気象観測データの外れ値検出方法の検討

1240147　松島一花

高知工科大学　システム工学群　建築・都市デザイン専攻

E-mail：240147b@ugs.kochi-tech.ac.jp

気象庁では様々な気象要素を観測しているが，度々異常値が記録される事例が発生している．社会気象工学研究室では佐岡地区にて気象要素を観測しているが，品質管理が行われておらず，また気象庁の観測とデータの取得条件が異なることから，同様の手法で品質管理を行うことが出来なかった．そこで，本研究では佐岡地区の観測データにおける新たな品質管理手法の検討を目的として，Local Outlier Factor や中心化移動平均を用いて月毎に外れ値検知を行い，異常値である可能性が高いデータを検出した．

Key Words： *気温，相対湿度，Local Outlier Factor，中心化移動平均*

1. 序論

(1)　背景

気象庁は，全国に配置されている地域気象観測システムAMeDASや地上気象観測所を用いて，気温や相対湿度，降水量や風速など，様々な気象要素を観測している．しかし，度々異常な値の観測データが記録されてしまう事例が発生している．異常値は，機器の故障や野焼き，植物の巻き付きや建造物の設置などが要因で発生している．気象観測データは多種多様な分野で使用されることから，観測データの品質管理を行う事が重要視されている．気象庁では，「急速に変化する気温に対する品質管理」ツールと「緩慢に変化する気温に対する品質管理」ツールを開発し，観測データから自然現象とそれ以外のノイズを分類する閾値を求め，観測データにおける異常値を検出することで，観測データの品質管理を行っている[1)]．

社会気象工学研究室では，高知県香美市土佐山田町佐岡地区にて気温や相対湿度などの気象要素を観測している．この観測データの信頼性を担保するため，気象庁と同様の手法で品質管理を行いたいが，データの取得条件が異なることから，気象庁の品質管理手法は適していない．

(2)　先行研究

2020年に社会気象工学研究室の有本が佐岡地区の観測データに対し，2つの手法を用いて品質管理手法の検討を行った[2)]．1つ目は，佐岡周辺の気象庁AMeDAS観測点(後免，高知)の観測データを使用して，佐岡地区の観測データとの差を求め，平均値と標準偏差(σ)を計算し，±3σの範囲外のデータを異常値とする手法である．2つ目は，佐岡地区の観測データをSavitzky-Golay(SG)法により平滑化したデータを使用して，佐岡地区の観測データとの差を求め，ホテリングのT^2法により自由度1のχ^2分布の99%点を異常値として検出する手法である．結果は，佐岡地区と後免や高知の周辺環境が大きく異なっていることや，ホテリングのT^2法の法則では99%点を異常値とすることから，異常値が多く検出されてすぎてしまうことが問題として挙げられ，適切に異常値検知が行われていない可能性が示唆された．そこで，これまでの手法とは異なる新たな異常値検知手法の検討が必要であると考えた．

(3) 目的

本研究は，佐岡地区における気象観測データの最適な品質管理手法の検討を目的とする．先行研究で挙げられた問題を踏まえ，本研究では近傍観測地点の観測データではなく，佐岡地区の観測データのみを用いて，観測データの変動として不可解なデータや局所的な変化点など，他の観測データと異なる変動を見せるデータを異常な可能性があるデータとして抽出する．本研究は異常値の検出ではなく，異常値である可能性が高い外れ値を検出することから，異常値検知ではなく外れ値検知と表記する．

2. 手法

(1) 使用するデータ・期間

佐岡地区には二カ所の観測地点があり，本研究では佐岡WS1を仮社殿，佐岡WS2を東屋と呼び，**図-1**に2地点の位置関係を示す．東屋と仮社殿では，気温と相対湿度，風や降水量などの気象要素を10分間隔で観測している．本研究では気温と相対湿度を対象に外れ値検知を行う．使用する観測データの期間は，2018年2月1日から2019年9月30日である．その中で欠損値が含まれる日時が複数存在する月は除外して，外れ値検知を行う．

図-1 「佐岡WS1：仮社殿」と「佐岡WS2：東屋」の位置関係[2), 4)]

(2) LOFによる外れ値の判別

外れ値検出手法は，局所外れ値因子法（Local Outlier Factor 通称：LOF）[3)]を用いる．LOFとは，ある任意の1つの点とｋ個の近傍点との距離から局所密度を推定し，自身と近傍点の局所密度の差が大きい点を外れ値とする手法である．膨大なデータがあり，かつ局所的なデータであれば，複雑な分布に対しても外れ値を検出することが出来ることから，時系列データなどにも利用されている．本研究で用いる気象観測データも時系列データの1つであるため，LOFを適用することが出来ると考えた．LOFの外れ値検知では，観測したデータを点で表す必要がある．本研究では，急激な気温変化を外れ値として検出することを目標としているため，観測データと観測データを平滑化したデータの２つのデータ値を点の座標とした散布図を作成する．そして，自身の近傍点の局所密度の比の平均を示すLOF値という値を求め，外れ値かどうかの判断を行う．LOF値は，自身の局所密度と近傍点の局所密度に差がないほど１に近づき，１より大きいほど他のデータから外れていることを示す．本研究では，PythonのScikit-learnを用いてLOFの外れ値検知を行うため，−1を基準として−1より小さくなるほど他のデータから外れているデータとなる．LOF値が定めた閾値よりも小さくなるデータを外れ値として検出する．

(3) 近傍点の個数kの決定

LOF値を計算するためには，近傍点の個数ｋを設定する必要があり，ｋの値を動かすことでそれぞれのデータ点の疎密の度合いが変わるため，適した値を見つける必要がある．本研究では以下の方法でｋの値を決める．

1) ｋの値を20から200まで10ずつ値を変えて，月毎に分けてLOFの外れ値検知を行う．
2) この中から最も多くの外れ値を検出したｋの値を上位3位まで割り出す．
3) 外れ値が1つ以上検出された月の数が最も多いｋの値を割り出す．
4) 割り出されたｋの値の中で，二つの条件に重なるものをｋの値として決定する．

(4) 中心化移動平均法による平滑化

外れ値検知を行うにあたって，急激に変動する観測データを抽出するために，中心化移動平均を用いて佐岡の観測データを平滑化したデータを作成する．中心化移動平均とは，ある任意の１つデータの直近

前後 n 個のデータを取って平均していく平滑化手法である．急激な上昇・下降による平滑化のズレを減らすため，時系列データにおいて付近のデータを参考にして平均を取る中心化移動平均を採用する．本研究では，前後3個ずつのデータを取って平均する７平均で観測データの平滑化を行い，元の観測データと比較する中で，局所的な変動を見せるデータを外れ値として検出する．外れ値検知はLOFと中心化移動平均を用いて，気温と相対湿度のそれぞれに対して適した方法を検討する．

(5)　気温に対する外れ値検知

10分間隔の気温データを中心化移動平均により平滑化したデータを作成し，観測データと平滑化データを用いた散布図から，月毎にLOFによる外れ値検知を行う．東屋，仮社殿の気温データに対して2章3項の方法を実施し，LOFの設定はk=70に決定した．外れ値とする閾値は，LOF値が－5.5以下のデータと－6.0以下のデータを外れ値として検出し，異常な可能性があるデータとして判別する．

(6)　相対湿度に対する外れ値検知

相対湿度は0%から100％の範囲があり，点の局所密度に偏りが生じてしまうことから，LOFを使用せずに外れ値検知を行う．

10分間隔の相対湿度データを中心化移動平均により平滑化したデータを作成し，観測データと平滑化データの差を計算して，東屋では差が±12%，仮社殿では差が±11%よりも大きいデータを外れ値として検出し，異常な可能性があるデータとして判別する．LOFの外れ値検知とは異なり，差から検出される外れ値検知は，観測データと平滑化したデータの差が大きいものが全て外れ値とされてしまうことから，同じような変動をしているデータが複数存在している場合を考慮することが出来ないため，外れ値の取り扱いに注意が必要である．

3. 結果・考察

(1)　気温に対する外れ値検知の結果

気温に対する外れ値検知では，局所的に変動する気温データを多く外れ値として検出した．**図-2**のような散布図上で見たとき，LOFの性質上，最高気温や最低気温付近のデータよりも，データが密集する中央から少し離れたデータの方がLOF値は高くなり，小さな変動でも気温帯によって外れ値として判別される傾向にあった．

東屋は**図-3**で示すように，LOF値が－5.5以下の条件では，気温が急速かつ急激に上昇することによって，平滑化データとズレが生じた箇所を外れ値として検出していた．これは，東屋が山の斜面付近で日陰の影響を受けやすい場所に設置されていることが要因であると推測される．日陰の時間が長く，午前9時から11時の間に日が当たりはじめ，気温が急速に上昇することにより，**図-3**のような結果が生じたと考えられるため，異常値である可能性は低い．そこでLOF値の条件を－6.0以下とすると，このようなデータは外れ値として検出されず，急速に変化する気温データのみを外れ値として検出することが出来た(**図-4**)．日当たりの良い仮社殿では，LOF値が－5.5以下の条件で，急速に変動する変化点のみを検出していたことから，観測地点の周辺環境によって，LOF値の大きさが異なることが分かった．

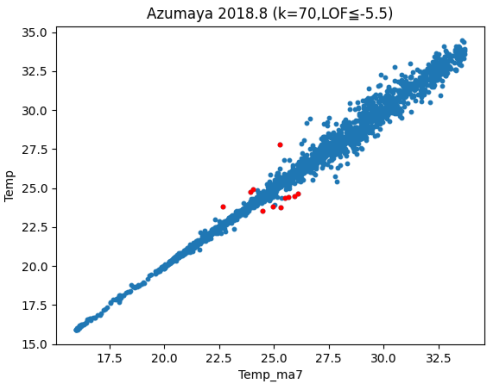

図-2 東屋2018年8月の気温データの散布図
（外れ値：赤，LOF≦－5.5）

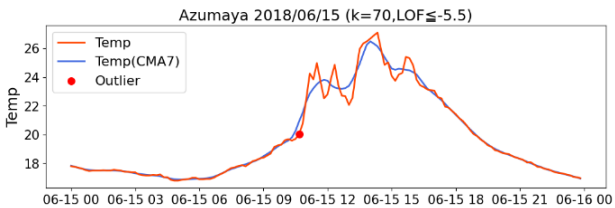

図-3 東屋2018年6月15日の気温データと平滑化した気温データのグラフ（LOF≦－5.5）

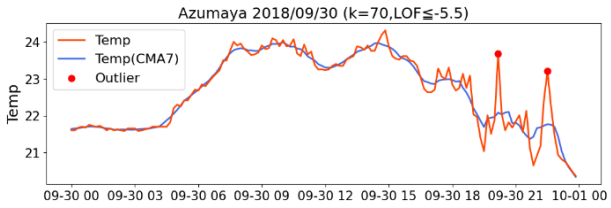

図-4 東屋2018年9月30日の気温データと平滑化した気温データのグラフ（LOF≦－6.0）

(2) 相対湿度に対する外れ値検知の結果

相対湿度の外れ値検知では，気温データよりもはるかに激しい変化点が外れ値として検出された．相対湿度は，気温の変動に合わせて大きく変動していたが，検出された外れ値の中には，気温はほとんど変動していないが，相対湿度は大きく変動しているデータも多く見られた(**図-5**，**図-6**)．そのため，気温の外れ値と相対湿度の外れ値は，異なる時間帯のデータが多く，関連性はあまり見られなかった．

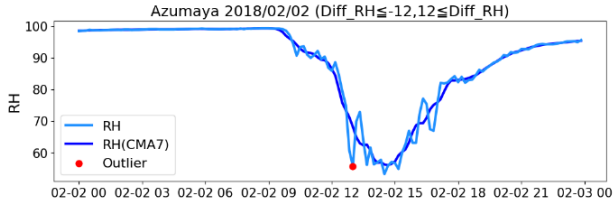

図-5 東屋2018年2月2日の相対湿度データと平滑化した相対湿度データのグラフ

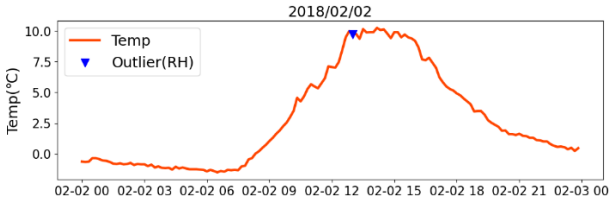

図-6 東屋2018年2月2日の気温データのグラフ

(3) 品質管理情報の作成

本研究で得られた結果を集計し，佐岡地区の気象観測データの品質管理情報を作成する．気温はLOF値が－5.5以下と－6.0以下，相対湿度は差が±12%と±11%の条件で，外れ値は1，条件を満たさないものは0として表に記した．表には，外れ値検知により外れ値として検出されたデータのみを記載し，日付，時間，気温，相対湿度，降水量，LOF値，外れ値判別の情報を表にまとめ，品質管理情報を作成した．

表-1 気温の品質管理情報

東屋			品質情報		
観測データ			外れ値判別		
日付	時間	気温(℃)	k=70	LOF≦-5.5	LOF≦-6
2018/3/7	11:00	14.218	-5.671214	1	0
2018/4/13	4:20	13.305	-5.705467	1	0
2018/5/4	17:00	18.366	-6.649399	1	1
2018/6/4	9:30	21.318	-6.815982	1	1
2018/6/15	10:40	20.031	-5.945448	1	0

4. まとめ

LOFと中心化移動平均を用いて東屋と仮社殿の外れ値検知を行った結果，急速に変化する気温，相対湿度を外れ値として検出した．先行研究で異常値とされたデータとの関連性は低く，また異常とされたデータの総数が大きく減少する結果となった．本研究では，急速に変動するデータに対する外れ値検知を行ったが，それ以外の変動を見せるデータに対しては外れ値検知が行われていないため，本研究では正常とされたデータの中に異常値が含まれている場合があり，取り扱いに注意が必要である．また，外れ値に対しても，外れ値として抽出された要因を明確にすることはできていない．そのため，本研究で検出された外れ値は異常値として断定することはできず，異常な可能性があるデータとして品質管理情報にまとめた．

5. 参考文献

1) 観測部計画課情報管理室：観測データの品質管理強化と高度化の現状 ―気温に関する自動品質管理機能，対話的品質管理ツール，観測所運用記録情報の業務への利用－，測候時報第82巻，2015

2) 有本陸矢：気象観測データの品質情報作成手法の検討 2020年度高知工科大学システム工学群卒業研究，2021

3) 井手剛：入門機械学習による異常検知: Rによる実践ガイド，2015

4) Google Map ： https://www.google.co.jp/maps/?hl=ja