



HOKKAIDO UNIVERSITY

Title	高度集約的シンボリックデータに対する解析手法に関する研究
Author(s)	松井, 佑介
Degree Grantor	北海道大学
Degree Name	博士(情報科学)
Dissertation Number	甲第11605号
Issue Date	2014-12-25
DOI	https://doi.org/10.14943/doctoral.k11605
Doc URL	https://hdl.handle.net/2115/57734
Type	doctoral thesis
File Information	Yusuke_Matsui.pdf



高度集約的シンボリックデータに対する
解析手法に関する研究

A Study on Analysis of
Advanced Aggregated Symbolic Data

北海道大学大学院情報科学研究科
複合情報学専攻 大規模情報システム学講座
松井 佑介

目次

第 1 章	はじめに	4
第 2 章	準備	7
2.1	シンボリックデータ解析について	7
2.2	基礎集約的シンボリックデータ	14
2.2.1	多値データ	14
2.2.2	区間値データ	15
2.2.3	モダル値データ	15
2.3	高度集約的シンボリックデータについて	16
2.3.1	分布値データ	17
2.3.2	関数値データ	17
2.3.3	尺度混在型データ	18
2.4	データセットについて	21
2.4.1	通信遅延時間データ	21
2.4.2	大規模環境データ	25
第 3 章	分布値データにおける外れ値検出法	38
3.1	はじめに	38
3.2	階層型シンボリッククラスタリングについて	39
3.3	提案手法	40
3.4	適用例	42
3.5	数値実験	43

3.6	おわりに	44
第 4 章	移動関数 k-means 法	51
4.1	はじめに	51
4.2	関数データに対するクラスタリング	52
4.3	提案手法	52
4.4	数値実験	54
4.5	適用例	57
4.6	おわりに	58
第 5 章	多次元関数値における L_0 正則化を用いた異常領域の同定法	66
5.1	はじめに	66
5.2	多次元値関数データと異常領域	67
5.2.1	関数データと多次元値関数データ	67
5.2.2	外れ値と異常領域	68
5.3	提案手法	68
5.3.1	準備	69
5.3.2	データの中心化とスケーリング	70
5.3.3	L_0 正則化による平滑化	70
5.3.4	閾値に基づく領域の分割	72
5.3.5	同時異常領域の検出	72
5.4	数値実験	73
5.5	実際の適用例	79
5.6	おわりに	83
第 6 章	尺度混在型データに対する解析手法	87
6.1	はじめに	87
6.1.1	ビッグデータと SDA	88

6.2	尺度混在型データの解析法	89
6.2.1	非類似度について	90
6.3	福島県における線量率データへの適用	91
6.4	おわりに	96
第7章 おわりに		99
参考文献		100

第1章 はじめに

情報技術の著しい発達に伴い、さまざまな分野における収集方法の多様化から、莫大なデータの蓄積が可能となっている。例えば、インターネットを流れる通信に関する膨大なログデータや、センサーネットワークから高頻度で得られるセンシングデータなどが代表的である。このようなデータは、容量、更新頻度、多様性などを主な特徴とするビッグデータとして、注目を集めている。しかし、統計科学において代表的な解析法をビッグデータへそのまま適用することを想定すると、時としてこれらの特徴が弊害となる可能性が示唆され、新たなアプローチが必要である。そこで、本論文では、これまでの統計解析において扱われることが少なかった、多様性に着目した新たなデータ表現とその解析法について提案する。データをどのような形で表現するかは、解析を行うために重要な要素の一つである。従来の多変量解析では、各解析対象を実数値多次元ベクトルによって表す記述が用いられてきたが、大規模データに対しては、そのまま用いることが適切ではない場合も多い。Diday が提唱したシンボリックデータ解析法 (SDA) において、解析対象とそれらを表す記述を一体として、内部変動を含むオブジェクトとして捉え、それを対象とする新たなデータ解析の枠組みが示されている。これを用いることで、内部変動を含みうる集約的な形にデータ表現を拡張することが可能となる。従来の SDA の研究で用いられるデータ表現として、主に多値や区間値、モダル値が扱われ、これらを対象とする回帰分析法や判別分析法、主成分分析法、クラスター分析法、多次元尺度構成法などの多変量解析手法が拡張されてきた。しかし、多様な大規模データに鑑み、より複雑な集約的データ表現に対する解析法が必要と考えられる。本論文では、従来の SDA で主に扱われてきた多値データ、区間値データなどのシンボリックデータを、基本的な集約的表現と位置づけ、基

基礎集約的シンボリックデータと呼ぶ。さらに、より複雑な集約的表現を、高度集約的シンボリックデータとして新たに提案し、それらに対する解析法を実現する。具体的には、分布値データ、関数データ、さらに尺度混在型データに対する高度集約的シンボリックデータ解析法を提案する。本論文の構成は以下の通りである。まず第 1 章では、本研究の背景とその目的について述べる。第 2 章では準備のため、SDA の基本的概念と基礎集約的シンボリックデータに対する解析法を概説するとともに、本論文を通じ、実解析例として具体的に扱う、複数のサイト間におけるインターネット上の通信遅延時間を測定したデータ、および、東京電力福島第一原子力発電所で発生した事故を原因として環境中に放出された放射性物質の空間線量率を測定したデータについて、概要を説明する。第 3 章では、解析対象毎に複数の測定値が観測されている場合に、統計的なばらつきを併せて表すことのできる分布値データを対象とした新たな解析法を提案する。具体的には、解析対象間の関係性を分布値によって表す分布値非類似度を導入し、関連性の高いデータ群を同定するための階層型シンボリッククラスター分析法を提案する。シミュレーションによる既知の手法との対比を通じて本手法の優位性を示すとともに、実際の通信遅延データへ本手法を適用し、その結果と解釈から本手法の有用性を示す。第 4 章では、Ramsay らにより提唱され、解析対象毎に連続的に測定される大量のデータを関数の形で表現する関数データを対象とした、解析法について論じる。具体的には、部分的な定義域を移動させながら関数クラスター分析を行う移動関数 k -means 法を提案し、従来は捉えにくかった部分的な特徴を反映した形で、クラスター構成の変化を分析できることをシミュレーションにより示す。さらに、第 2 章で触れた、複数のセンサーにより一定期間、短い時間間隔で測定された放射性物質の空間線量率データに対して提案手法を適用した結果から、クラスター構成が大きく変化するような期間の同定が可能となることを示した上で、地理的情報を考慮した解釈の検討を通じ、本手法の有用性を示す。第 5 章では、値域が多次元である関数データを多次元値関数データとして定義し、同時に異常とみなしうる領域を同定する解析法を提案する。各関数データにおける、部分的な領域での異常に対

して、 L_0 正則化を用いた平滑化処理を施すことで効率的に検出する方法を示し、局所的なノイズに影響されにくく、大域的な異常領域を検出できるなど、シミュレーションを通じて本手法の有効性を示す。さらに、第 2 章で触れた実際の大規模環境データにおいて、多数のセンサーで同時に発生した異常領域を同定する問題に本手法を適用し、同定された期間と気象データを用いた解釈を通じて、本手法の有用性を示す。第 6 章では、各解析対象の特徴が複数の記述により表される尺度混在型データを、高度集約的シンボリックデータとして扱う。異なる集約的表現で記述されたデータを直接比較することは一般に難しいが、様々な集約的表現に対して定義が可能である非類似度に着目し、特徴毎に多次元尺度構成法を適用することによって、集約的シンボリックデータを同一の空間に布置し、その上で、データの特徴間の関連性を分析する手法を提案する。応用例として、第 2 章で触れたデータを地域毎に構成した高度集約的シンボリックデータを扱う。データセットが複数の測定方法毎に得られているため、それらを考慮した形で改めて高度集約化データを再構成し、その高度集約化シンボリックデータ間の関連性について分析する。得られた結果の解釈を通じて、本手法が有用であることを示す。第 7 章では、2 章から 6 章で得られた研究成果を総括するとともに、今後の展開について述べる。

第2章 準備

2.1 シンボリックデータ解析について

データの規模が大きくなるにつれ、適度にデータを要約することが解析する上で重要となってくる。例えば、環境データでは、個々の地点毎の観測データよりもブロックや地域、行政体を対象にした解析に主な関心があり、各個のデータを地域毎にグルーピングした値の「集まり」を対象として解析するのが望ましい場合がある。

Diday により提唱されたシンボリックデータ解析 (Symbolic Data Analysis; SDA) では、「集まり」を対象とした解析が可能である。

SDA の基礎となる概念に「2 レベル・パラダイム」がある。これは、解析対象を、ミクロレベルとマクロレベルの二つの階層により捉える考え方であり、先ほどの環境データの例で考えれば、各観測地点はミクロレベルでの対象であり、ブロックや地域は、複数の各観測地点を含んだマクロレベルにおける対象として捉えることができる。SDA ではしばしば、ミクロレベルの個体を「第 1 レベル・オブジェクト」と呼び、マクロレベルの個体を「第 2 レベル・オブジェクト」という。

定義 1 $\Omega = \{\omega_1, \omega_2, \dots, \omega_m\}$ を第 1 レベル・オブジェクトからなる集合とする。このとき、 2^Ω の元からなる集合を第 2 レベル・オブジェクトと呼び、 $\mathcal{B}(\Omega) = \{B_1, B_2, \dots, B_n; B_i \in 2^\Omega\}$ と記す。

補足 1 環境データにおける地点、地域、行政体、国... というように、複数のレベルによって対象を捉える場合には、2 レベル・パラダイムの概念を再帰的に用いること

で第3レベル・オブジェクト、第4レベル・オブジェクト...を構成することも可能である。

第1レベル・オブジェクトや第2レベル・オブジェクトが具体的にどのようなものであるかを表すために、その特徴を記述しなければならない。オブジェクトの特徴を示す量の集合を特徴集合と呼ぶ。

例1 ある気象観測センサーがあり、1時間毎に気象観測値（気温（℃）、湿度（%）、風速（m/s）、雨量（mm/h）など）が1年間を通して測定されている状況を考える。解析者が、仮に、夏（ここでは、3～5月（春）、6月～8月（夏）、9月～11月、12月～2月とする）の気候に関心があり、その特徴を気象観測値のレンジにより捉えようとする（注：気象とは大気の状態を（短期の）定量データとして捉える、気候はより長期的な定性的な傾向を捉える）。観測時刻（月日時）を第1レベル・オブジェクトと考えると、各季節は第2レベル・オブジェクトと捉えることができ、

$$\Omega = \{1月1日0時, \dots, 1月31日23時, \dots, 12月1日0時, \dots, 12月31日23時\}$$

に対して、

$$\begin{aligned} B(\Omega) &= \{B_1 = \{3月1日0時, \dots, 5月31日23時\}, \\ &\quad B_2 = \{6月1日0時, \dots, 8月31日23時\}, \\ &\quad B_3 = \{9月1日0時, \dots, 11月30日23時\}, \\ &\quad B_4 = \{12月1日0時, \dots, 2月29日23時\}\} \\ &= \{B_1 = \{\text{春}\}, B_2 = \{\text{夏}\}, B_3 = \{\text{秋}\}, B_4 = \{\text{冬}\}\} \end{aligned}$$

となる。また、夏の気候における特徴を考えると、例えば、

$$(\text{気温} = [20, 34], \text{湿度} = [60, 87], \text{風速} = [1, 15], \text{雨量} = [0, 14])$$

のように記述することが考えられる。すなわち、この例では特徴集合は $\mathbf{R}^4$ における区間の集合である。

例 2 多変量解析では、解析対象の特徴を、多次元実数値ベクトルにより記述していると考えられる。例えば例 1 の状況を考えて、夏の気候を、要約した気象観測値（平均値や最大値、最小値など）により捉える場合などが考えられる。すなわち、

$$(\text{気温} = 27.5, \text{湿度} = 73.5, \text{風速} = 8, \text{雨量} = 7)$$

のように記述する。この例では、特徴集合は $\mathbf{R}^4$ における点の集合である。

定義 2 第 1 レベル・オブジェクトまたは第 2 レベル・オブジェクトの空間から、特徴集合への写像を記述 (Description) と呼ぶ。

例 3 例 2 において、4 次元実数空間における区間の集合を $\mathbf{IR}^4$ とおくと、記述を Y とおくと、

$$Y : \text{季節} \rightarrow \mathbf{IR}^4$$

と書くことができる。例えば

$$Y(\text{夏}) = (\text{気温} = [20, 34], \text{湿度} = [60, 87], \text{風速} = [1, 15], \text{雨量} = [0, 14])$$

と表すことができる。

例 4 例 2 の多変量データの例を考える。各地域の気候の特徴は、4 次元実数空間の点として考えられるので、ある地域（例えば北海道）の気候を表す記述を Y とおくと、

$$Y : \text{季節} \rightarrow \mathbf{R}^4$$

と表すことができ、例えば

$$Y(\text{夏}) = (\text{気温} = 27.5, \text{湿度} = 73.5, \text{風速} = 8, \text{雨量} = 7)$$

と表される。

次に第2レベル・オブジェクト自体がどのようなプロセスを通じて得られるのかについて述べる。Bock and Diday(2000), Billard and Diday(2005) は、第1レベル・オブジェクトと記述を用いて、それに対する形式的な定義を与えた。そのためにまず、「関係」と呼ばれる概念（または演算子）と「アサーション」と呼ばれる関数を定義する必要がある。

定義 3 二つの要素 A, B を関連づける形式を関係 (Relation) と呼び、 ARB と表記する。

典型的な例として、以下のようなものがある。

例 5 $\mathcal{R} \in \{<, \leq, >, \geq, =, \neq, \in, \notin, \subset, \subseteq, \prec, \preceq, \succ, \succeq, \rightarrow, \leftarrow\}$.

例 6 1 km グリッド毎に、ある 1 時点における放射性物質の空間線量率 ($\mu\text{Sv/h}$) が得られている状況を考える。特定の地域における空間線量率に関心があるとする。仮に、そのような地域を、ある基準値（年間被曝量 $0 \text{ mSv} \sim 20 \text{ mSv}$ ）を基にして考える。各 1km グリッドを第1レベル・オブジェクトとしたとき、各地域は第2レベル・オブジェクトとして捉えられる。 Y を各グリッド ω_i ; $i = 1, 2, \dots, m$ における、空間線量率から算出した年間被曝量 ($\text{mSv} = \mu\text{Sv/h} \times 365 \times 24 \times 10^3$) とおくと、

$$\text{基準値の地域} = \{\omega_i | Y(\omega_i)(\mathcal{R} = " \in ") [0, 20]; i = 1, 2, \dots, m\}$$

と表すことができる。

例 6 において、「基準値の地域」は、各グリッドの被曝量を表す記述 $Y(\omega_i)$ と、ある基準に基づく年間被曝量 (の記述) $[0, 20]$ が、関係 $\mathcal{R} = " \in "$ によって関連づけられることで得られる。

定義 3 による関係を評価する関数を導入する。単純な評価関数としては、関係が「有」「無」を評価する 2 値関数であるが、より一般的に $[0, 1]$ の間の値により、段階的に評価する関数も定義される。

定義 4 関係を評価する関数を、アサーション (Assertion) と呼ぶ。特に、二値で評価するアサーションをブーリアン・アサーション (Boolean assertion) と呼び、 $A\mathcal{R}B$ が真であれば 1, 偽であれば 0 となる関数に対して、

$$a(x) = \begin{cases} 1 & \text{if } x \text{ is True} \\ 0 & \text{if } x \text{ is False} \end{cases} \quad (2.1)$$

と定義する。また、 $[0, 1]$ の間の値で、関係を評価する関数を、モダル・アサーション (Modal assertion) と呼び、

$$a : x \mapsto I \in [0, 1] \quad (2.2)$$

と定義する。

関係とアサーションの定義を用いて、第 1 レベル・オブジェクト、記述、第 2 レベル・オブジェクトの間にある関係性を定式化する。

定義 5 $Y(\omega_i)$ を第 1 レベル・オブジェクト $\omega_i \in \Omega; i = 1, 2, \dots, m$ に対する記述とし、 $\mathcal{D}(B_j)$ を第 2 レベル・オブジェクト $B_j \in \mathcal{B}(\Omega); j = 1, 2, \dots, n$ に対する記述とおく。また、 $a^{(b)}(x)$ をブーリアン・アサーションとする。このとき、

$$\{\omega_k | a^{(b)}(Y(\omega_i)\mathcal{R}\mathcal{D}(B_j)) = 1, k \leq m\}$$

を外延 (Extension) と呼ぶ。また、 $a^{(m)}(x)$ をモダル・アサーションとすると、 $\alpha \in [0, 1]$ に対して、

$$\{\omega_k | a^{(m)}(Y(\omega_i)\mathcal{RD}(B_j)) \geq \alpha, k \leq m\}$$

を α レベルの外延 (Extension of level α) と呼ぶ。

定義 5 より、第 2 レベル・オブジェクトとは、ある記述 $\mathcal{D}(B_j)$ と記述 $Y(\omega_i)$ との関連性を通じて構築されるような第 1 レベル・オブジェクトの集まりである。言い換えれば、ある性質（あるいは特徴） $\mathcal{D}(B_j)$ を少なくとも一定以上満たすような第 1 レベル・オブジェクトの集まりが、第 2 レベル・オブジェクトであると言える。

定義 6 アサーション、関係、記述を用いて表される第 2 レベル・オブジェクトをシンボリック・オブジェクト (Symbolic object) と呼び、 $s := (a, \mathcal{R}, \mathcal{D})$ と表す。

表記 1 本論文では以降、特に断りがない限り、第 1 レベル・オブジェクトを「個体」と呼び、またシンボリック・オブジェクトを「オブジェクト」と呼ぶ。

補足 3 Diday はシンボリック・オブジェクトと「概念」(Concept) の関連について議論している。Billard and Diday(2005) では、概念について、その性質を表す「内包」(Intent) と、性質を満たすような対象の集まりを表す「外延」(Extent) を考慮して、二つのタイプの概念を議論している。一つは、内包を満たすような全ての外延を実世界の中で列挙できるタイプである。例えば、北海道という概念を考えると、実世界の中に（すべて）存在している。もう一つは、内包のみで、明確な外延を持たないようなタイプである。例えば、交通事故という概念を考えると、実際の全ての交通事故を列挙することは（恐らく）できない。Diday によると、シンボリック・オブジェクトはその二つの側面を持っており、いわば「概念」を近似するモデルである。この場合、記述は内包に相当し、また個体は外延に相当する。

例 8 例 6 を再び考える。1km 四方のグリッド毎に空間線量率の記述(観測値) $Y(\omega_i)$ ($i = 1, 2, \dots, m$) が得られており、「標準的な地域」とは被曝線量に基づき、 $\mathcal{D}(B) = [0, 20]$ として記述できることが知られているとする。各グリッドが「標準的な地域」かどうかは、 $Y(\omega_i)$ が $[0, 20]$ に含まれるかどうかで判断できるため $\mathcal{R} = "\in"$ と書くことができ、またアサーションとして定義 5 によるブーリアン・アサーションを用いることで、オブジェクト「標準的な地域」とは

$$\text{標準的な地域} = (a^{(b)}, \in, [0, 20])$$

とモデル化できる。

具体的なオブジェクトを構成する方法の一つとして、個体の集約 (aggregation) が考えられる。例えば、1km グリッドごとに空間線量率が測定された状況で、分析者の関心が市区町村ごとの空間線量率にあるような場合を考えると、各グリッドは個体に相当し、市区町村はオブジェクトに相当すると考えられる。このとき、グリッドを緯度・経度に従って集約して得られた市区町村に対して、空間線量率の集まりを得る。ここで、分析者が仮に、市区町村の空間線量率の範囲に関心があれば、値の集まりを「区間」によって記述すればよい。

定義 7 オブジェクト $B_j \in \mathcal{B}(\Omega)$ における個体を $\omega^{(j)} \in B_j$ とおく。すべての $\omega^{(j)}$ について、記述 $Y(\omega^{(j)})$ が得られているとき、 $\{Y(\omega^{(j)}); \omega^{(j)} \in B_j\}$ を利用して構成できる各オブジェクトに関する記述 $\mathcal{D}(B_j)$ よりなるデータを、集約的シンボリックデータ (Aggregated Symbolic Data) と呼ぶ。

集約的シンボリックデータには、各オブジェクトの内部における個体のばらつきと、オブジェクト間のばらつき (いわゆる、分散) がある。

定義 8 集約的シンボリックデータにおいて、各オブジェクト内に含まれる内部のばらつきを内部変動 (Internal variation) と呼ぶ。

内部変動をどのような形で表現するかは、集約的シンボリックデータを解析した結果にも関わるため、重要である。

2.2 基礎集約的シンボリックデータ

従来の SDA が扱ってきた記述は主に、集合である。実際、Bock and Diday(2000), Billard and Diday(2005) でも、まずはじめに、有限個の要素を元にもつ集合を、シンボリックデータとして導入している。SDA では集合をベースにして、区間に拡張したり、集合の各要素に測度を導入するなど、次第に集まりの内部の状態を表す表現へと拡張される。これまでの SDA では、多値、区間値、モダル値が主な解析対象として扱われ、殆どの解析手法の研究はこれらの記述に対して行われてきたといっても過言ではない。これらは、少ないパラメータでオブジェクトを表すことができる。例えば、(一次元の) 区間値は、最小値と最大値の二つのパラメータにより表される。その意味で、本論文ではこれらを基礎的な集約的シンボリックデータと位置づける。ここでは、少数のパラメータにより記述できる集約的シンボリックデータを基礎集約的シンボリックデータと呼ぶ。

オブジェクトの集合を $E = \{B_1, B_2, \dots, B_n\}$, 記述を

$$Y(B_i) = (Y_1(B_i), Y_2(B_i), \dots, Y_p(B_i)) \subseteq \mathcal{Y} = (\mathcal{Y}_1, \mathcal{Y}_2, \dots, \mathcal{Y}_p)$$

とおく。 $\mathcal{Y}_j$ は Y_j の値域を表す。

2.2.1 多値データ

多値データとは、有限個の要素を元に持つ集合値である。要素として、数値あるいはカテゴリのどちらの場合も考えられる。

定義 9 有限な $m_i; i = 1, 2, \dots, n$ に対して

$$Y_j(B_i) = \{x_{j,i,1}, x_{j,i,2}, \dots, x_{j,i,m_i}\}; j = 1, 2, \dots, p \quad (2.3)$$

を多値データと定義する。

2.2.2 区間値データ

多値データでは、有限個の要素を元に持つ集合を考えたが、より一般的に、区間で表すことが可能である。

定義 10 p 次元実数空間における部分集合を $\mathcal{Y}_j; j = 1, 2, \dots, p$ とおく。このとき、

$$Y_j(B_i) = [y_{j,i}^{(L)}, y_{j,i}^{(U)}]; j = 1, 2, \dots, p \text{ where } y_{j,i}^{(L)} = \min \mathcal{I}_j, y_{j,i}^{(U)} = \max \mathcal{I}_j \quad (2.4)$$

を区間値と定義する。

2.2.3 モダル値データ

モダル値は集合の各要素に対して測度を導入することで、定義される。

定義 11 $\mathcal{Y}_j$ における測度を p_j とおき、台 (support) を $U_j \subseteq \mathcal{Y}_j$ とおく。このとき、

$$Y(B_i) = (U(B_i), \pi_i) \text{ for } B_i \in E. \quad (2.5)$$

をモダル値と定義する。

定義 9 において、 $U(B_i)$ として多値データを定義すれば、 p_j は各要素が出現する相対頻度により表される。また、区間値データを $U(B_i)$ として定義すれば、 p_j は相対密度により表される。

定義 12 (モダル・多値データ)

$$Y(B_i) = \{(y_{j,i,k}, \pi_{j,i,k}); k = 1, 2, \dots, m_i\} \quad (2.6)$$

をモダル・多値データと定義する。

定義 13 (モダル・区間値データ) 区間値データ $Y_j(B_i)$ を $l_k; k = 1, 2, \dots, K_i$ 個の部分区間に分割する。このとき、

$$Y(B_i) = \{[y_{j,i,k}^{(L)}, y_{j,i,k}^{(U)}], \pi_{j,i,k}; k = 1, 2, \dots, m_i\} \quad (2.7)$$

をモダル・区間値データと定義する。

2.3 高度集約的シンボリックデータについて

基礎集約的シンボリックデータにより、従来までの解析対象である記述ベクトル（多次元実数値ベクトル）で表現された個体を対象とした解析から、オブジェクトを対象とする解析へと拡張された。

しかし、オブジェクトの概念は、非常に広範な解析対象を含んでおり、積極的にその概念を適用していくことで、基礎集約的シンボリックデータのみに限らず、他の集約的シンボリックデータを考慮した解析手法を開発していくことも可能である。

また、基礎集約的シンボリックデータは、少数のパラメータにより記述できるメリットがあるが、解析の背景や目的によっては、より多くのパラメータを用いて、詳細な解析を行うことも考えられる。例えば、（一次元の）区間値は二つのパラメータにより表されるが、それ以上のパラメータを考えることはできない。

ここでは、基礎集約的シンボリックデータに対して、多くのパラメータにより記述されるものを高度集約的シンボリックデータと呼ぶ。

2.3.1 分布値データ

モダル・区間値データは、区間ごとに相対頻度が与えられるため、ヒストグラム値とも呼ばれ、確率分布の近似表現とみなせる。分布表現を考えることにより、値の集合の内部における歪度や尖度のばらつきの詳細を表現することが可能である。

定義 14 $\mathcal{Y} = [0, 1]$, $X = (X_{i,1}, X_{i,2}, \dots, X_{i,p}) \in \mathbf{R}^p (i = 1, 2, \dots, n)$ とする。このとき、

$$Y(B_i) = \Pr(X_{1,i} \leq x_1, X_{2,i} \leq x_2, \dots, X_{p,i} \leq x_p) := F_i(x_1, x_2, \dots, x_p) \quad (2.8)$$

を分布値データと定義する。ただし、 $F_i : (x_1, x_2, \dots, x_p) \mapsto u \in [0, 1]$ として定義される単調増加関数とする。

補足 4 分布値データは、定義 11 のモダル値データにおいて、 $U(B_i) = \mathbf{R}$ に対する確率測度 π_i を導入して得られる。

2.3.2 関数値データ

関数データは、観測値が関数となるようなデータである。例えば、センシングデータにおいて、高頻度な測定を行うような場合は、その観測値は時間に対する連続的な関数だとみなすことができる。

定義 15 実直線の部分集合 $\mathcal{T} \subseteq \mathbf{R}$ に関する写像

$$F : t \rightarrow \mathbf{R}; t \in \mathcal{T} \quad (2.9)$$

を関数値データと定義する。

実際の解析で用いるデータは、そのままでは関数と見なすことが難しいため、B スプライン関数やフーリエ級数などの基底関数を用いた平滑化手法を適用することで元データから滑らかな連続関数にし、関数データを得る。

関数 $x(t)$ は、直交基底関数を $\phi_i (i = 1, 2, \dots, p)$; $\langle \phi_i, \phi_i \rangle = 1$; $\langle \phi_i, \phi_j \rangle = 0 (i \neq j)$ とおくと

$$x(t) = \sum_i^p c_i \phi_i \quad (2.10)$$

と表すことができる。ただし、 $\langle \phi_i, \phi_j \rangle := \int_{-\infty}^{\infty} \phi_i(t) \phi_j(t) dt$ である。係数 c_i はデータから最小二乗基準などにより決定する量である。関数データは定義により連続な関数であるため、時系列データとして考えたときには無限次元の時系列データと見なすこともできる。計算機で扱う際には関数は離散化されるため、次元があまりに大きいと関数そのものを扱うよりも、推定された係数 $\{c_i\}$ を代わりに用いて分析する方が、係数はスカラーであり、従来の多変量解析の枠組みで扱うことができるゆえに、効率的である。

関数データ解析法 (Ramsay and Silverman, 2005) による手法の多くは、係数 c_i からなる行列を、関数の代わりに用いることで、計算量を節約し、また従来の多変量解析手法を拡張した手法を開発している。これまでに関数主成分分析、関数正準相関分析、関数回帰分析、関数クラスタリングなどの有益な手法が開発されている。

また、関数データは分布値データにも関連する。分位関数は $[0, 1] \rightarrow \mathbf{R}$ の関数データである。共通した定義域 $[0, 1]$ に対する関数値として扱える利点などがある。

2.3.3 尺度混在型データ

実際のデータ解析では、目的やデータが測定された背景などを考慮した集約的データを扱う場合がある。定義 4 から明らかなように、記述空間では、異なる記述の組み合わせを考えることが可能であり、例えば、分布値と関数値からなるような記述を構成できる。このようにして構成されるデータを本論文では、尺度混在型データと呼ぶ。

定義 16 記述を $Y_j; j = 1, 2, \dots, p$ とおく。このとき、 $k \neq l; k, l = 1, 2, \dots, p$ に対して、少なくとも一つの記述に組み合わせについて、

$$Y_k \neq Y_l \quad (2.11)$$

となるような記述 $Y = (Y_1, Y_2, \dots, Y_p)$ により表されるデータを、尺度混在型データと定義する。

尺度混在型データに関連した研究は、これまでに数多く扱われており、特に、量的データと質的データが混在したデータを中心に、その解析手法が研究されてきた。Leon and Chough(2013) では、量的変数と質的変数から多変量データに対して、潜在変数をベースとして同時分布を定義し、それに基づき、正準相関分析などといった従来の多変量解析手法を拡張した。

また、尺度混在型データは、SDA の分野でも研究されており、例えば、Ichino and Yaguchi (1994) は、区間値とカテゴリカル・モダルが混在したデータのクラスタリング手法を提案し、さらに、de Carvalho(2009) はヒストグラム値も加えた混在型データのクラスタリング手法を示した。他にも、尺度混在型データに対するクラスタリング手法は数多く提案されている。

しかし、記述間の関連性を分析する方法は非常に少ない。Verde and Irpino(2008) では、尺度混在型データに対する一般化正準相関分析法を提案しているが、そこでは、異なる集約表現をファジィ表現により、従来の多変量データの形にコーディングしなおしたデータに対して、従来の一般化正準相関分析法を適用している。しかし、大量のカテゴリを集約したモダル値に対しては、コーディングする過程で次元数が非常に大きくなることや、適用可能なデータは基礎集約的シンボリックデータに限られるなど、応用範囲は限られている。

Ramsay and Silverman(2005) は、関数データと実数値ベクトルが混在したデータに対する主成分分析法を提案し、関数データを直交基底関数で展開したときの係数を、

実数値ベクトルと組み合わせて多変量データに見立てることで、従来の主成分分析法を適用している。

2.4 データセットについて

本論文では、二つのデータセットを扱う。一つは、ネットワーク上におけるサイト間の通信遅延時間を測定したデータセットであり、もう一つは、福島第一原子力発電所で起きた事故により拡散された環境中における放射性物質を測定した環境データである。

2.4.1 通信遅延時間データ

サイト間の通信遅延を測定したデータは、帯域幅やトラフィック量などといったネットワークに関する現況を知る上では重要な手がかりである。通信遅延を測定するための一つの方法として、ICMP パケットによるラウンド・トリップ・タイム (Round Trip Time, RTT) がある。TCP / IP では、送信元のノードから送信されたパケットに対して、宛先のノードが受信を確認し、受信が完了したことを知らせるパケットを送信元に送り返す操作が行われる。このとき、往復に要する大凡の時間が RTT に相当する。もしも途中でパケットが失われたり、異常な遅延が生じた場合には、通常、パケット損失を知らせるパケットが送信元のノードに送られる。

ICMP パケットの送受信を確認するツールも数多く知られており、例えば、Ping などがある。実際に Ping を実行した例を図 2.1 に示す。図 2.1 の例では 4 個の ICMP パケットが送信され、パケットの損失はなく、RTT の要約量として 4 回の試行における最小値、最大値、平均値が出力されている。このようにして、ICMP により通信遅延時間を測定することができる。

Ping を用いて通信遅延を測定し続けているプロジェクト (<http://www-iepm.slac.stanford.edu/pinger/>) など知られており、通信遅延から算出される通信スループットを基に、年次別に世界中の通信状況を記した白書を公開している。

日本では Resilent / Distix によるプロジェクトが知られており、本論文で用いるデータもその一部を用いて、手法の適用例を示した。本データは、2006 年から全国 30

```
C:\Users\yusuke>ping www.google.co.jp
www-cctld.l.google.com [74.125.235.151]に ping を送信しています 32 バイトのデータ:
74.125.235.151 からの応答: バイト数 =32 時間 =25ms TTL=51
74.125.235.151 からの応答: バイト数 =32 時間 =24ms TTL=51
74.125.235.151 からの応答: バイト数 =32 時間 =19ms TTL=52
74.125.235.151 からの応答: バイト数 =32 時間 =24ms TTL=51

74.125.235.151 の ping 統計:
    パケット数: 送信 = 4, 受信 = 4, 損失 = 0 (0% の損失),
ラウンド トリップの概算時間 (ミリ秒):
    最小 = 19ms, 最大 = 25ms, 平均 = 23ms
```

宛先ノード

4回中の測定値

(4回中の) RTTの最小, 最大, 平均

図 2.1: Ping の実行例

箇所に設置されたサイト間の通信遅延時間を測定したものである。5 分毎に各サイト間で送受信される 50 の ICMP パケットを用いて RTT を計算し、ログデータとして蓄積された。

本論文では、欠損が含まれていない完全なデータセットとして、9 つのサイト間で 2006 年 1 月 18 日から 2006 年 1 月 21 日までの 4 日間で測定されたデータを用い、RTT における往路と復路を平均した値を通信遅延時間とみなした。各サイトが設置された場所を図 2.2 に示し、9 つのサイト間で測定された通信遅延時間の概略を図 2.3 に示す。

いくつかのサイト間では、途中、大きく変化する期間が含まれていることが、予備分析によりわかった (図 2.4, 図 2.5)。これらは 2006 年 1 月 19 日付近に集中していた。この期間、日本ではちょうどライブドアショックがメディアで大きく報じられ、また株価取引が集中したことで東京証券取引所のシステムがダウンした日付とちょうど重なり、それらに関連したトラフィックの異常が疑われる。



図 2.2: 設置されたノードの場所

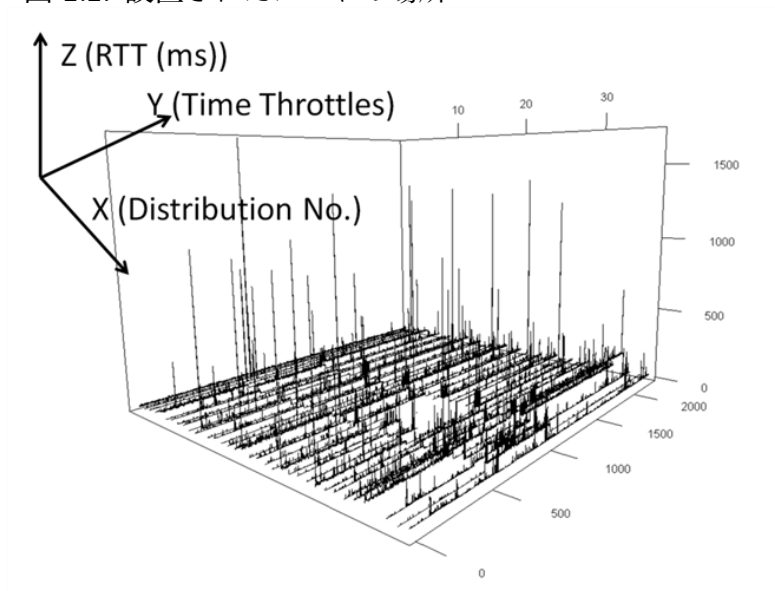


図 2.3: 全データのプロット

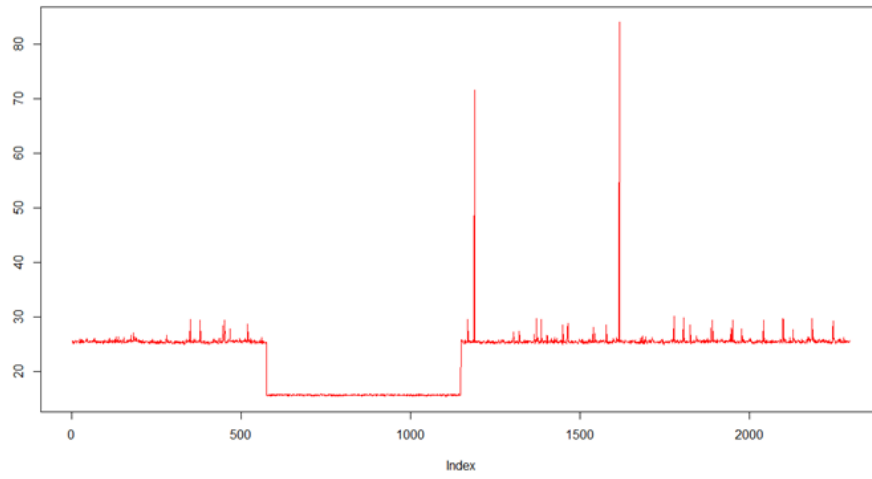


図 2.4: 通信遅延の異常な増加例

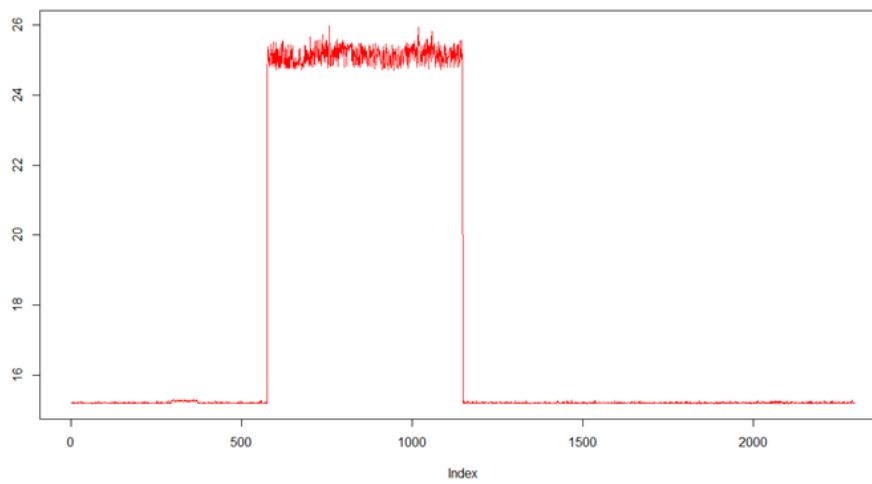


図 2.5: 通信遅延の異常な減少例

2.4.2 大規模環境データ

2011.3.11 の東日本大震災の後、震災時に起きた東京電力福島原子力発電所メルトダウン事故による、その後の環境への影響を評価するため、様々な大規模調査が繰り返し行われている。それらに関するデータセットは主に環境省、原子力規制委員会により公開されておりオープンに入手することが可能である。特に、空間線量率に関する大規模な調査は比較的事故直後から継続的に行われており、原発周辺の線量率の空間的、時間的变化について詳細に調べることが可能である。本論文で用いるデータは、それらのうち、航空機モニタリング、走行サーベイ、モニタリングポストの3つを扱い、検討する。

航空機モニタリングは航空機（ヘリコプター）に搭載された計測器（NaI シンチエーションサーベイメーター）により高度約 1500 メートルから半径 300 メートルにおける、地上 1 メートルの空間線量率を測定している。2011 年 4 月から不定期に実施され、主に福島原子力第一発電所を中心に福島県と近隣の地域および全国にわたり複数回の調査を実施している（表 2.1）。測定データには、空間線量率に加え、航空機の軌跡を記録した座標値、飛行した座標値から変換された細分メッシュコード、福島原発第一原発からの距離が主に記録されている。また、測定値の下限限界値は $0.1\mu\text{Sv/h}$ であり、それ以下の値は欠損扱いになっている。

表 2.1: 航空機モニタリングの概要

測定日時	実施名
平成 23 年 4 月 29 日	第 1 次航空機モニタリング
平成 23 年 5 月 26 日	第 2 次航空機モニタリング
平成 23 年 7 月 2 日	第 3 次航空機モニタリング
平成 23 年 11 月 5 日	第 4 次航空機モニタリング
平成 24 年 2 月 10 日	警戒区域と計画的避難区域の航空機モニタリング
平成 24 年 6 月 28 日	第 5 次航空機モニタリング
平成 24 年 12 月 28 日	第 6 次航空機モニタリング
平成 25 年 3 月 11 日	避難指示区域の航空機モニタリングの空間線量率の測定結果

航空機モニタリングデータの具体的な項目は以下の通りである。

- 航空機サーベイノード ID
- 補正基準日
- 都道府県
- 市町村
- 測定座標 北緯
- 測定座標 東経
- 測定座標 北緯 (10 進法)
- 測定座標 東経 (10 進法)
- 福島第一原発からの距離 (km)
- 空間線量率 (μ Sv/h)
- 積雪 (有・無)

航空機サーベイデータについて、もう少し詳細に予備分析を試み、線量率の時間的な減衰率を実際に分析してみる。メッシュサイズは 1 km メッシュのメッシュコードで公開されているため、それに合わせてメッシュサイズを 1 km とする。また、第 1 回航空機モニタリングサーベイから第 3 回モニタリングサーベイでは、測定時の補正方法が逆加重法と言われる手法により空間補正が成されているが、それ以降は別の一定の方法により補正されている。補正前の原データは入手困難であるため、ここでは、補正方法が統一されている第 4 回以降のデータを用いる。

経時的な（クラスター）データに対して、線形モデルを用いて、減衰率を分析した。メッシュ i ; $i = 1, 2, \dots, 1398$ に対して、測定値を Y_i とおき、また測定時刻を t とおく。このとき、モデルを

$$\begin{aligned} y_i &= \beta_{i,0} + \beta_{i,1}t + \varepsilon_i \\ \varepsilon_i &\sim N(\mu_i, \sigma_i^2) \end{aligned} \tag{2.12}$$

と仮定する。そのうえで、以下の検定により減衰率がゼロ以上か否かに関する検定を行う。

$$\begin{aligned} H_0 &: \beta_{i,1} = 0 \\ H_1 &: \beta_{i,1} \neq 0. \end{aligned} \tag{2.13}$$

検定 (2.13) により 1392 個のメッシュにおいて仮説 H_0 は有意水準 5% 棄却され、残り 6 個のメッシュでは棄却できなかった。図 2.6 に帰無仮説を棄却できなかった 6 つの

メッシュにおいて当てはめた線形モデルを図示する。放射性物質には物理崩壊による

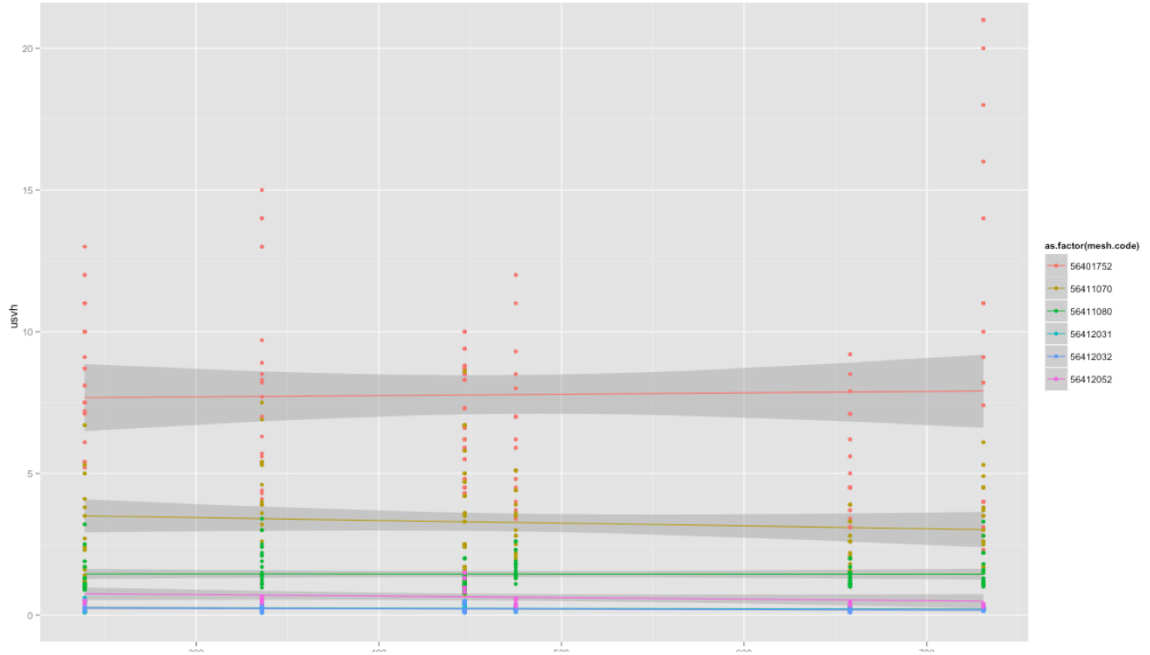


図 2.6: 棄却できなかったメッシュの線形モデル

減衰と環境減衰によるものが知られている。環境減衰（増加もあり得る）には、気象の変化の影響を受けるウェザリング効果や土砂の移動、空中や人口移動の際に付着した塵に伴った線量性物質の移動、除染による減少など、複合的な要因が放射性物質の変化の原因として考えられる。仮に、物理崩壊のみであれば放射性物質は減衰し続けるはずである。しかし、線型モデルも示しているように、単純な減衰を示しているわけではないことがわかる。松井 他 (2013) は、物理崩壊で減衰した線量とそれ以外の原因で変化した線量をモデル化し、それぞれを推測するモデルを提案した。現状で、重要な放射性物質はセシウム 134(Cs 134) とセシウム 137(Cs 137) である。Cs134 の半減期は 2.065 年で Cs137 の半減期は約 30.167 年であることが知られている。この半減期を考慮すると、Cs134 と Cs137 の混合核種の半減期は、

$$\frac{Y_i(t+\gamma)}{Y_i(t)} = p_1 \left(\frac{1}{2}\right)^{\frac{\gamma}{2.065}} + p_2 \left(\frac{1}{2}\right)^{\frac{\gamma}{30.167}}, \quad (2.14)$$

where $p_1 + p_2 = 1$

と書ける。例えば 2012 年時点では、 $p_1 = 0.29$, $p_2 = 0.71$ （原子力研究開発機構調べ）などが参考値として知られている。ここで、物理崩壊で減少した線量率の割合を δ_i ; $0 \leq \delta_i \leq 1$, それ以外の要因で変化した線量率の割合を $1 - \delta_i$ とおく。このとき、(2.14) を用いて、

$$\begin{aligned} Y_i(t + \gamma) &= Y_i(t) (1 - \delta_{i,j}) \left(p_1 \left(\frac{1}{2} \right)^{\frac{\gamma}{T_1}} + p_2 \left(\frac{1}{2} \right)^{\frac{\gamma}{T_2}} \right) \\ \Leftrightarrow \delta_i(t) &= 1 - \frac{Y_i(t+\gamma)}{Y_i(t)} \left(p_1 \left(\frac{1}{2} \right)^{\frac{\gamma}{T_1}} + p_2 \left(\frac{1}{2} \right)^{\frac{\gamma}{T_2}} \right) \end{aligned} \quad (2.15)$$

となる。式 (2.15) より、環境的な減衰割合を求めることができる。環境変化率 $\delta_i(t)$ の解釈は以下の通りである。

$$\begin{aligned} \delta_i(t) > 0 &: \text{環境的に減少} \\ \delta_i(t) = 0 &: \text{環境的な変化なし} \\ \delta_i(t) < 0 &: \text{環境的に増加} \end{aligned} \quad (2.16)$$

式 (2.16) を用いて、(2.13) における帰無仮説を棄却できなかった 6 つのメッシュを調べた。図 2.7 から明らかなように、棄却できなかったメッシュでは、物理減衰による変化分が大きい。したがって、より詳細に他の要因を調べる必要がある。

環境的な変化分については既述の通り、様々な要因が考えられるが、その中で特に、周辺メッシュとの線量率における相対量の大きさに着目する。例えば、周辺よりも高い線量率の地域は時間経過とともに低くなり、逆に周辺地域の線量率は高くなることが予想できる。そこで、メッシュ i に隣接する 4 つのメッシュにおける線量率を $M_j^{(i)}$; $j = 1, 2, \dots, 4$ とおく。また、メッシュ i における線量率と隣接するメッシュにおける線量率との差を Δ_i とおき、以下のような線型モデルを仮定する。

$$\begin{aligned} \delta_i(t) Y_i(t) &= \alpha_0 + \alpha_1 \Delta_i(t), \\ \text{where } \Delta_i(t) &= \frac{1}{m_i} \sum_{j=1}^4 \left(Y_i(t) - M_j^{(i)}(t) \right) \end{aligned} \quad (2.17)$$

式 (2.17) により推定したモデルと観測値の関係を図 2.8 に示す。航空機モニタリン

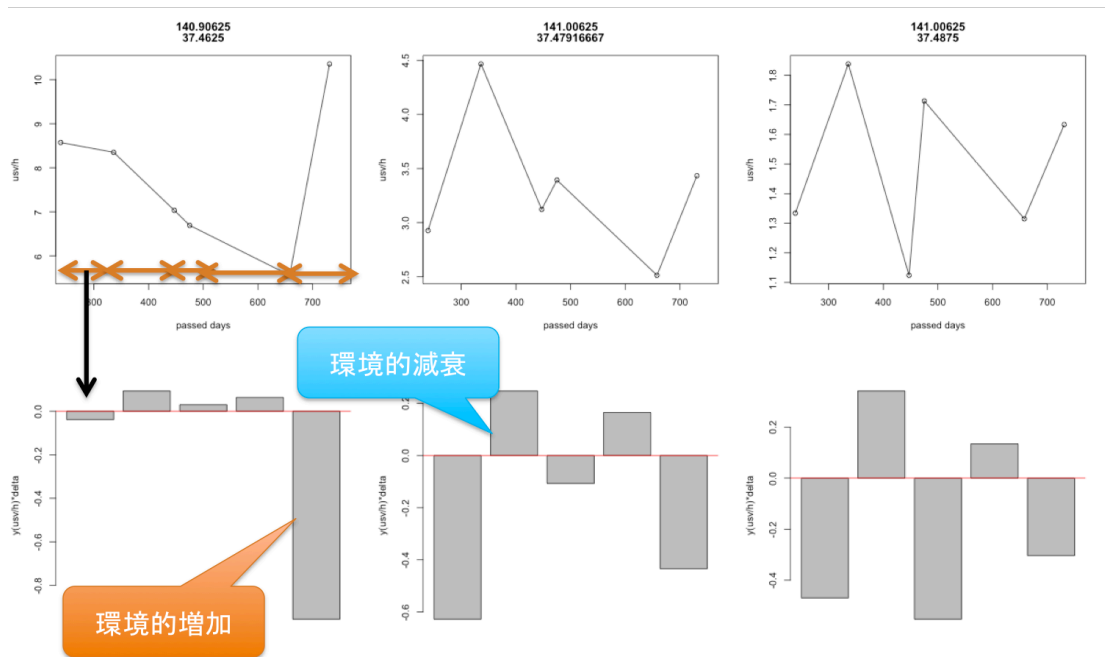


図 2.7: 各時間で環境的に変化した割合

グデータについて、詳細な予備分析を行った。この分析例から、環境的な要因が空間線量率に与える影響は単純ではないと思われる。そのため、空間線量率は一様に減少するとは言えず、今後の線量を予測することは容易ではない。このことは、航空機モニタリングデータに限らず、他の測定方法にも言えることである。

次に走行車モニタリングデータを概説する。走行車モニタリングは小型のセンサーを搭載した車両（バイク、バスなど）で、該当地域の道路を走行しながら、周囲約 100m の空間における線量率を測定している。航空機モニタリングと同様に、2011 年 4 月以降、原発周辺地域を重点的に、不定期な大規模測定が実施されている（表 2.2）。航空機モニタリングとは違い、 $0.1\mu\text{Sv/h}$ 以下の値も測定されている。測定地域は走行可能な地域に限定されるため、空間を網羅的に測定できる航空機モニタリングに比べると、測定範囲は狭い。また、通過する道筋は、調査毎に異なるため同じ地点の経時的変化を追うことは困難であることや、線量計を積載する走行車両毎に設置する高さが異なるなどの複合的な要因も重なっているため、測定データがどの程度に実際の空間

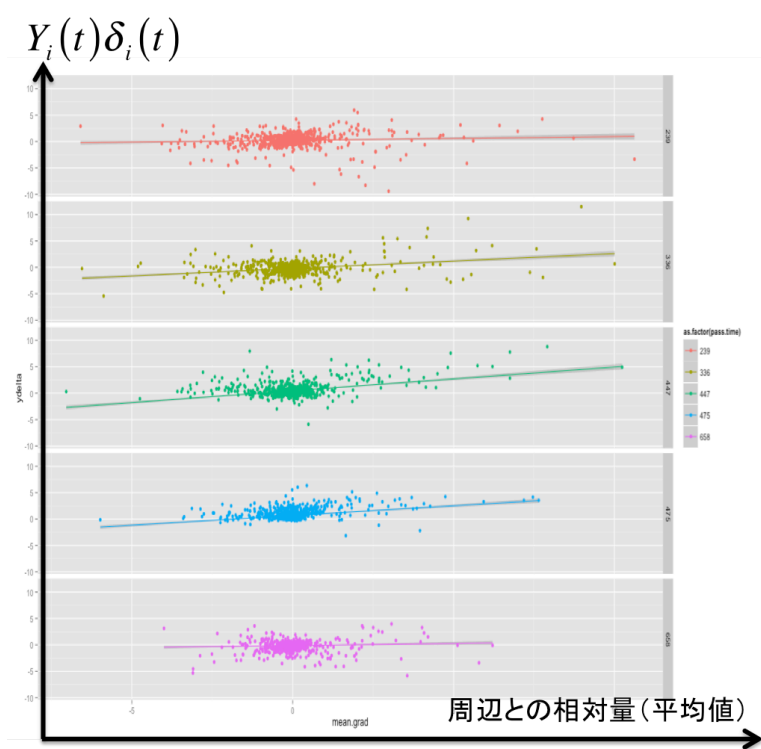


図 2.8: 周辺メッシュとの線量率の差（横軸）と環境的変化量（縦軸）の関係

線量率を反映しているかは未知である。

表 2.2: 走行サーベイの概要

測定日時	実施名
平成 23 年 6 月 4 日～13 日	第 1 次走行サーベイ
平成 23 年 12 月 5 日～28 日	第 2 次走行サーベイ
平成 24 年 3 月 13 日～30 日	第 3 次走行サーベイ
平成 24 年 8 月 20 日	第 4 次走行サーベイ
平成 24 年 11 月 5 日～12 月 10 日	第 5 次走行サーベイ
平成 25 年 6 月 12 日～8 月 8 日	第 6 次走行サーベイ
平成 25 年 11 月 5 日～12 月 12 日	第 7 次走行サーベイ

最後にモニタリングポストデータについて概説する。モニタリングポストによる測定では、全国各地に計 3921 箇所に設置されたセンサーにより、各地点における空間線量率の定点観測を実施している。10 分おきの観測データが逐次更新されるため、時間軸における詳細なデータを手に入れることができる。

これらの 3 つの測定方法はそれぞれ異なることから、それらの測定値が同等であることについては調べる必要がある。しかし、得られたデータの各測定値が観測された日時、地点（あるいはメッシュ）は異なり、比較に用いられるデータはごく少数である。その点に関しては、第 6 章においても言及する。図 2.15 は観測地点（1km メッシュ）が同じであるような測定値（278 個）のみを抽出して、時間に沿ってプロットしたものである。

また、同様にして、2013 年 6 月の走行モニタリングで観測された測定値、2013 年 09 月の航空機モニタリングで観測された測定値、モニタリングポストで 2013 年 06 月 1 日に観測された測定値について、各測定方法を軸にとった対散布図が図 2.16 である。ごく限られた範囲のデータではあるが、走行モニタリングによる測定値と航空機モニタリングによる測定値が高い相関を示す傾向があり、一方、モニタリングポストは他のどの方法とも相関が見られなかった。

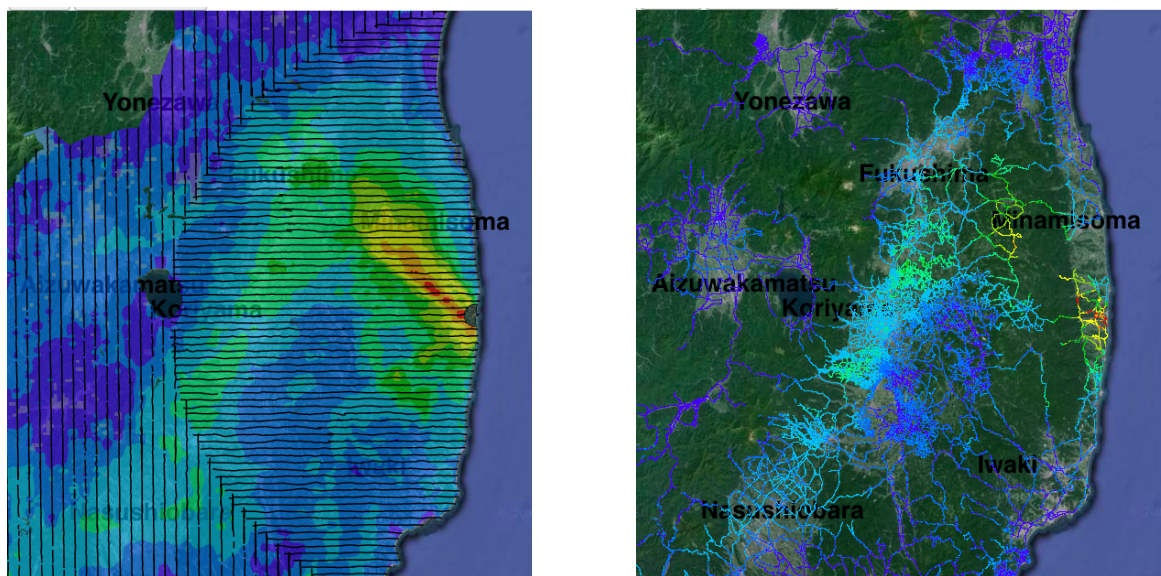


図 2.9: 空間線量率の空間分布。左図は航空機により 2012 年 12 月頃に観測された測定値を示し、右図は走行車により 2012 年 11 月頃に観測された測定値を表す。左図中、黒線は航空機が飛行した軌跡を表している

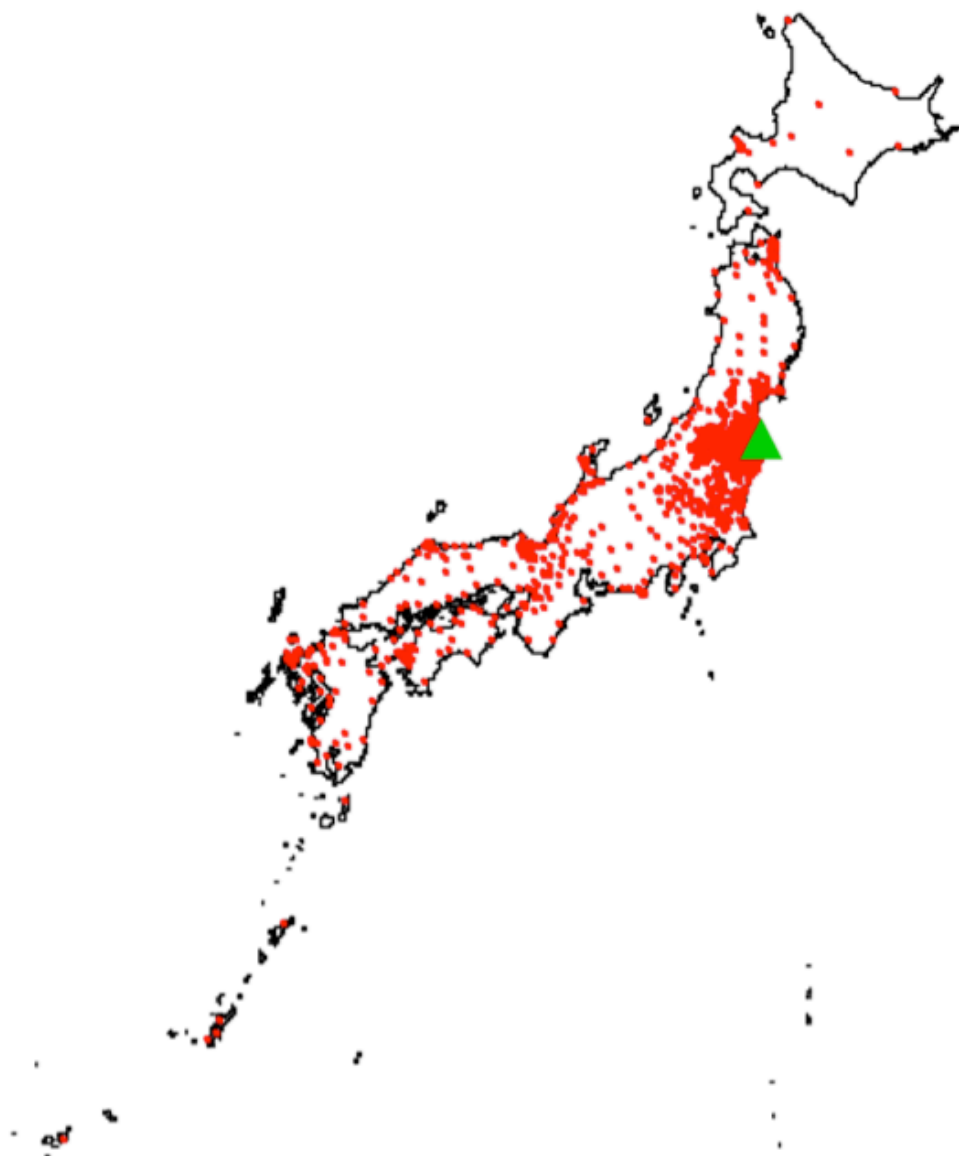


図 2.10: モニタリングポストの設置状況

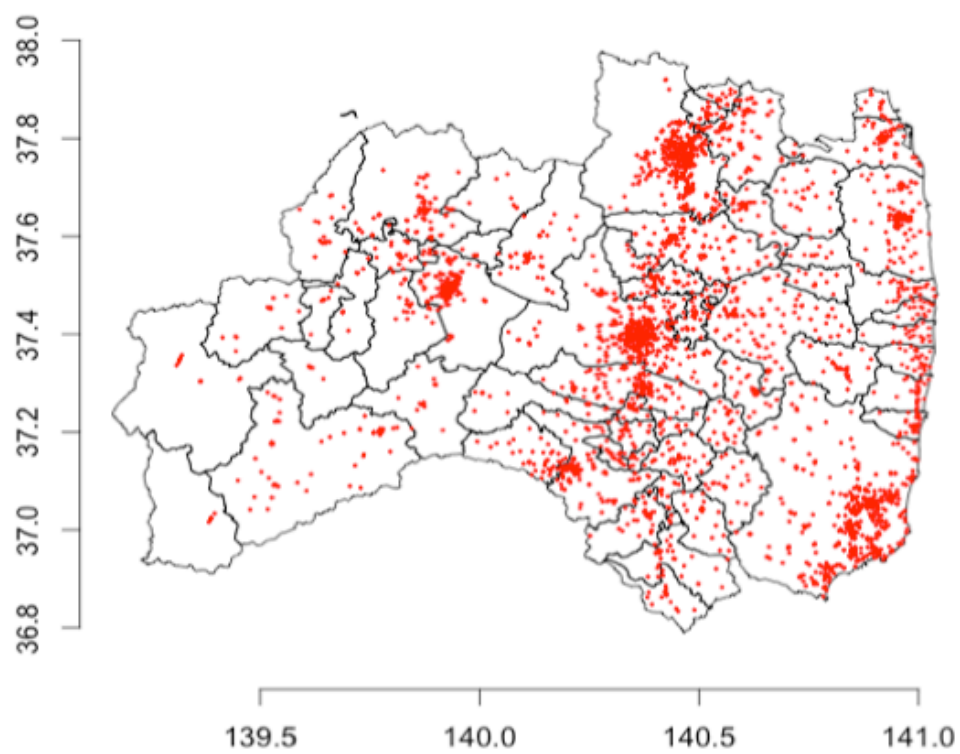


図 2.11: モニタリングポストの設置状況（福島県のみ）

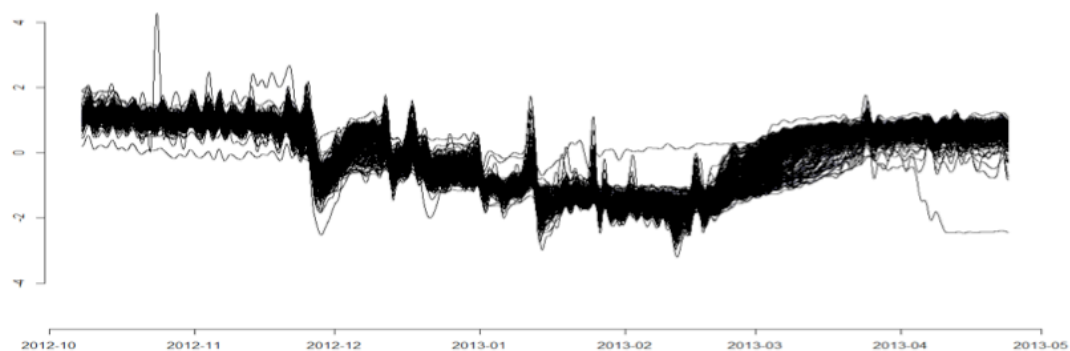


図 2.12: 会津地域における全測定値（2012 年 10 月～2013 年 5 月）

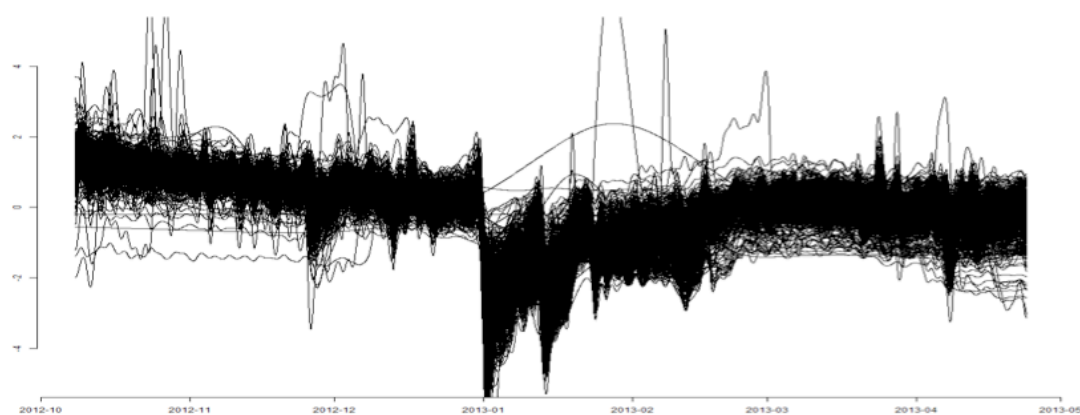


図 2.13: 県中地域における全測定値（2012 年 10 月～2013 年 5 月）

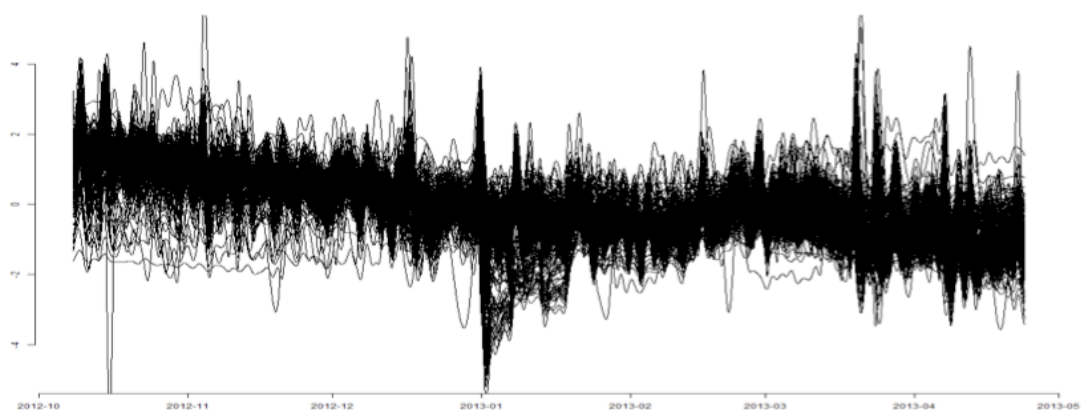


図 2.14: いわき地域における全測定値（2012 年 10 月～2013 年 5 月）

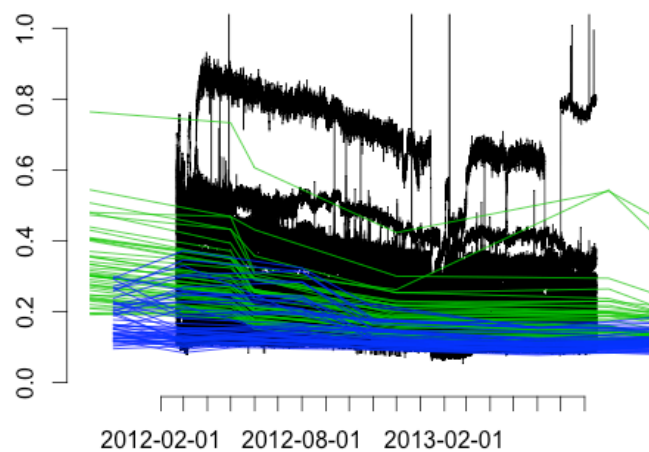


図 2.15: 各測定方法で同じ地点を観測した測定値

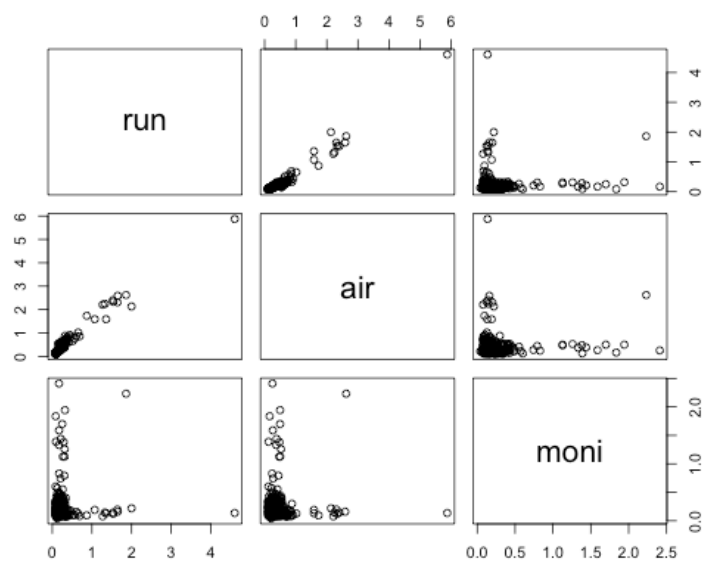


図 2.16: 各測定方法で同じ地点を観測した測定値の対散布図

第3章 分布値データにおける外れ値検出法

3.1 はじめに

近年、大規模な複雑データの解析手法が注目を集めている。SDA では、データマイニングのような巨大データベースからの「知識」や「パターン」の抽出にとどまらず、それらの知識に基づいた「概念 (concepts)」を引き出すことを視野に入れている。そのため、観測の対象（オブジェクト）として、カテゴリやクラス、コンセプトなどを扱う。

典型例としては、多値データや区間値データ、モダルデータ、分布値データなどがあり、とりわけ分布値データについては、Diday and Vrac(2005) が ”Distributions are the numbers of the future” という Schweitzer(1968) の論文を引用して分布値に対する解析手法を提案し、近年注目を集めている。

シンボリックデータに対して多くの手法が提案されている (Bock and Diday, 2000; Billard and Diday, 2006; Diday and Noirhomme-Fraiture, 2008)。クラスタリング法は、観測値の背後にある未知のサブグループを同定するために有用な手法である (Huh, 2002)。シンボリッククラスタリングは、シンボリックデータにおいて類似したオブジェクトのグループを同定するための手法として開発され、これまでに多値データ、区間値データ、モダルデータに対するクラスタリング手法が提案されている。

分布値に対するクラスタリングは近年の重要なテーマの一つである。Diday and Vrac (2005) は分布値データに対する混合分布の分解によるクラスタリング法を提案した。Katayama *et al.* (2009) は階層的シンボリッククラスタリング法を提案し、Terada

and Yadohisa (2010) は分布値に対する k -means 法を提案した。これらの手法は、入力データにおけるオブジェクトが分布値で表現されていると仮定した手法である。

非類似度の分析では、一般的に多次元尺度構成法 (MDS) や階層的クラスタリング法が用いられる。Mizuta and Minami (2012) は分布値非類似度に対する MDS を提案した。本論文では、分布値非類似度に対する階層的クラスタリング手法を提案する。また、実際のデータおよび人工データに対して、提案手法を適用した例を示す。

3.2 階層型シンボリッククラスタリングについて

階層的クラスタリングでは、二つの入力データが考え得る。一つは、入力データが p 変量からなる n 個のベクトルである。この場合、ベクトル間の非類似度を定義する必要がある。もう一つは、あるオブジェクトの組に対する非類似度 s_{ij} が直接、データとして得られる場合である。

階層的シンボリッククラスタリングにおいて、 p 変量をもつ n 個のオブジェクトが入力データに対しては多くの研究があり、様々な非類似度が提案されている (Billard and Diday, 2000)。シンボリックデータに対して非類似度が定義された後は、多くの場合、通常のクラスタリング手法を適用することで結果を得る。

Katayama *et al.* (2009) は n 個のオブジェクトが p 次元同時分布によって表されるシンボリックデータに対する階層的手法を提案し、非類似度として対称的カルバックライブラー情報量を導入することで、通常の階層的クラスタリングの手続きを経てクラスターを得る方法を示した。

しかし、非類似度がシンボリックデータとして得られるような場合はあまり研究されていない。この場合、通常の階層的クラスタリング法の手続きはそのまま適用することができないため、非類似度がシンボリックデータであることを想定した新たな手続きを考える必要がある。

3.3 提案手法

N 個のオブジェクトがあり、オブジェクト i, j の非類似度を S_{ij} とおく。 S_{ij} は分布であると仮定した場合の階層的クラスタリング法を示す。クラスターを $\{C_k; k = 1, 2, \dots, K\}$ と表す。

初期化

$K = N$, $C_i = \{i\}$ ($i = 1, 2, \dots, N$) として初期化する。

最も近いクラスターの同定

最も近いオブジェクトの組 (i, j) を以下のように定義する。

$$(i, j) := \operatorname{argmax}_{q, r} \{ \Pr \{ S(C_i, C_j) \leq S(C_q, C_r); (i, j) \neq (q, r), q < r, \text{ for all } q, r \} \}. \quad (3.1)$$

相対頻度により $\Pr \{ S(C_i, C_j) \leq S(C_q, C_r) \}$ を評価する。

$\{S(C_i, C_j); i, j = 1, 2, \dots, K, i < j, \text{ for all } i, j\}$ の実現値が M 個あると仮定し、 b ($b = 1, 2, \dots, m$) 番目の実現値を $\{s_{ij,b}\}$ とおく。

$\{S(C_i, C_j) \leq S(C_q, C_r); (i, j) \neq (q, r), q < r, \text{ for all } q, r\}$ に対する相対頻度は、

$$\frac{\sum_{b=1}^m \mathbf{1}_{ij,b}}{m}, \quad (3.2)$$

ただし、 $\mathbf{1}_{ij,b}$ は

$$\mathbf{1}_{ij,b} := \begin{cases} 1 & \text{if } (i, j) = \operatorname{argmin}_{q, r} \{s_{qr,b}\} \\ 0 & \text{otherwise} \end{cases}. \quad (3.3)$$

とする。

そのようにして、最も近いペアを

$$(i, j) = \operatorname{argmax}_{i, j} \left\{ \frac{\sum_{b=1}^m \mathbf{1}_{ij,b}}{m} \right\}. \quad (3.4)$$

により得る。

非類似度の更新

ここで、クラスター C_i と C_j が新たなクラスターに併合されたとする; $C_{new} = C_i \cup C_j$ 。次に新たに得たクラスターとその他のクラスターとの非類似度を更新する。分布値非類似度を混合させ、混合比 p で $S(C_i, C_q)$ と $S(C_j, C_q)$ を混合させる。 p の値は、 C_i と C_j に含まれるクラスター数を考慮して、

$$p = \frac{N_{C_i}}{N_{C_i} + N_{C_j}}, \quad (3.5)$$

と定義する。ただし、 N_{C_i} は C_i に含まれるクラスター数である。 $S(C_i, C_j)$ の累積分布を $F_{C_i C_j}(s)$ of $S(C_i, C_j)$ とおく。このとき、新たに更新された分布値非類似度 $S(C_{new}, C_q)$ を、

$$F_{C_{new} C_q}(s) = p F_{C_i C_q}(s) + (1 - p) F_{C_j C_q}(s). \quad (3.6)$$

とする。 K の値を 1 つ減じて、 $K = 1$ になるまで以上のステップを繰り返す。

上記のアルゴリズムを R により実装した。通常の階層型クラスタリング法では、非類似度行列を入力を用いる。しかし、本手法では分布値非類似度を扱うことから、非類似度を 3 次元の配列 $\mathbf{s}[i, j, m]$ ($m = 1, 2, \dots, M$) により表し、3 つ目の次元を確率変数 S_{ij} における m 番目の実現値に対応させる。最も近いクラスターを同定するステップでは、配列 $\mathbf{s}[i, j,]$ よりランダムに選んだ M' ($M' \leq M$) 個の値に対して、(3.3) 式および (3.4) 式を適用して最も近いクラスター (i, j) を選択する。非類似度の更新ステップにおける非類似度 $S(C_i \cup C_j, C_q)$ の再計算では、(3.5) で定義した p を用いて、

$s[i, q,]$ より $M' \times p$ 個の値、また $s[j, q,]$ より $M' \times (1 - p)$ 個の値をランダムに選び混合させる。そのようにして新たな非類似度の配列を得る。非類似度が全く同じになった場合には、どちらかをランダムに選択する。

3.4 適用例

提案手法を通信遅延に関する実際のデータに適用する。インターネット上におけるサイト間の仮想的な距離は Round Trip Time (RTT) により推測することができる。本データにおける RTT は 2 年間にわたり 35 箇所のサイト間で測定されたもので、それぞれの測定は 5 分おきに行われている。通信遅延に基づいたサイトの分類情報は、通信状態を把握する際の重要な手がかりとなる。

データセットには大量の欠損値が含まれており、最終的に完全なデータ（双方向の通信遅延データ）として得られたのは、9 つのサイト間での 4 日間の測定値であった。

図 3.1 は、9 つのサイト間における分布値非類似度の配列を表している。多くの非類似度には二つのピークがあることがわかる。ピークの要因として、この時期に大きく報道された、巨大 IT ベンチャー企業によるスキャンダルとの関連性が考えられる。そのスキャンダルに対する国民的関心は高く、通信量を大きくした要因の一つである可能性がある。

提案手法をデータに適用した結果を図 3.2 に示す。従来手法との比較のために、各分布値非類似度の平均をとり、最短距離法、最長距離法、平均距離法、重心法を適用した結果を図 3.3 に示す。

図 3.2 および図 3.3 について解釈を試みる。図 3.2 より、二つのクラスターを仮定して、サイトを 2, 4, 5, 6 と 1, 3, 7, 8, 9 に分けた。二つのクラスターの差異は通信帯域である。サイト 2, 5, 6 は東日本に設置されており、サイト 1, 3, 7, 8, 9 は西日本にある（ただし、サイト 4 はインターネットエクスチェンジであり、他と比べて特殊であるため例外とした）。この解釈より、東日本における通信帯域は西日本における通

信帯域よりも狭いことが、再確認された。図 3.3 では、いくつかのクラスター数を試したが、解釈が可能なものはなかった。

3.5 数値実験

人工データを用いて提案手法を検証する。9つのオブジェクトを仮定し、それらの非類似度はガンマ分布で与えられているとする。パラメータは分布の平均を固定した状態で変化させてデータセットを生成させる。すなわち、分布の平均を用いて従来の階層的クラスター分析法を適用する方法では、いずれのデータセットでも結果は同じとなる。二次元空間に9つの点を正規乱数により生成させた。正規分布のパラメータは、

$$N\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix}, 5\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$$

(表 3.1) とした。得られた9つの座標値に対してユークリッド距離 $\{s_{ij}^{(init)}; i < j\} (i, j = 1, 2, \dots, 9)$ (表 3.2) を計算した。それらの点を平均値として固定してガンマ分布を生成させる。すなわち、 $\alpha_{ij}\beta_{ij} = s_{ij}^{(init)}$ の条件下で、形状パラメータを $\{\alpha_{ij}; i < j\} (i, j = 1, 2, \dots, 9)$ 、尺度パラメータを $\{\beta_{ij}; i < j\} (i, j = 1, 2, \dots, 9)$ とする (表 3.3)。

従来の階層的クラスター法（最短距離法、最長距離法、平均距離法、重心法）を $\{s_{ij}^{(init)}; i < j\} (i, j = 1, 2, \dots, 9)$ に適用した結果を示す (図 3.4)。

次に形状パラメータ α_{ij} を一様分布 $U(1,3)$ より生成させ、それから尺度パラメータを $\beta_{ij} = s_{ij}^{(init)}/\alpha_{ij}$ により得た。それらのパラメータに対応するガンマ分布を用いて、それぞれ乱数を 1000 個生成させ ($\{s_{ij,b}; b = 1, 2, \dots, 1000\}$)、提案手法を適用した (図 3.5)。4つのデンドログラムの違いは、非類似度のばらつきの差異を反映させている。平均値を用いて従来手法を用いるといずれの結果も同じとなるが、非類似度の歪度や尖度などを考慮して分析することで、異なる結果が得られる。

例えば、シミュレーション 1 において、オブジェクト 5, 9 は同じクラスターに併合されるが、シミュレーション 2 においては、オブジェクト 5, 7 が併合される。そ

の結果の違いを調べるために、 $s_{5,9}$ と $s_{5,7}$ の四分位数 ($q_{0.25}, q_{0.5}, q_{0.75}$) を考える。添え字はパーセンテージを表している。シミュレーション 1 では $s_{5,9}$ の四分位数は (1.787, 3.266, 5.415) で、 $s_{5,7}$ は (2.210, 3.536, 5.315) である。したがって、四分位数の観点では、オブジェクト 5, 7 が類似していると考えるのが妥当である。同様な観点により、シミュレーション 2 では、 $s_{5,9}$ の四分位数は、(2.290, 3.554, 5.220) で、 $s_{5,7}$ は (2.151, 3.504, 5.340) であり、オブジェクト 5, 7 を併合するのは妥当であると考えられる。以上より、提案手法により分布のばらつきを考慮したクラスターを得ることができる。

3.6 おわりに

本論文では、分布値非類似度に対する階層的クラスタリング法を提案した。実際の適用例では、平均値を用いて従来の階層的クラスタリング法を適用する場合に比べて、提案手法により解釈可能なクラスターが得られることを示した。今回は、入力データとして、分布値非類似度が直接得られていることを想定したが、実際の場面では、分布値としてデータを得られる場合は少ない。分布値をデータとして得るには多くの観測値を必要とするためである。

ビッグデータの一つの特徴が多様性であるのに対して、SDA では多様なデータ記述を備えているため、ビッグデータに対して極めて有用であると考えられる。すなわち提案手法は、ビッグデータに対して適していると考えられる。この考え方は、分布値非類似度を扱うような他の手法（例えば多次元尺度構成法）に対しても拡張できると考えられる。

表 3.1: 座標点

	x_1	x_2
1	-1.80	2.18
2	-2.07	3.02
3	1.64	1.84
4	1.16	5.39
5	0.16	-2.36
6	-0.15	1.04
7	-3.60	-0.90
8	0.45	1.01
9	2.82	-5.33

表 3.2: 座標点間の距離

	1	2	3	4	5	6	7	8
2	0.882							
3	3.457	3.893						
4	4.366	4.006	3.582					
5	4.945	5.824	4.453	7.814				
6	2.006	2.758	1.961	4.543	3.414			
7	3.567	4.208	5.913	7.888	4.034	3.958		
8	2.536	3.223	1.451	4.437	3.382	0.601	4.478	
9	8.817	9.676	7.266	10.848	3.987	7.028	7.800	6.768

表 3.3: 4 回のシミュレーションにおけるガンマ分布のパラメータ

	Simulation 1		Simulation 2		Simulation 3		Simulation 4	
	α_{ij}	β_{ij}	α_{ij}	β_{ij}	α_{ij}	β_{ij}	α_{ij}	β_{ij}
$s_{1,2}$	2.581	0.341	2.748	0.320	1.673	0.526	1.257	0.700
$s_{1,3}$	1.543	2.236	2.214	1.559	2.595	1.330	2.674	1.290
$s_{1,4}$	1.025	4.257	1.476	2.956	2.127	2.051	1.012	4.311
$s_{1,5}$	1.002	4.929	1.743	2.834	2.008	2.460	2.467	2.003
$s_{1,6}$	1.913	1.049	1.303	1.540	1.513	1.326	2.908	0.690
$s_{1,7}$	1.173	3.043	1.955	1.827	1.951	1.830	2.214	1.613
$s_{1,8}$	1.085	2.339	2.220	1.143	2.130	1.191	2.506	1.012
$s_{1,9}$	2.402	3.670	2.216	3.979	2.660	3.315	1.832	4.812
$s_{2,3}$	2.647	1.471	1.319	2.951	1.703	2.286	1.479	2.632
$s_{2,4}$	2.346	1.710	1.480	2.711	2.513	1.596	2.879	1.393
$s_{2,5}$	2.223	2.617	2.661	2.187	1.294	4.496	1.243	4.682
$s_{2,6}$	2.963	0.932	1.993	1.386	1.084	2.548	2.527	1.092
$s_{2,7}$	2.754	1.526	2.885	1.457	1.027	4.092	1.704	2.467
$s_{2,8}$	2.384	1.354	1.104	2.924	2.607	1.238	2.969	1.088
$s_{2,9}$	2.913	3.322	2.975	3.252	1.723	5.616	1.672	5.786
$s_{3,4}$	1.422	2.521	2.162	1.658	2.232	1.606	2.802	1.279
$s_{3,5}$	2.158	2.064	2.797	1.592	1.802	2.470	1.166	3.818
$s_{3,6}$	1.781	1.097	1.281	1.526	1.234	1.583	2.645	0.739
$s_{3,7}$	1.257	4.702	2.605	2.269	2.472	2.391	1.874	3.154
$s_{3,8}$	2.614	0.553	2.247	0.644	2.982	0.485	1.529	0.946
$s_{3,9}$	1.341	5.423	1.509	4.817	1.238	5.870	1.284	5.662
$s_{4,5}$	2.549	3.066	1.777	4.397	1.190	6.570	1.129	6.923
$s_{4,6}$	2.513	1.808	2.963	1.533	2.078	2.187	1.750	2.596
$s_{4,7}$	1.403	5.624	1.834	4.300	2.341	3.369	2.951	2.673
$s_{4,8}$	1.239	3.583	1.282	3.464	2.553	1.739	1.970	2.254
$s_{4,9}$	1.769	6.137	1.436	7.559	1.946	5.576	2.086	5.204
$s_{5,6}$	2.090	1.632	1.895	1.800	1.981	1.722	1.800	1.895
$s_{5,7}$	2.654	1.518	2.490	1.618	1.815	2.219	2.172	1.854
$s_{5,8}$	1.914	1.765	2.657	1.272	1.411	2.394	1.101	3.069
$s_{5,9}$	1.763	2.263	2.972	1.343	1.048	3.807	1.943	2.054
$s_{6,7}$	2.291	1.729	2.874	1.378	2.132	1.858	1.747	2.267
$s_{6,8}$	1.747	0.344	2.755	0.218	2.815	0.213	1.726	0.348
$s_{6,9}$	1.650	4.258	2.887	2.434	1.097	6.404	1.387	5.068
$s_{7,8}$	1.862	2.406	1.174	3.816	1.027	4.364	2.122	2.111
$s_{7,9}$	2.213	3.524	1.789	4.359	2.217	3.516	2.356	3.309
$s_{8,9}$	1.306	5.179	2.368	2.858	2.538	2.666	1.351	5.009

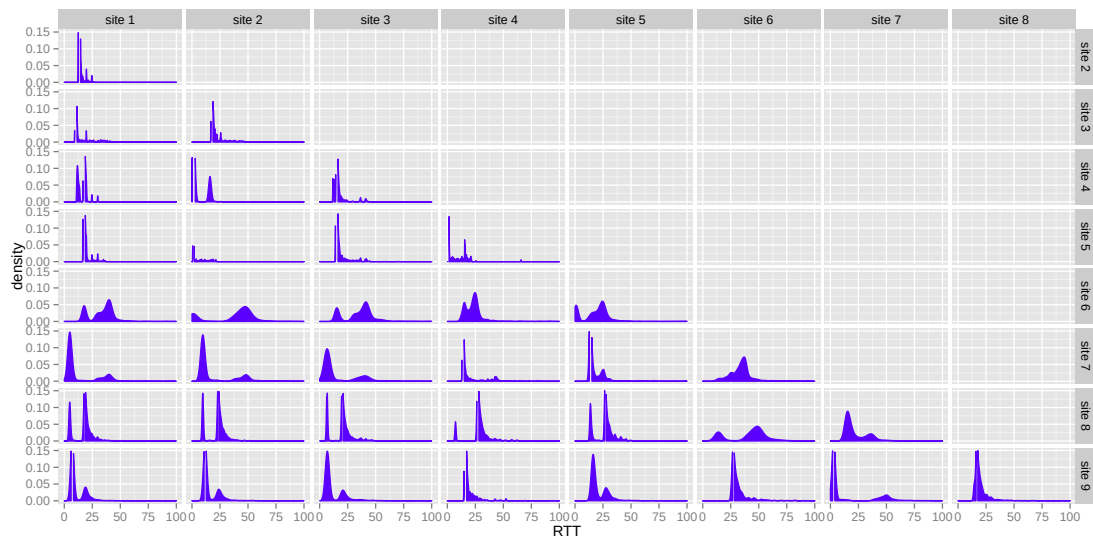


図 3.1: 分布値非類似度の配列 (RTT データ)

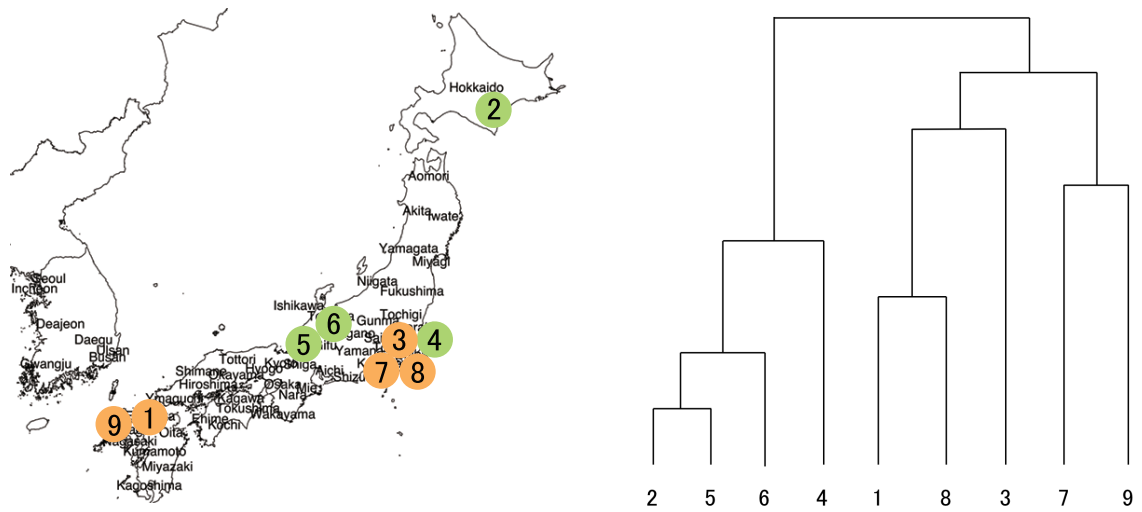
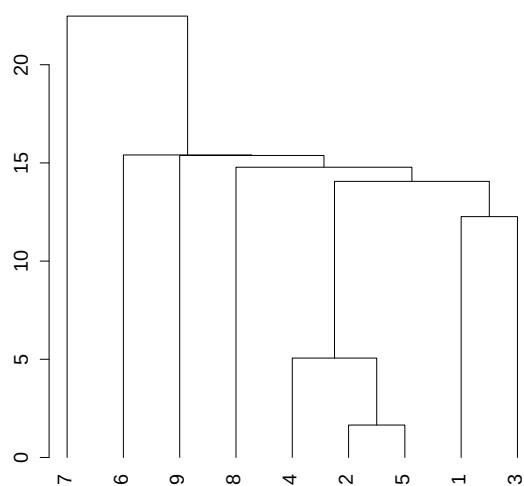
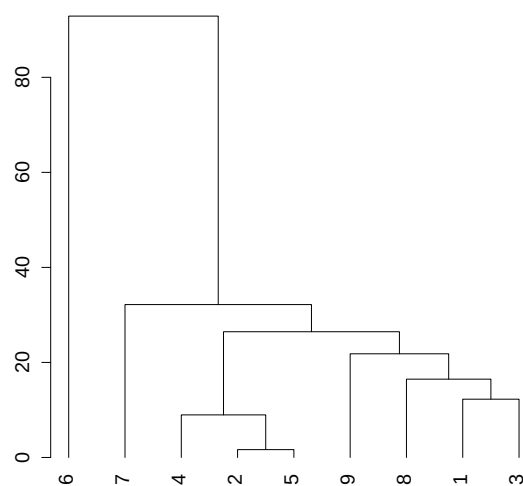


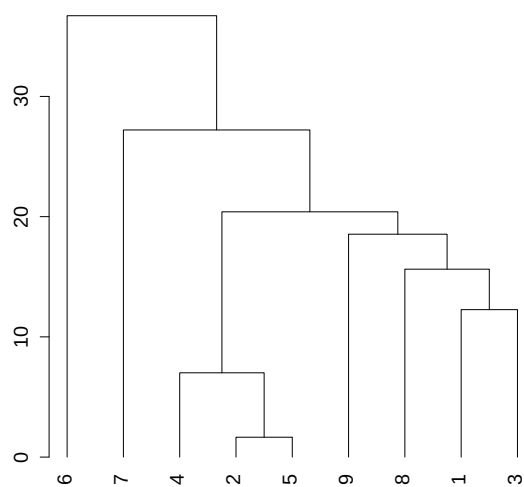
図 3.2: 提案手法を RTT データに適用した結果



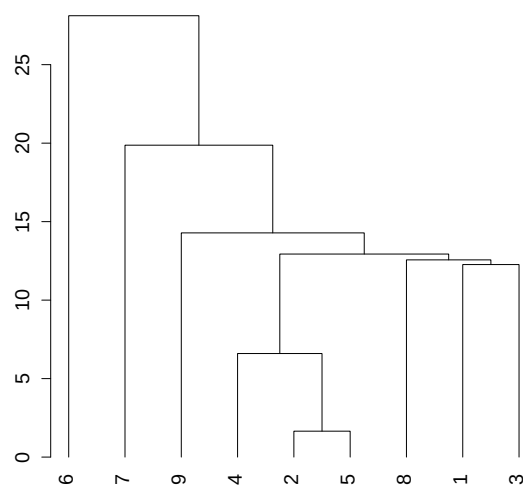
a: single linkage



b: complete linkage

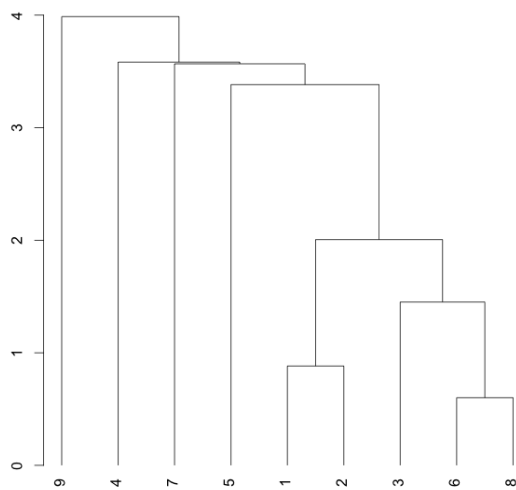


c: average linkage

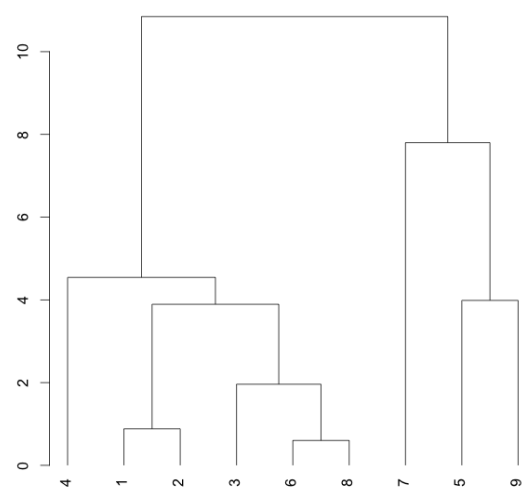


d: centroid linkage

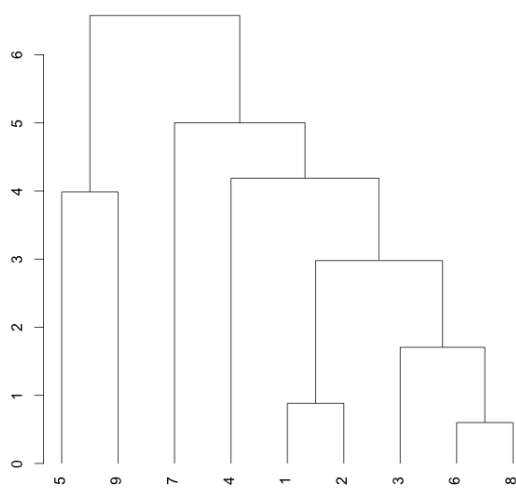
図 3.3: RTT の平均値を用いて従来の階層型クラスタリング法を適用した結果



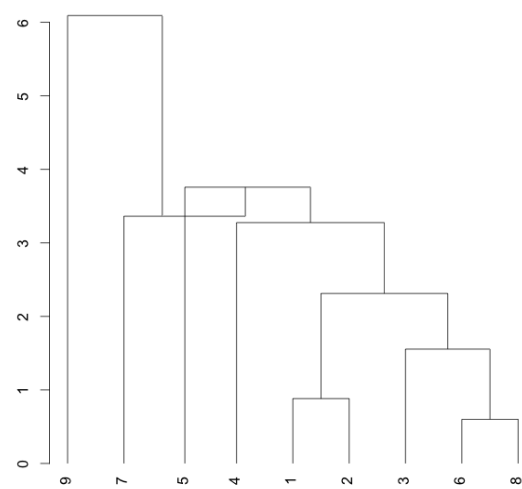
a: single linkage



b: complete linkage



c: average linkage



d: centroid linkage

図 3.4: 分布値非類似度の平均値を用いて従来の階層型クラスタリング法を適用した結果

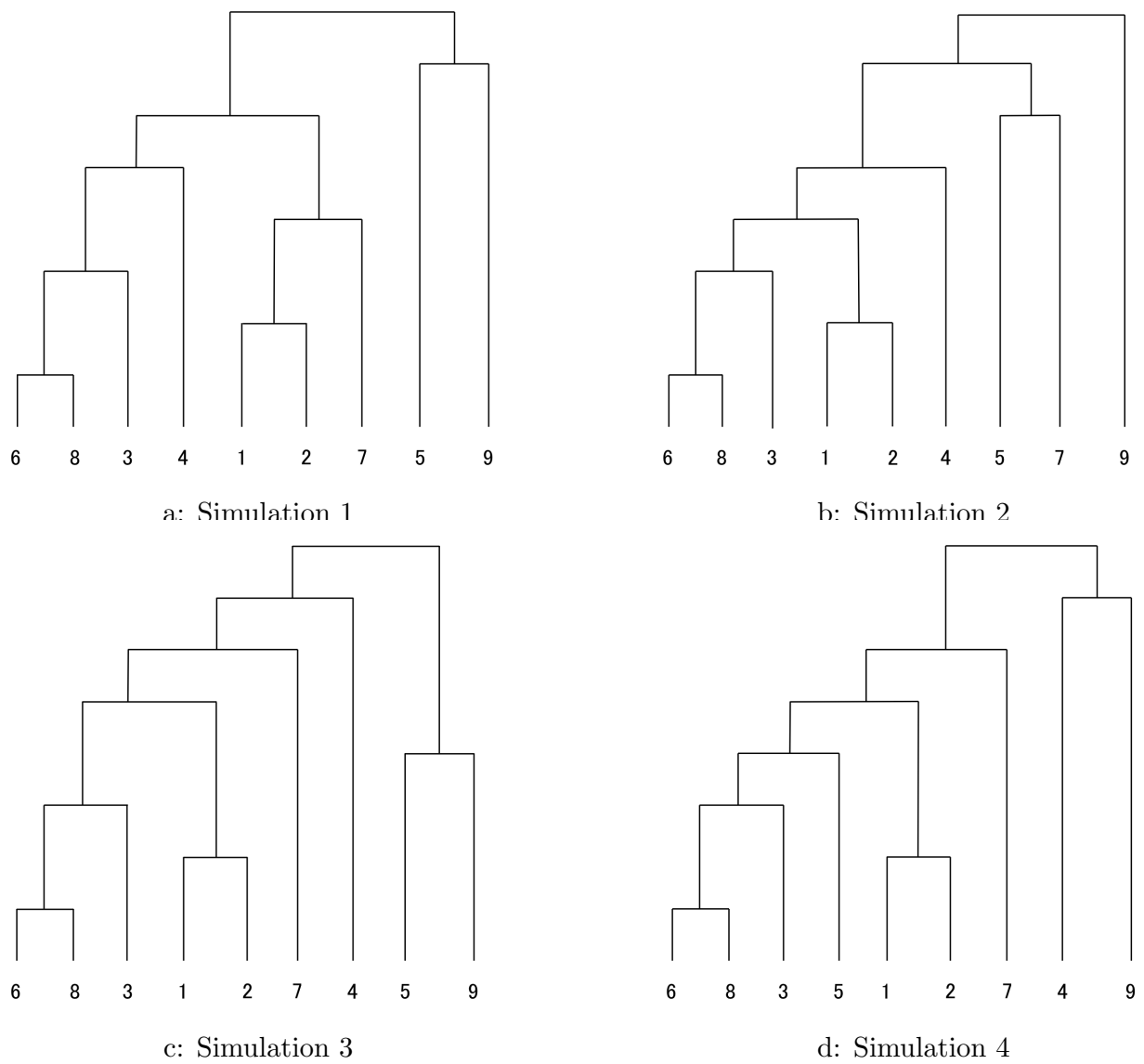


図 3.5: 提案手法を用いたシミュレーション結果

第4章 移動関数 k -means 法

4.1 はじめに

近年のデバイス技術の著しい進歩によって、以前までは得られなかったような非常に高い頻度で、大量のデータを得られる場面が増えている。特に時間に依存したデータが大量に収集され、その解析が社会的に望まれている。このようなデータは、時間帯により特異な挙動を示すことがあり、従来の統計解析手法のように各データ点に基づいて解析を行うのではなく、データの集まりを一つの要素と考えて解析する手法が提案されている。その中で、連続的なデータを「関数」と見て、関数として表されたデータの集まりを一つの要素として扱う関数データ解析 (Functional Data Analysis, FDA) が注目されている。FDA は、1990 年代に Ramsay らによって提唱され、これまでの成果をまとめた成書も出版されている (Ramsay & Silverman 2002, 2005; Ramsay, Hooker & Graves 2009)。関数データ解析では、従来の多変量解析法を関数に対応した手法に拡張して用いることが多く、これまでに関数回帰分析や関数主成分分析、関数クラスター分析など多くの有用な手法が開発・応用されている (Ramsay et al., 2005)。

ところで、これらの関数データ解析法においては、定義域をあらかじめ固定している。しかし、実際のデータ解析では部分的な定義域に限定された特徴を導出するなど、定義域を予め仮定しない関数データを分析する場合も考えられる。例えば、Mizuta (2004) は定義域を固定しない階層的なクラスタリングについて提案した。本論文では、定義域を考慮した関数データに対する非階層的なクラスタリング法を提案する。また実際の適用例として、福島県における空間線量率を測定したセンシングデータを用いて、本手法の有効性を示す。

4.2 関数データに対するクラスタリング

関数クラスタリングは、似ている関数同士をグループに分類するための一連の手法であり、これまでに多くの手法が開発されている (Ferraty & Vieu, 2006; Jacques & Preda, 2013)。以下では n 個のデータを $x_i(t); i = 1, 2, \dots, n; t \in [a, b] (a, b \in R)$ とおき、 n 個の個体を、互いに素な集合 $C_1, C_2, \dots, C_K$ に分割することを考える。

階層的関数クラスタリング i と j の間の非類似度を定義することで、それに基づく手法として階層的関数クラスタリングが提案されている。例えば、

$$s_{ij} \equiv \int_a^b (x_i(t) - x_j(t))^2 dt \quad (4.1)$$

により非類似度を定義することで、通常の階層的クラスター解析法のアルゴリズムを用いることができる。

非階層的関数クラスタリング 代表的な非階層的手法として関数 k-means 法がある。これはクラスター C_k の重心となる関数を $\mu_k(t) = \frac{1}{\#C_k} \sum_{i \in C_k} x_i(t); t \in [a, b]$ とおき、目的関数

$$E = \sum_{k=1}^K \sum_{i \in C_k} \int_a^b (x_i(t) - \mu_k(t))^2 dt \quad (4.2)$$

を最大化するようなクラスターを求める手法である。この他に、関数ファジィクラスタリングなども提案されている (Tokushige, Yadohisa & Inada, 2007)。

4.3 提案手法

経時的なデータを扱う分野では、適切な期間に対する移動平均を用いて季節変動などを調整する手法がしばしば用いられ、それにより大きなトレンドを見つけることができる。移動平均と同様に、期間に対応して関数データの定義域を柔軟に扱うことは重要である。一般に、2 節で述べた関数クラスタリングでは、関数の定義域は固定して扱わ

れる。そこで、定義域を固定せずに、部分的な定義域を逐次移動させながら関数データをクラスタリングする移動関数 k -means 法 (Moving Functional k -means Clustering) を提案する。 r に依存して変化する定義域を $I(r) = [r, r + w] \subseteq [a, b]$; $r, w \in R$ とおき、その上で定義される関数データ $\{x_1(t), x_2(t), \dots, x_n(t)\}; t \in I(r)$ を考える。 $I(r)$ ごとに、関数 k -means 法を適用する。すなわち、 n 個の個体を r に依存した K 個の互いに素な集合 $C_1^{(r)}, C_2^{(r)}, \dots, C_K^{(r)}$ に分割する。

$C_k^{(r)}$ で重心となる関数を $\mu_k^{(r)}(t) = \frac{1}{\#C_k} \sum_{i \in C_k} x_i(t); t \in I(r)$ とおいて、目的関数

$$E^{(r)} = \sum_{k=1}^K \sum_{i \in C_k} \int_r^{r+w} \left(x_i(t) - \mu_k^{(r)}(t) \right)^2 dt \quad (4.3)$$

各 r ごとに独立してクラスタリングを行うため、各 r でつけられたラベル $C^{(r)}$ には一意性が担保されないが、できるだけ整合するよう再ラベリングする。具体的には、以下の指標

$$G\left(C^{(r)}, C^{(s)}\right) = \sum_{i=1}^K \#C_i^{(r)} \cap C_i^{(s)} \quad (4.4)$$

が最大となるように $C^{(s)}$ のラベルを換えて、新たなクラスターラベル $C'^{(r)} = C'_1{}^{(r)}, C'_2{}^{(r)}, \dots, C'_K{}^{(r)}$ を得る。

表 4.1: 各クラスターに属する個体の推移

	$C_1^{r_{j+1}}$	$C_2^{r_{j+1}}$	$C_3^{r_{j+1}}$	合計
$C_1^{r_j}$	415	0	140	555
$C_2^{r_j}$	0	716	300	1016
$C_3^{r_j}$	0	58	128	186
合計	415	774	568	1757

4.4 数値実験

提案手法を人工データに適用して、その有効性を確かめる。まず、150 個の関数データ $x_i(t); i = 1, 2, \dots, 150$ を生成させる。本実験では B-スプライン基底関数を用いた。全体の定義域を $I = [1, 5000]$ とし、節点 (ノット) を等間隔に 50 個おき、また次数は 5 とした。基底関数 $\phi_1(t), \phi_2(t), \dots, \phi_{24}(t)$ に対する係数行列 $c_{ib}; i = 1, 2, \dots, 150; b = 1, 2, \dots, 24$ を正規分布 $c_{ib} \sim N(\mu_b, 1)$ により生成させた (表 4.4)。 t を $t_j = j; j = 1, 2, \dots, 5000$ とし、関数データ $x_i(t_j) = \sum_{b=1}^{24} c_{ib} \phi_b(t_j)$ を得る。図 4.1 は生成した関数データを示しており、係数を発生させた分布ごとに色分けされている。図 4.2 に従来の関数 k -means 法を適用した結果を示す。また、図 4.3, 4.4 に提案手法を、それぞれ $w = 400$ および $w = 200$ とし適用した結果を示す。 $t_j = 1000$ から $t_j = 3000$ 付近の間で、クラスターの構成が大きく変化するが、提案手法ではその変化を捉えていることがわかる。一方、従来の関数 k -means 法では、部分的な区間におけるクラスター構成の変化は捉えられていない。定義域の幅である w については、 $w = 200$ は $w = 400$ に比べ、局所的な変化を反映したクラスター構成の変化を表せている (例えば $t_j = 3000$ 付近)。

表 4.2: 実験で用いた分布のパラメータ

	μ_b	乱数の個数
$1 \leq b \leq 7$	0	40
$15 \leq b \leq 24$	2	50
	-2	60
$8 \leq b \leq 14$	3	10
	6	90
	1	50

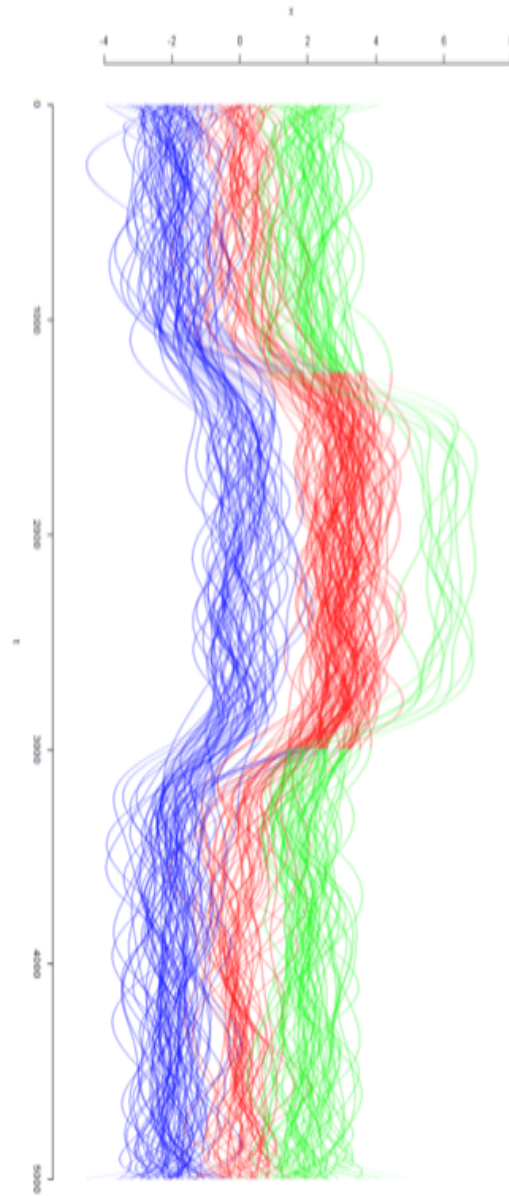


図 4.1: 実験に用いた関数データ

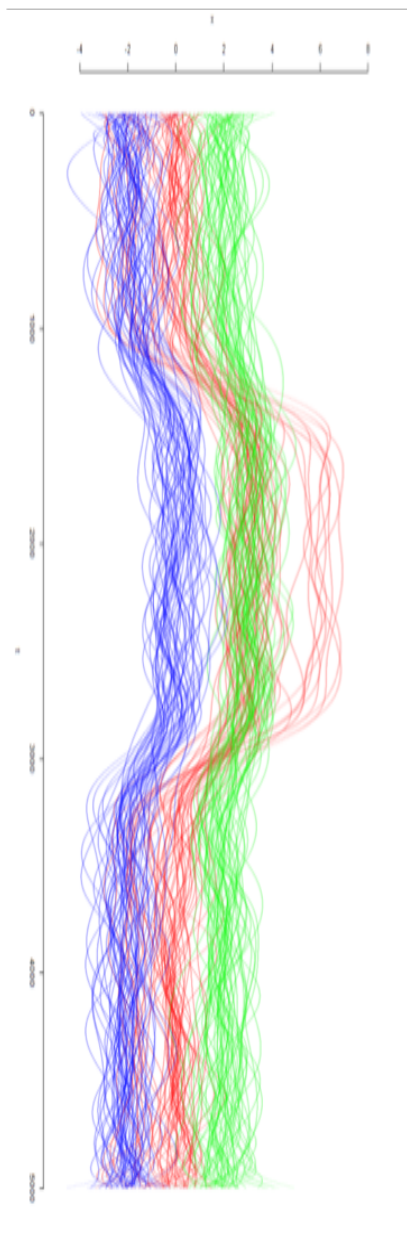


図 4.2: 従来の関数 k-means 法

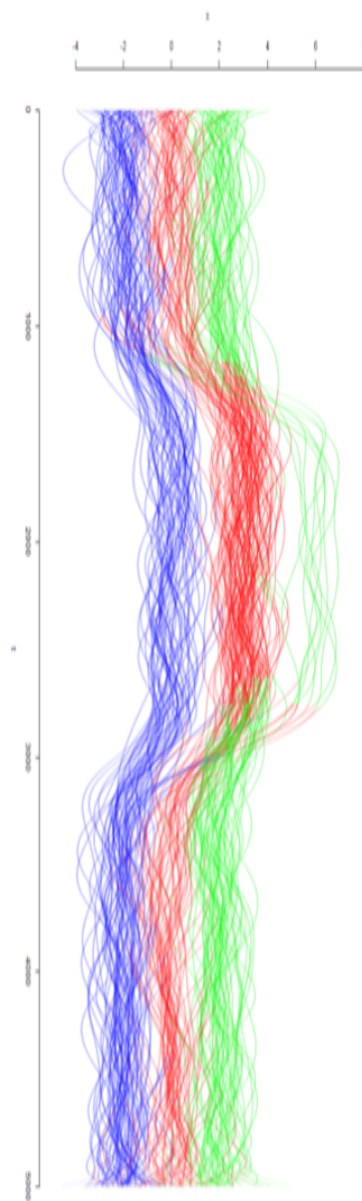


図 4.3: 提案手法 ($w=400$)

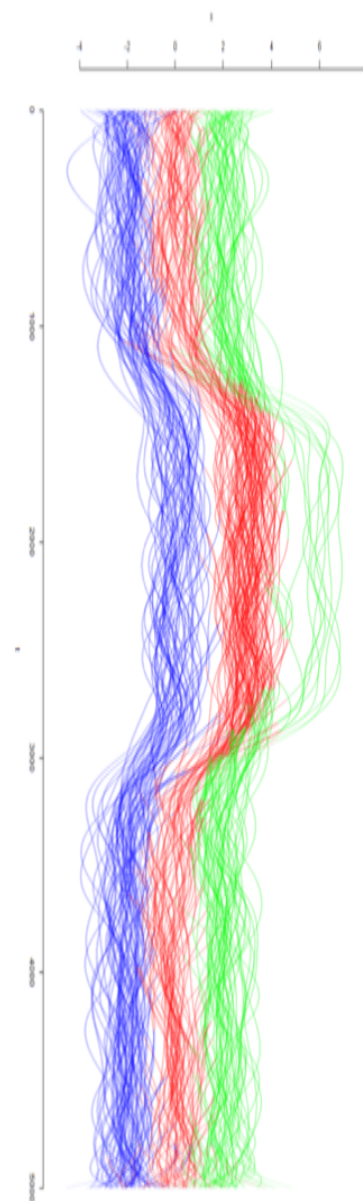


図 4.4: 提案手法 ($w=200$)

4.5 適用例

提案した移動関数クラスタリング法を、福島県の空間線量率を測定したセンシングデータに対して適用する。本データは原子力規制委員会によって公開されているモニタリングデータ (<http://radioactivity.nsr.go.jp/map/ja/>) である。このデータは、2011 年 3 月 11 日に起きた東北地方太平洋沖地震の影響で、東京電力福島第一原子力発電所で発生した事故による環境への影響を調べるためであり、全国 3921 箇所 に順次設置されたセンサー（線量計）を用いて、空間線量率 ($\mu\text{Sv/h}$) を 10 分間隔で測定している。福島県には最も多くのセンサーが設置され (2013 年 8 月 15 日時点で 2380 箇所)、各地域ごとにラベル（会津、南会津、いわき、相双、県北、県南、県中）がつけられている。本論文では、福島県の会津地域およびいわき市を合わせた計 872 箇所のセンサーにおいて、2012 年 10 月 1 日から 2013 年 5 月 1 日までに測定されたデータを用いる。測定は 10 分ごとになされているため、欠損データがなければ各センサーで 30529 個のデータが得られている。

ここではそれらの原データを $x_i^*(t_j); i = 1, 2, \dots, 872; j = 1, 2, \dots, 30529$ とおく。なお、1 ヶ月以上にわたり連続して欠損しているセンサーは 7 か所あったが、872 箇所には含めずに予め除外した。関数データ解析の前処理として、B-スプラインによる平滑化を施し、データが欠損している時刻では補間を行った。平滑化基準には一般化クロスバリデーションを用い (Hastie et al., 2009)、また各関数データに対して、 $x_i(t_j) = \frac{x_i^*(t_j) - E[x_i^*(t_j)]}{\text{Var}[x_i^*(t_j)]}$ によって標準化を施した。 $x_i(t_j); i = 1, 2, \dots, 872; j = 1, 2, \dots, 30529$ に対し、提案手法を適用する。定義域の大きさ w は 2 週間 ($6 \times 24 \times 14 = 2016$) とし、10 分ごとに定義域を移動させて $I(r) = [r, r + w] (r = 1, 2, \dots, 30529)$ とした。また 10 分ごとに再ラベリングを行った。図 4.5 に従来の関数 k-means 法 (クラスター数, $K = 2$) を適用した結果を示す。図 4.6, 4.8, 4.10, 4.12 に、提案手法である移動関数 k-means 法 ($K=2,3,4,5$) を適用した結果を示す。図 4.5 では、12 月から 4 月にかけて大きく変化している関数とそうではない関数とのクラスターの特徴を捉えているが、10 月から 12 月および 4 月から 5 月におけるクラスターの特徴は解釈できなかった。一方、移動

関数 k -means 法 (図 4.6) では、12 月から 4 月では従来の関数 k -means 法と同様なクラスターを検出しているが、10 月から 12 月および 4 月から 5 月の期間においても、解析に用いた全期間における平均値 ($x_i(t_j) = 0$) よりも高いクラスターと低いクラスターに大きく分かれていることが分かる。図 4.5 と図 4.6 の比較から、12 月から 4 月におけるクラスターの構成と 10 月から 12 月および 4 月から 5 月におけるクラスターの構成は異なるものであると考えられる。さらに 12 月から 4 月の期間について詳細に調べると、この期間周辺では、会津地方は積雪量が多くなる反面、いわき市ではほとんどない (気象庁データベース、2012 年 12 月から 2013 年 5 月にかけての「最深積雪」データを参照: <http://www.data.jma.go.jp/obd/stats/etrn/>)。クラスターの構成がこの期間で大きく変化した要因に、積雪の影響が考えられる。同様に、 K が 3 以上の場合でも 12 月から 5 月におけるクラスターの構成が変化している様子が表れている。また、その期間の前半と後半でクラスターが大きく変化したり、異常に低い外れ値なども見出される。ただし、クラスターが急激に変わるところなど、解釈が困難な部分も残されている (図 4.8, 4.10, 4.12)。

4.6 おわりに

本論文では、関数クラスタリングにおいて定義域を逐次的に移動させて解析する移動関数 k -means 法を提案した。従来の関数クラスタリングにおいて、定義域についてはあまり考慮されていなかったが、移動平均における期間のように、部分的な区間ごとの関数データに着目して分析することで、従来のクラスタリングでは検出がしにくいようなクラスターの特徴を検出することができる。本手法を適用した、福島県における空間線量率を測定したセンシングデータの例では、積雪のある冬期間にクラスターの構成が変化する様子を捉えることができた。関連した研究として、Mizuta (2004), Mizuta and Kato (2007), Tokushige et al. (2007) では、個体間の非類似度を関数 $s_{ij}(t)$ として定義し、各 t ごとにクラスタリングを行う関数クラスタリング法や多次元尺度構成法を行う関数多次元尺度構成法を提案している。それらに対して本

手法では、関数データの部分的な変化を捉えることができる点に特徴がある。最近の関数データ解析法の研究では、多次元の関数データに対する新たな手法も提案されている (Berrendero, Justel & Svarc, 2011; Jacques et al., 2014)。適用例で用いたデータの分析では、空間線量率の他に気温や降雨量、積雪量などの複数事象に基づく関数データも合わせて、より詳細な分析を進めていくことが必要で、その手段として多次元関数データの解析手法は重要と考えられる。今後の課題として、本手法の拡張を含め、多次元関数データ解析における具体的手法の開発をしていきたい。

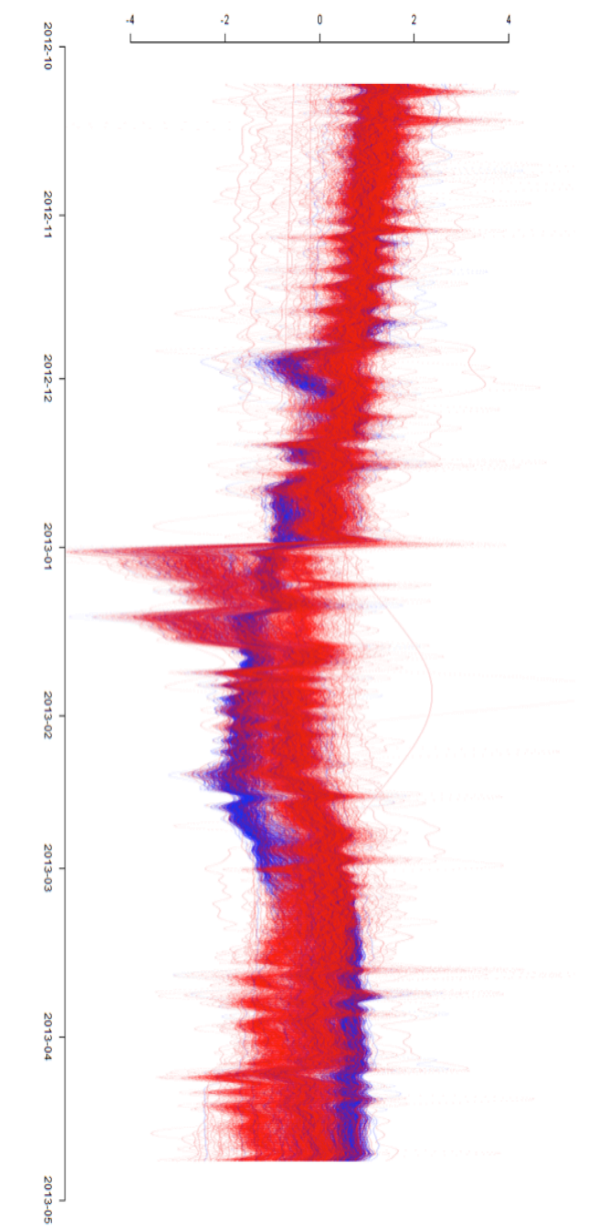


図 4.5: 従来の数 k -means 法を適用した結果 ($K=2$)

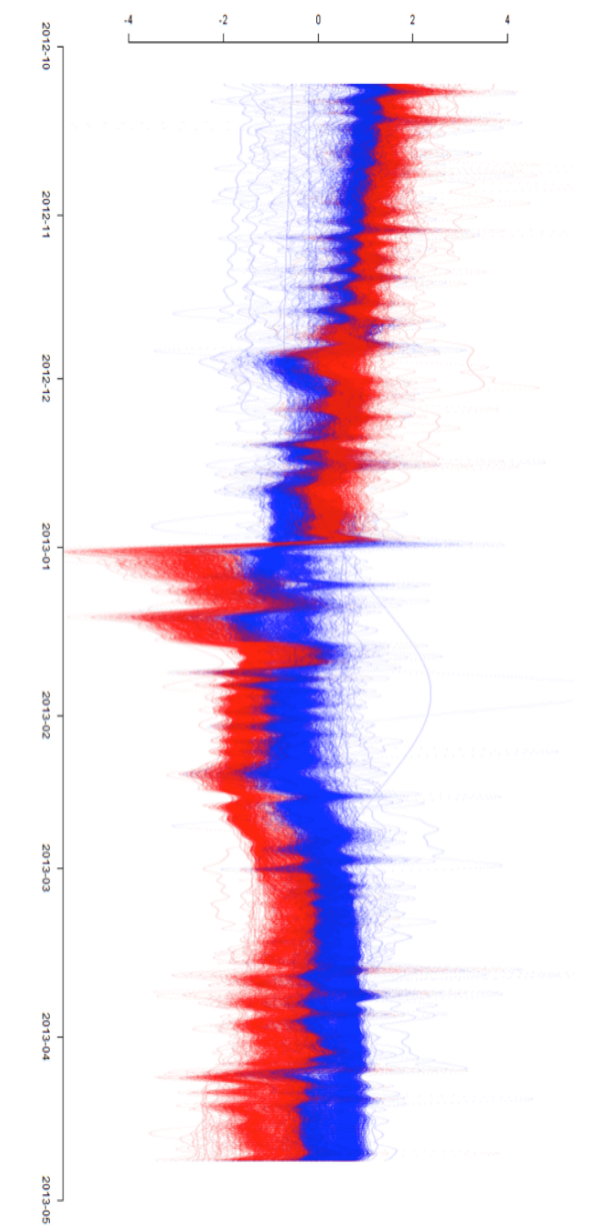


図 4.6: 移動関数 k -means 法を適用した結果 ($K=2$)

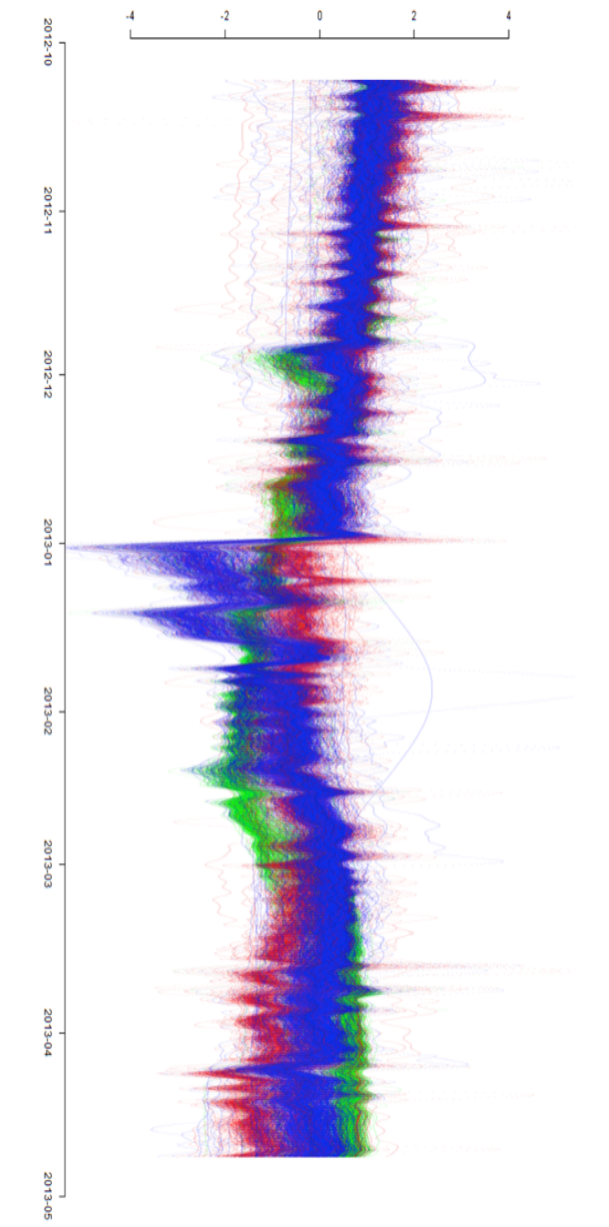


図 4.7: 従来の関数 k-means 法を適用した結果 ($K=3$)

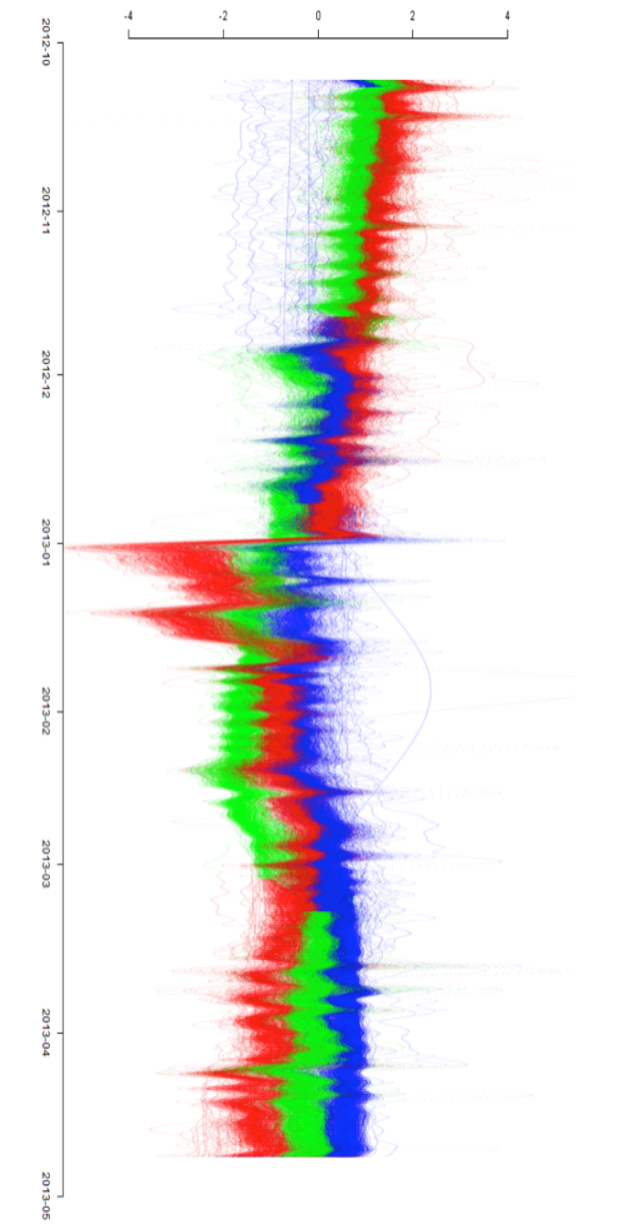


図 4.8: 移動関数 k-means 法を適用した結果 ($K=3$)

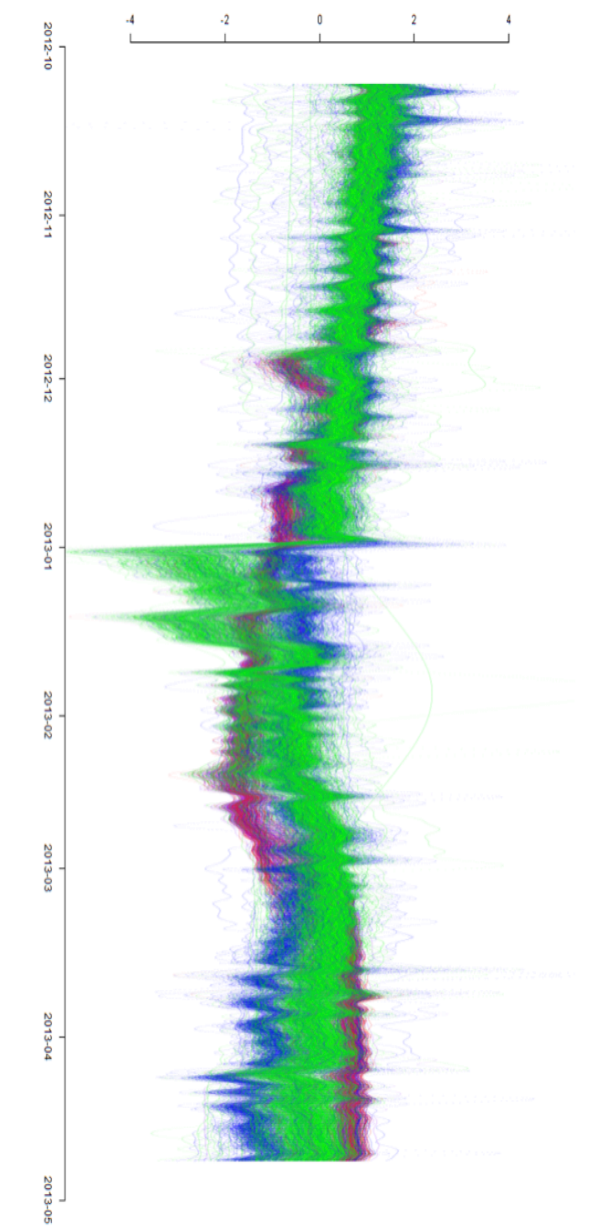


図 4.9: 従来の関数 k-means 法を適用した結果 (K=4)

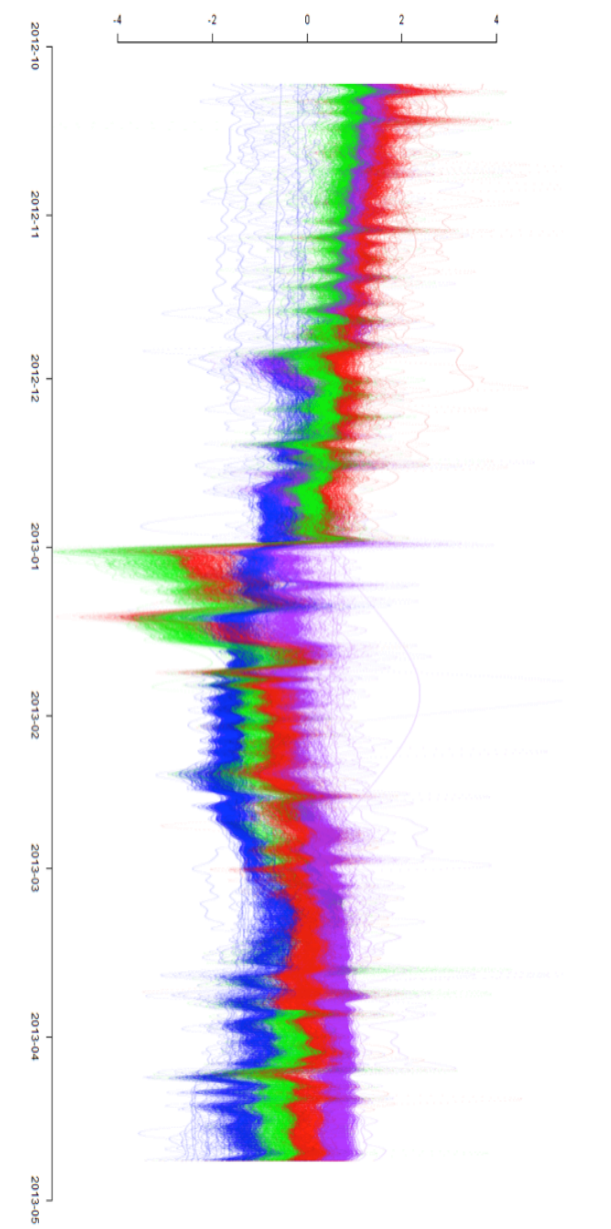


図 4.10: 移動関数 k-means 法を適用した結果 (K=4)

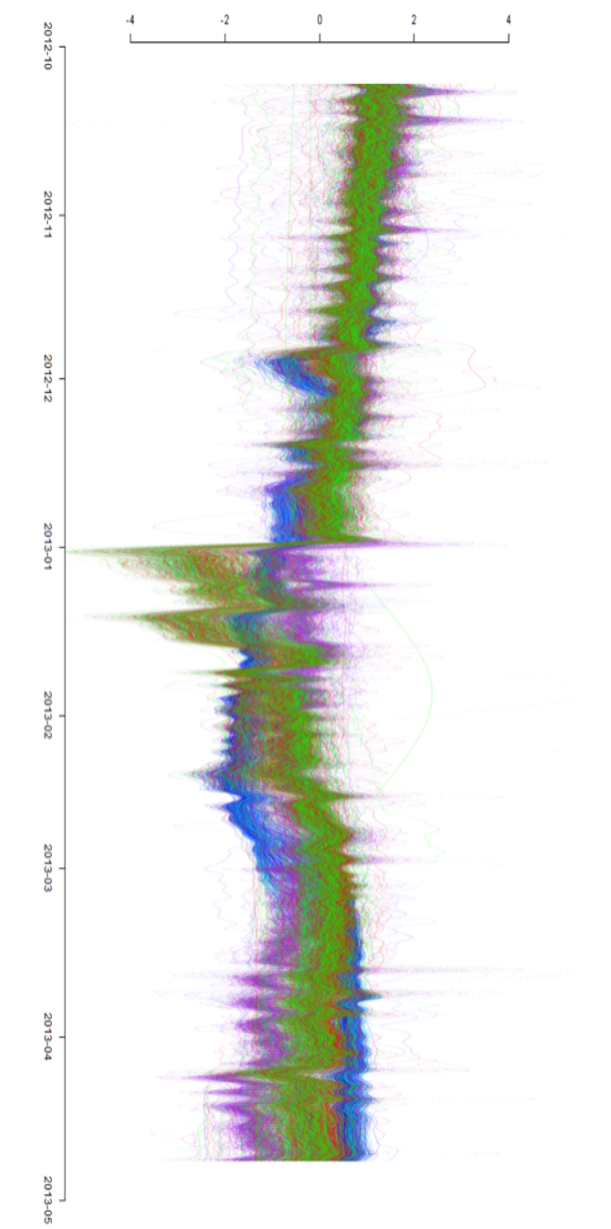


図 4.11: 移動関数 k-means 法を適用した結果 ($K=5$)

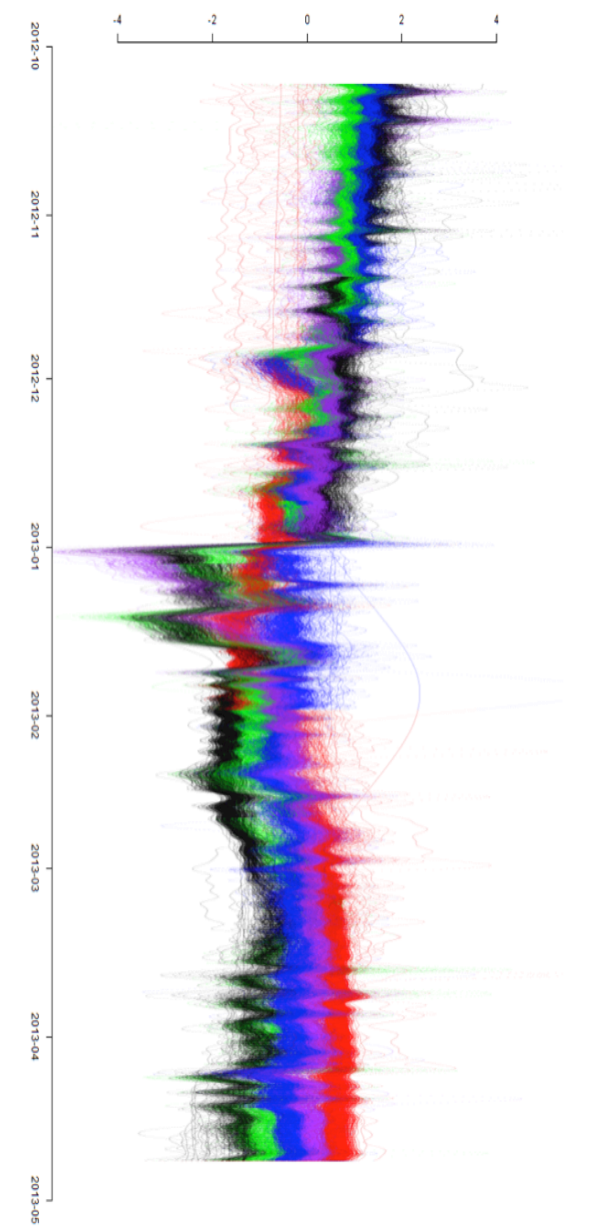


図 4.12: 移動関数 k-means 法を適用した結果 ($K=5$)

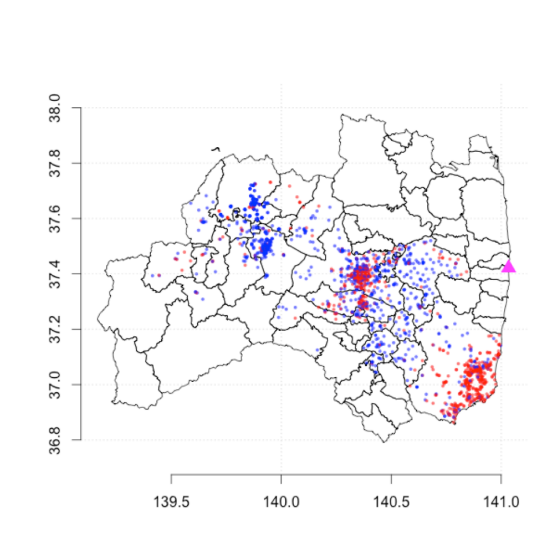


図 4.13: 提案手法によるクラスターと地図との対応関係 ($K=2$, 2012 年 11 月)

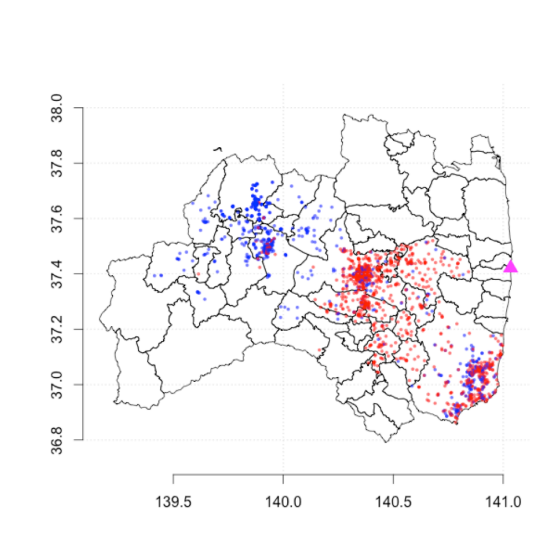


図 4.14: 提案手法によるクラスターと地図との対応関係 ($K=2$, 2013 年 1 月)

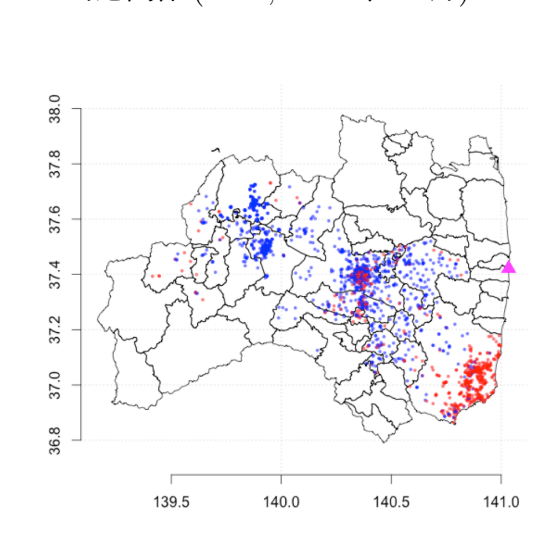


図 4.15: 提案手法によるクラスターと地図との対応関係 ($K=2$, 2013 年 4 月)

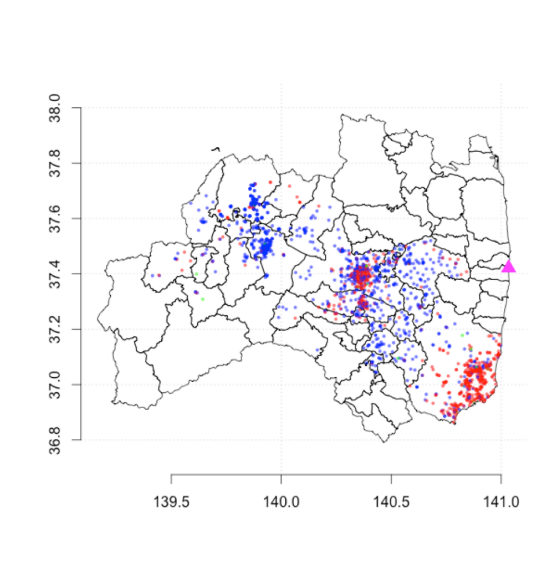


図 4.16: 提案手法によるクラスターと地図との対応関係 (K=3, 2012 年 11 月)

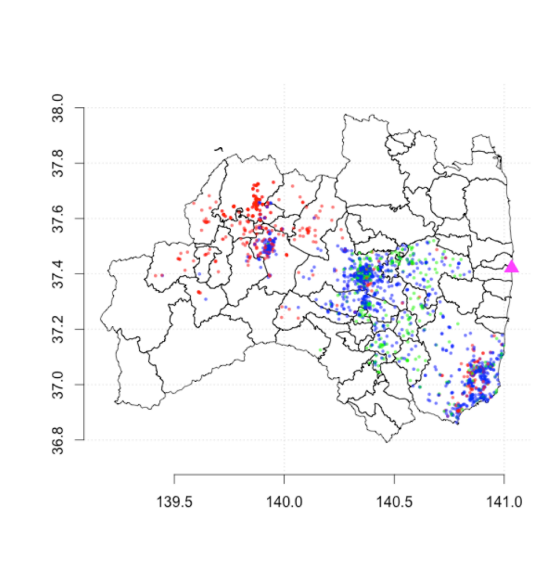


図 4.17: 提案手法によるクラスターと地図との対応関係 (K=3, 2013 年 1 月)

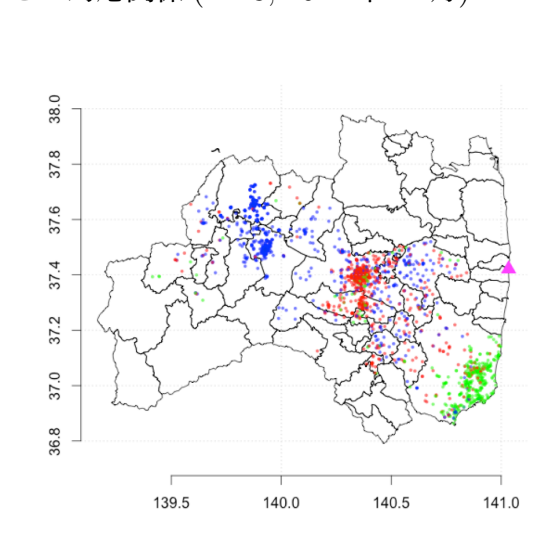


図 4.18: 提案手法によるクラスターと地図との対応関係 (K=3, 2013 年 4 月)

第5章 多次元関数値における L_0 正則化を用いた異常領域の同定法

5.1 はじめに

近年のモバイル技術やクラウド技術等の著しい発達に伴って、ビッグデータが注目されている。ビッグデータの特徴は主に、3V – Volume (容量)・Velocity (更新頻度)・Variety (多様性) – であると言われることが多い。また、「従来のアプローチでは対応できないデータ」がビッグデータであるとする考え方もある (Manyika *et al.*, 2011)。ビッグデータが関連する分野は科学、産業、公共政策など多岐に渡っており社会に対する影響が大きく、データ解析技術への必要性が高まっている。

センシングデータは多数のセンサーにより大量のデータが連続的に観測される典型的なビッグデータの例である。センサー技術が向上し広く普及している現在、その解析法の開発は社会的に重要な課題である。近年では、多数のセンサーを用いて構築されたワイヤレスセンシングネットワークより得られるデータに対して、その解析手法などが注目を集めている (Yang *et al.*, 2010; O'Reilly *et al.*, 2014)。

多数のセンサーにより測定されるデータを扱う際の一つの重要な問題は、データに含まれる異常値への対応である。センサーにはバッテリーなどの物理的制約が伴うため、常に理想的な状態で測定できるわけではない。さらに通信機能があるセンサーでは、通信時に電磁波の外乱が生じる。また、センサーは、観測者が介在しにくいような厳しい環境で用いられることも多い。そのため、予め想定し得ないような環境の変化も起こり得り、そのような場合には、平常時とは全く異なるような測定値が観測されることがある (O'Reilly *et al.*, 2014)。

このように、センシングデータには外れ値や連続的な異常などが含まれている。そのようなデータから何らかの推測をする際には、異常値を区別して扱うべきであり、したがって、異常値の検出は実用上重要である。中でも多数のセンサーによる測定では、複数のセンサーで同時に「連続的な異常」が発生した期間（異常領域）を検出しなければならないことがある。しかし一般的に異常領域の検出は、従来の外れ値検知手法だけでは十分に対応することが難しく、新たなアプローチが必要である。

センシングデータは連続的に得られるため、観測時点ごとに離散的に扱うことは、必ずしも適切とは言えない。一方で、センシングデータは時刻に対して測定値を与える「関数」と見なすことができる。関数をデータとして解析する手法として近年、関数データ解析 (Ramsay & Silverman, 2005) が注目されている。関数データ解析では、時刻に対して一つの値が与えられるデータを扱うことが多いが、複数のセンサーで測定される場合は、時刻に対して各次元ごとに値を返すデータとみなすことができる。

本論文では、値域が多次元である関数データの異常領域を検出する手法を提案する。また福島県における空間線量率を測定したセンシングデータに本手法を適用し、その有効性を示す。

5.2 多次元値関数データと異常領域

5.2.1 関数データと多次元値関数データ

関数データ解析 (Functional Data Analysis) は 1990 年代に Ramsay らによって提案された、関数をデータとして解析する手法であり、これまでに数多くの研究がなされている (Ramsay & Silverman, 2002, 2005; Ramsay *et al.*, 2009)。

関数データ解析が有効である場面はいくつか考えられる。例えばデータが関数として入手できる場合や、多変量データなどに平滑化などを施すことでデータを関数として扱える場合である (水田, 2005)。実際の応用ではデータが関数として得られる場面は少なく、データに平滑化を施すことで関数データとみなすことが多い。Ramsay &

Silverman (2005) でも平滑化に多くの説明が割かれている。

データを関数として扱うことの利点として、微分が可能になることがあげられる。Ramsay & Silverman (2002) でも微分を用いた解析が有効であるような例が示されている。センシングデータのように連続的に観測されるデータは、関数データ解析が適用できる典型的な例である。

以下では、値域が多次元となる関数データ、すなわち $t \in \mathbf{R}$ に対して $\mathbf{y}(t) \in \mathbf{R}^p$ ($p \geq 1$) が定義される関数データを扱う。このような関数データを「多次元値関数データ (Multi Dimensional valued Functional Data)」と呼ぶことにする。

5.2.2 外れ値と異常領域

一般に、測定値には外れ値や異常領域が含まれており、それらに対しての留意が必要である。これまでに外れ値について多くの研究がなされているが (Aggarwal, 2013; Hawkins, 1980; Kriegel *et al.*, 2010)、定義の一例として、

An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism.

がよく知られている (Hawkins, 1980)。この定義では、「1 つ」の外れ値に重きが置かれている。しかし、関数データでは異常領域が問題となる。ここでは、複数の次元において同時発生する異常領域を「同時異常領域」と呼ぶことにする。

以下では、同時異常領域を同定する手法を提案する。

5.3 提案手法

同時異常領域の同定において、各次元における異常領域を同定してから、ある一定の区間で異常領域である次元を数え上げることで同定する方法を考える。

センサーから得られたデータには異常との区別が困難なノイズが多く含まれているため、最初にそれらを取り除く必要がある。ノイズをフィルタリングするために用いられている方法は多く存在しており (Henry, 1988; Jansen, 2001)、分析の目的に応じて選択するのが一般的である。関数データ解析では、原データから関数を近似して得るための主要な手段として平滑化を用いることが多く、ここでも平滑化によってノイズをフィルタリングするアプローチをとる。

平滑化についても多くの手法が存在するが (de Boor, 2001; Hastie *et al.*, 2009)、本論文の主な目的である、異常領域の検出をできるだけ明確に行えるような手法が求められる。近年、 L_0 正則化による平滑化を用いることで、エッジを強調した平滑化を行う手法が提案された (Rippe *et al.*, 2012; Xu *et al.*, 2011)。これを応用することで、ノイズのフィルタリングと同時に、異常領域が強調された平滑化を施すことが可能である。そこで本論文では Rippe *et al.* (2012) によって提案された手法を応用し、 L_0 正則化による平滑化をデータに施す。

ここまでで得られる、平滑化を施しノイズを除去したデータから、共通したルールによって各次元の異常領域を特定するために、閾値に基づいて3つの領域（「異常に高い」「異常に低い」「その他」）に区分し、一定数以上の複数次元で同時に「異常領域」とされた値域を、「同時異常領域」として同定する。

5.3.1 準備

次元 i ($i = 1, 2, \dots, N$) で時刻 t に観測されたデータを $y_i(t)$ とおき、多次元値関数データを $\mathbf{y}(t) = (y_1(t), y_2(t), \dots, y_N(t))^T$ とおく。また実際の計算では t を離散値として扱うので、離散的な表現を $\mathbf{y}(t_k) = (y_1(t_k), y_2(t_k), \dots, y_N(t_k))^T$ ($k = 1, 2, \dots, K; t_1 < t_2 < \dots < t_K$) とする。

ここで、時刻に対する線形的なトレンドを考慮する必要がある場合には、 $\{t_k, y_i(t_k); k = 1, 2, \dots, K\}$ に対し、 t_k を説明変数、 $y_i(t_k)$ を目的変数とした単回帰により得られた係

数を a_i, b_i とおき、

$$y_i(t_k) - (a_i t_k + b_i) \quad (5.1)$$

を新たな $y_i(t_k)$ とする。

5.3.2 データの中心化とスケーリング

3つの領域の境界付近の値は、実際には非常に不安定となることが多く、大域的に捉えにくい。そこで、 $\{y_i(t_k)\} (k = 1, 2, \dots, K)$ における両側 $\alpha\%$ によるトリム平均を $\bar{y}_i^{(\alpha)}$ 、スケーリングの定数を $h_i := \frac{4}{\max_k(y_i(t_k))}$ とおき、

$$x_i(t_k) = h_i \left(y_i(t_k) - \bar{y}_i^{(\alpha)} \right) \quad (5.2)$$

とし、領域の境界を明確にする。

5.3.3 L_0 正則化による平滑化

$x_i(t_k)$ に対して L_0 正則化を用いた平滑化により、エッジのみを強調した関数 $z_i^*(t_k)$ を得る。すなわち目的関数

$$\begin{aligned} R(\mathbf{z}_i^*) &= \sum_{k=1}^K (x_i(t_k) - z_i^*(t_k))^2 \\ &+ \lambda_i \sum_{k=2}^K |z_i^*(t_k) - z_i^*(t_{k-1})|_0 \end{aligned} \quad (5.3)$$

を最小にする $\mathbf{z}_i^* = (z_i^*(t_1), z_i^*(t_2), \dots, z_i^*(t_K))^T$ を求める。ただし、 λ_i は平滑化パラメーターで、

$$|a|_0 = \begin{cases} 0 & (a = 0) \\ 1 & (a \neq 0) \end{cases}$$

と定義する。(5.3) の第 1 項はデータへの当てはまりを、第 2 項は全体の滑らかさを表している。ここでの「滑らかさ」とは、各時刻 t_k の前後で値の変化回数が全体として少ないことを指す。

一般に、 L_0 正則化の最適化は難しい。そこで、繰り返しによって近似解に収束させる方法が研究されている。以下では Rippe *et al.*(2012) を応用した計算方法を示す。

(5.2) に基づいて $\mathbf{x}(t_k) = (x_1(t_k), x_2(t_k), \dots, x_N(t_k))^T$ とおき、また、あるベクトル $\mathbf{a}$ の 0-ノルムを $\|\mathbf{a}\|_0$ で表すと、(5.3) の行列表記は、

$$R(\mathbf{z}_i^*) = \|\mathbf{x}_i - \mathbf{z}_i^*\|^2 + \lambda_i \|D\mathbf{z}_i^*\|_0 \quad (5.4)$$

となる。ただし、 D は $(K-1) \times K$ の行列で

$$D = \begin{pmatrix} -1 & 1 & 0 & \dots & 0 & 0 \\ 0 & -1 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & \dots & -1 & 1 \end{pmatrix} \quad (5.5)$$

である。

繰り返し回数を r として $\mathbf{z}_i^{*(r)} = (z_i^{*(r)}(t_1), z_i^{*(r)}(t_2), \dots, z_i^{*(r)}(t_K))^T$ とおく。 $\mathbf{z}_i^{*(1)}$ と $\lambda_i^{(1)}$ に適切な初期値を与え、

$$\mathbf{z}_i^{*(r)} \rightarrow (I + \lambda_i^{(r-1)} D^T V_i^{(r-1)} D)^{-1} \mathbf{x}_i \quad (5.6)$$

により、(5.3) の値に基づく収束条件を満たすまで、 $\mathbf{z}_i^{*(r)}$ を更新し $\mathbf{z}_i^*$ を得る。ただし、 V_i は $(K-1) \times (K-1)$ の行列で

$$V_i^{(r)} = \text{diag}(-z_i^{*(r)}(t_k) + z_i^{*(r)}(t_{k-1}); \\ k = 2, \dots, K) \quad (5.7)$$

とする。

また $\lambda_i^{(r)}$ については、 $\lambda_i^{(r)} = \delta \lambda_i^{(r-1)}$ ($0 < \delta < 1$) として単調減少させる。

5.3.4 閾値に基づく領域の分割

3.3 節で得られた $\mathbf{z}_i^*$ に対して、以下の規則を適用して 3 つの領域に分ける。

$$z_i(t_k) = \begin{cases} 1 & (z_i^*(t_k) > 0.5) \\ 0 & (|z_i^*(t_k)| \leq 0.5) \\ -1 & (z_i^*(t_k) < -0.5) \end{cases} \quad (5.8)$$

以上により多次元値関数データを 3 つの領域に区分した $\mathbf{z}(t_k) = (z_1(t_k), z_2(t_k), \dots, z_N(t_k))^T$ を得る。

本論文では異常値の存在を仮定して、平常値および異常値（説明の都合上、高い方）の分布を考える。平常値の平均を 0、異常値の実現値の平均を ν_i 、分散を s_i^2 とおく。平常値と異常値を分離できると仮定して、その平均の差を $\nu_i - 0 = 4s_i$ 、異常値の最大値を、分布を考慮して $\max_k(y_i^*(t_k)) = \nu_i + 4s_i$ と設定する。異常値の閾値を c_i とおくと、 $\frac{4|x_i|}{\max_k(y_i^*(t_k))} \geq c_i$ より $|x_i| \geq \frac{c_i(\nu_i + 4s_i)}{4} = 2c_i s_i$ となる。従って、平均 $\pm\sigma$ に対応させ、閾値を 0.5 とした。

5.3.5 同時異常領域の検出

3 つの領域に分けられた多次元値関数データに対し、同時異常領域の検出を行う。すなわち一定期間以上（例えば、 $\{t_{l+1}, t_{l+2}, \dots, t_{l+m}\}$ ）にわたり、 β ($\beta \geq 2$) 個以上の次元で発生している同時異常領域を求める。ただし、 $l \geq 0$, $K \geq l + m$ である。

$\{-1, 1\}$ のいずれかの値をとる定数を u とおく。ここで、 $k = 1, 2, \dots, K$ に対して、

$$T_{i,k}^{(u)} = \begin{cases} 1 & (z_i(t_k) = u) \\ 0 & (z_i(t_k) \neq u) \end{cases} \quad (5.9)$$

であるベクトル $T_i^{(u)} = (T_{i,1}^{(u)}, T_{i,2}^{(u)}, \dots, T_{i,K}^{(u)})^T$ を考える。さらに、すべての k に対して $S_k^{(u)} = \sum_{i=1}^N T_{i,k}^{(u)}$ を計算し、それらを要素に持つベクトルを

$$S^{(u)} = (S_1^{(u)}, S_2^{(u)}, \dots, S_K^{(u)})^T$$

とおく。 $S^{(u)}$ は各時刻で値 u をとる次元の個数である。同時異常が観測された時刻を得るには、 $S_k^{(u)}$ が β 以上となる時刻の添え字 k のみを抽出すればよいから、

$$\arg_k \{S_k^{(u)} \geq \beta\} \quad (5.10)$$

とし、必要に応じて小さい順に並べ変えて、同時に異常が発生した時刻の添え字 k を要素とするベクトル $\mathbf{b} = (b_1, b_2, \dots, b_p)^T$ ($p \leq K$) を得る。一定期間以上にわたる同時異常領域を同定するには、 $\mathbf{b}$ から m 個以上の連続した列の集合を全て抽出すればよい。

また、 $\{t_{l+1}, t_{l+2}, \dots, t_{l+m}\}$ で異常領域が生じている次元 i は

$$\arg_i \left\{ \sum_{k=l+1}^{l+m} T_{i,k}^{(u)} = m \right\} \quad (5.11)$$

として特定する。

5.4 数値実験

人工データに対して提案手法を適用して評価する。真のデータ系列点を 1000 個生成させ、それぞれを平均とした分散 1 の正規乱数をノイズとして加える。実データにおけるセンサー数を意図して、3 つのセットを作成した。図 5.1, 5.4, 5.4 にノイズを加える前のデータ点と加えた後のデータをプロットした図を示す。横軸には系列点の番号 ($1 \leq k \leq 1000$)、また縦軸は生成した数値を表している。

本実験では、異常に低い領域を検出する。ここで真の異常領域を下がり始めの時点から上がり始めの時点までの期間とする。表 5.4 に、3 つのセットにおける真の異常

領域を示す。

表 5.1: 人工データの異常領域

	真の異常領域
セット 1	$401 \leq k \leq 700$
セット 2	$601 \leq k \leq 1000$
セット 3	$401 \leq k \leq 700$

これらのデータに対して、提案手法を適用した。ただし、 k に対する線形関係は含まれていないと仮定する。また (5.2) におけるトリム平均は両側 10% で計算した。 L_0 正則化項の係数は初期値に $\lambda_i^{(1)} = \frac{2}{\sqrt{\text{var}(x_i(t_k))}}$ を与え、係数の変化率を $\delta = 0.999$ とし平滑化を施した。

それぞれのセットに対する適用結果を図 5.4, 5.4, 5.4 に示す。点線は実線に対して L_0 正則化項による平滑化を施した結果を表している。また、破線は点線に対して (5.8) の閾値に基づいて 3 つの領域に分割した結果を表している。また検出した異常領域を表 5.4 に示す。セット 1 および 3 における $k = 401$ は、明らかに異常な変化点であるが、提案手法により正しく検出できている。また、検出しにくいような、緩やかに変化する領域 (図 5.4 $600 \leq k \leq 700$ の領域など) についても概ね検出できていることがわかる。図 5.4 の $k = 500$ 付近ではごく一時的に異常ではない値をバースト的なノイズとして含めたが、そのような場合でも、大域的な異常領域を検出する点において、対応できていることがわかる。

最後に、二つ以上のセットで異常が起こった場合の、同時異常領域の検出を行った結果を図 5.4 に示す。横軸には系列点の番号、縦軸には異常が検出されたセットの数を表している。このときの同時異常領域は $401 \leq k \leq 677$ である。

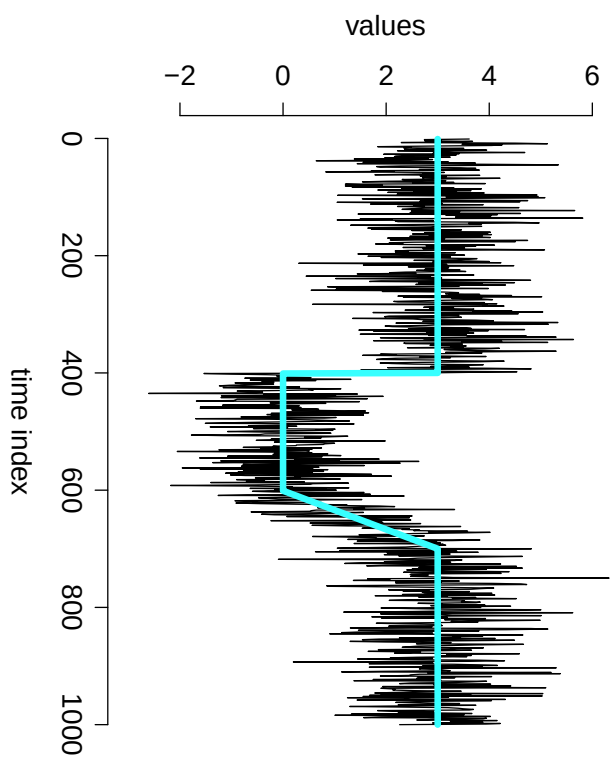


図 5.1: 実験に用いた人工データ (セット 1)

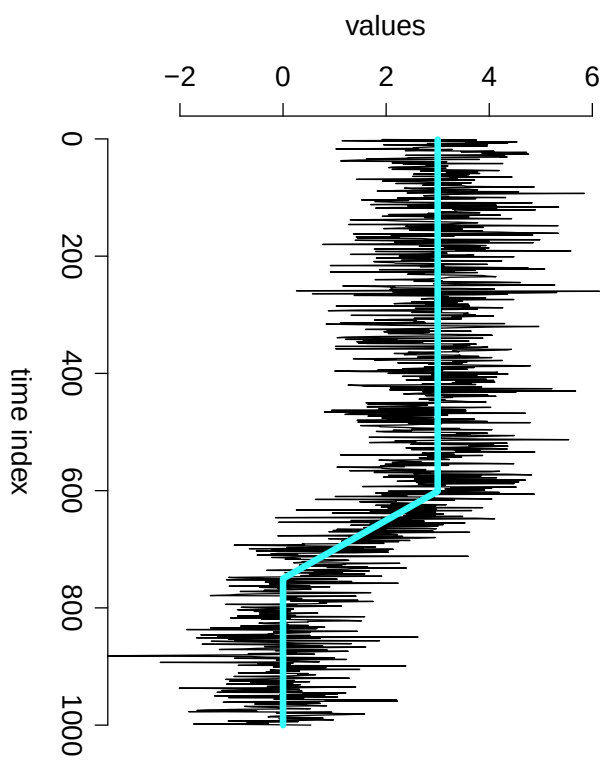


図 5.2: 実験に用いた人工データ (セット 2)

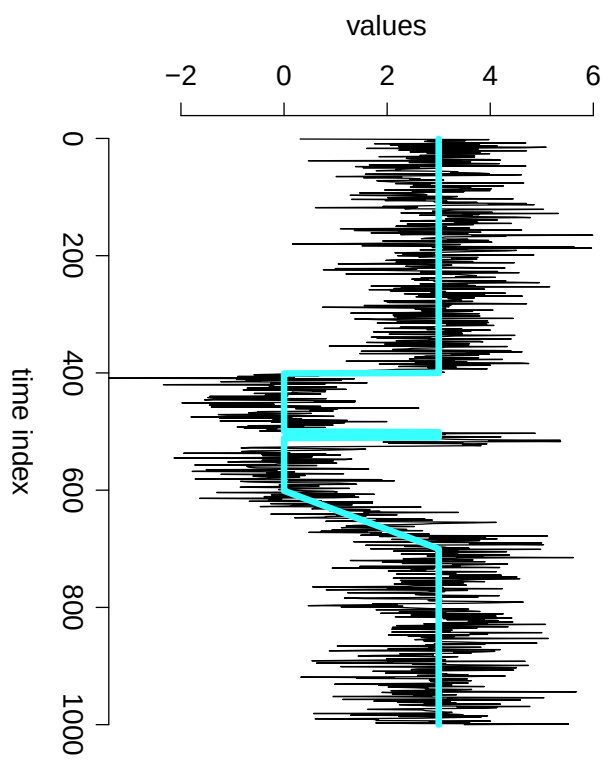


図 5.3: 実験に用いた人工データ (セット 3)

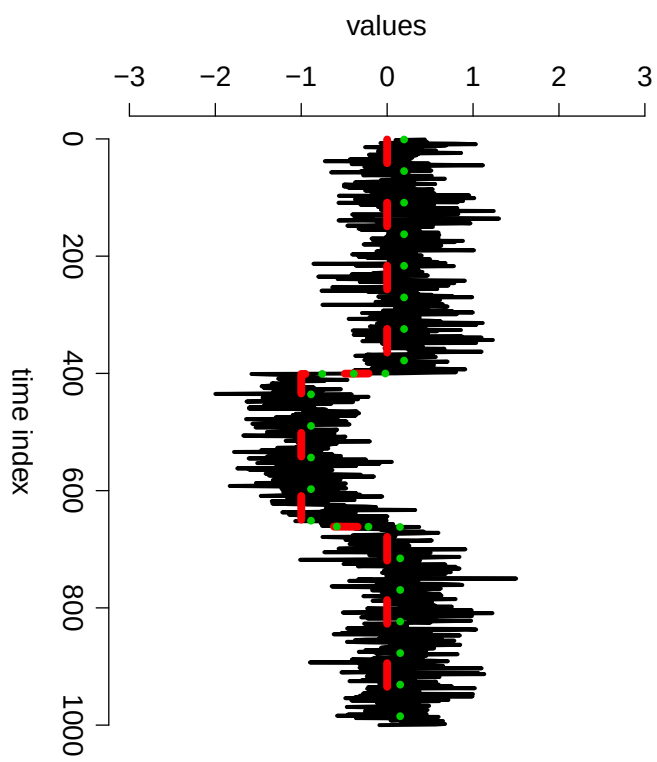


図 5.4: 適用結果 (セット 1)

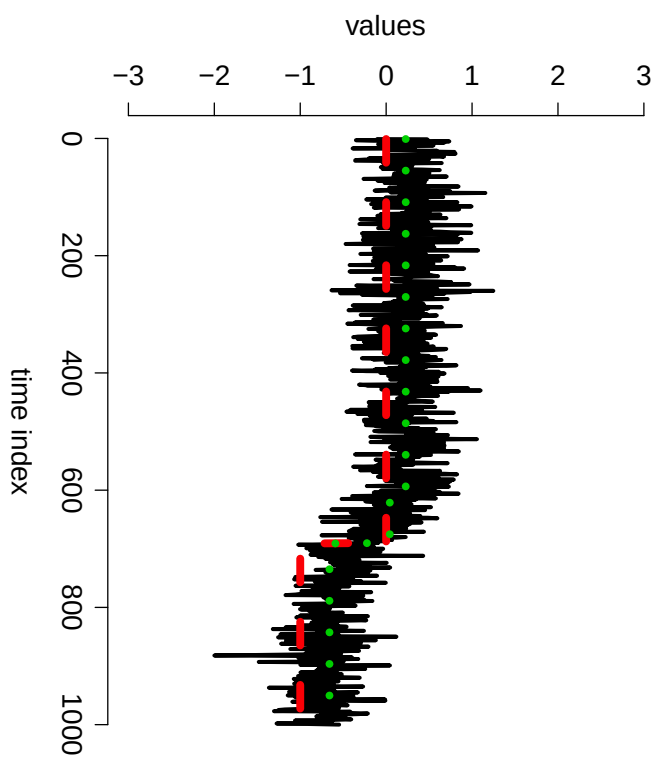


図 5.5: 適用結果 (セット 2)

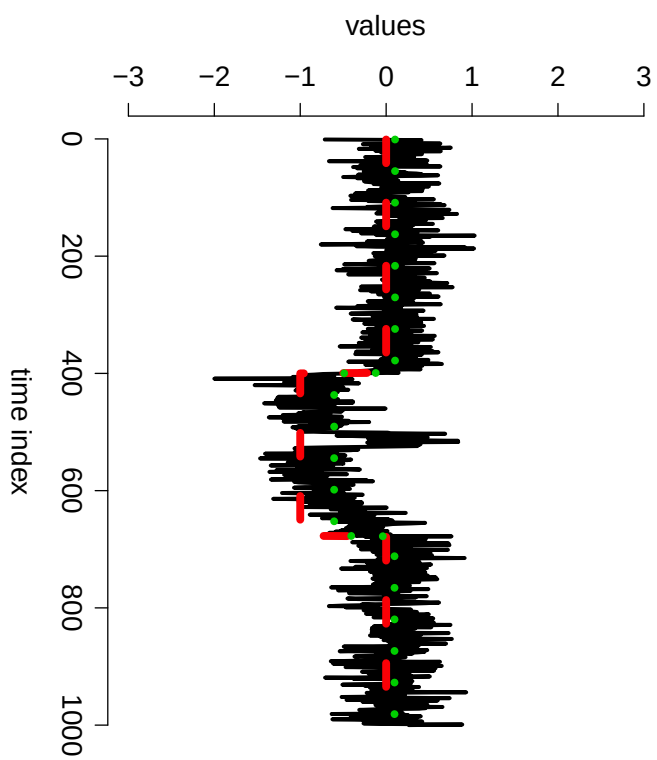


図 5.6: 適用結果 (セット 3)

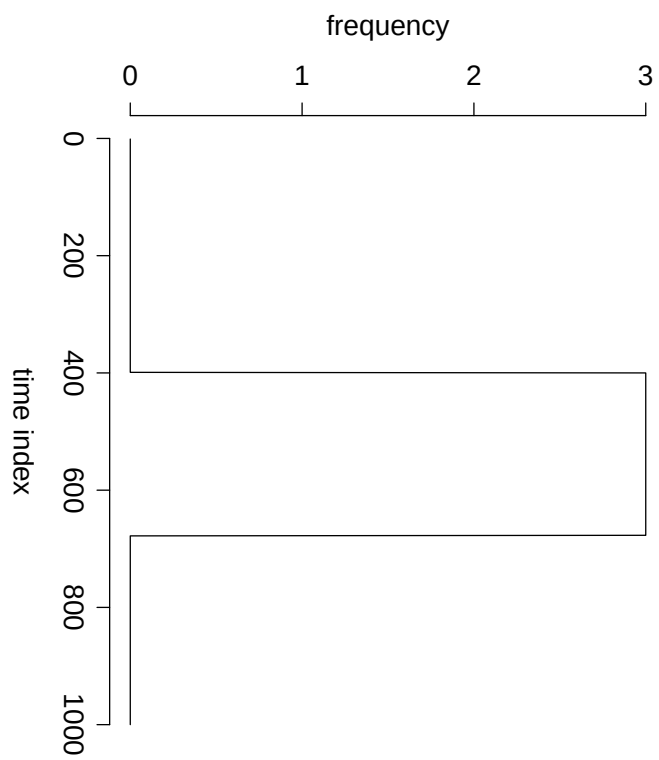


図 5.7: 異常が検出されたセットの数

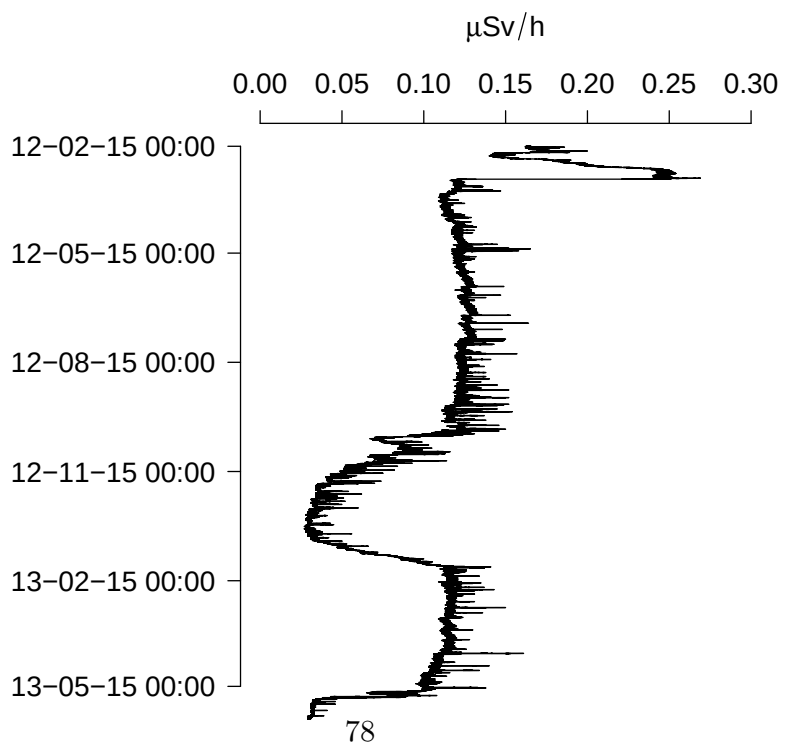


図 5.8: 異常領域が含まれていると疑われる例

表 5.2: 異常領域の検出結果

	検出した異常領域
セット 1	$400 \leq k \leq 601$
セット 2	$691 \leq k \leq 1000$
セット 3	$400 \leq k \leq 677$

5.5 実際の適用例

原子力規制委員会により公開されている「東京電力福島第一原子力発電所事故による環境モニタリング等データベース」の一部として公開されているものを例として用いる。データベースには航空機サーベイによる測定データやカーサーベイによる測定データ、土壌深度分布に関するデータなどが多数公開されており (<https://mapdb.jaea.go.jp/mapdb/>)、近年、多角的な観点から統計的手法を用いた研究が進められている (石岡 他, 2013; 鈴木 他, 2013; 松井 他, 2013)。ここでは、データベースの一部として公開されている、福島県における空間線量率をセンシングしたモニタリングポストデータに提案手法を適用した例を示す。

モニタリングポストデータは東京電力福島第一原発事故後から全国 3921 か所に順次設置されたセンサー（線量計）により、空間線量率 ($\mu\text{Sv/h}$) を 10 分間隔で測定したデータである。これらのデータはリアルタイムで公開されており、ウェブサイト (<http://radioactivity.nsr.go.jp/map/ja/>) より地点ごとのデータを確認したり、CSV ファイルをダウンロードすることができる。全データの数データベースのレコード数に換算すると 3 億程度である (2013 年 8 月 15 日時点)。福島県には最も多くのセンサーが設置されており (時点で 2380 箇所)、各地区ごとにラベル (会津、南会津、いわき、相双、県北、県南、県中) がつけられている。

数十か所のセンサーで測定されたデータについて予備分析を行ったところ、大量のノイズに加えて、途中、連続的に高い (または低い) 値を観測している期間が散見された (図 5.8)。詳細な分析に先立ち、そのような異常領域がいつ、どこで生じているのかを把握する必要がある。

本適用例では、2012 年 02 月 15 日 00 時 00 分～2013 年 08 月 15 日 00 時 00 分 ($K = 78769$) において福島県の会津地域 ($N = 510$) およびいわき市 ($N = 473$) で観測されたデータを対象に異常領域の同定を行い、さらに異常領域が検出されたセンサーを特定する。会津地域とiwaki市のセンサーで観測されたデータはそれぞれ、510 次元および 473 次元の多次元値関数データとして扱う。

放射性物質には物理崩壊があるため、空間線量は長期的に減少することが予想できる。そこで、はじめに (5.1) により線形的なトレンドを除く。

(5.2) における中心化では、両側 50% に当たる $\alpha = 50$ としてトリム平均を用いて中心化し、また (5.3) から (5.7) における L_0 による平滑化では、正則化項の係数について初期値に $\lambda_i^{(1)} = \frac{2}{\sqrt{\text{var}(x_i(t_k))}}$ を与え、係数の変化率を $\delta = 0.999$ として平滑化を施した。

ある二つのセンサーにおける典型的な結果を図 5.9, 5.10 に示す。実線は、時刻に対する線形関係を除いて、トリム平均に対して中心化およびスケーリングしたものを表している。点線は実線に対して L_0 正則化項による平滑化を施した結果を表している。また、破線は点線に対して (5.8) の閾値に基づいて 3 つの領域に分割した結果を表している。

図 5.9 では 2012 年 11 月 15 日付近から 2013 年 05 月 15 日付近まで連続して平均よりもかなり低い値が表れており、全体としても 2012 年 11 月 15 日付近から 2013 年 05 月 15 日付近までが、異常領域として検出されていることを示している。図 5.10 の 2013 年 02 月 15 日付近では、連続してやや平均よりも低い傾向が見られるが全体としては異常領域はないことを示している。

次に同時異常領域を検出した。(5.9), (5.10) で用いるパラメーターは $\beta = 50$, $m = 4320 (= 6 \times 24 \times 30)$, $u = -1$ とし、50 個以上のセンサーで同時に 1 ヶ月以上にわたって異常に低い値が観測された領域を検出した。

各時刻において同時異常が生じたセンサー数を図 5.11, 5.12 に示す。横軸は時刻、縦軸は同時異常が観測されたセンサーの数である。

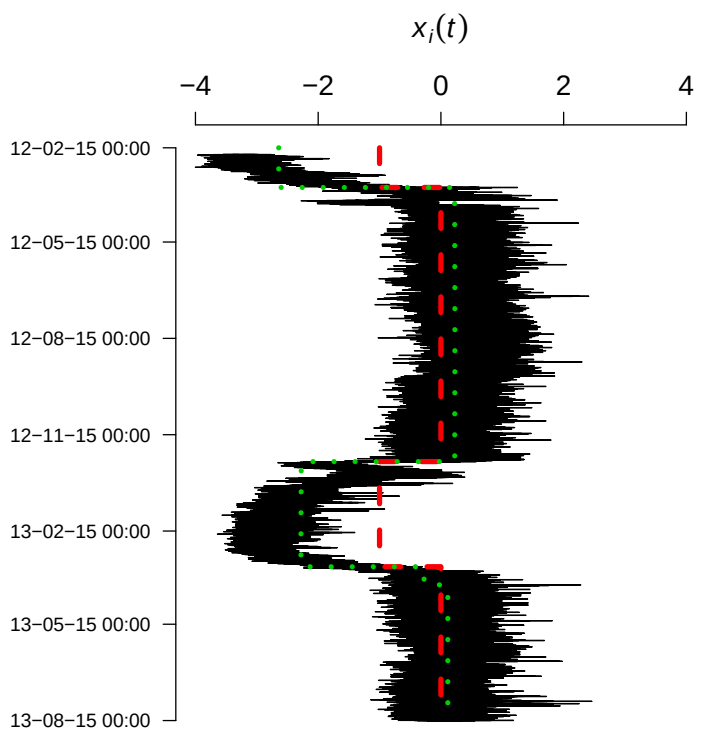


図 5.9: 3つの領域に区分した例 (連続的に低い領域が明確な例)

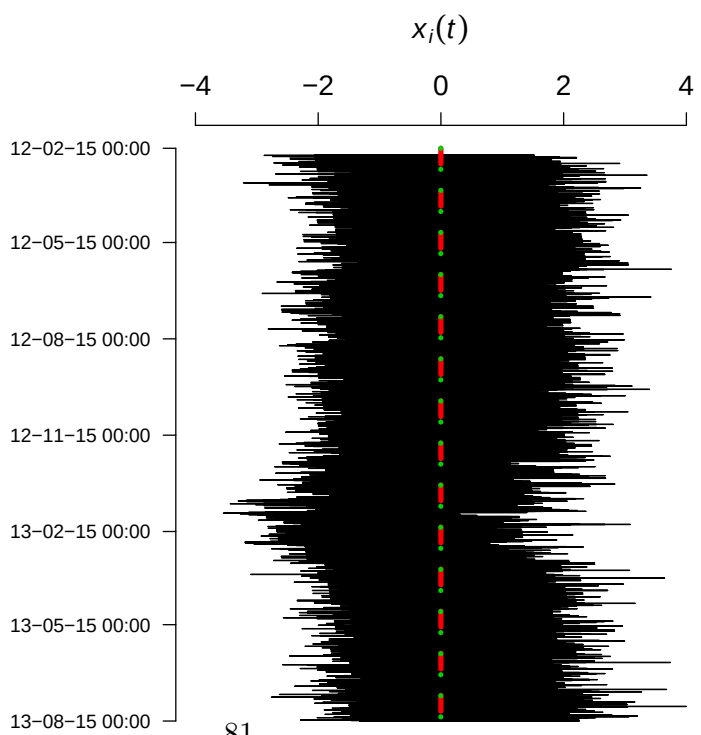


図 5.10: 3つの領域に区分した例 (連続的に低い領域が明確でない例)

図 5.11: 同時異常領域が観測された期間とセンサー数(会津
地域)

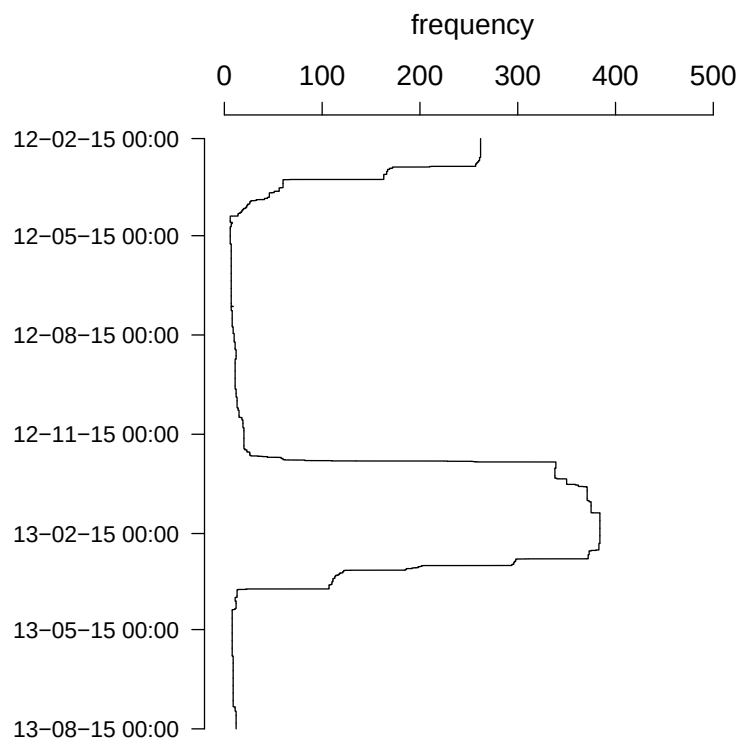
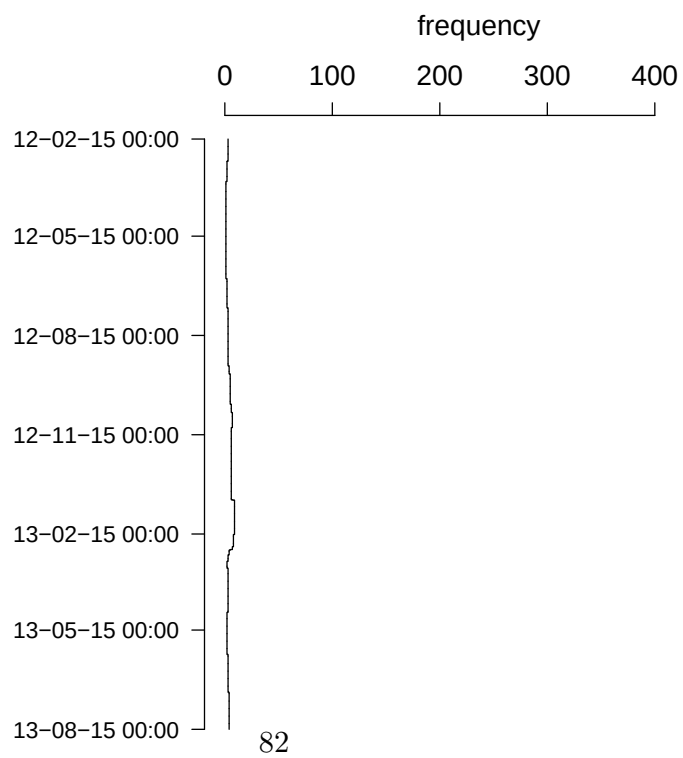


図 5.12: 同時異常領域が観測された期間とセンサー数(いわ
き市)



これらより、会津地域では 2012 年 02 月 15 日 00 時 00 分～2012 年 03 月 23 日 21 時 50 分および 2012 年 12 月 09 日 10 時 40 分～2013 年 04 月 07 日 08 時 00 分が同時異常領域として検出された。しかし一方で、いわき市では同時異常領域が一つも検出されなかった。図 5.12 を見てわかるように、いわき市では同時異常の発生数そのものが極端に少なく、したがって同時異常領域もほとんど発生していないためである。

また同時異常領域が検出された会津地域のデータに対して、2012 年 02 月 15 日 00 時 00 分～2012 年 03 月 23 日 21 時 50 分および 2012 年 12 月 09 日 10 時 40 分～2013 年 04 月 07 日 08 時 00 分で同時異常領域が発生したセンサーを同定した結果を図 5.13, 5.14 に示す。縦軸は経度（東経）、横軸は緯度（北緯）を表す。

ところで、検出された異常領域はちょうど冬期間にあたり、福島県では積雪する地域とそうではない地域がある。そこで気象庁の過去の気象データを用いて、会津地域の 4 地点（若松、猪苗代、金山、西会津）およびいわき市の小名浜における 2012 年 2 月 15 日～2013 年 8 月 15 日の 1 日ごとの最深積雪 (cm) との関連を調べた（図 5.15）。図 5.15 の各実線は 1 日ごとの最深積雪 (cm) を示し、網掛された領域は先ほど検出した同時異常領域を表す。

図 5.15 より、会津地域では、提案手法により検出された異常領域と、積雪の多い期間がほぼ重なることがわかる。一方、小名浜では最深積雪データは観測されていないか欠損していたため、代わりに小名浜の過去 10 年の積雪量を調べたところ、最も多い日でも 7cm 程度で、この地域における積雪はほとんどないと考えられる。以上から、会津地域といわき市で観測されたモニタリングポストデータに関しては、異常領域と積雪量との間に関連性があると考えられる。

5.6 おわりに

本論文では、多次元値関数データに対する異常領域の検出手法を提案し、また福島県の空間線量率を測定したセンシングデータを用いて提案手法の有効性を示した。

(5.8) における閾値は適用する分野によって調整する必要性が生じうる。異常値の

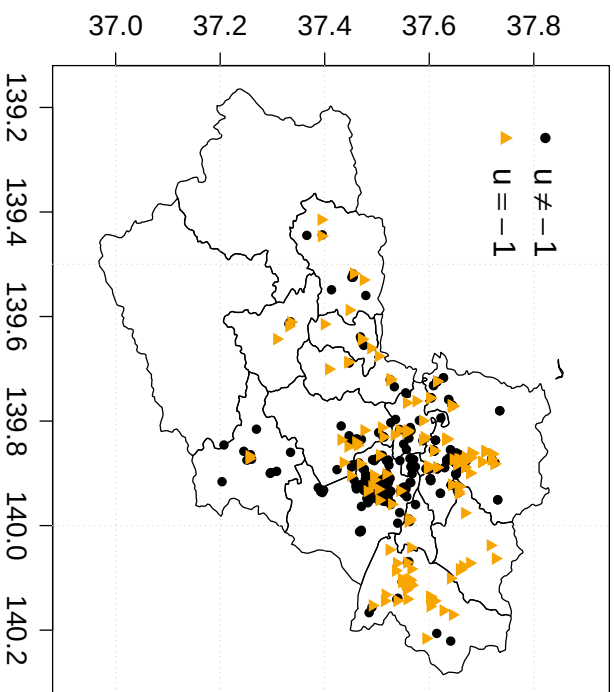


図 5.13: 会津地域において同時異常領域が1か月以上観測されたセンサー (2012 年 02 月 15 日 00 時 00 分-2012 年 03 月 23 日 21 時 50 分)

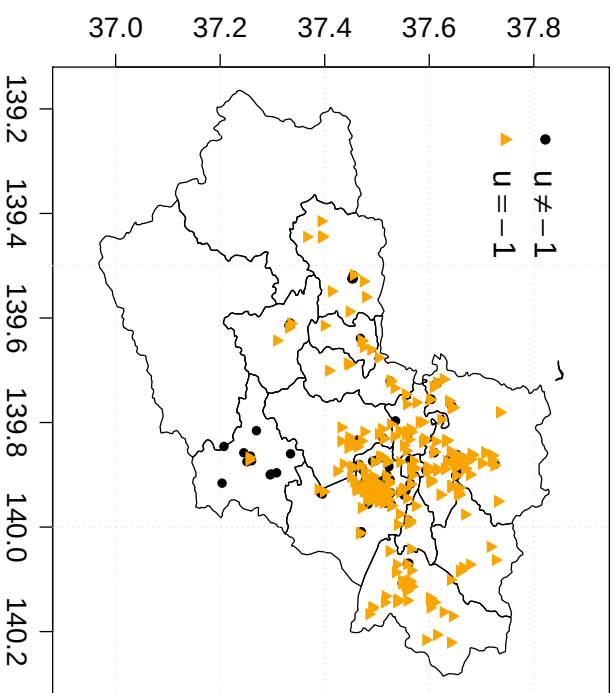


図 5.14: 会津地域において同時異常領域が1か月以上観測されたセンサー (2012 年 12 月 09 日 10 時 40 分-2013 年 04 月 07 日 08 時 00 分)

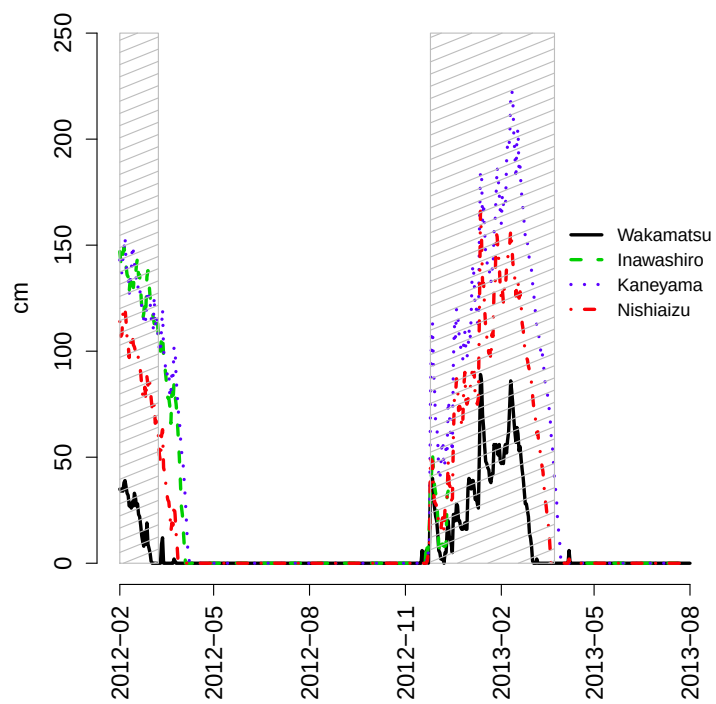


図 5.15: 会津地域の 4 箇所（若松、猪苗代、金山、西会津）における 1 日ごとの最深積雪 (cm) と検出された同時異常領域

範囲を広く（狭く）設定する場合には閾値を大きく（小さく）する。適用例では 0.5 以外にも 0.25 や 1.0, 1.5 等の閾値を用いた分析も行った。2012 年 12 月～2013 年 4 月付近で検出される同時異常領域について、0.25 とした場合には、検出される同時異常領域は約 3 日長くなり、1.0, 1.5 とした場合には、約 2 週間短くなる。積雪期間との関連では、0.25 および 0.5 がともによい解釈を与えた。しかし、0.25 の場合には、 β に対して不安定であったため、本論文では 0.5 とした。

センシングデータでは、多くのセンサーによって測定される。その場合、センサー全体におけるデータの特徴をわかりやすく把握することが可能になれば、実際の環境で起こっている現象を深く理解することができる。本適用例では、多くのセンサーにおいて異常に低い領域と積雪期間との関連を考察した。なお最近の関連研究では、放射性物質と大気（気温や降雨など）との関連なども指摘されており（青山 他, 2012）、今後、気象データを用いたより詳細な分析が必要である。

ビッグデータの解析では結果のわかりやすさが重視されることが多い。本手法により同時異常領域を検出し、結果を可視化することができる。また、これまでの関数データ解析の研究は、関数主成分分析 (Ramsay & Silverman, 2005) や関数クラスター分析 (Jacques & Preda, 2013) などといった、関数データの定義域全体に対する特徴を捉える手法が中心であったと考えられる。それに対して、今回のアプローチは関数データの部分的な特徴を捉えようとするものである。定義域が非常に大きな関数データの場合には、いくつかに分割して分析するアプローチも有効と考えられる。具体的な手法の開発と応用例については今後の課題としたい。

第6章 尺度混在型データに対する解析手法

6.1 はじめに

近年の情報技術の発達により、巨大かつ複雑なデータを扱う場面が増えつつある。生命科学から人文科学まで非常に多くの研究者や技術者がそのようなデータに注目している。ビッグデータはしばしば”3V”によって特徴づけられる。すなわち容量 (Volume)、更新速度 (Velocity)、多様性 (Variety) である。容量に対しては多くのストレージ容量とデータへの高速なアクセスが要求される。更新速度では、効率的なデータ処理方法と分析のアルゴリズムが必要である。また多様性に対しては、データベースの観点での効果的なデータ構造や解析手法が必要とされる。

データ解析の観点では、多様性という特徴は新たな問題を含んでいると考えられる。従来の解析手法では、スカラー値よりなるデータを主に扱っていた。しかし、データの種類が増えるに従って、関数や区間値、モダル、分布値などのシンボリックデータのような、構造化されたデータを扱わなければならない場面も出てきた。さらに最近では、位相幾何学的なデータに対する統計解析手法なども研究されている。

実際の場面では、データの収集方法や目的に応じて、解析対象を表すデータの記述は異なる。とりわけ、同じ解析対象に対して、異なる記述が混在しているような場合は、行列操作などの通常の操作を行えないため、その取り扱い是非常に困難である。そのような問題に対して、本論文では非類似度に着目して解析する方法を示す。非類似度は、多様なデータの記述について定義することができる。

本論文では、データの記述間の関連性に着目し、それらの線形関係性を分析する手

法を提案する。また、提案手法を、福島第一原子力発電所の事故に関連した、環境中の放射線レベルのデータに適用した例を示す。

6.1.1 ビッグデータと SDA

ここでは、ビッグデータに関連して SDA を簡潔に導入し、後に用いる用語を定義する。ビッグデータにおいて多様性が注目される所以には、非構造データの存在がある。すなわち、一般的な p 次元の n 個のベクトルでは表現できないデータである。実際、そのようなデータは従来の関係データベースの形式に収まらず、それに加えて欠損や十分にデータの形式が整えられていないことなども多い。従来の解析手法を用いるには、データの変換や除外などにより単純なデータに整形する、いわゆる”クリーニング”をしなければならない。

SDA の枠組みでは、 p 次元の n 個のベクトル形式に当てはまらないようなデータについて、自然に扱うことができる。観測対象は第一レベルの個体と第二レベルの個体の、二つのレベルに分けて考えられる。第一レベルの個体は最も基本的な観測レベルで、各観測値が p 次元ユークリッド空間における各点に対応させて考えられる。しかし、区間値のように観測値の内部に構造をもつような場合も考え得る。そのような観測対象は第二レベルの個体と呼ばれる。

SDA では、カテゴリやクラス、コンセプトなどの第二レベルの個体を扱うことが多く、その個体の集まりをコンセプト空間と呼ぶ。各コンセプトに対して、区間やモデル、分布や関数などへのマッピングに対応する操作は記述と呼ばれる。

そのような枠組みにより、複雑なデータを自然に扱えるだけでなく、またビッグデータに見られるような価値密度の低い第一レベルの個体を抽象化して、価値密度の高いデータを得ることができる。

本論文では、解析対象をコンセプトとする。また、各コンセプトに対して様々な記述が混在しているデータを尺度混在型データと呼び、その解析手法を提案する。

6.2 尺度混在型データの解析法

コンセプト空間 C を考え、その要素を $c_i \in C (i = 1, 2, \dots, n)$ とおく。記述を $Y_m : C \rightarrow S_m (m = 1, 2, \dots, p)$ と定義する。ただし、 S_m は記述の空間である。各コンセプトに対する記述を $Y(c_i) = (Y_1(c_i), Y_2(c_i), \dots, Y_p(c_i))$ とおく。

各記述に対して、非類似度を $d_m : C \times C \rightarrow \mathbf{R}^+$ と定義する。 Y_m におけるコンセプト間の非類似度を $D^{(m)} = \{d_{k,l}^{(m)}; k, l = 1, 2, \dots, n\}$ とおく。 $D^{(m)}$ に基づいたユークリッド空間における付置を考える。

一般に、非類似度からユークリッド空間における付置を構成する方法として多次元尺度構成法 (MDS) が用いられる。ここでは、MDS の中でも計量的 MDS を用いる。付置の次元数を k_m とし、その座標値を $X^{(m)} = (X_1^{(m)}, \dots, X_{k_m}^{(m)})$ とおく。各記述に対して MDS を適用して、全体として得られる空間は $\mathbf{R}^M = \mathbf{R}^{\sum_{m=1}^p k_m}$ 次元における付置

$$\begin{aligned} X &:= (X^{(1)}, X^{(2)}, \dots, X^{(p)}) \\ &= (X_1^{(1)}, X_2^{(1)}, \dots, X_{k_1}^{(1)}, X_1^{(2)}, X_2^{(2)}, \dots, X_{k_2}^{(2)}, \dots, X_1^{(p)}, X_2^{(p)}, \dots, X_{k_m}^{(p)}) \end{aligned}$$

を得る。付置された点 X を尺度混在型データにおけるプローブとして用いることにより、記述間の関連性を分析する。

付置された点間の関連性を分析する方法にはいくつかアプローチがあるが、ここでは、主成分分析法を用いる。主成分分析法は、元々の空間における点間の分散をできる限り保つようにして、初めの次元数よりも少ない次元数の直交空間に射影する方法である。結果として得られる空間は共分散行列の固有ベクトルの張る空間となる。

$(\sum_{m=1}^p k_m) \times (\sum_{m=1}^p k_m)$ の共分散行列を

$$\Sigma = (X - E[X])^T (X - E[X]) \quad (6.1)$$

と定義する。固有方程式は、

$$\gamma_l \Sigma = \gamma_l \mathbf{a}_l, \quad (6.2)$$

であり、固有値を γ_l ($1 \leq l \leq \sum_{m=1}^p k_m$; $\gamma_1 \geq \gamma_2 \geq \dots \geq \gamma_{k_m}$), 固有ベクトルを $\mathbf{a}_l$ とおく。 N 個の固有ベクトルによる累積寄与率を、

$$\nu_N = \frac{\sum_{l=1}^N \gamma_l}{\sum_{l=1}^M \gamma_l}, \quad (6.3)$$

とおく ($0 \leq \nu_N \leq 1$)。

6.2.1 非類似度について

前節における手続きに先立って、各記述に対する非類似度を選択しなければならない。汎用性の高い重要な非類似度として、Gowda-Diday 非類似度 (Gowda and Diday, 1999) と Ichino-Yaguchi 非類似度 (Ichino and Yaguchi, 1994) がある。Gowda and Diday(1999) は、シンボリックデータの特徴を位置、長さ、コンテンツに分けて考え、非類似度を定義した。Ichino and Yaguchi(1994) では、シンボリックデータをカルテシアン空間モデルにより表現することを提案し、一般化ミンコフスキ距離を導入した。区間値やモダル、ヒストグラム値、木における非類似度は、これらの定義で対応できる。さらに一般的な非類似度として一般化ハウスドルフ距離などもある (Bock and Diday 2000)。以下では、典型的な例を示す。

区間値 A, B を区間値とする。すなわち、区間の下限を l_a, l_b とし、上限を u_a, u_b とおくとき、 $A = [l_a, u_a]$, $B = [l_b, u_b]$ とおく。このとき、非類似度を

$$d_{A,B} = (|l_a - l_b| + |u_a - u_b| - 2|\max(l_a, l_b) - \min(u_a, u_b)|)$$

と定義する。

分布値 A, B を分布値とする。ワッサーシュタイン距離を適用する。 Y_i と Y_j を A, B の確率変数とおき、分布関数をそれぞれ F_i と F_j とおく。このとき、非類似度を $d(A, B) = \int_0^1 |F_i^{-1}(q) - F_j^{-1}(q)|^2 dq$ と定義する。

関数値 関数間の非類似度として、 L_2 ノルムがしばしば用いられる。すなわち、関数が $L = [a, b] (a, b \in \mathbf{R}, a < b)$ 上で $t \in L$ に対して定義されるとき、非類似度を $d(f_i, f_j) = \int_a^b |f_i(t) - f_j(t)|^2 dt$ と定義する。

6.3 福島県における線量率データへの適用

東日本大震災において発生した福島第一原子力発電所事故後より、文部科学省と環境省が中心となり環境における放射能レベルの測定が行われた (<http://radioactivity.nsr.go.jp/en/>)。様々な観点から測定が行われており、空間的に網羅的に測定したデータや、中長期的な放射能レベルの推移を測定したデータなどがある。

このようにして得られたデータセットの種類は非常に多いため、複数のデータセットによる情報をまとめた上で意思決定をしなければならない。しかし、それぞれのデータセットにおけるデータの特徴は、それぞれ微妙に異なっており、測定座標やメッシュ区分、測定時刻がずれている場合や、互いに同じ地域・期間での放射能レベルを測定したものであっても測定値が異なる場合などがある。また、データの記述について、あるデータセットでは分布地データが適切であるが、別のデータセットでは関数データが適切である場合もある。このようなデータに対して、データセット間で同様な測定結果が得られているのかを調べる必要がある。

ここでは、3つのデータセット：走行車サーベイ、航空機サーベイ、モニタリングポストを用いて、提案手法を適用する。いずれのデータセットも、福島県における空間線量率を測定したデータである (http://radb.jaea.go.jp/mapdb_prev/en/)。

走行車サーベイは車道に沿って線量率を測定している。航空機サーベイはヘリコプターにより上空からセンサーにより地上 1m の線量率を測定している。またモニタリ

ングポストは地上 3263 箇所に設置されたセンサーにより 10 分ごとに線量率を測定している。

コンセプトとして各市町村を考え ($B_j \in \Omega, j = 1, 2, \dots, 55$)、それらのコンセプトに 3 つの記述を与える。すなわち $Y_1(B_j)$ =走行車サーベイによる測定結果, $Y_2(B_j)$ = 航空機サーベイによる測定結果 and $Y_3(B_j)$ = モニタリングポストによる測定結果である。

図 6.1 は 2012 年 12 月頃における福島県の放射線レベルを表したヒートマップである (<http://ramap.jmc.or.jp/map/eng/>)。図 6.1 左の黒線は、航空機が測定時に飛行した軌跡を表している。

図 6.2 はそれぞれの記述に対して、集約したデータを示している。 Y_1 と Y_2 を分布値として、 Y_3 を関数として表した。モニタリングポストについては、2012 年 11 月から 2012 年 12 月における測定値を用いた。各地域より 10 個のセンサーをランダムに抽出して、平均関数を用いた。これらの記述間の関連性を分析する。

一つの方法は、各コンセプトの記述毎に数値の平均をとって、その相関を調べることである。その結果を表 6.3 に示す。しかし、分布のばらつきや時間的変動の情報は失われる。

次に提案手法を適用する。 Y_1 および Y_2 における非類似度をワッサーシュタイン距離によって定義し、 Y_3 における非類似度を L_2 ノルム距離によって定義する。表 6.3 と図 6.3 に MDS の結果を示す。固有値の結果から各記述のユークリッド空間における次元数は $k_1 = 1, k_2 = 1$ and $k_3 = 2$ とした。得られた付置における関連性を調べるために主成分分析を適用する (表 6.3, 図 6.4)。結果からは、記述間に明らかな関連性は見られなかった。すなわち、各データセット間に明確な線形性は見られなかった。一方、楢葉町 (図 6.4 右上) と郡山市 (図 6.4 左下) および川俣町 (左上) などの外れ値が見て取れる。それらに対応する原データを詳しく調べると、楢葉町は航空機サーベイにおいて異常に大きな測定値が含まれ、郡山市は走行車サーベイにおいて大きな測定値が、また川俣町はモニタリングポストにおいて異常に大きな測定値が含まれていた。

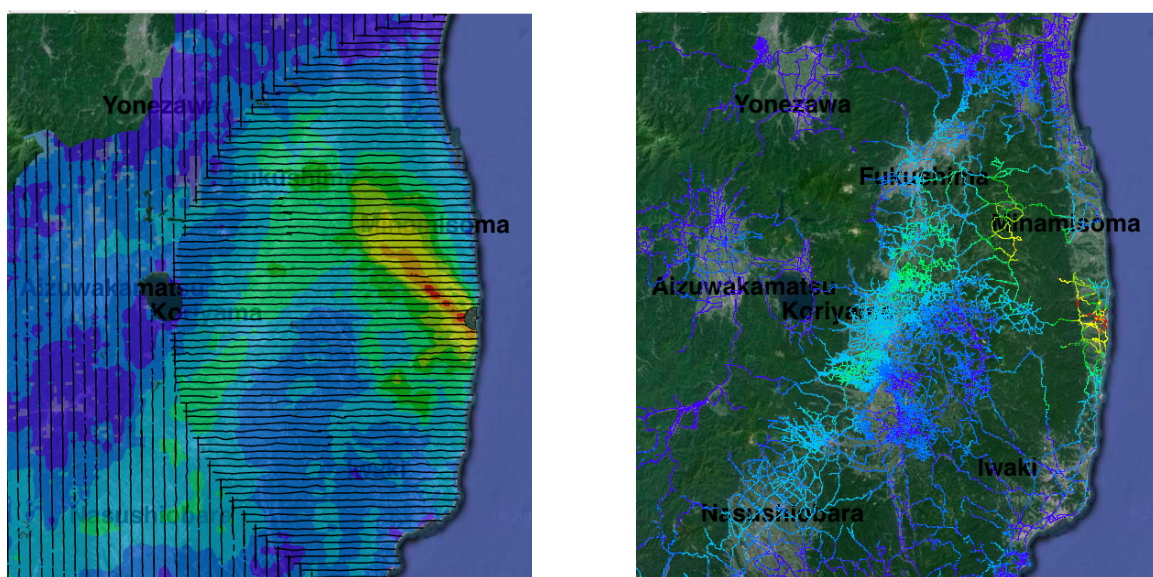


図 6.1: 空間線量率の空間分布。左図は航空機により 2012 年 12 月頃に観測された測定値を示し、右図は走行車により 2012 年 11 月頃に観測された測定値を表す。左図中、黒線は航空機が飛行した軌跡を表している

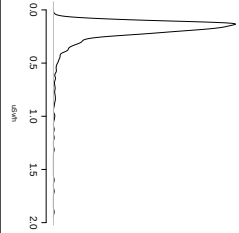
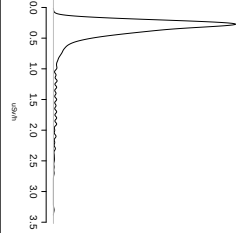
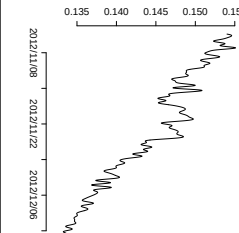
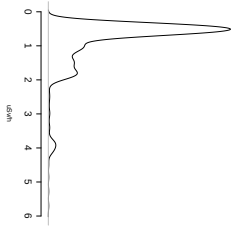
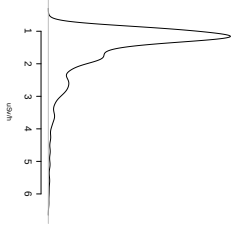
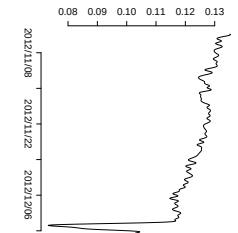
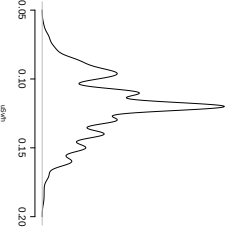
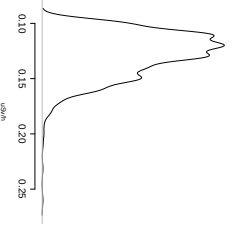
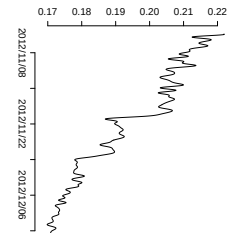
	Vehicle borne survey (Y_1)	Air borne survey (Y_2)	Monitoring post (Y_3)
Concept 1			
Concept 2			
...	...	...	...
Concept 55			

図 6.2: 3 つの Description からなるシンボリックデータテーブル

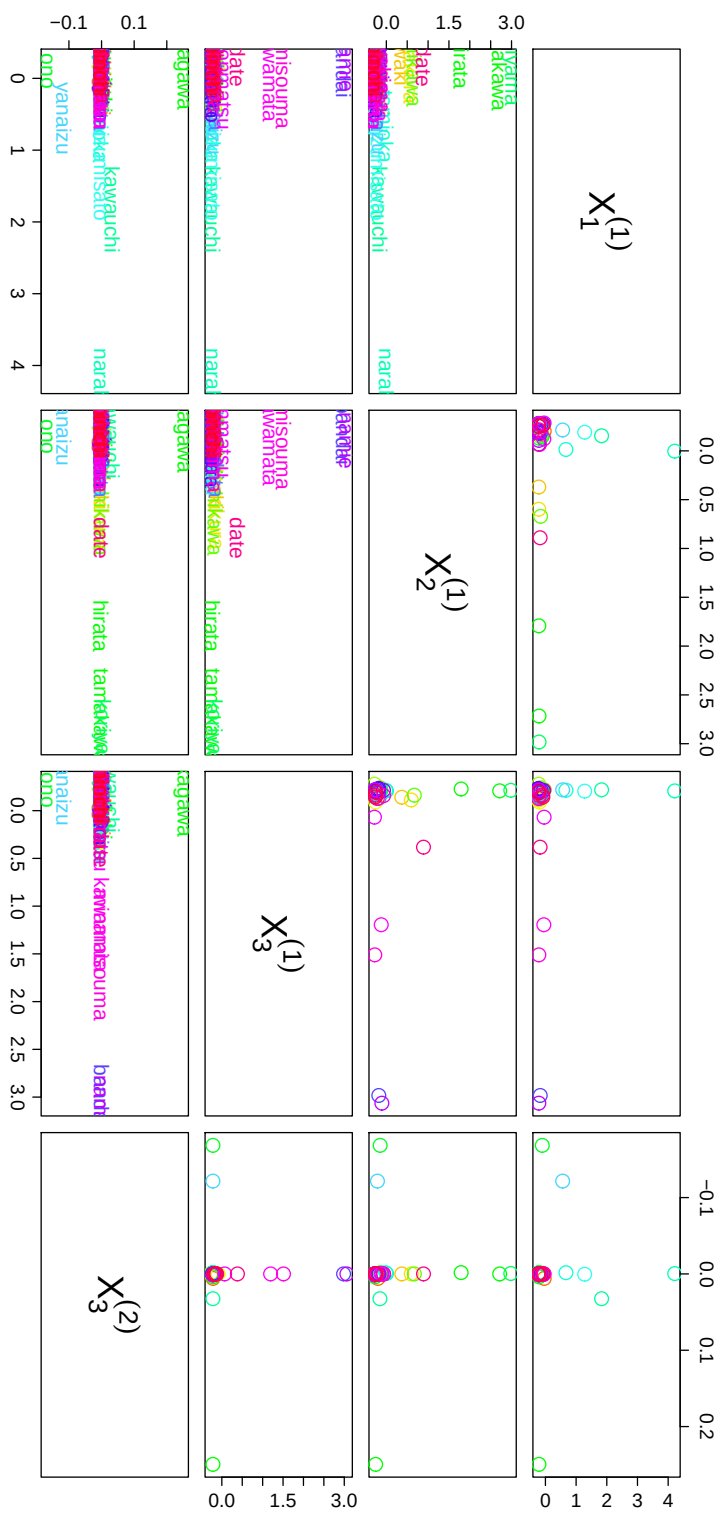


図 6.3: MDS による付置

表 6.1: 平均値による線形相関係数

	Vehicle borne	Air borne	Monitering post
VehicleBorne	1.000	-0.041	-0.082
AirBorne		1.000	-0.046
MoniteringPost			1.000

表 6.2: MDS による固有値

	Vehicle borne survey	Air borne survey	Monitoring post
$X_1^{(m)}$	25.63	24.11	16.42
$X_2^{(m)}$	0.10	0.19	7.15
$X_3^{(m)}$	0.10	0.16	0.11
$X_4^{(m)}$	0.10	0.16	0.11
$X_5^{(m)}$	0.10	0.16	0.11

6.4 おわりに

本論文では、尺度混在型データに対して記述間の関連性を分析するために手法を提案し、とりわけ関数データと分布値データの混在型データに対する解析手法を示した。

尺度混在型データに対して直接、関連性を分析するのは困難であるため、各記述において定義が可能なコンセプト間の非類似度に着目し、MDS を用いることでユークリッド空間における付置を構成してその関連性を分析するアプローチをとった。

コンセプト内における平均値を用いて関連性を調べることができるが、提案手法は各記述に対応した非類似度を用いて分析するため、結果として得られる記述間の関連性には、元データにおけるコンセプト内における内部変動が情報として含まれている。

実際の適用例では、分布値と関数値のみの例であった。しかし、非類似度は様々なデータの記述に対応できるため、その点において、様々な尺度混在型データに対して有効である。

MDS については、今回は計量的 MDS を用いたが、他にも様々な手法が提案されている (Borg and Goronen, 2005)。

表 6.3: PCA の結果

	Comp.1	Comp.2	Comp.3	Comp.4
$X_1^{(1)}$	0.571	0.201	0.314	0.000
$X_2^{(1)}$	-0.043	-0.549	0.366	0.000
$X_3^{(1)}$	-0.406	0.341	0.404	0.000
$X_3^{(2)}$	-0.001	0.000	-0.001	0.044
Eigenvalues	0.493	0.459	0.395	0.002
Proportion	36.535	34.000	29.319	0.146
Cumulative prop.	36.535	70.535	99.854	100.000

あらゆるビッグデータは、たとえ非構造データであれ、解析するためには最終的には構造化されなければならない。データが収集された目的や特徴に応じて、解析対象が多様な記述により表現されることは十分に考えられ、非類似度に着目したアプローチはそのような場合に有効であると思われる。他にも様々な記述が混在するケースも考えられ、それに関しては今後の課題としたい。

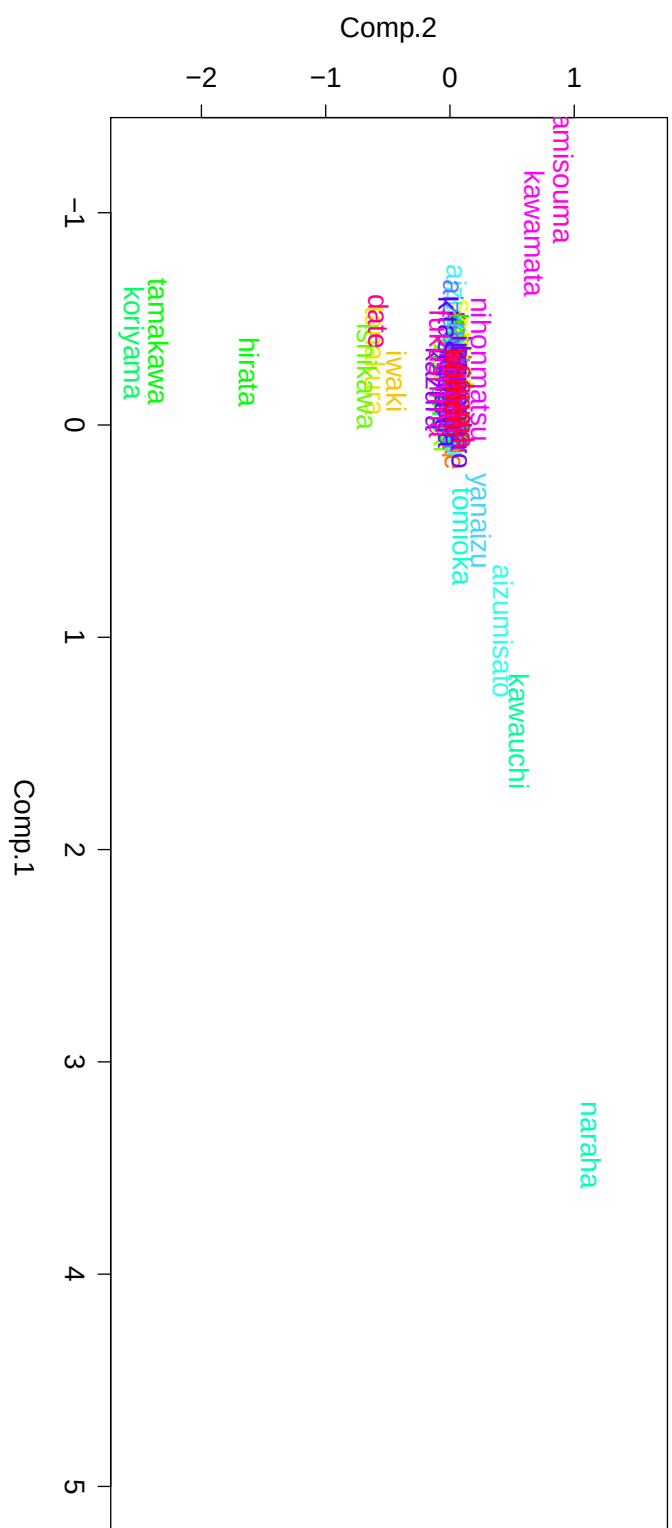


図 6.4: 主成分スコア

第7章 おわりに

本論文では、オブジェクトに対する記述として、新たに分布値データ、関数値データ、尺度混在型データを導入するとともに、それらの解析手法を提案し、従来のSDAにおける記述を拡張した。また、通信遅延時間データおよび大規模環境データに対する本手法の適用結果とその解釈を通じて、提案手法の有効性を示した。

高度集約的シンボリックデータを用いることで、膨大な個体の集まりに対して、確率的なばらつきや、時間に依存したばらつきを考慮したオブジェクトを記述することができる。また、提案手法により、そのようなオブジェクトの集まりに対して、特徴的なパターンを分析することができる。

オブジェクトにおける複雑な記述を考えることにより、ビッグデータの重要な要素である多様性を捉えることが可能である。本研究の成果により、ビッグデータに対して有効な解析手法を提示することができた。

今後の展開として、空間データ解析への応用があげられる。センサーの普及に伴って、様々な位置情報付きの大規模データが手に入るようになりつつあり、空間に依存した観測値間の関連性や、空間的な観測値のパターンを考慮した解析が重要となっている。通常空間データ解析法では、規則的に空間を分割したグリッド（あるいは、格子）ごとに代表値で要約された値を対象に解析を行う。

しかし、グリッド内に本来含まれる複数の観測値による内部変動についても、各グリッドに関する重要な情報となり得、それも考慮することで、さらに複雑な、グリッド間の関連性や空間的なパターンを見つけることが期待できる。現状では、空間的なシンボリックデータ解析法の研究は、殆ど行われていない。第4章や第5章の適用例で扱った環境データに見られるように、各グリッドが分布値や関数値により表されう

るような例は、ますます増えていくと考えられる。今後、本研究による高度集約的シンボリックデータを空間データ解析に導入し、空間的な特徴を捉えられる新たな解析手法を発展させていくことが重要だと思われる。

参考文献

- [1] Aggarwal, C. C. (2013). *Outlier Analysis*. Springer.
- [2] Berrendero, J.R., Justel, A. & Svarc, M. (2011). Principal components for multivariate functional data. *Computational Statistics & Data Analysis*, 55(9), 2619-2634. Elsevier.
- [3] Billard, L. and Diday, E. (2006). *Symbolic Data Analysis: Conceptual Statistics and Data Mining*, Wiley, Chichester.
- [4] Bock, H. H. and Diday, E. (2000). *Analysis of Symbolic Data*, Springer, Berlin Heidelberg.
- [5] Borg, I., Groenen, P. J. F. (2005). *Modern Multidimensional Scaling* (2nd edition), Springer.
- [6] de Boor, C. (2001) *A Practical Guide to Splines*, Revised Edition, Springer.
- [7] Diday, E. and Brito, M. P. (1989). Symbolic Cluster Analysis, In: Optiz, Otto (eds.), *Conceptual and Numerical Analysis of Data*, pp.45–84, Springer, Berlin Heidelberg.
- [8] Diday, E. and Vrac, M. (2005). Mixture decomposition of distributions by copulas in the symbolic data analysis framework, *Discrete Appl. Math.*, **147**(1), pp.27–41, Elsevier Science Publishers B. V, Amsterdam.

- [9] Diday, E. and Noirhomme-Fraiture, M. (2008). *Symbolic Data Analysis and the SODAS Software*, Wiley-Interscience.
- [10] Ferraty, F. & Vieu, P. (2006). *Nonparametric Functional Data Analysis*. Springer.
- [11] Gowda, K. C. and Diday, E. (1999). Symbolic clustering using a new dissimilarity measure, *Pattern Recognition*, **24**(6), 567–578.
- [12] Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning* (2nd edition). Springer.
- [13] Hawkins, D. M. (1980). *Identification of Outliers*. Chapman and Hall.
- [14] Henry, W. Ott. (1988). *Noise Reduction Techniques in Electronic System* (2nd edition). Wiley.
- [15] Huh, M. H. (2002). Setting the Number of Clusters in K-Means Clustering, In: Baba, Y., Hayter, A. J., Kanefuji, K. and Kuriki, S. (eds.), *Recent Advances in Statistical Research and Data Analysis*, pp.115–124, Springer, Tokyo.
- [16] Ichino, M; Yaguchi, H. (1994). Generalized Minkowski metrics for mixed feature-type data analysis, Systems, Man and Cybernetics, *IEEE Transactions on*, **24**(4), pp.698–708.
- [17] Jacques, J. & Preda, C. (2013). Functional data clustering: a survey. *Advances in Data Analysis and Classification*, 1-25. Springer.
- [18] Jacques, J. & Preda, C. (2014). Model-based clustering for multivariate functional data. *Computational Statistics & Data Analysis*, **71**, 92-106. Elsevier.
- [19] Jansen, M. (2001). *Noise Reduction by Wavelet Thresholding*, Lecture Notes in Statistics, **161**. Springer.

- [20] Katayama, K., Minami, H. and Mizuta, M. (2009). Hierarchical Symbolic Clustering for Distribution valued Data, *Journal of the Japanese Society of Computational Statistics*, **22**(2), pp.83–89 (In Japanese).
- [21] Kriegel, H.-P., Kröger, P. & Zimek, A. (2010). Outlier Detection Techniques: *The 2010 SIAM International Conference on Data Mining*.
- [22] McKinsey Global Institute(2011). *Big data: The next frontier for innovation, competition, and productivity*. http://www.mckinsey.com/verb|~/media/McKinsey/dotcom/Insights/%20and/%20pubs/MGI/Research/Technology/%20and/%20Innovation/Big/%20Data/MGI_big_data_full_report.as\hx.
- [23] Matsui, Y., Komiya, Y., Minami, H. and Mizuta, M. (2013). Comparison of Two Distribution Valued Dissimilarities and Its Application for Symbolic Clustering, In: Gaul, W., Geyer-Schulz, A., Baba, Y. and Okada, A. (eds.), *German-Japanese Interchange of Data Analysis Results*. Studies in Classification, Data Analysis, and Knowledge Organization. (to appear), Springer, Heidelberg.
- [24] Matsui, Y., Minami, H. and Mizuta, M. (2013). Symbolic Cluster Analysis for Distribution Valued Data, In: Cho, S. H. (eds.), *Proceedings of Joint Meeting of the IASC Satellite Conference and the 8th Conference of the Asian Regional Section of the IASC*, pp.305–310, Aug. 22–23, 2013, Yonsei University, Seoul, Korea.
- [25] Mizuta, M. (2004). Clustering methods for functional data. *Proceedings in Computational Statistics*, 1503-1510. Springer.
- [26] Mizuta, M. & Kato, J. (2007). Functional Data Analysis and Its Application. *Lecture Notes in Artificial Intelligence* 4481, 228-235, Springer.

- [27] Mizuta, M. and Minami, H. (2012). Analysis of Distribution Valued Dissimilarity Data. In: Gaul, W. A., Geyer-Schulz, A., Schmidt-Thieme, L. and Kunze, J., *Challenges at the Interface of Data Analysis, Computer Science, and Optimization*, Studies in Classification, Data Analysis, and Knowledge Organization, pp.23–28, Springer, Heidelberg.
- [28] O'Reilly, C., Gluhak, A., Imran & M. Rajasegarar, S. (2014) Anomaly Detection in Wireless Sensor Networks in a Non-Stationary Environment, *Communications Surveys & Tutorials, IEEE*, PP(99), pp.1–20.
- [29] Ramsay, J. O. & Silverman, B. W. (2002). *Applied Functional Data Analysis*. Springer.
- [30] Ramsay, J. O. & Silverman, B. W. (2005). *Functional Data Analysis* (2nd edition). Springer.
- [31] Ramsay, J. O., Hooker, G. & Graves, S. (2009). *Functional Data Analysis with R and MATLAB*. Springer.
- [32] Rippe, R. C. A., Meulman, J. J. & Eilers, P. H. C. (2012). Visualization of Genomic Changes by Segmented Smoothing Using an L_0 Penalty. *PLoS ONE*, **7**(6), e38230.
- [33] Schweizer, B. (1968). Distributions are the numbers of the future, *Proc. sec. Napoli Meeting on "The mathematics of fuzzy systems"*, pp.137–149, Istituto di Mathematica delle Faculta di Achitectura, Universita degli studi di Napoli.
- [34] Terada2010 Terada, Y. and Yadohisa, H. (2010). Non-hierarchical clustering for distribution-valued data. *COMPSTAT 2010: Proceedings in Computational Statistics*, pp.1653–1660, Psysica-Verlag, Heidelberg.

- [35] Tokushige, S., Yadohisa, H. & Inada, K. (2007). Crisp and fuzzy k-means clustering algorithms for multivariate functional data. *Computational Statistics*, 22(1), 1-16. Springer.
- [36] Xu, L. Lu, C., Xu, Y. & Jia, J. (2011). Image Smoothing via L_0 Gradient Minimization. *Proceedings of ACM SIGGRAPH Asia 2011*, **30**(6), 174, 1–12.
- [37] Yang Z., Meratnia, N. & Havinga, P. (2010). Outlier Detection Techniques for Wireless Sensor Networks: A Survey, *Communications Surveys & Tutorials*, *IEEE*, **12**(2), pp.159–170.
- [38] 青山智夫・神部順子・中山榮子・長嶋雲兵 (2012). 放射性物質の滞留半減期. *Journal of Computer Chemistry, Japan*, **11**(1), 49–77.
- [39] 石岡文生・栗原孝次 (2013). 福島第一原子力発電所事故に伴う放射性物質の集積性の検討. 日本計算機統計学会第 27 回シンポジウム講演報告集, 27, pp.3-6.
- [40] 鈴木和之・小宮由里子・南 弘征・水田正弘 (2013) 土壌中における放射性セシウムの深度分布に関する解析. 日本計算機統計学会第 27 回シンポジウム講演報告集, 27, pp.11-14.
- [41] 松井佑介・南 弘征・水田正弘 (2013). 放射性物質の分布状況等調査データベースを用いた空間線量率分布に関する考察. 2013 年度統計関連学会連合大会講演報告集, 145.
- [42] 水田正弘 (2005) 関数データとその解析法. 知能と情報, **17**(4), pp.413–417.

謝辞

水田正弘 教授、南 弘征 准教授、小宮 由里子 助教には多岐にわたり、多大なるご支援を賜りました。この場をお借りして、深く御礼を申し上げます。また、研究室で、一緒に頑張ってきた仲間たちにもお礼を申し上げます。一人だけの環境になることなく、励ましや良い刺激を受けて頑張ることができました。

そして、最後まで公私にわたり支えてくれた私の家族にも、感謝の気持ちをここに記します。本当に、ありがとうございました。