GAP法：外れ値除外の新手法と野生動物GPSデータへの適用

GAP method: a new technique for removing data outliers and its application to wildlife GPS data

平川 浩文
（無所属）

村松 大輔
（京大WRC／JST/JICA, SATREPS／奈教自然セ）

Marcelo Gordo
（UFAM／Brazil）

瀧井 暁子・泉山 茂之
（信州大学 山岳科学研究拠点）

2022年3月14-19日　日本生態学会69回大会

要旨

　データセット内の外れ値は分析者にとってノイズあるいは信号として気になる存在である。外れ値を検出するには、各データ点がデータ空間内のどこに位置しているかを数値化し、外れ値とそれ以外を分けるための閾値を設定する必要がある。しかし、後者はこれまで実質的に分析者の主観に委ねられてきた。演者らは、野生動物GPSデータから外れ値を除外するため、この技術的な空白を埋める手法を開発した。野生動物GPSデータは測定に複数の要因が関与し、それぞれが状況により複雑に変化するため、測定精度が不安定で非常に広い範囲の外れ値を含むことが少なくない。こうした外れ値の閾値を探すため、この手法はデータが値の昇順に並べられたデータセットにおける隣接データ間の値のギャップを利用する。核となるアイデアは、データセット上位集団の中で最大のギャップを見つけ、そのギャップから上を外れ値とみなして除外することであり、この除外を除外後のデータセットでも繰り返す。この反復をいつやめるかは分析者に委ねられる。しかし、本手法によって分析者は意味ある少数の候補を得ることができる。時系列データや多次元データの場合でも問題となる特異性を数値化して単変量のデータセットを作成すれば、この手法を利用できる。手法の適用例をブラジル、マナウスの熱帯林に生息するナマケモノと日本の中部山岳地帯に生息するシカのGPSデータで示す。この外れ値検出の手法は、野生動物GPSに限らず、幅広い分野のデータで応用可能である。

はじめに

外れ値検出

二つの過程、**数値化**と**二分化**、が必要。

従来研究はほぼすべて**数値化**に関連。

二分化（閾値設定）は分析者の**主観任せ**。

野生動物GPSデータの特徴

測定精度に影響する要因が多数。

それぞれが状況により複雑に変化。

その結果、精度が安定せず、広範囲の外れ値を含む。

外れ値検出の**二分化**（閾値設定）手法 GAP法 を開発

GAPの定義とGAP法の原理

GAP法はギャップ度**GAP**を外れ値の閾値設定に利用する。
GAPの定義は次の通り。

正値のデータが昇順に並んだデータセットにおいて、
隣接データの値をA, Bとし、 A ≤ B のとき、
GAP = B / A　（定義により **GAP** ≥ 1）。

GAP法の原理は、データ上位集団で**最大GAP**を持つ組を見つけ、その間を境に上位のデータを外れ値として除外することであり、これを除外後のデータセットでも**最大GAP**が一定レベルを下回るまで繰り返すこと。

GAPの具体例

なぜこの原理が働くのか、**GAP**の具体例で検討したい。

10個の正数からなる、次のデータセットを仮定しよう。

(1, 2, 3, 4, 5, 6, 7, 8, 9, 10)

このデータセットは次の9個の**GAP**を持つ。

(2.00, 1.50, 1.33, 1.25, 1.20, 1.17, 1.14, 1.13, 1.11)

このとき、**最大GAP**は2.00でデータ1と2の間にある。

この例に限らず、正の等差数列における**最大GAP**はつねに、

最小とその次の値の間にある。

一方、等比数列の**GAP**は定義によりすべて等しい。

GAPの性格

一般に、広い範囲の正の実数が昇順に並ぶデータセットにおいて
下端近くでは、僅かな値の差でも大きな**GAP**をとりうる。
しかし、上端近くで大きな**GAP**をとるには、
値の差が非常に大きくなければならない。

GAPのこの性格から、次の３点が示唆される。
１）**GAP**は外れ値の閾値を探る道具として、データ上部でのみ有効。２）データ最上位部で大きな**GAP**を期待できるのは、そこのデータ分布が疎なときだけで、まさに外れ値がある状況。データ範囲が下がりデータ分布が密になってくると、閾値を示す道具として**GAP**の有効性は低下。３）**GAP**は正値の比として定義されるので、比例尺度のデータセットで有効。

GAP法の基本手続き

GAP法の基本手続きは次の通り。

1）データセットのデータを昇順に並べる。
2）値が上位のデータ集団の中で**最大GAP**を持つ組を見つけ、その間を境に上位のデータをすべて外れ値とする。
3）データセットから外れ値を除外する。
4）除外後のデータセットを用いて、２）に戻る。

この反復をいつやめるかは分析者に委ねられるが、反復で急速に低下する**最大GAP**が良い指標。経験的にはこれが**1.10**を下回ると継続可否の判断が必要。実際の運用で**最大GAP**を探す上位集団の割合をデータサイズの**5%**とした。除外率はこれよりはるかに低かったので、結果的にこの範囲で十分と判断された。

時系列や多次元データの場合

基本手続きは単変量で順番に意味がないデータセットでのみ有効。

時系列データセットでは、個々のデータの値ではなく、前後のデータの値と関係の特異性が問題となったり、多変量データセットでは、個々の変量の値ではなく、多変量の合成値の特異性が問題となったりする。

この場合、GAP法の原理を適用の前に、元のデータセットからこうした特異性を数値化して新たなデータセットを作成する。これを「特異度データセット」と呼ぶ。

特異度算出の方法は一意には定まらない。
問題の性格に合わせて個別に考案する必要がある。

GAP法の一般化手続き

一般化手続きは次のとおり（色文字が基本手続きからの変更点）。

0）元データセットから特異度データセットを作成する。
1）特異度データセットのデータを昇順に並べる。
2）値が上位のデータ集団の中で**最大GAP**を持つ組を見つけ、
　その間を境に上位のデータをすべて外れ値とする。
3）元データセットからこれに対応するデータを除外する。
4）除外後の元データセットを用いて、0）に戻る。

基本手続きと異なるのは、0）が加わり、さらに0–1）が反復に加わったこと。0）で特異度データセット作成が毎回必要になるのは、3）で元データセットが間引かれてデータの構造・相互関係が変化するため。3）では、2）で示された特異度データセット内の外れ値を元データセット内のデータ除外のために利用。元データセットと特異度データセットが1対1対応する場合、3）は上記のように単純。1対1対応しない場合、一般化が困難なため適用例2で具体例を提示する。

野生動物GPSデータへの適用

適用例１
ニホンジカの標高データ
目的：突出標高の除外
場所：北アルプス
装置：首輪型，測位間隔：2時間

GAP法
個別の標高値：基本手続き
標高値の変化：一般化手続き

メモ：
１）適用例では、標高値の変化による除外だけでもすべて除外可。ただ、これでは異常標高の連続に対処不能なため、個別標高値による除外を用意。そのGAP度下限値が高い(1.5)のはそのため。
２）標高の変化では下方突出も除外可。

適用例２
ミユビナマケモノの経緯度データ
目的：孤立測位点（群）の除外
場所：ブラジル・マナウス
装置：背負子型，測位間隔：15分

GAP法
元と特異度のデータセットが１対１対応しない場合の一般化手続き

メモ：
１）GAP法で直接扱うのはXY座標で経緯度データは毎回XY座標化。
２）座標化は正射図法による。
３）反復ごとに空間スケールが急激に縮小することに注目。
４）GAP法以外による除外も次に用意。

適用例１：突出標高除外

２つの手続きによる段階的除外

１）基本手続きにより個別標高値（元データセットをそのまま使用）

２）一般化手続きにより標高変化（特異度３種類を定義。下記参照）

元データセット（標高の時系列）

$(H_1, H_2, H_3, \ldots, H_{i-2}, H_{i-1}, H_i, H_{i+1}, H_{i+2}, \ldots)$

特異度データセット（標高変化の時系列）

$(S_1, S_2, S_3, \ldots, S_{i-2}, S_{i-1}, S_i, S_{i+1}, S_{i+2}, \ldots)$

特異度定義（分析者が選択：どれでも両データセットは１対１対応）

a）$S_i = (H_i - H_{i-1}) \times (H_i - H_{i+1})$

b）$S_i = (2H_i - H_{i-1} - H_{i-2}) \times (2H_i - H_{i+1} - H_{i+2})$

c）S_i ＝次の３つのうちの最大値　$\{(H_i - H_{i-2}) \times (H_i - H_{i+1}),$

$(H_i - H_{i-1}) \times (H_i - H_{i+1}), (H_i - H_{i-1}) \times (H_i - H_{i+2})\}$

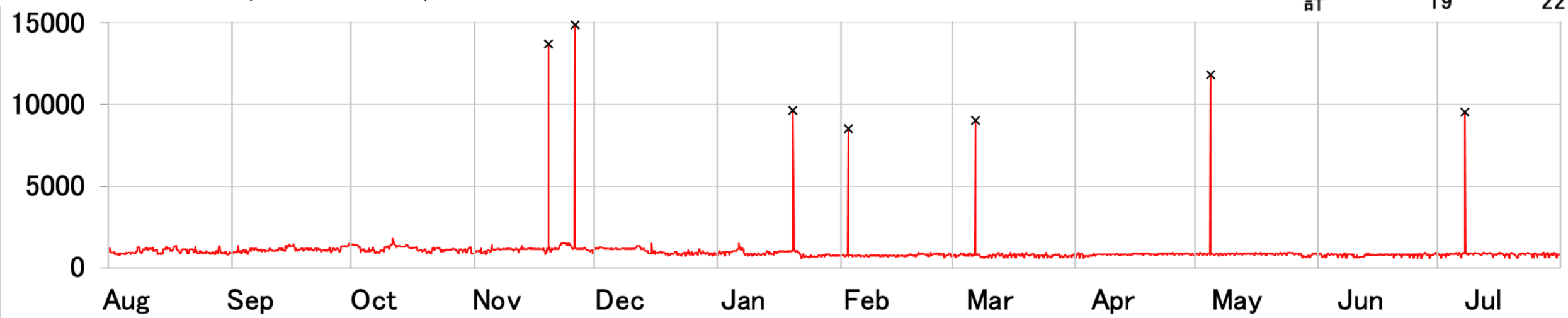
適用例１：突出標高除外

ニホンジカ；期間365日(2012年８月–2014年７月)；当初データ数4254

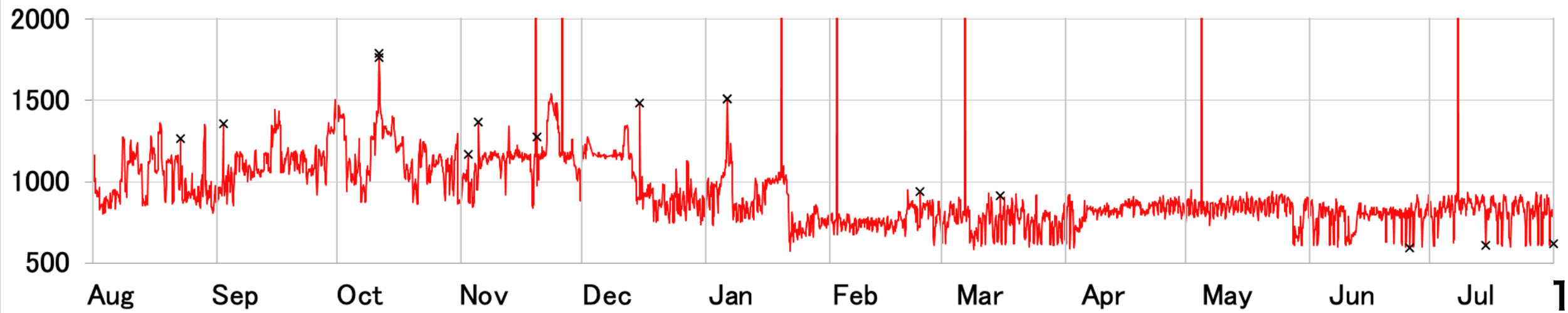
2段階で除外	最大GAP 設定下限値	結果 反復数	最終 最大GAP	除外数
個別値：	G=1.5	１回	G=4.765	7
変化値：	G=1.1	４回	G=1.159	15

標高突出の除外				除外候補		除外	
定義	反復	最大GAP	しきい値	上頂点数	上下計	上頂点数	上下計
個別標高値	段階 1	4.765	/ 1788	7	7	7	7
	停止	1.098	/ 1609	2	2		
標高値変化	段階 1	3.178	/ 442133	1	1	1	1
	段階 2	1.484	/ 297920	1	1	1	1
	段階 3	1.411	/ 211214	1	1	1	1
	段階 4	1.159	/ 109984	9	12	9	12
	停止	1.068	/ 42261	16	46		
					計	19	22

元データセット(個別標高値)による除外



特異度データセット(標高値の変化)による除外　数値化：前後との標高差の積；両セットが１対１対応



適用例２：孤立点(群)の除外

各測位点の経緯度から測位点間ユークリッド距離からなる特異度データセットを作成：二つのデータセットは１対１対応せず

元データセット（各測位点の経緯度）

$((Lt_1,Ln_1), (Lt_2,Ln_2), (Lt_3,Ln_3), (Lt_4,Ln_4), (Lt_5,Ln_5), (Lt_6,Ln_6), ...)$

中間データセット（各測位点のXY座標）

$((X_1,Y_1), (X_2,Y_2), (X_3,Y_3), (X_4,Y_4), (X_5,Y_5), (X_6,Y_6), ...)$

特異度データセット（連続測位点間のユークリッド距離）

$(D_{12}, D_{23}, D_{34}, D_{45}, D_{56}, ...)$

一般化手続き３）の具体例

3-1）元データセットを外れ値の距離があるところで分割

3-2）分割群中、単独あるいは少数点群を抽出して除外候補に

3-3）分割群を色分けし、除外候補がわかるように地図に表示

3-4）地図空間上で孤立が明確な除外候補を分析者が選んで除外

適用例2

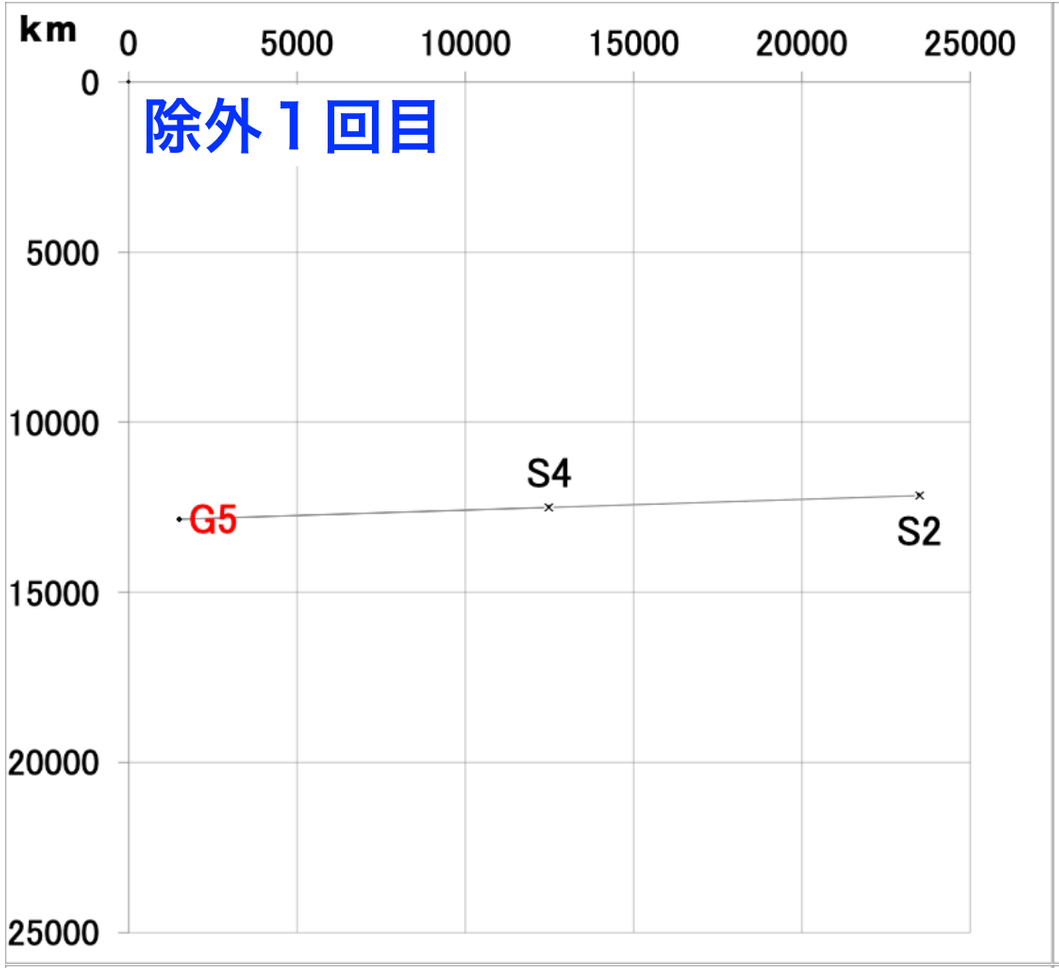
孤立点群の除外

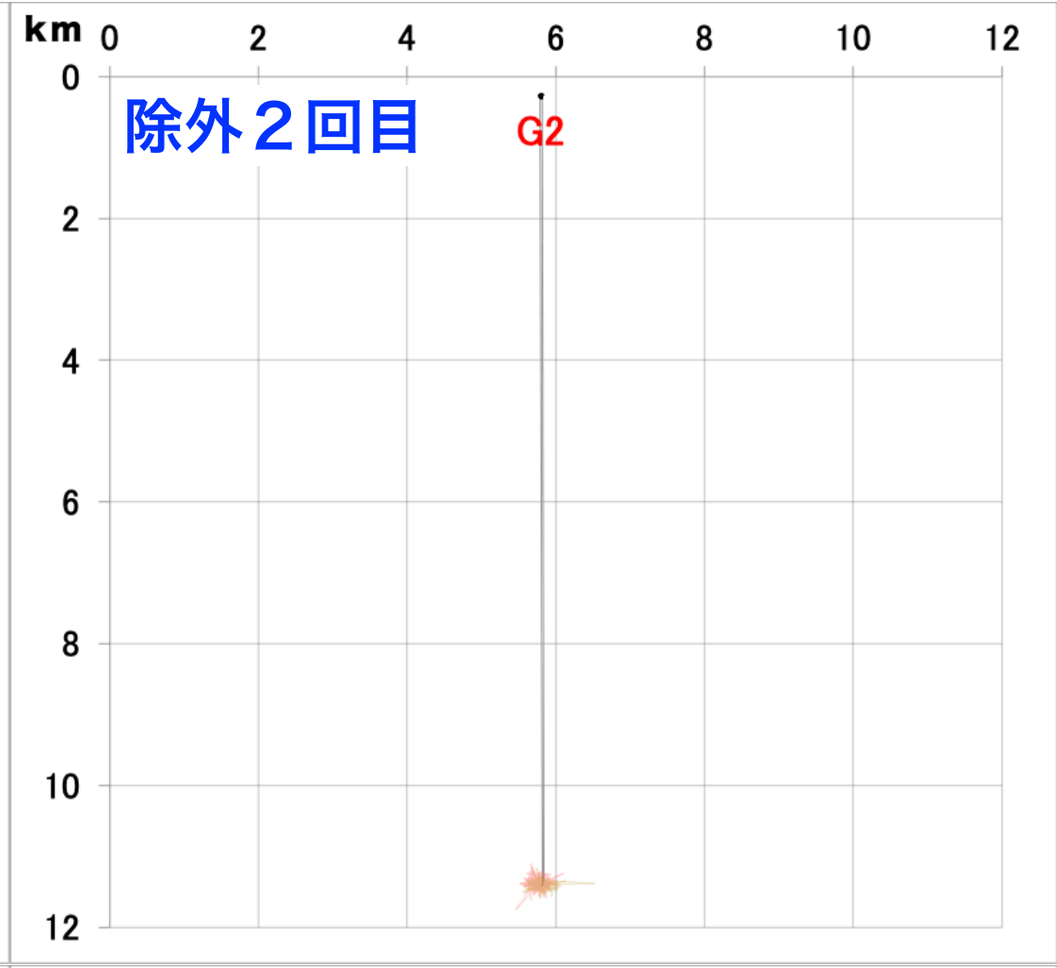
ナマケモノ

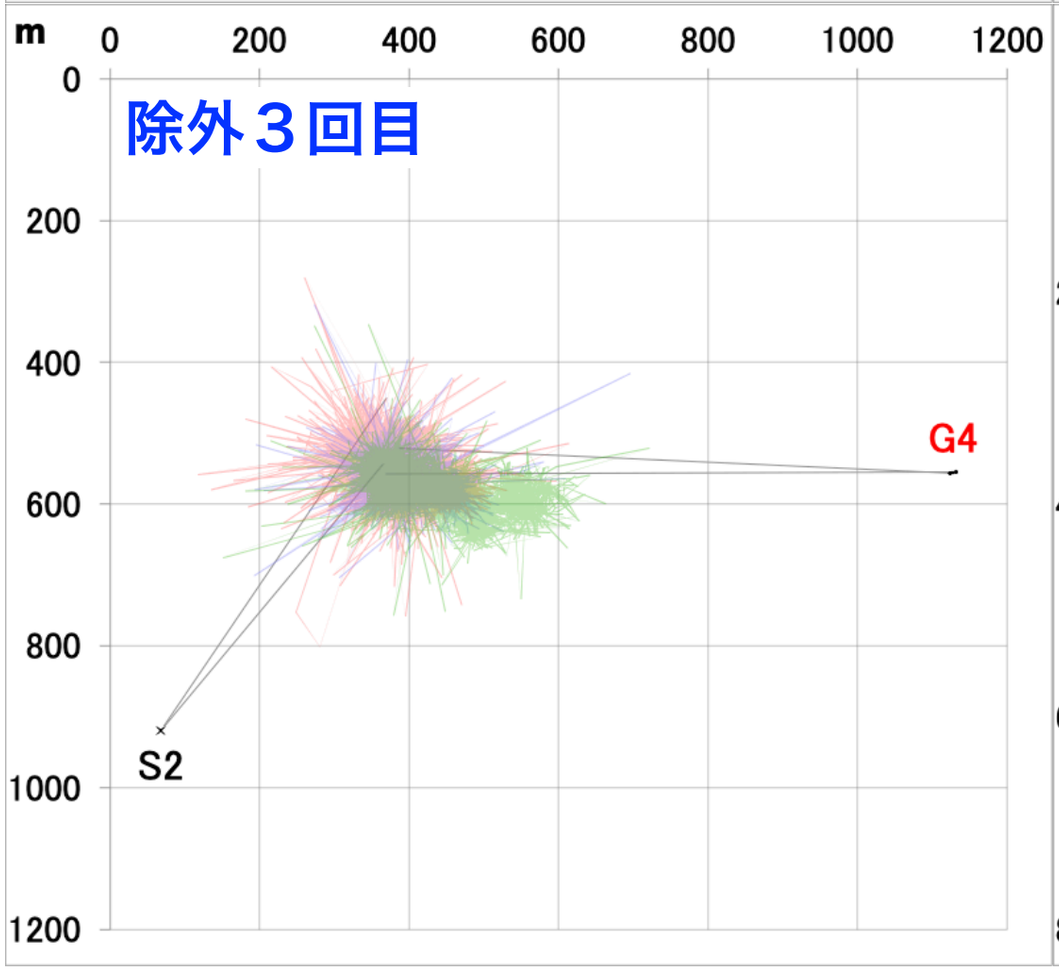
期間243日
2019年10月
–2020年6月

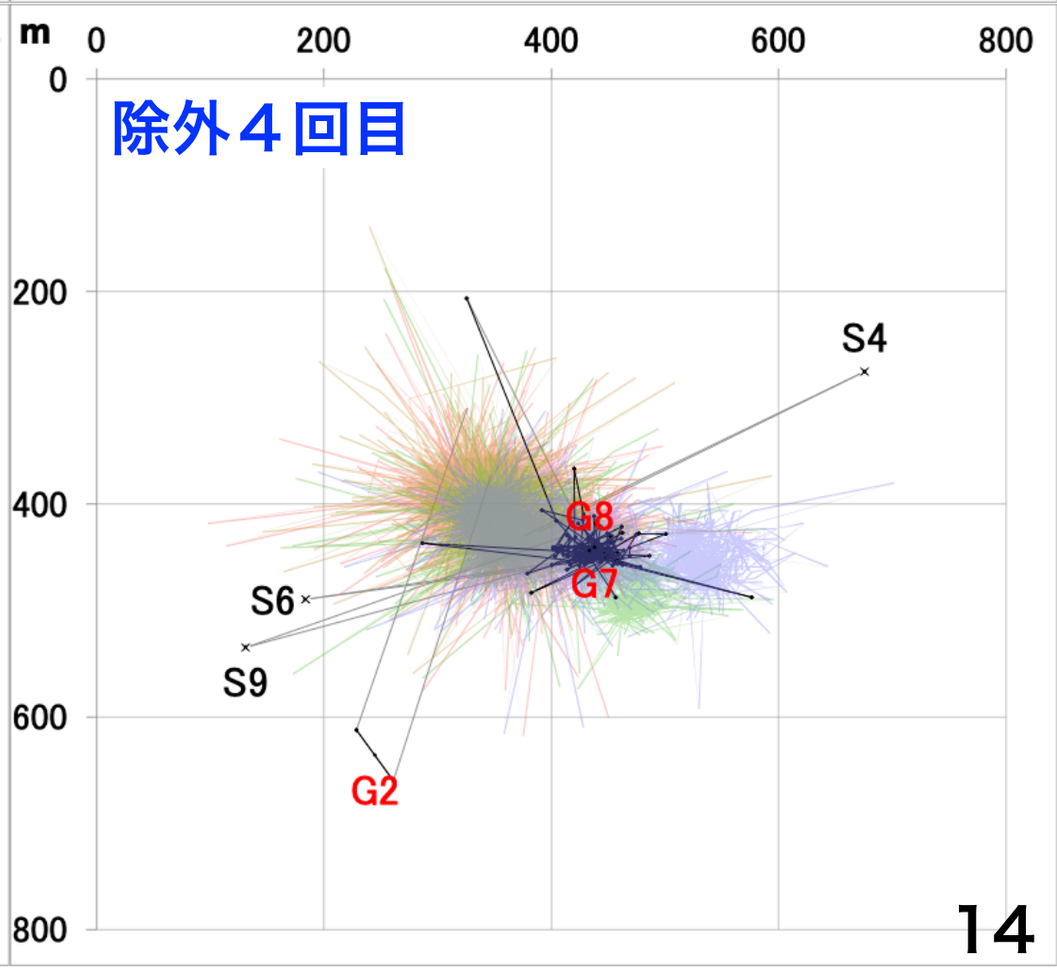
当初データ数
22618

次ページに
各段階の
詳細データ







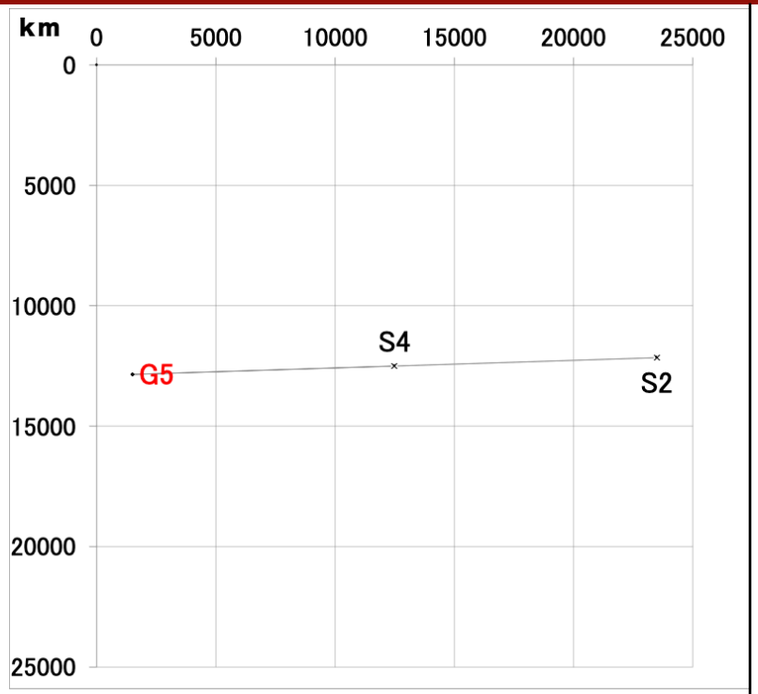


適用例２：孤立点(群)の除外

除外１回目

最大GAP: 979
閾値: /11243
分割群数: 5
除外群候補数: 3

グループ	サイズ	除外候補
G1	16620	
S2	1	1
G3	5993	
S4	1	1
G5	3	3
計	22618	5

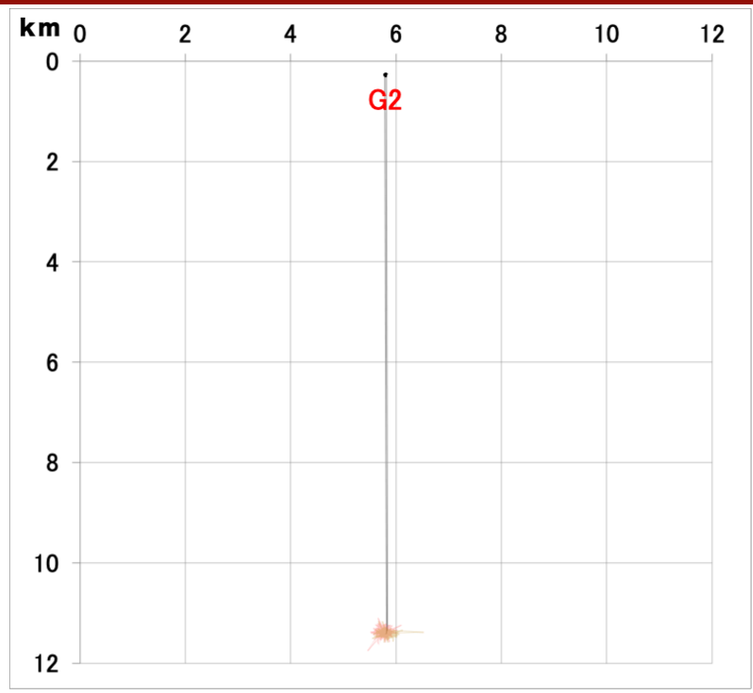


S2,S4を除外、G5は除外せず

除外２回目

最大GAP: 14.5
閾値: /763
分割群数: 3
除外群候補数: 1

グループ	サイズ	除外候補
G1	16286	
G2	4	4
G3	6326	
計	22616	4

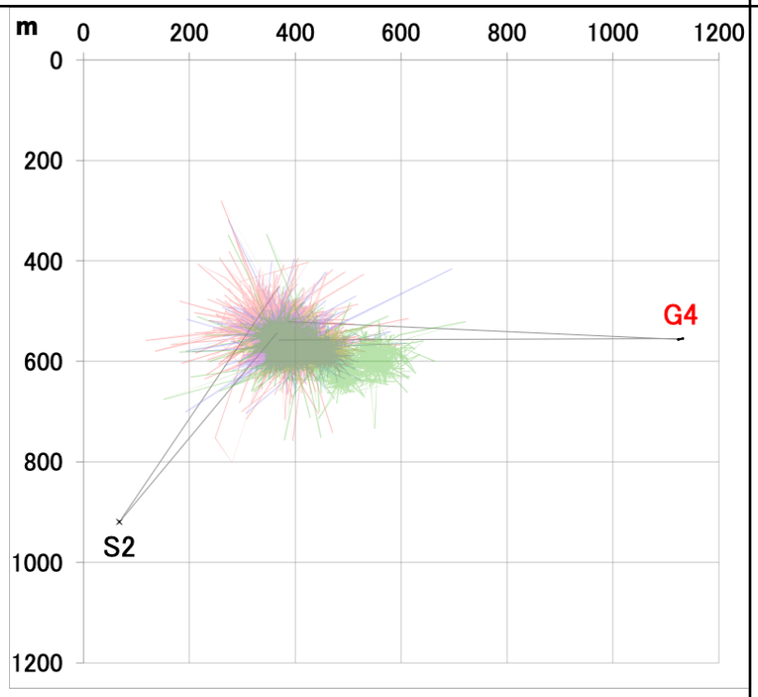


G2を除外

除外３回目

最大GAP: 1.42
閾値: /337
分割群数: 5
除外群候補数: 2

グループ	サイズ	除外候補
G1	11533	
S2	1	1
G3	6214	
G4	2	2
G5	4862	
計	22612	3

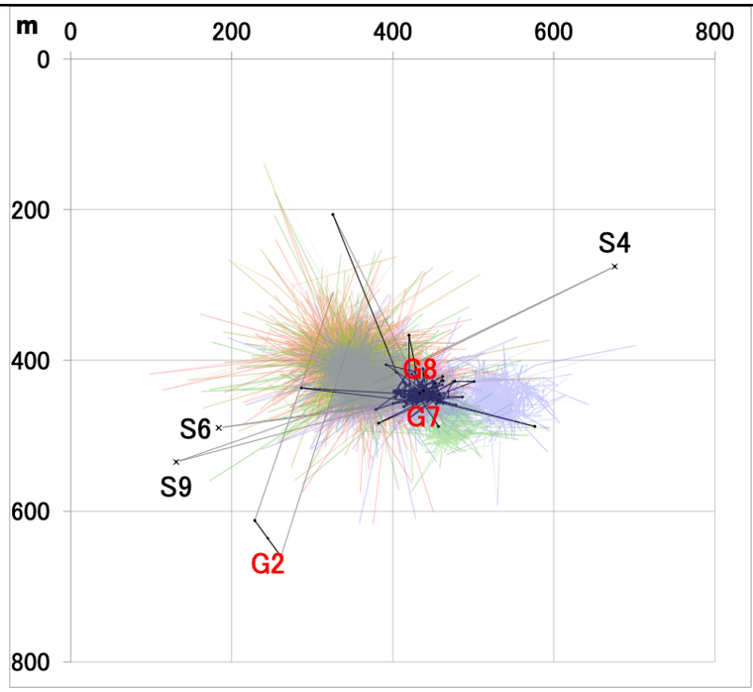


S2,G4を除外

除外４回目

最大GAP: 1.09
閾値: /265
分割群数: 10
除外群候補数: 6

グループ	サイズ	除外候補
G1	8148	
G2	2	2
G3	4465	
S4	1	1
G5	5937	
S6	1	1
G7	78	78
G8	69	69
S9	1	1
G10	3907	
計	22609	152



G2,S4,S6,S9を除外
G7,G8は除外せず

GAP法提案の意味

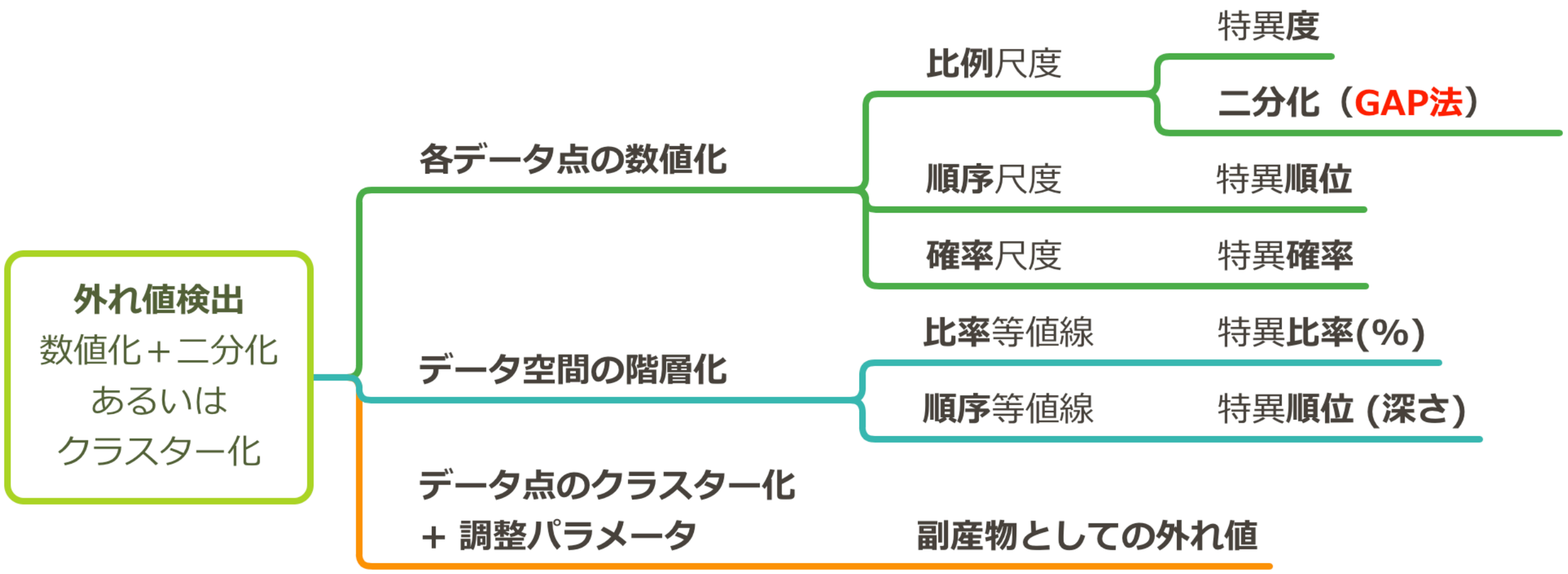
外れ値検出の分野（統計学・データサイエンス）で
学習を要しない二分化手法として初。

原理（値の大きいギャップで外れ値を区分）は直感的。
これまでも分析者が無意識に利用してきた可能性あり。
これを初めて手法化。

最大の特長：
主観を抑えて少数の意味ある閾値候補を得られること。

最終判断は分析者に委ねる。ただし、**最大GAP1.10**が
目安。これを切ると、手続きの反復停止の考慮が必要。

外れ値検出におけるGAP法



数値化に基づく外れ値検出の手法分類と
GAP法の位置づけ

除外の包括性・階層性

手続きに反復性があり、広範囲の外れ値を除外可。装置やシステムのエラーに伴う異常値の認知は容易だが、その他の外れ値も併せて包括的に除外可能。

除外が段階的に行われるため、外れ値の階層化が可能。外れ値の原因解明などに有用。

汎用性と簡易性

GAP法は原理として、順番に意味がない単変量の比例尺度データで有効（基本手続き）。

しかし、時系列データでも多変量データでも、問題とする特異性を比例尺度で数値化すれば適用可能（一般化手続き）。
そうした数値化手法はすでに多数存在する。
ただ、GAP法を前提にすると手間が軽減される。

例えば、GAP法は比例を使うため、標準偏差や四分位数間領域といった測定の単位には意味がなく、考慮不要。
また、測定の原点は、データ分布の平均、中央値、最頻値ほか使いやすい点で良い。GAP法はこの選択に堅固robustだから。

外れ値とは何か？

外れ値検出の研究：
長年、データの特異性の**数値化**（特異度算出）に腐心。

次の段階、**二分化**（閾値設定）で想定されている問いかけ。
「値がどれだけ大きければ、外れ値とみなすべきか？」

これに代わるGAP法の問いかけ。
「値のギャップがどれだけ大きければ、外れ値を区分すべきか？」
問いの根本的シフト。この方が人間の外れ値認識に近い。

GAP法：
外れ値とは何か、外れ値検出のために外れ値の影響をどう抑えるか
といった、積年の混迷した議論に対する一つの解。