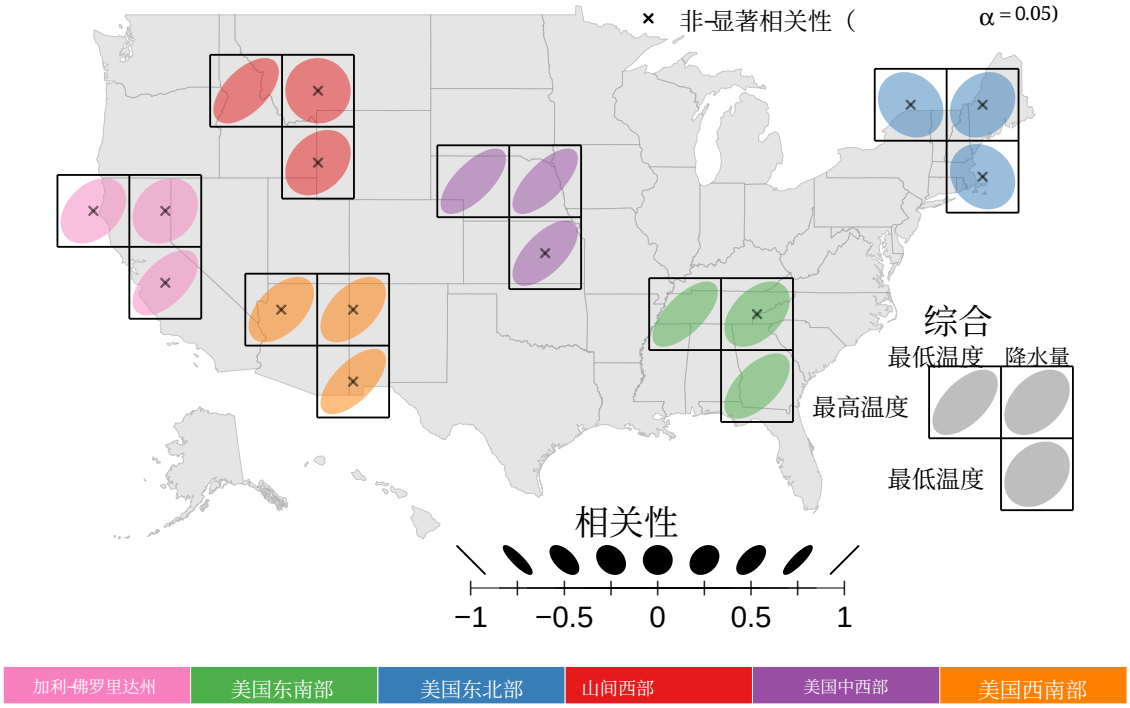


图4：以椭圆表示的预报误差变量之间的斯皮尔曼相关系数，叠加在美国地图上。每个相关系数的 p 值均与 0.05 的显著性水平进行比较

2.2 误差散点图

散点图揭示了天气区域内及不同预报滞后期下的异常值和整体趋势。预报滞后定义为预报日期与所预报日期之间相隔的天数。因此，当日预报的滞后期为 0，提前一天的预报滞后期为 1，以此类推。由于我们要对美国范围内三个变量在空间和时间上的情况进行比较，静态图表并不适合用于评估所有关注的关系。我们构建了一个交互式散点图应用，便于考察汇总所有预报滞后期或针对单个预报滞后期的三个预报误差变量之间的趋势。图5 (a-c) 展示了交互式应用中的示例图。图中显示了汇总所有预报滞后期的数据散点图，以及滞后期为 5、3 和 1 的散点图，以说明预报准确度如何随预报滞后期变化。

图5(a)比较了最低气温预报准确度与降水准确度。最低气温预测最差的气象站位于新英格兰和山间西部。新英格兰以极端冬季天气闻名，极端天气事件的发生频率似乎正在上升 [12]。这可能是这些气象站难以预测最低气温的原因之一。最低气温预测最差的地点是内华达州奥斯汀。图5(c)将进一步讨论这一地点。加利福尼亚—佛罗里达的最低气温预测始终最佳。不过，在考察各个滞后期时，加利福尼亚—佛罗里达的降水预测准确度也呈现出最大的变异性。

图5(b)比较了最高气温预报准确度与降水准确度。五大湖地区有四个气象站的降水预报在数据集中最差。该地区降水预报准确度较差，说明预报湖效应雪十分困难。这一现象将在第2.3节中深入讨论。随着预报滞后缩短，五大湖地区的降水预报准确度大幅提高；预报滞后为 1 时，其准确度与全国其他地区相当。

图5(c)展示了最低气温与最高气温预报准确度之间的关系。这些散点图中有三个异常值格外突出，分别是佛罗里达州基韦斯特、内华达州奥斯汀和加利福尼亚州旧金山。基韦斯特对最低气温和最高气温的预报都比其他任何气象站更准确。基韦斯特在八个天气变量中有六项的变异性排名也位居最低的前五名，这或许解释了其预报为何如此准确。奥斯汀对这两项指标的预报表现最差。沿着“美国最孤独的公路”行驶七十英里，便可从奥斯汀到达

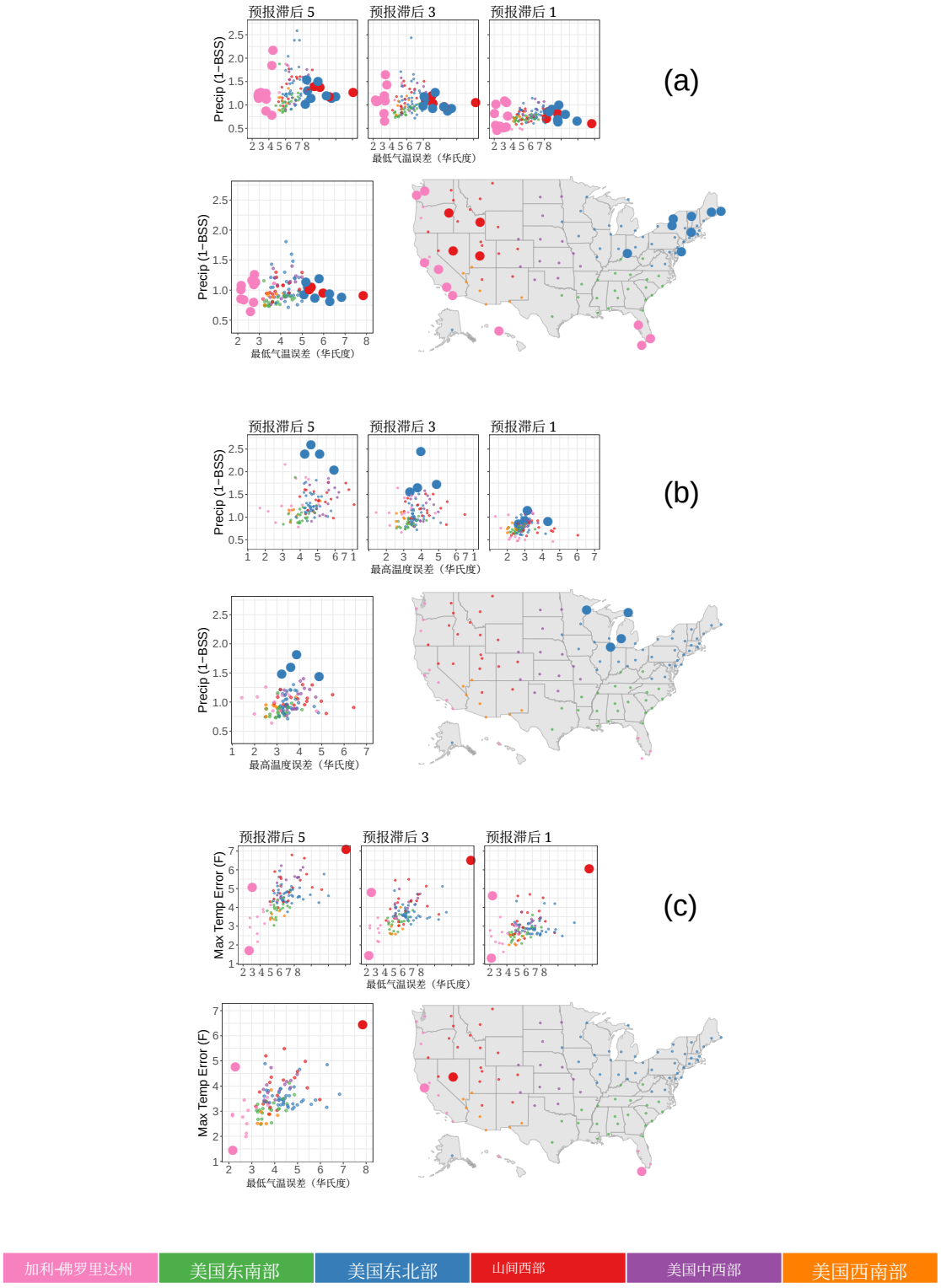


图5：比较三个预报误差变量的散点图。地图左侧的散点图汇总了所有预报滞后期的数据。文中讨论的关注点已在相应的图中突出显示

其天气测量数据采集自内华达州尤里卡。最高气温和最低气温的预测效果不佳，可以用如此长距离上的气候变化来解释。这反映为最高气温存在约 5°F 的负预测偏差，而最低气温存在约 7°F 的正预测偏差。旧金山的最低气温预测效果良好，但最高气温预测效果不佳。第2.3节将进一步解释这一现象。

与本项目配套开发的交互式应用可用于进一步调查预报准确度趋势。第3节将介绍该应用。

2.3 季节性趋势

美国位于北半球，因此该国大部分地区都有明显的天气季节。美国北部的季节特征最为显著。我们假设，预报误差的变化与季节性密不可分。我们通过一系列时空图对此进行探究。对空间和时间同时建模会形成一个三维问题，通常用小多图来可视化。小多图是“一系列图形，展示相同的变量组合 [例如，纬度和经度]，，并根据另一个变量 [例如，时间]的变化进行索引” [14]。这种方法的问题在于，除了图与图之间最显著的变化外，其他变化都很难通过视觉理解。一种能够同时可视化空间和时间的替代方法，是在空间情境中使用字形或符号，从而实现多维可视化 [15][16]。

图6展示了随时间变化的季节性预报误差字形图。预报误差以从中心点到多边形边缘的缩放距离呈现；多边形包含12个观测值，从12点钟位置的一月开始，按顺时针方向排列。字形相对于中心点的不对称性，体现了预报误差如何随时间和空间变化。例如，美国东北部各地在冬季的降水预报不如夏季准确，而美国东南部各地全年降水预报的准确度则相当。

除了凸显预报方面的不对称性外，图6还揭示了地点特异性异常。例如，加利福尼亚州旧金山全年最低气温预报都很准确，但只有在冬季月份才能准确预报最高气温。这可能是因为该地区全年经常出现寒冷的沿海雾，导致短距离内温差显著 [17]。鉴于这些事实，温度预报存在困难似乎合情合理，因为测量地点距离预报地点超过11英里。冬季这一问题可能不那么明显，因为内陆与沿海地区的温差较小。

内华达州奥斯汀的夏季最高气温预测尤其不准确。目前尚不清楚为什么夏季的预测会比冬季更差。

另一个值得注意的位置特异性异常，是图6所示五大湖周边地区降水预测的季节性变化十分显著。图5(b)中的误差散点图显示，该地区的降水准确度较低，但散点图无法呈现预测结果的季节性。冬季预测异常糟糕，很可能是由于该地区常见的湖效应雪所致。冬季苏必利尔湖下风向的降雪量比没有湖效应时预计的多出最多 100% [18]。此前已有研究指出，该地区的降水模式是全美最难预测的 [19]。以上示例表明，利用这些基于图元的图表，可以轻松地对不同空间和时间的数据进行比较。附录B中介绍了如何生成这些图元。

2.4 变量重要性

不同区域的预测误差模式存在差异，因此需要找出对预测误差最重要的气候测量指标。我们使用随机森林 [20]来确定哪些天气变量对预报误差的影响最大。数据按预报滞后和月份进行汇总。我们以预报误差变量为响应变量，针对每个天气区域分别建立了三个随机森林模型。表1所列各天气变量的均值和标准差，以及预报滞后，均作为预测变量。图7包含三个平行坐标图，展示了各区域中每个预报误差变量对应的变量重要性度量。我们将随机森林得出的重要性度量减去最小重要性度量，使其重新居中；然后，以各天气聚类与预报误差变量组合中重新居中后的最大重要性度量为除数，将数值缩放到区间 $(0, 100)$，最后乘以100。因此，对于每种误差度量，每个天气区域内最重要的变量取值为100，最不重要的变量取值为0。这使我们能够直接比较不同天气区域之间以及不同误差度量之间的重要性。

图7显示，无论天气区域如何，对降水误差影响最大的变量都是预报滞后。与预报滞后相比，其他变量的重要性都不高。美国东南部的最低露点（DP）和

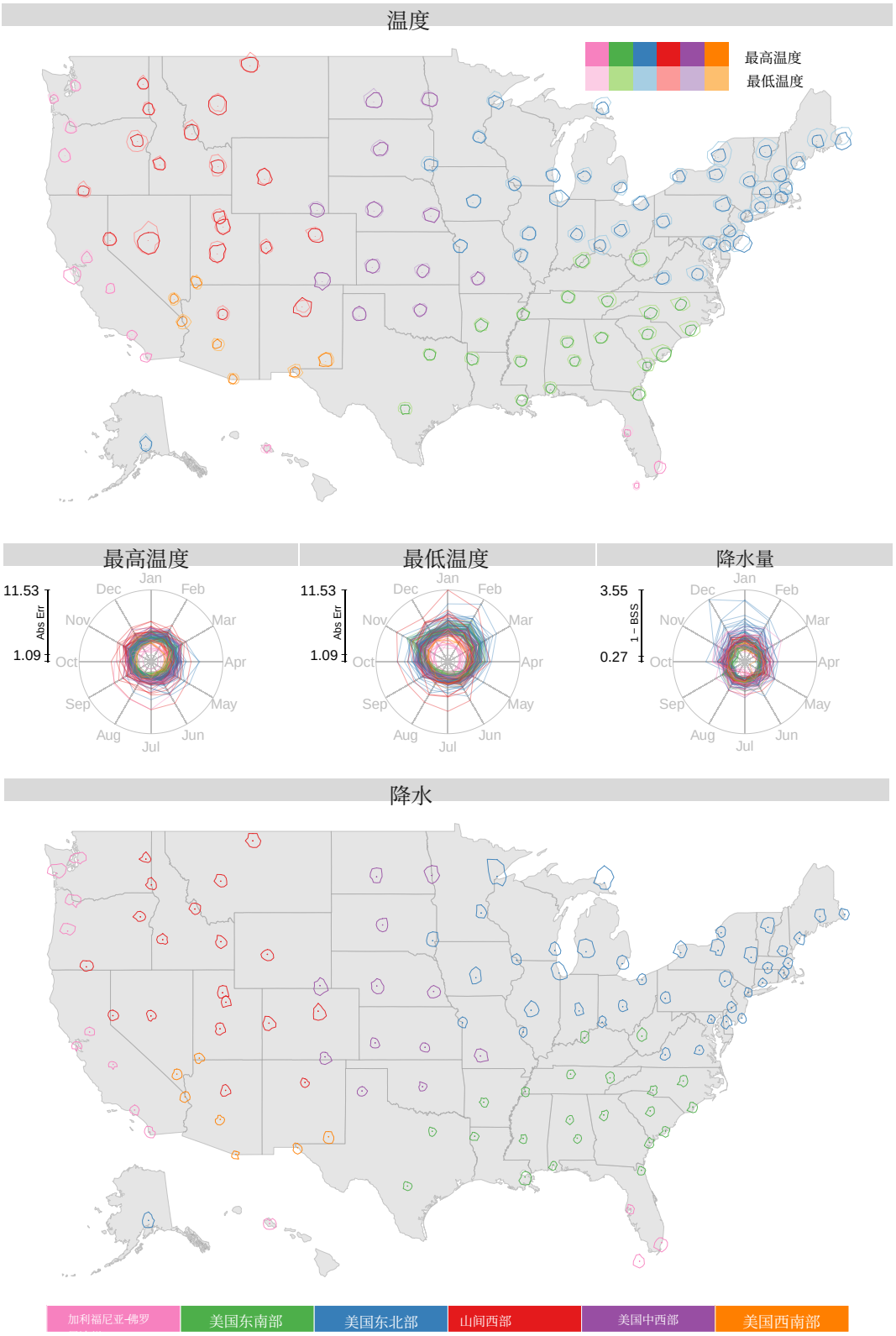


图6：按月份平均的天气预报准确率字形图。误差以从中心点到多边形边缘的缩放距离表示，多边形从12点钟位置的一月开始，按顺时针方向排列

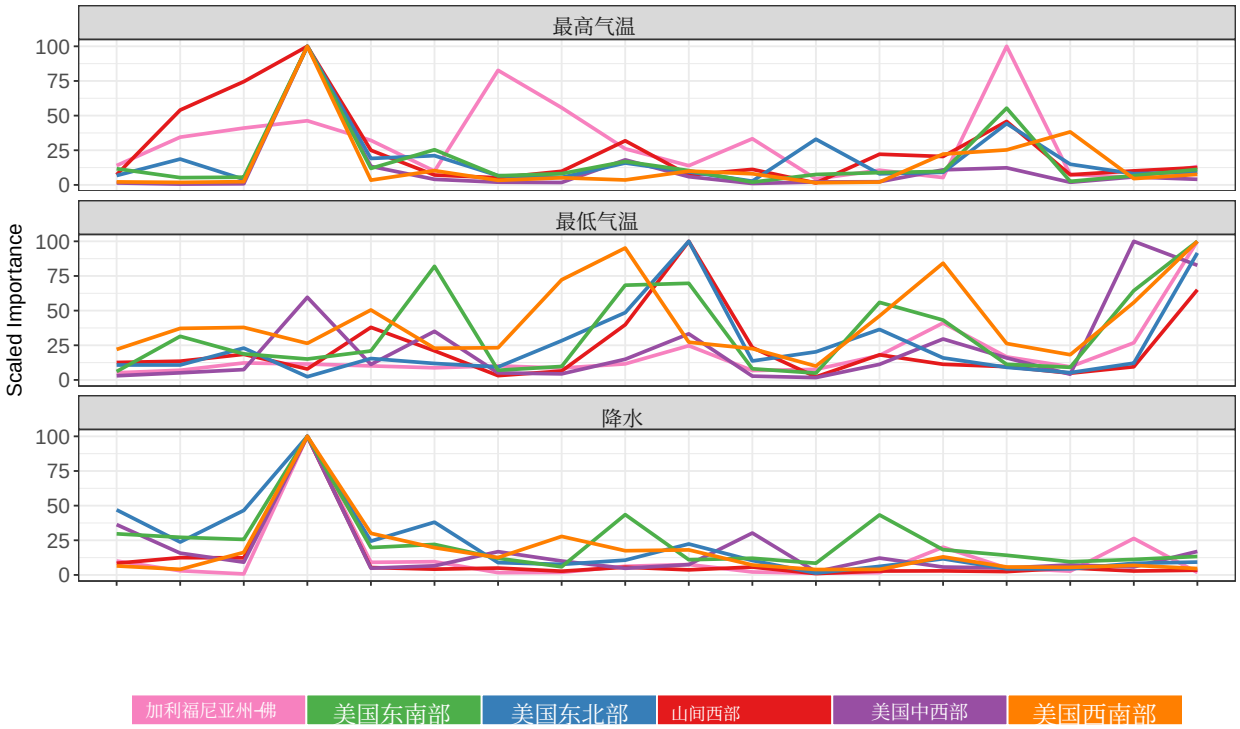


图7：三个预报准确度指标各自的变量重要性。变量重要性度量经过重新缩放，使其能够在不同天气区域和准确度指标之间直接比较

最高露点的标准差具有一定的重要性。云量对美国东北部的降水误差很重要。

除加利福尼亚—佛罗里达外，预报滞后也是所有天气区域中影响最高气温误差的最重要变量。在加利福尼亚—佛罗里达，最高气温的标准差和最大风速（WS）比滞后更重要。最高气温的变异性对美国东南部、美国东北部和山间西部也很重要。距海岸距离（Dist2Coast）和海拔对山间西部的最高气温误差很重要。

影响最低气温误差的重要变量在各天气区域之间差异很大。最低气温的变异性对所有区域都很重要，但其他重要变量因区域而异。最低气温对美国东北部和山间西部最重要，而最高气温对美国东南部很重要。最低露点和最高海平面气压（SLP）的变异性对美国西南部很重要，而最低海平面气压的变异性则对美国中西部、美国东南部和美国西南部最重要。除美国中西部外，预报滞后对其他区域都不算特别重要。

3 交互式应用程序

仅通过静态可视化，很难识别所有天气区域的气候测量值和预报误差中的模式。我们开发了一款交互式 Shiny 应用，以便更深入地探索天气数据。可通过以下网址访问此应用：

`https://jilllundell.shinyapps.io/finaldataexpoapp/.`

应用的第一个选项卡是图3所示平行坐标图的交互式版本。用户可以选择一个天气区域，该区域会在图中高亮显示。选定区域的特征一目了然，并且可以轻松地与所有其他观测值进行比较。

应用的第二个选项卡是交互式散点图。图5 (a-c) 展示了此选项卡生成的图表示例。用户最多可以选择三个预报误差变量中的两个，将其显示在坐标轴上。用户还可以选择预报滞后。用户可以在散点图上刷选或点击数据点，选中的点会显示在美国地图上。图表下方的表格会列出所选气象站的信息。Monmonier [21] 首次提出了散点图与地图之间联动刷选的概念。与静态图表相比，此应用可以更全面地探索不同预报滞后期以及不同误差变量之间数据中的异常值和趋势。

4 结论

通过聚类分析，利用表1所列天气变量的均值和标准差，可以将美国的气候模式清晰地划分为六个易于识别的区域。我们使用图4所示的相关椭圆，探讨了三种天气预报变量（即最低气温、最高气温和降水量）之间的关系。我们发现，除东北部聚类外，所有聚类中的误差变量都呈现正相关迹象。

我们通过图5中涵盖所有预报滞后期的一系列散点图，展示了预报误差之间的两两关系。这些图突显了加利福尼亚—佛罗里达地区在所有滞后期预测最低气温方面的优势，也表明五大湖地区对降水量的预测较差，这种情况大多仅出现在滞后期大于2的预报中。最后，内华达州奥斯汀的误差异常偏高，可能是由于预报地点与测量地点之间距离较远所致。
我们在图6中探讨了预报误差的季节差异，并观察到，与南部聚类相比，北部内陆聚类的预报误差季节差异往往更为明显。我们还指出，某些特定地点的异常现象，例如旧金山最高气温预报误差的季节性不对称，以及五大湖附近的降水量预报误差，在文献中都有合理的解释。

接下来，我们使用图7中经过缩放的随机森林变量重要性度量，比较了各聚类中决定预报误差的重要变量。这些度量表明，预报滞后对最高气温和降水量预报误差的影响最大，但对预测最低气温预报误差并不重要。许多聚类对少数几个变量赋予了相近的重要性，但也有些变量只在单个聚类中重要，例如，在加利福尼亚—佛罗里达地区，最大风速对于预测最高气温预报误差的重要性。

若想进一步了解这六个聚类中的预报误差特征，请读者参阅上一节介绍的 R Shiny 应用。该应用的当前版本可通过以下网址访问:

https://jilllundell.shinyapps.io/finaldataexpoapp/

该应用与本文中的可视化结果相结合，进一步印证了这样一种观点：美国清晰地划分为若干特征鲜明的天气区域，而预报误差模式与各区域特有的气候密切相关。

本文中的可视化，无论是交互式还是静态的，均旨在适用于规模更大的天气数据集。我们预计未来将通过更多气象站展示这一能力。扩大分析范围也有助于验证本文观察并描述的区域模式。此外，我们预计将把本文中的若干静态字形图改编为交互式图表。增强交互性将有助于更深入地探索美国天气模式的时间与空间变化。

5 致谢

作者谨此感谢美国统计学会统计计算与统计图形分会提供本分析所用的数据。本分析主要使用 R [22] 作为分析工具。weatherData R 包 [23]提供了有关特定测量地点的补充数据。距离测算和空间计算使用了 fields [24], 、geosphere [25], 、mapproj [26], 、rgdal [27], 和 sp [28] R 包。其他数据处理和可视化则使用了 tidyverse [29], , 以及 ggforce [30], 、latex2exp [31],、RColorBrewer [32], 和 reshape2 [33] R 包。变量重要性模型使用了 randomForest [34] R 包。另

数据清理

我们主要使用 Data Expo 提供的数据集进行文章中所述的分析。我们在所提供的位置信息基础上，补充了海拔和距最近主要海岸的距离信息。我们通过 rgbif R 包 [36]，利用 Google 的 API 服务器 [35] 获取各位置的海拔信息。距海岸距离按每个测量地点与美国中等比例尺海岸线数据集 [37]，中某个顶点之间的最短地理距离计算；该数据集涵盖美国本土 48 州的所有海洋海岸和五大湖海岸。由于该数据集不包括阿拉斯加和夏威夷的海岸线，这些地区的距海岸距离是根据从 NOAA 海岸线数据浏览器 [38] 手动提取的海岸线计算的。我们承认，这种距离计算方法存在局限性：对于亚利桑那州（弗拉格斯塔夫、诺加利斯和菲尼克斯）等一些地点，计算出的距离略长于使用墨西哥加利福尼亚湾海岸线信息时的结果。尽管如此，这些测量结果仍能有效区分内陆气象站和沿海气象站。

表1展示了最终分析中包含的天气变量。我们排除了温度、降水、露点、湿度和海平面气压的日均测量值，因为这些测量值与相应的最低值和最高值几乎呈完美线性组合。我们还从分析中排除了最大能见度，因为在所有记录的测量值中，超过 97% 的最大能见度均为 10 英里。最后，我们去除异常值后，保留最大风速和最大阵风风速两个测量值中较低的一个，从而合并了这两项信息。决定合并这两个风力变量的信息，主要是因为所有最大阵风风速值中有 13% 缺失。此外，很难将异常偏高但有效的最大阵风风速和风速测量值与真正的异常值区分开来。

有些气象站没有记录相关气候变量。在可能的情况下，我们通过美国国家气候数据中心（NCDC） [39]获取美国国家气象局（NWS）一等气象站的相应测量值，以替换这些缺失的观测值。缺失值包括马里兰州巴尔的摩的风速、科罗拉多州丹佛的降水量，以及多个地点被替换的异常降水测量值。如果无法通过 NCDC 轻易获取替代数据，则使用数据集中地理位置最近邻站点的相应观测值来替换系统性缺失的观测值。例如，马里兰州巴尔的摩的能见度和云量数据由特拉华州多佛的测量值替换，内华达州奥斯汀的数据则由内华达州里诺的测量值替换。

表1还列出了所纳入各变量的观测范围。这些测量范围有些是定义性的，例如湿度的上下限；有些则只是基于实际情况设定的，例如温度的上下限。所有超出表1所示范围的测量值均在分析前剔除。我们还根据特定位置的不一致情况，剔除或替换了若干个别异常值，包括

- 剔除 of one 异常的 low 最低 温度 测量 in 檀香山，夏威夷（< 10°F），以及加利福尼亚州旧金山的两个（< 20°F）；
- 将以下异常偏高的降水读数替换为附近气象站的降水读数 [39]:

  – 俄克拉何马城（俄克拉何马州），2017年8月10日（38.33英寸 $\rightarrow$ 0.8英寸）– 爱达荷州萨蒙，2015年4月21日、2016年5月2日及2017年5月3日至4日（10.02英寸 $\rightarrow$ 0英寸）– 亚利桑那州弗拉格斯塔夫，2016年12月24日（7.48英寸 $\rightarrow$ 0.97英寸）– 印第安纳州印第安纳波利斯，2015年7月15日（9.99英寸 $\rightarrow$ 0英寸）；

- 移除一个异常偏低的最低露点测量值：夏威夷檀香山（$<$ 40°F）、华盛顿州霍奎厄姆的两个（$<$ 0°F）、内华达州拉斯维加斯的四个（< -15°F），以及科罗拉多州丹佛的两个（< -20°F）。

预报变量仅限于最低气温、最高气温和降水概率。我们在天气预报中没有发现明显的异常值。这是合理的，因为预报不会受到实际测量时不可避免的传感器技术故障的影响。相反，预报数据中最低气温和降水概率存在大量重复值。对于每份预报，我们保留了最低气温的最低预报值和降水概率的最高预报值。

预报滞后期为六天或七天时，缺失值数量较多。我们删除了所有预报滞后期超过 5 天的预报。我们还删除了所有包含负滞后期的预报（即在实际观测之后 作出的预报）。

B 地理地图的极坐标注意事项

图4和图6中的字形图依赖于从极坐标到地理坐标或笛卡尔坐标的正确转换。这样就可以将图元直接绘制在底图上，而不必在图像中嵌入极坐标子图。避免使用子图能更精确地放置图元，也省去了创建和嵌入多个图形所带来的计算负担。这种直接绘图方法在绘制地理地图时需要特别注意，因为将地理坐标投影到笛卡尔平面时，极坐标图元会发生变形。例如，在北半球，地理坐标中的完美圆形投影后会呈现纵向拉长的形状。解决这一问题的一种方法，是在构建图元之前，先将所有地理坐标投影到笛卡尔平面上。使用 mapproj R 包中的 mapproject() 函数 [26] 即可轻松完成此操作。

极坐标以半径 r 和角度 θ来定义。图6将 $r \in [0,1]$ 定义为预测温度与实际温度之间经过缩放的平均绝对误差，并将 $\theta = {}_{(4-m)\ 6}{}^{\pi}$ 定义为其中 m 表示数字月份。我们将每个图元以笛卡尔坐标为基准，中心置于 0

$$(x, y) = (r\cos\theta, r\sin\theta)$$

令 $(\mathbf{x}_i,\ \mathbf{y}_i)$ 表示一组以原点为中心、构成与位置 i相关联图元的笛卡尔坐标。这些坐标使用的单位与底层地图投影相同。渲染后字形的最终坐标定义为

$$\alpha \cdot (\mathbf{x}_i + u_x, \mathbf{y}_i + u_y)$$

其中，$(u_x,\ u_y)$ 表示位置 i 的坐标，α 表示一个全局缩放参数，用于调整地图上渲染后图元的大小。在位置 $(u_x,\ u_y)$ 绘制一个点，作为图元的参考点。点 $(u_x,\ u_y)$ 周围的不对称性揭示了预报误差的季节性模式。

我们构造图4中的相关椭圆，其焦点 F_1，F_2 位于半长轴 $y = x\,(\theta = \frac{\pi}{4})$ 上（正相关时）和 $y = -x\,(\theta = -\frac{\pi}{4})$ 上（负相关时）。我们将 F_1 固定在原点，并将 r 定义为从 F_1 延伸至椭圆边缘的半径，如图 8 所示。Knisley 和 Shirley [40] 概述了这种椭圆构造方法，我们在此对其进行了改编，将 r 定义为 $\theta \in [0, 2\pi]$ 的

$$r = \frac{(1-|\rho|)^2}{1 - \sqrt{|p|(2-|p|)}\cos(\theta - \frac{\text{sign}(\rho)\pi}{4})},$$

其中，$\rho \in (-1,\ 1)\backslash\{0\}$ 表示预报误差之间期望的相关性。当 $\rho = -1$，1，或 0 时，我们在创建椭圆时使用 $\rho \pm \epsilon (\epsilon > 0)$，以避免数值精度误差。随后将椭圆转换为笛卡尔坐标，并以原点为中心，如下所示

$$\left(r\cos(\theta) - \frac{|\rho|(2-|\rho|)}{\sqrt{2}}, r\sin(\theta) - \text{sign}(\rho)\frac{|\rho|(2-|\rho|)}{\sqrt{2}}\right).$$

每个椭圆都会缩放至外切于 $[-0.5, 0.5] \times [-0.5, 0.5]$ 正方形。这样的缩放使我们能够使用共同网格尺寸创建椭圆矩阵，还能缩小椭圆之间的面积差异，从而便于比较形状。该缩放定义如下

$$(\mathbf{x}_i', \mathbf{y}_i') = \left(\frac{\mathbf{x}_i}{2 \cdot \max(|\mathbf{x}_i|)}, \frac{\mathbf{y}_i}{2 \cdot \max(|\mathbf{y}_i|)}\right).$$

请注意，每个位置都有三个椭圆。我们定义一个以格点的公共顶点为中心的椭圆矩阵，该顶点表示为（$u_x,\ u_y$）。令（$\mathbf{x}_i,\ \mathbf{y}_i$）表示以此位置为中心的三个椭圆之一的坐标。每个椭圆在地图上经过居中和缩放后，如下所示

$$\alpha \cdot (\mathbf{x}_i' + u_x + o_1, \mathbf{y}_i' + u_y + o_2)$$

其中，o_1 和 o_2 表示偏移量，用于分隔为每个位置定义的矩阵中的三个椭圆的中心。

这种直接绘制椭圆的绘图方法便于进行绘图定制，因为无需协调独立创建的子图之间的格式差异。这种方法还可推广用于在地理地图上绘制其他几何形状。对于需要根据动态输入快速渲染图像的交互式应用，也很有帮助。

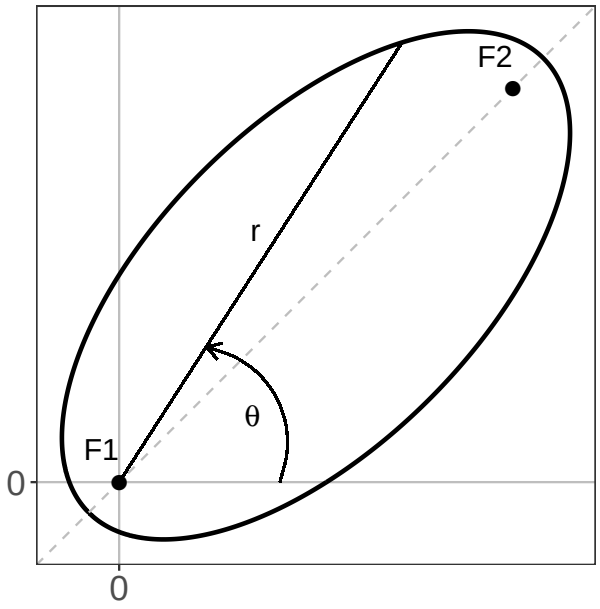


图8：以 F_1 为中心的椭圆示例

参考文献

[1] Jill Lundell、Brennan Bean 和 Juergen Symanzik。《来谈谈天气》。载于*JSM*会议论文集，统计计算分部，第1944–1958页。美国统计学会，2018年。[2] Antony Unwin 和 Pedro Valero-Mora。集成图形。《计算与图形统计学杂志》，27(1):157–165，2018年。[3]环境合作委员会。北美生态区域：迈向共同视角。http://www3.cec.org/islandora/en/item/1701-ecological-regions-north-america-toward-common-perspective-en.pdf，1997年。[4] Robert S Briggs、Robert G Lucas 和 Z Todd Taylor。建筑能耗规范与标准的气候分类：第2部分——分区定义、地图和比较。*ASHRAE*汇刊，109:122，2003年。[5] Trevor Hastie、Robert Tibshirani 和 Jerome Friedman。统计学习要素，第1卷。施普林格，纽约（纽约州），2001年。[6] Fionn Murtagh 和 Pierre Legendre。Ward层次凝聚聚类法：哪些算法实现了Ward准则？《分类学杂志》，31(3):274–295，2014年。[7] Thomas Karl 和 Walter James Koss。1895—1983年按面积加权的区域和全国月度、季节及年度气温。https://repository.library.noaa.gov/view/noaa/10238，1984年。[8] Henry F Diaz。美国干旱。《气候与应用气象学杂志》，22(1):3–16，1983年。[9] JE Hathaway、TC Pulsipher、Jeremiah Rounds 和 JA Dirks。利用NLDAS-2强迫数据定义气象站周围气候相似区域的统计方法。技术报告，PNNL-SA-98705。美国华盛顿州里奇兰：太平洋西北国家实验室，2013年。[10] Andreas P Weigel、Mark A Liniger 和 Christof Appenzeller。离散Brier评分和排序概率技巧评分。《月度天气评论》，135(1):118–124，2007年。[11] DJ Murdoch 和 ED Chow。大型相关矩阵的图形显示。《美国统计学家》，50(2):178–180，1996年。[12] Judah Cohen、Karl Pfeiffer 和 Jennifer A Francis。北极暖期与美国极端冬季天气发生频率上升相关。《自然·通讯》，9(1):869，2018年。[13] 大奥斯汀商会。内华达州奥斯汀：精彩纷呈。http://austinnevada.com/ visitor-information/chamber-information/，2018年。[14] E. R. Tufte。定量信息的视觉呈现。Graphics Press，康涅狄格州柴郡，第二版，2002年。

[15] Daniel B Carr、Anthony R Olsen 和 Denis White。用于展示单变量和双变量地理数据的六边形镶嵌地图。《制图学与地理信息系统》，19(4):228–236，1992。[16] Hadley Wickham、Heike Hofmann、Charlotte Wickham 和 Dianne Cook。用于直观探索气候数据及模型中时间模式的符号地图。《环境计量学》，23(5):382–393，2012。[17] Carl Nolte。旧金山的夏日故事有些雾蒙蒙。https://www.sfchronicle.com/bayarea/nativeson/article/Summer-fog-can-be-here-today-gone-tomorrow-9141028.php，2016 年八月。[18] Robert W Scott 和 Floyd A Huff。五大湖对区域气候条件的影响。《五大湖研究杂志》，22(4):845–863，1996。[19] Nate Silver 和 Reuben Fischer-Baum。哪个城市的天气最难预测？https://fivethirtyeight.com/features/which-city-has-the-most-unpredictable-weather/，2014 年十二月。[20] Leo Breiman。随机森林。《机器学习》，45(1):5–32，2001。[21] Mark Monmonier。地理刷选：增强对散点图矩阵的探索性分析。《地理分析》，21(1):81–84，1989。[22] R Core Team。R：用于统计计算的语言和环境。https://www.R-project.org/，2022。[23] Ram Narasimhan。weatherData：从网络获取天气数据。https://cran.R-project.org/package=weatherData，2017。R 软件包版本 0.5.0。[24] Douglas Nychka、Reinhard Furrer、John Paige 和 Stephan Sain。fields：空间数据工具。https: //cran.R-project.org/package=fields，2015。R 软件包版本 9.0。[25] Robert J. Hijmans。geosphere：球面三角学。https://cran.R-project.org/package=geosphere，2016。R 软件包版本 1.5-5。[26] Doug McIlroy、Ray Brownrigg、Thomas P Minka 和 Roger Bivand。mapproj：地图投影。https: //cran.R-project.org/package=mapproj，2017。R 软件包版本 1.2-5。[27] Roger Bivand、Tim Keitt 和 Barry Rowlingson。rgdal：用于“地理空间”数据抽象库的绑定。https://cran.R-project.org/package=rgdal，2018。R 软件包版本 1.2-18。[28] Roger S. Bivand、Edzer Pebesma 和 Virgilio Gomez-Rubio。《使用 R 进行空间数据分析》。http: //www.asdar-book.org/，2013。[29] Hadley Wickham。tidyverse：轻松安装并加载 tidyverse。https://cran.R-project.org/package=tidyverse，2017。R 软件包版本 1.2.1。[30] Thomas Lin Pedersen。ggforce：加速 ggplot2。https://cran.R-project.org/package=ggforce，2018。R 软件包版本 0.1.2。[31] Stefano Meschiari。latex2exp：在绘图中使用 LaTeX 表达式。https://cran.R-project.org/package=latex2exp，2015。R 软件包版本 0.4.0。[32]Erich Neuwirth。Rcolorbrewer：ColorBrewer 调色板。https://cran.R-project.org/package=RColorBrewer，2014。R 软件包版本 1.1-2。[33] Hadley Wickham。使用 reshape 包重塑数据。《统计软件杂志》，21(12):1–20，2007。http://www.jstatsoft.org/v21/i12/。[34] Andy Liaw 和 Matthew Wiener。使用 randomForest 进行分类与回归。《R News》，2(3):18–22，2002。https://cran.R-project.org/doc/Rnews/。[35] Google。入门。https://developers.google.com/maps/documentation/elevation/start，2018。[36] Scott Chamberlain。rgbif：全球“生物多样性”信息设施 API 的接口。https://cran. R-project.org/package=rgbif，2017。R 软件包版本 0.9.9。[37] 美国国家海洋和大气管理局。NOAA 中分辨率海岸线。https://shoreline. noaa.gov/data/datasheets/medres.html，2018。[38] 美国国家大地测量局。NOAA 海岸线数据浏览器。https://www.ngs.noaa.gov/CUSP/，2018。[39] 美国国家海洋和大气管理局。在线气候数据。https://www.ncdc.noaa.gov/ cdo-web/，2018。[40] Jeff Knisley 和 Kevin Shirley。在线多变量微积分。http://math.etsu.edu/multicalc/ prealpha/，2001。