气温传感器网络的建模与异常数据检测

Yujin Wang
Department of Electronic Engineering, Tsinghua University
Haidian District, Beijing
yujin-wa20@mails.tsinghua.edu.cn

Abstract

本文对所提供的的美国气温资料进行了统计与分析，并使用图网络来建模美国的气温特征，在此基础上建立了一个异常数据的检测系统。我们的检测系统具有较强的鲁棒性，可以实现最高。

***K*eywords** 图网络 · 异常检测 · 气温

1 Introduction

异常温度的检测可以看做一个辅以空间信息的时间序列检测问题，根据地理常识所得到的一些先验知识也可以加以判断。本文的思路是通过统计得到温度的统计分布特征，然后使用图建模温度模型，并使用空间上相近的点对错误数据进行挖掘。本文参考了 [1, 2, 3, 4]。

2 Preliminaries

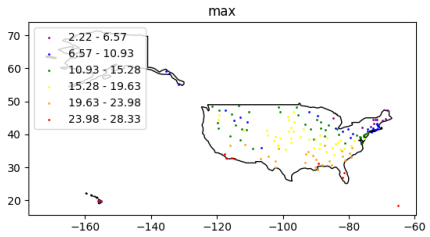
2.1 Region Domain Data Analyses

我们使用 `pandas` 对数据进行了简单的分析，并使用 `geopandas` 这一可视化工具对数据进行了可视化分析。我们分析了气温数据的最大值、最小值、平均值、标准差、最大温差值、3 月 1 日的气温和 3 月 31 日的气温，结果如图 1所示：

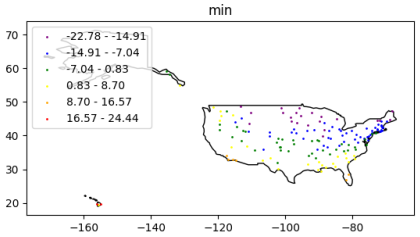
针对这些图像，我们可以得出一系列基本结论：

1. 温度基本上呈现出“北低南高”，“内陆低沿海高”的分布。
2. 内陆地区温度波动幅度大，沿海地区温度波动幅度小。

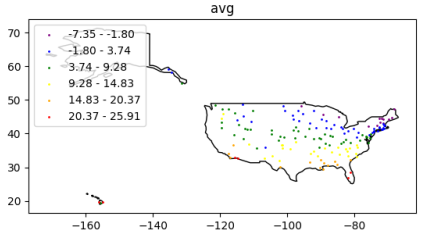
值得一提的是，我们在数据中也可以挑选出一些“异常值”，例如在夏威夷地区存在一个均温不超过 10 度的检测点，与附近的若干检测点统计规律明显不符，可能与海拔、地势等原因有关，我们在判断时需要对其进行特殊处理。



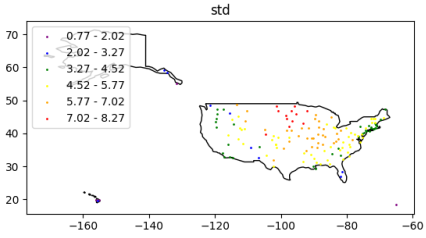
(a) 最大值



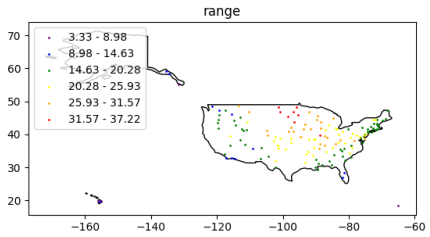
(b) 最小值



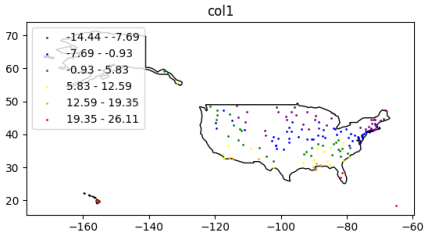
(c) 平均值



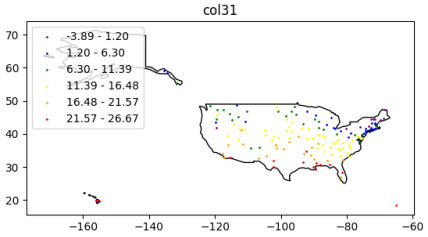
(d) 标准差



(e) 最大温差值



(f) 3 月 1 日



(g) 3 月 31 日

图 1: 数据可视化

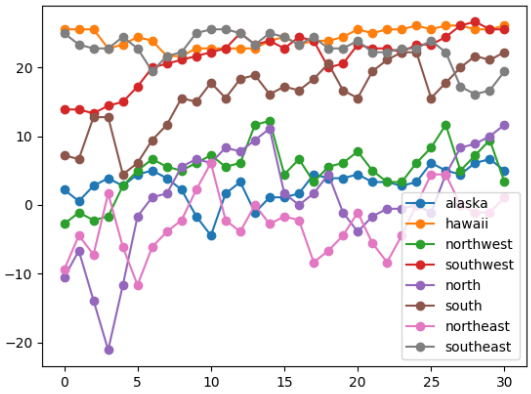


图 2: 3.1-3.31 气温变化示意图

2.2 Time Domain Data Analyses

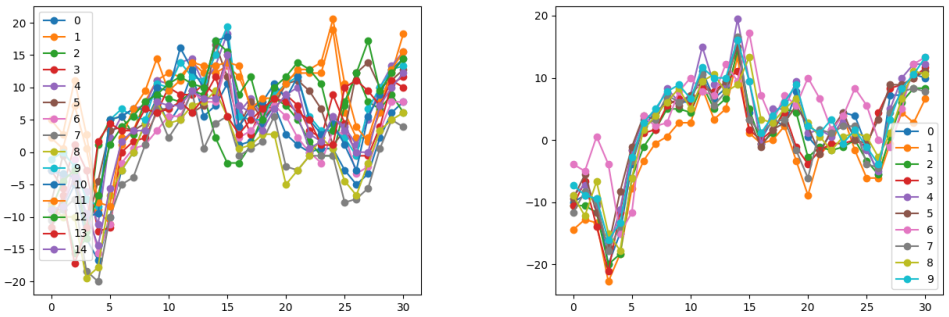
之后，我们选择若干个代表地区（阿拉斯加、夏威夷、西北、西南、中北、中南、东北、东南），记录其 3.1-3.31 之间温度的趋势，如图 2。

针对该图像，我们可以看出，夏威夷等地的气温变化较为平稳，异常值较小；而中北部、东北部地区的气温变化较为剧烈，且经常有异常值出现。

根据之前的温差分析可以得出，大约有 50% 的检测站在整个 3 月内的温差值在 20 度之内，大多分布于沿海地区，这些检测点的异常值较容易检出；至于剩下的 50% 温差值较大的点，可以看到它们大多分布于内陆地区，我们温差的来源主要有两种：

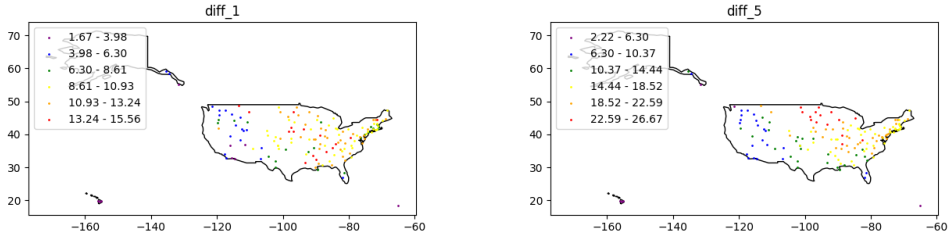
1. 随时间推移气温迅速上升。
2. 特殊天气影响下使得气温出现剧烈波动。

针对 1，我们猜想可以用一个函数模型对气温随时间的变化进行建模。我们抽取若干个月温差大于 27 度的点进行可视化，如图 3。



(a) 月温差在 27 ∼ 30 摄氏度之间的监测点气温变化　　(b) 月温差在 30 摄氏度以上的监测点气温变化

图 3: 温差值大于 27 摄氏度的检测站可视化



(a) 1 日差值　　(b) 5 日差值

图 4: 差值可视化

据此我们可以得出以下结论：

1. 温度和时间并非简单的单调递增关系，如果直接使用幂函数进行建模可能会导致较大的误差。
2. 相邻 2 天内的温差一般不超过 10 摄氏度。
3. 地域相邻的检测站，其温度变化规律有着一定的相似性。

针对猜想 2，我们引入差分算子 $y[n] = x[n+k] - x[n]$，观察在较小时间范围内温度的变化，并可视化其最值，如图 4所示，可见判断基本成立。

3 Method

3.1 Time Domain Detection

我们首先从时域信号中探测异常值而不依赖于空间信息。在此之前，我们需要引入若干先验知识：

1. 随时间推移气温整体上呈现出上升趋势。
2. 2 日内气温变化一般不超过 10 摄氏度，5 日内气温变化一般不超过 20 摄氏度。

对此，为了处理时序序列的异常，我们引入了 Hampel 滤波器，其工作原理如下：该过滤器是一个宽度为 k 的滑动窗口，对于每一个窗口，该滤波器计算其中值，并使用中值绝对偏差（MAD，$\text{MAD} = \text{median}(|X_i - \text{median}(X)|)$）来估计窗口的标准偏差 $\sigma = 1.4826\text{MAD}$。如果窗口中某点与其中值的距离大于 $n\sigma$，则该值为异常值。

而在我们的方法中，我们不仅需要找出所有可能的异常值，而且还需要评估它们 “哪一个是最可能的”。我们考虑以下场景：当一个气温变化通常较为缓和的地区和气温变化通常较为剧烈的地区均出现气温波动，且异常值区间的 MAD 大小相同时，我们有理由相信气温变化较为缓和的地区数据出差错的可能性更大。据此我们引入系数 γ 来判定区间内是否存在异常值：

$$\gamma = \frac{x - \text{median}(\vec{v})}{\sigma_{\text{w/o max min}}}$$

其中，$\vec{v}$ 为与 x 在同一个 Hampel 窗中的点，$\sigma_{\text{w/o max min}}$ 为一地的数据去除最大和最小值之后计算得到的标准差。可以看到，x 在局部时域中距离中值的差距越大，且数据本身的方差越小，γ 的值就越大，也就意味着该数据出差错的可能性越大。

这里需要特别说明为什么要使用“去除最大和最小值之后计算得到的标准差”，这是因为记录出错的温度有较大概率成为时序温度序列中的最大值和最小值（特别是对于气温变化平缓的地区）。去除最大值和最小值能让数据在一定程度上减轻错误值为统计特性带来的负面影响。

检测准确率的算法描述如下：我们首先计算每一条无差错温度序列的 γ 值，其中的最大值记为 $\gamma_{\max}$；之后引入人为误差后，计算该条温度序列对应位置的 γ 值，记为 γ_{error}。若 $\gamma_{\text{error}} > \gamma_{\max}$，则认为判定成功，否则判定失败。

当然，此种方法效果较差。具体分析见实验分析一节。

3.2 Region Domain Detection

我们引入空间信息来进一步提高鲁棒性，该做法基于这样的先验知识：两个检测站之间的距离越近，它们之间的气温越相近。我么可以用“图”来建立这种关系，各点之间的权重可以记为：

$$W_{ij} = \exp(-\frac{[\text{dist}(i,j)]^2}{2\theta^2}) \text{ if dist} \leq \kappa$$

其中 θ 为超参数，κ 是与实际问题情境有关的阈值。为了确定最佳阈值，我们需要确定不同地点之间气温的关联有多大。我们使用 MSE loss 来衡量它们之间的关联，以北部、南部、西北和东北为例 5：

我们基本可以断定，距离相近的点之间气温分布更相似。因此我们可以找出地图上与某点距离最近的 n 个点，用这些点的加权均值来给决策提供信息。点之间的距离通过以下公式给出：

$$d_{AB} = R \arccos(\cos(\alpha_A - \alpha_B)\cos\beta_A \cos\beta_B + \sin\beta_A \sin\beta_B)$$

随后我们对气温进行加权，得到一个参考值：

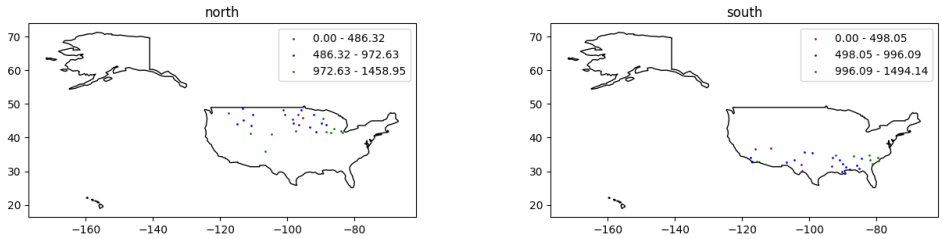
$$T_{ref} = \frac{\sum_{i=1}^{n} \exp(-\frac{[\text{dist}(i,j)]^2}{2\theta^2})T_i}{\sum_{i=1}^{n} \exp(-\frac{[\text{dist}(i,j)]^2}{2\theta^2})}$$

或：

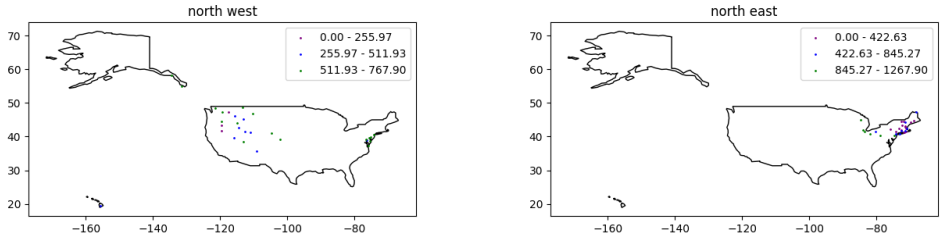
$$T_{ref} = \frac{1}{n}\sum_{i=1}^{n} T_i$$

得到参考点之后并计算出 T_{ref} 后，我们需要按照如下方式计算“偏差度” ξ：

$$\begin{aligned}
T_{diff} &= T_{ref} - T \\
T_{diff-norm} &= T_{diff} - \text{mean}(T_{diff}) \\
maxval &= \max(\text{abs}(T_{diff-norm})) \\
\xi &= maxval/\text{mean}(T_{diff-norm})
\end{aligned}$$



(a) 北部　　(b) 南部



(c) 西北　　(d) 东北

图 5: 温度分布相似度

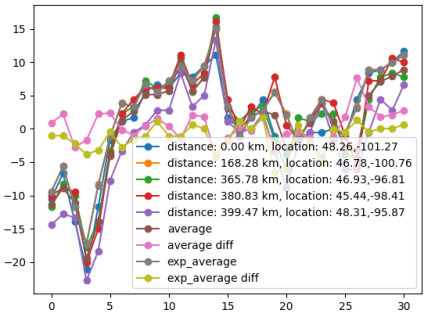


图 6: 温度重建及其误差

ξ 的值越大，说明该处出现错误数据的概率越大，我们将 ξ 的最大值处记为出错处。

我们通过一个简单的可视化实验来验证这一方法的有效性。如图所示，我们随机抽取北部（北部被认为是温度变化最为剧烈的地区）某地，并计算其 T_{ref}。我们使用两种方法（普通平均、指数平均）重建温度，结果如图 6所示。可以看出，重建之后的温度与实际温度的差距基本在 5 摄氏度之内，而且指数平均相对于普通平均效果更好，证明了这一思路的有效性。

这里有两种情况需要特别注意：1. 目标点发生差错，而参与重建的点全部正确；2. 目标点无差错，而参与重建的点可能有差错。

对于情况（1），我们将重建后的温度走势与原本的温度作比较，如果某点的温度差值最大且大于某阈值，则判定记录错误点，最终挑选差值最大的错误点作为最终错误点。

对于情况（2），我们力求我们的方法对分辨情况（1）和情况（2）具有一定的鲁棒性，因此我们对参与计算的参考点使用 Hampel 滤波器进行平滑处理，在不被可能存在的误差点影响的前提下进行重建。

4 Experiment

4.1 Time Domain Detection

实验结果记录如表 1：

表 1: Time Domain Detection

delta temperature	window length	abnormal threshold	acc
20	11	1σ	63.91%
20	9	1σ	**68.32%**
20	7	1σ	55.01%
10	11	1σ	10.17%
10	9	1σ	**11.27%**
10	7	1σ	6.96%

可以看出，仅从时域入手正确率并不算理想。对此，我们可以分析仅使用时域方法时，哪些地点的检测正确率较高，哪些地点的检测正确率较低，结果如图 7所示。

可以看出，沿海地区准确率较高，而内陆地区准确率较低，这与之前的猜想一致。原因在于，内陆地区的气温波动本身较大，时域方法无法较好地辨别人为误差和自然波动。

4.2 Region Domain Detection

我们使用指数加权重建温度，超参 $\theta = 100$，Hampel 滤波器的窗长 11，检测结果如表 2：

可以看到，region based 方法对温度误差较小时的准确率有较大的提升。

参考文献

[1] Kaggle——rain in australia (predict rain tomorrow in australia). https://blog.csdn.net/weixin_44441131/article/details/106505160.

[2] article2. https://saruagithub.github.io/2020/02/10/20191210%E5%BC%82%E5%B8%B8%E6%A3%80%E6%B5%8B%E6%96%B9%E6%B3%95/.

[3] article3. https://blog.csdn.net/xsdxs/article/details/71608157.

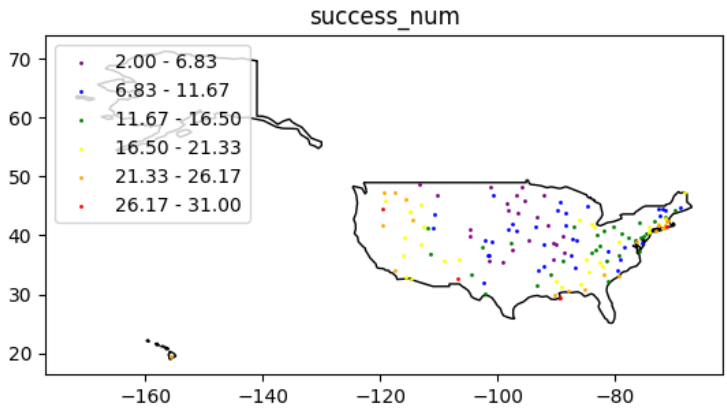


图 7: 成功次数

表 2: Region Domain Detection

delta temperature	ref point num	acc
20	2	69.05%
20	3	**70.75%**
20	4	69.96%
10	2	27.72%
10	3	**35.08%**
10	4	24.02%

[4] Outlier detection with hampel filter. `https://towardsdatascience.com/outlier-detection-with-hampel-filter-85ddf523c73d`.