

NMF による自然災害被害統計の構造分析

奥村 誠¹・水谷 大二郎²

¹正会員 東北大学教授 災害科学国際研究所 (〒980-0845 仙台市青葉区荒巻字青葉468-1)

E-mail:makoto.okumura.b6@tohoku.ac.jp (Corresponding Author)

³正会員 東北大学助教 大学院工学研究科 (〒980-0845 仙台市青葉区荒巻字青葉468-1)

E-mail:mizutani@irides.tohoku.ac.jp

近年、深刻な自然災害が毎年のように発生し、地域防災や緊急対応体制の強化が唱えられている。災害の原因となるハザードの種類や被災地域の特性によって、死傷者数、インフラの損傷、家屋の損傷、農地被害などの被害項目への影響は異なる。今後の防災や体制強化を効率的に進めるために、どのような被害の組み合わせが、どのようなハザードの災害で、どの地域で発生しやすいかを理解しておくことが有用である。本研究では、最近の十数年間の日本の異常気象別・都道府県別の災害統計データに、非負値行列因子分解 (NMF; Non-negative Matrix Factorization) 法を適用し、被害項目の典型的な発生パターンを抽出する。この時、ベイズ統計に基礎をおく推定法を用いると、スパースなパターンが抽出され、意味の明確化と信頼性の向上が期待できる。

Key Words : *natural disaster, impact of disaster, non-negative matrix factorization, Bayesian statistic*

1. はじめに

近年、深刻な自然災害が毎年のように発生し、地域防災や緊急対応体制の強化が唱えられている。災害の原因となるハザードの種類や被災地域の特性によって、死傷者数、インフラの損傷、家屋の損傷、農地被害などの被害項目への影響は異なるものとなる。今後の防災や体制強化を効率的に進めるために、どのような被害の組み合わせが、どのようなハザードを原因として、どの地域で発生しやすいかのパターンを理解しておくことが役立つ。

このような問題意識から、日本国内の災害統計データ、あるいは外国のデータを収集して内在するパターンを分析する試みがなされてきた^{1,4)}。これらの多くは、自然ハザードの時空間的な分布特性の把握を主目的としており、日本の気候区分に対応したハザードごとの空間分布などの興味深い結果が得られている。一方で、災害の社会的影響や緊急対応の観点からは、どのような被害が同時に起きやすいかの理解が有用であると考えられる。

多くの観測データの中に共通する変動パターンを、少数の非負の要因の加算として把握する手法として、非負値行列因子分解 (NMF; Non-negative Matrix Factorization) 法が開発され²⁾、最近では土木計画学分野での適用が進んできている。この手法は、これまで一般的に使われることが多かった一般的な因子分析や主成分分析などの多変

量解析手法に比べて、発生率に関する確率的特性を変えずにパターンを抽出できる特徴がある。筆者らは地域の転出、転入、残留者数は、測定誤差や記録時の誤差を考慮しなければ必ず非負の値を取ることから、NMFに基づいた人口移動特性の把握のための方法論を提案して、災害を挟む期間の都道府県間の人口移動分析に適用した³⁾。

本研究では、最近の十数年間の日本の都道府県別・異常気象別の被害項目別データにNMFを適用して、被害項目の典型的な組み合わせ発生パターンを抽出するとともに、地域別・ハザード別の発生量を分析する。この時、ベイズ統計に基礎をおく推定法を用いると、スパースなパターンが抽出されるため、意味の明確化と信頼性の向上が期待できる⁷⁾。

2. ベイジアンNMF⁷⁾

NMFについては数多くの既往研究や成書が存在する。本研究ではNMFの中でもベイズ統計学に基礎をおく計算手法を用いる。

(1) 非負値行列因子分解

主成分分析や因子分析など、与えられたデータを複数の成分に分解する手法は様々あるが、特に NMF は非負値のデータを対象とした多変量解析手法である⁵⁾。現実

世界に存在するデータの多くは非負値で表されるため、NMF によって合奏曲や顔画像、文書内の単語の出現回数といった非負値のデータの構成成分を抽出し、音源分離や顔画像処理、文書の自動分類をする研究が進められている。

ここで、分析したいデータをサイズ $K \times N$ の観測行列 \mathbf{X} とする。観測行列 \mathbf{X} を

$$\mathbf{V} \cdot \mathbf{W} + \mathbf{E} \approx \mathbf{X} \quad (1)$$

を満たすようなサイズ $K \times M$ の因子得点の行列 \mathbf{V} とサイズ $M \times N$ の因子負荷量の行列 \mathbf{W} とサイズ $K \times N$ 残差行列 \mathbf{E} に分解する。なお、分解後の因子数 M は解析する者が事前に決めておき、

$$M \leq \min(K, N) \quad (2)$$

を満たす必要がある。一般的に M の数が多ければ多いほど残差行列 \mathbf{E} が小さくなり、内積 $\mathbf{V} \cdot \mathbf{W}$ は観測行列 \mathbf{X} に近づいていく。

(2) ベイズ統計学に基礎をおく非負値行列因子分解

ベイズ統計学の考え方を取り入れることで、得られる結果の不確実性評価が可能になり、説明力が向上する⁷⁾。以下ではベイズ統計学に基礎を置く NMF の推計手順を引用しておく。

まず残差行列 \mathbf{E} に尤度関数を設定し、行列 \mathbf{V} , \mathbf{W} のそれぞれの行列値について事前分布を仮定する。一般的に、尤度関数と事前分布の組合せとして正規分布と指数分布を選ぶことが多い。まず誤差行列 \mathbf{E} に平均 0 、分散 σ^2 の正規分布を仮定して

$$p(\mathbf{X}|\mathbf{V}, \mathbf{W}, \sigma^2) = \prod_{k,n} \mathcal{N}(X_{k,n}; (\mathbf{VW})_{k,n}, \sigma^2) \quad (3)$$

のように尤度関数を設定する。また、行列 \mathbf{V} , \mathbf{W} のそれぞれの行列値については以下の事前分布を仮定する。

$$\begin{aligned} p(\mathbf{V}) &= \prod_{k,m} \mathcal{E}(V_{k,m}; \varphi_{k,m}) \\ p(\mathbf{W}) &= \prod_{m,n} \mathcal{E}(W_{m,n}; \omega_{m,n}) \end{aligned} \quad (4)$$

ここで、

$$\mathcal{N}(x; \mu, \sigma^2) = \frac{\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}{\sqrt{2\pi\sigma^2}} \quad (5)$$

は正規分布であり、

$$\mathcal{E}(x; \lambda) = \lambda e^{-\lambda x} \cdot u(x) \quad (6)$$

は指数分布を表す。なお、

$$u(x) = \begin{cases} 0 & (x < 0) \\ 1 & (x \geq 0) \end{cases} \quad (7)$$

は、非負の値を持つ x が表れた場合、 0 を返す単位ステップ関数であり、非負条件を満たすために用いる。

また、残差行列の分散は、形状パラメータ k と規模 ρ

ラメータ θ によって記述される逆ガンマ分布に従って発生すると仮定する。

$$\begin{aligned} p(\sigma^2) &= \mathcal{G}^{-1}(\sigma^2; k, \theta) = \frac{\theta^k}{\Gamma(k)} (\sigma^2)^{-k-1} \exp\left(-\frac{\theta}{\sigma^2}\right) \end{aligned} \quad (8)$$

以上の設定により計算できる事後確率を最大化することによって、行列 \mathbf{V} や \mathbf{W} の推定値が得られる。

(3) ギブスサンプリングのための条件付確率設定

本研究では、事後確率を最大化するためにマルコフ連鎖モンテカルロ法のギブスサンプリングの手順に従ってうち行列 \mathbf{V} , \mathbf{W} の推定を行う。

ギブスサンプリングでは、モデルのパラメータに適切な条件付確率を設定することでサンプルを生成するが、それを適切な回数繰り返し、サンプルを更新していく。ここでは、行列 \mathbf{V} , \mathbf{W} の値に関する条件付確率として、正規分布と指数分布を掛け合わせたものに比例した、式(9)のような調整済み正規分布を仮定し、次々と値を更新する。

$$\mathcal{R}(x; \mu, \sigma^2, \lambda) \propto \mathcal{N}(x; \mu, \sigma^2) \mathcal{E}(x; \lambda) \quad (9)$$

行列 \mathbf{V} の条件付確率はこの調整済み正規分布を仮定し、式(10)のようにおける。

$$\begin{aligned} p(\mathbf{V}_{k,m} | \mathbf{X}, \mathbf{V}_{\setminus(k,m)}, \mathbf{W}, \sigma^2) &= \mathcal{R}(\mathbf{V}_{k,m}; \mu_{V_{k,m}}, \sigma_{V_{k,m}}^2, \varphi_{k,m}) \end{aligned} \quad (10)$$

なお、記号 $\mathbf{V}_{\setminus(k,m)}$ は行列 \mathbf{V} の k 行 m 列以外の要素全てを指す。対称性を考慮し、 $\mathbf{W}_{m,n}$ についても同様に、調整済み正規分布を設定できる。

また、分散 σ^2 の条件付確率は式(12)のように逆ガンマ分布を仮定する。

$$\begin{aligned} \mu_{V_{k,m}} &= \frac{\sum_n (X_{k,n} - \sum_{m' \neq m} V_{k,m'} W_{m',n}) W_{m,n}}{\sum_n W_{m,n}^2} \\ \sigma_{V_{k,m}}^2 &= \frac{\sigma^2}{\sum_n W_{m,n}^2} \end{aligned} \quad (11)$$

$$p(\sigma^2 | \mathbf{X}, \mathbf{V}, \mathbf{W}) = \mathcal{G}^{-1}(\sigma^2; k_{\sigma^2}, \theta_{\sigma^2}) \quad (12)$$

$$\begin{aligned} k_{\sigma^2} &= \frac{KN}{2} + 1 + k \\ \theta_{\sigma^2} &= \frac{1}{2} \sum_{i,j} (X - \mathbf{VW})_{i,j}^2 + \theta \end{aligned} \quad (13)$$

以上の条件付確率の設定により、ギブスサンプリングを行いながら、得られた事後分布と照らし合わせ確認することもできる。

3. 被害統計データの分解

(1) 被害統計データ

本研究では、消防庁が集約し、毎年の消防白書の付表として公開されている、各年の異常気象別被害統計、お

表1 異常気象別データから得られた被災パターン

	2	3	4	5	6	7	8	9	10	11
		施設罹災	施設罹災	破壊死亡罹災	破壊死亡罹災	破壊死亡罹災	破壊死亡罹災	破壊死亡罹災	破壊死亡罹災	破壊死亡罹災
通信全壊罹災	通信全壊	通信全壊	通信施設	通信施設	通信施設	通信施設	通信施設	通信施設	通信施設	通信施設
			洪水破壊	洪水破壊	洪水破壊	洪水破壊	洪水破壊	洪水破壊	洪水破壊	洪水破壊
								半壊浸水罹災	半壊浸水罹災	半壊浸水罹災
							施設罹災者	施設罹災者	施設罹災者	施設罹災者
										罹災斜面橋梁
					全壊世帯	全壊農地世帯	全壊農地世帯		全壊農地世帯	全壊農地世帯
								畑浸水	半壊罹災	施設罹災者
							罹災道路河川	罹災道路河川	罹災道路河川	罹災道路河川
半壊小浸水	半壊小浸水	小浸水	小浸水	小浸水	小浸水	小浸水	小浸水	小浸水	小浸水	小浸水
		農地床下浸水	農地浸水	農地浸水	農地浸水	農地浸水	農地浸水	水田浸水	農地浸水	農地浸水

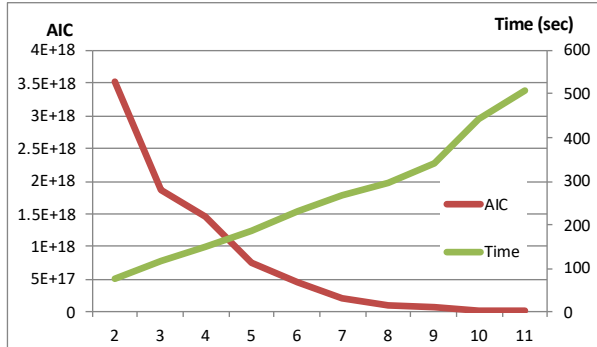


図1 因子の次元数に伴うAICと計算時間の変化

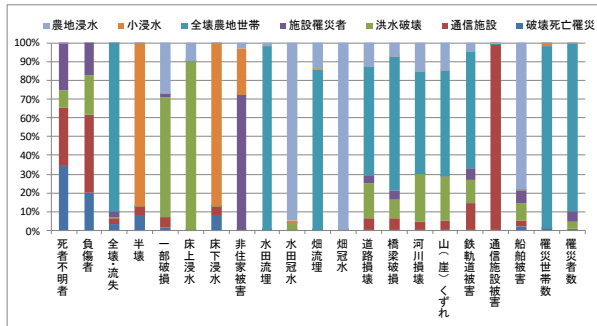


図2 因子数7の場合の因子負荷量 (異常気象データ)

および各年の都道府県別被害統計のデータを分析する。なお、最新年のデータは、消防白書よりも先に、国土交通省の災害統計の中の付表として公表されている。本研究では、両者ともに、調査項目が同じである2007年から2018年の12年間のデータを用いる。

前者のデータでは、異常気象として、1)冬風浪、2)風浪、3)融雪、4)豪雨、5)台風、6)地震、7)地すべり、8)その他の自然災害に区分しており、大きな台風や地震は個別に取り扱われる。豪雨は発生期間を数日～10日間ぐらいの長さで区切って区分しており、同期間内に来襲した台風と併せて集計されている場合も多い。異常気象のケース数は188である。

一方、被害項目として、a)死者、b)不明者、c)負傷者、d)住宅の全壊・流失、e)住宅の半壊、f)住宅の一部破損、g)住宅の床上浸水、h)住宅の床下浸水、i)非住家建物被害の各件数、j)水田の流失・埋没、k)水田の冠水、l)畑の流失・埋没、m)畑の冠水の各面積、n)道路損壊、o)橋梁破損、p)河川損壊、q)山(崖)くずれ、r)鉄軌道被害、s)通信施設被害、t)船舶被害の各件数、u)罹災世帯数、v)罹災

者数の22項目が表章されている。また、w)自然災害に起因した火災の発生数も得られる。

後者のデータでは47都道府県別に、上記のa)-v)の22項目に加えて、x)被害総額、y)都道府県の災害対策本部、z)市町村の災害対策本部の各設置回数、aa)災害救助法適用回数、ab)消防隊員、ac)消防団員の延べ出勤人数のデータがある。さらに、c)負傷者を重傷者と軽傷者に区分して、i)非住家被害件数を公共建物、民間建物に区分して知ることができる。また学校の被害件数も表章されている。ケース数は47×12=564である。

以下の分析では、2種類のデータから抽出されるパターンの比較を行うために、両社に共通する項目を対象に分析を行う。またa)死者とb)不明者をまとめて扱うこととし、a+b、c)～v)の合計22の項目についての分析を行う。

なお、因子得点行列Vおよび因子負荷量行列Wの解釈を行うにあたり、因子負荷量行列Wの行和が1となるように基準化する。つまり、因子負荷量行列Wの各行は、M個ある被害パターンのそれぞれにおいて、異なる被害項目での値が含まれる構成比を表すこととなる。また因子得点行列Vの各行は、観測行列Xの各異常気象において、異なる被害パターンがどのような頻度で発生したか、あるいは各都道府県において発生した、異なる被害パターンの組み合わせ頻度を表すこととなる。

今回用いたベイズモデルでは、因子負荷量行列Wの各要素は式(9)のように半正規分布に従う平均値を持つ指数分布によって与えている。指数分布では実現値が0を取る確率が大きくなるため、計算の途中である要素の値が0を取り、その残差を他の要素を用いて表現するように調整されるため、一旦0の値を取った要素が0に固定される可能性が高い。その結果、因子負荷量行列Wが疎になり、因子が互いに独立になるように、スパース性の高い疎なパターンが求まる傾向にあり、分析結果の解釈はより容易になる。

(2) 異常気象別被害統計データの分析結果

分析にあたり、抽出する成分の次元数を指定する必要がある。図1は、次元数を2から11まで変化させて与えた場合のAICと計算時間との変化を表している。AICの減少は徐々に効果が小さくなる一方で、計算時間は次元数

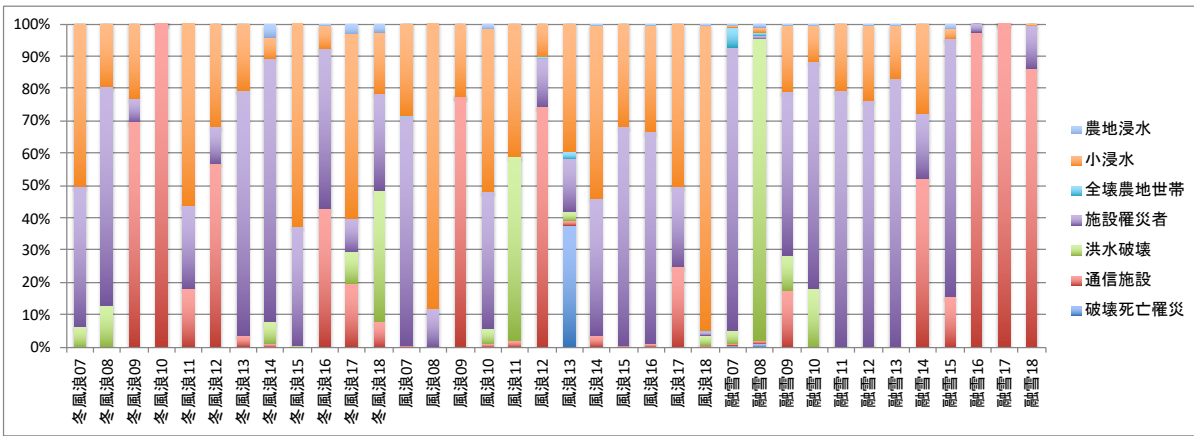


図3 異常気象別の因子得点の計算値 (その1: 風浪及融雪)

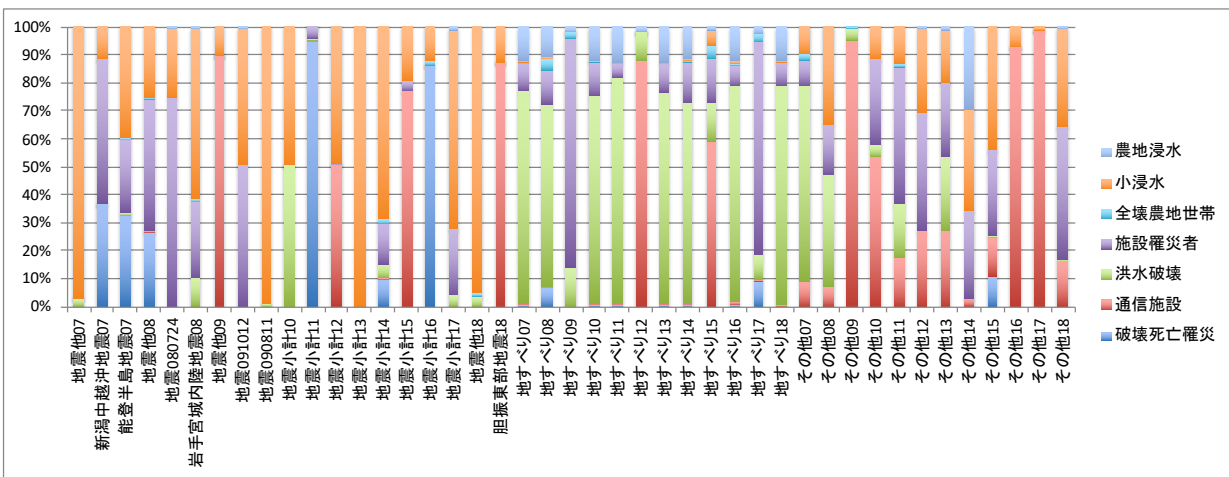


図4 異常気象別の因子得点の計算値 (その2: 地震, 地すべり, その他の災害)

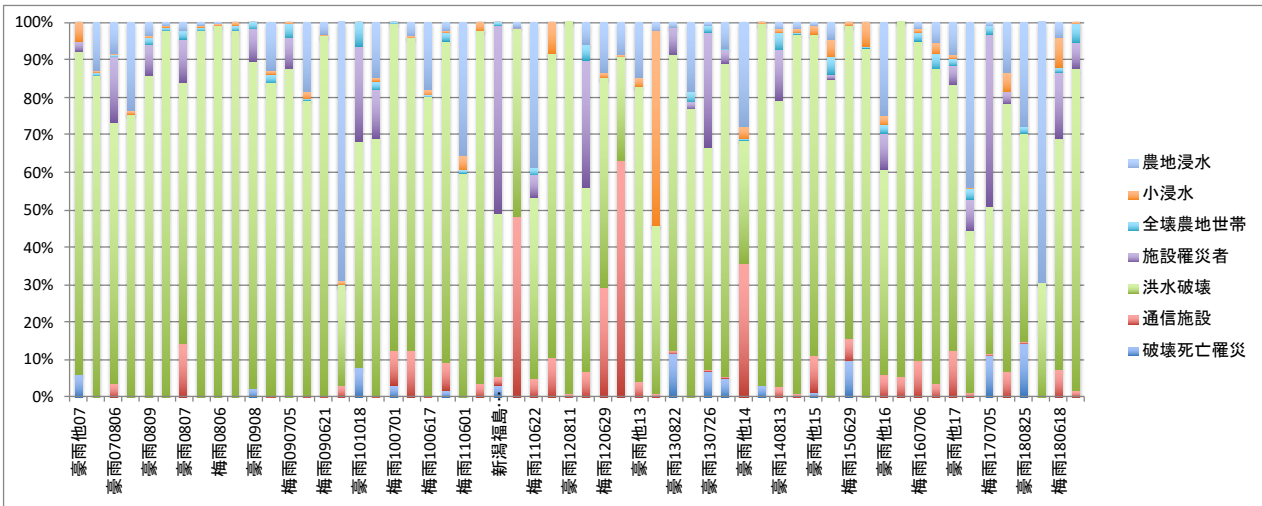


図5 異常気象別の因子得点の計算値 (その3: 豪雨災害)

に比例する以上に増大する。AICの傾きから、およそ6から7次元を用いることが適切であると判断した。

表1は、次元数ごとに、抽出された成分がどの被害項目との相関が強いかによって、成分の名前づけを行ったものである。これより最初に、強風や地震で構造物の被害が起こり、それが人的被害や罹災者の発生につながる成分と、高頻度で浸水被害を中心とする成分に大きく分かれ、さらにそれぞれの中で違いを持つ成分に細分化されていく様子がわかる。

図2は7次元の成分について、各成分の因子負荷量の値を示している。これより、通信施設被害が特有の成分で説明され、他の項目との共通性が少ないことがわかる。ついで、住宅の半壊と床下浸水が共通の成分により発生すること、住宅の床上浸水も別の共通性分により発生し、その時には非住家被害も懸念されることがわかる。農地の流失や埋没は、住宅の全壊・流失をもたらす成分で発生し、罹災世帯数や罹災者数も同時に大きくなること、同時に道路、橋梁、河川、鉄軌道などのインフラにも被

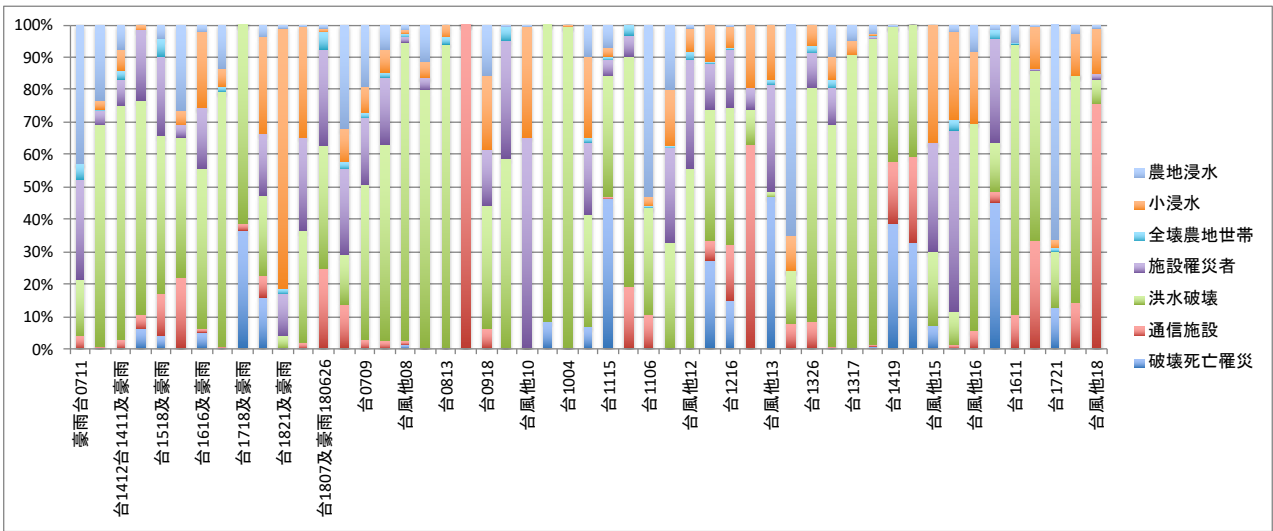


図6 異常気象別の因子得点の計算値（その4：台風）

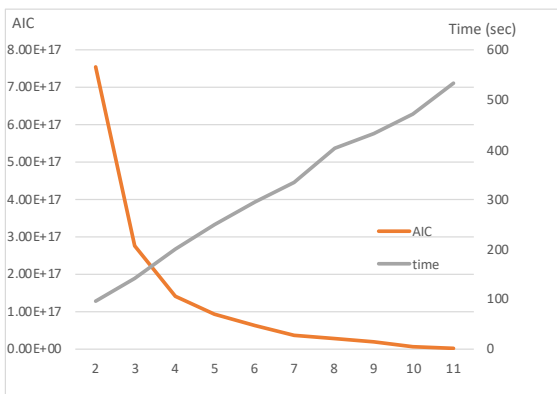


図7 因子の次元数に伴うAICと計算時間の変化

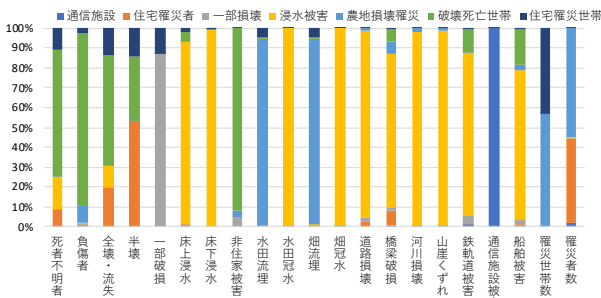


図8 因子数7の場合の因子負荷量（都道府県データ）

害を与える可能性が大きいことが確認できる。これに比べると、農地の冠水は異なる成分で説明され、同時に船舶への被害が懸念されるものの、他の被害項目への影響は大きくないことがわかる。

図3から図6では、元データにおける異常気象ごとの因子得点の値をプロットしている。これらより、異常気象の種類によって、どのタイプの成分の発生につながることが大きく異なることがわかる。例えば図3から、風浪では小規模の浸水や施設の破損による罹災者の発生につながりやすい。融雪では最近になり、通信施設への影響が顕著になってきていることがわかる。図4では地震が小

規模浸水や施設の破壊につながるケースが多いこと、2011年と2016年の地震、すなわち東日本大震災と熊本地震では破壊的な影響が死亡者や罹災者の発生につながったことが明確に現れている。一方、地すべり災害は破壊的な影響をもたらす洪水の発生につながっていること、年によっては通信施設への影響が大きいことがわかる。豪雨時の成分を示す図5からは、ここで集計されている規模の豪雨が発生すると小規模な浸水では済まず、洪水の発生につながるような状況が発生していることが確認できる。最後に台風の影響を示す図6から、台風により洪水をもたらす成分が頻出するほか、農地の浸水、施設の破壊、さらには通信施設の被災など、多様なタイプの被害をもたらす成分が混合して発生していることが理解できる。

(3) 都道府県別被害統計データの分析結果

都道府県別の被害データに対しても同様の分析を行なった。図7は、次元数を2から11まで変化させた場合のAICと計算時間の変化を表している。AICの減少から事前数増加の効果は徐々に小さくなり、先のケースと同じく、AICの傾きから、およそ6から7次元を用いることが適切であると判断できる。計算時間は次元数の増加にほぼ比例して増加している。

表2は、次元数ごとに、抽出された成分がどの被害項目との相関が強いかによって、成分の名前づけを行ったものである。これより、最初に通信施設などのインフラに影響を与え、罹災者数の発生につながる成分と、住宅の破壊をきっかけに罹災世帯や死亡者を生み出すという、いずれもハードな被害の発生を意味する2つの成分が検出された。次元数を増やすと、先の異常気象別データに基づく場合とは異なり、先に現れていた成分が検出されなくなりながら、検出される成分が変化していく。

表2 都道府県別データから得られた被災パターン

2	3	4	5	6	7	8	9	10	11
通信罹災者	通信罹災者	通信施設	通信施設	通信施設	通信施設	通信施設	通信施設	通信施設	通信施設
	破壊罹災者	破壊罹災者	破壊罹災者	住宅罹災者	住宅罹災者	施設負傷者	施設負傷者	施設負傷者	道路河川浸水
		一部損壊	一部損壊	一部損壊	一部損壊	一部損壊	一部損壊		一部損壊
			半壊浸水	半壊浸水世帯		半壊罹災者	半壊罹災者	小半壊	小半壊
				浸水被害	浸水被害	浸水被害	浸水被害	農宅地浸水	農地浸水
								小浸水	小浸水
				農地損壊罹災	農地損壊罹災	農地損壊罹災	農地損壊罹災	農地損壊罹災	農地損壊罹災
破壊世帯	破壊世帯	破壊死亡世帯	破壊死亡世帯		破壊死亡世帯	全壊死亡	全壊死亡	全壊死亡	全壊死亡
								破壊浸水死亡	罹災者
					住宅罹災世帯	一部損壊	一部損壊	部分損壊世帯	部分損壊世帯
							半壊施設罹災	半壊施設罹災	半壊施設罹災

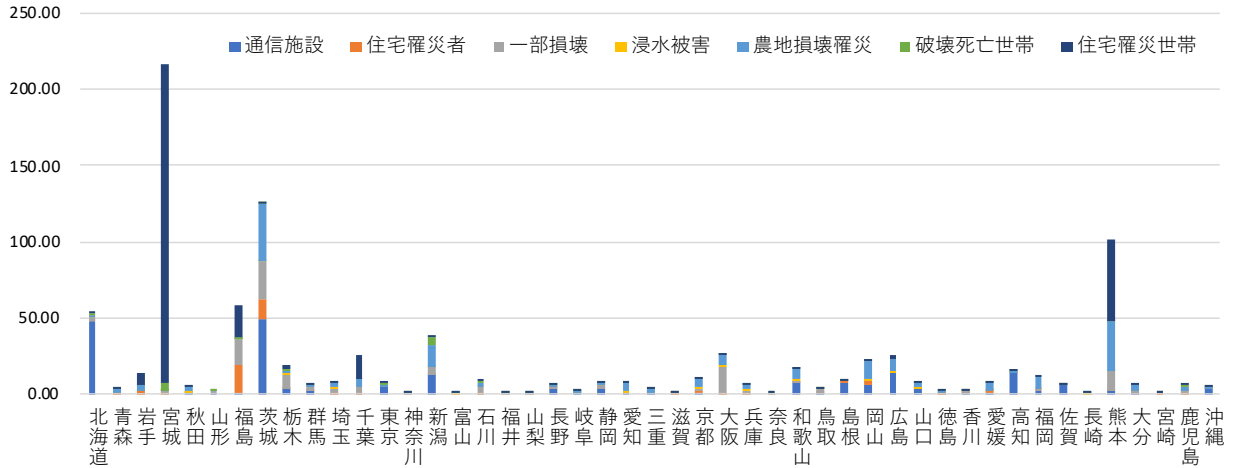


図9 因子数7の場合の因子得点の分布（都道府県データ）

因子数を7としたときの因子負荷量の分布を図8に示す。項目ごとの構成は、先に図2で説明したものとほぼ同様である。

また、都道府県ごとに年次別の因子得点が算出されるが、その平均値を図9に示した。これより、東日本大震災の影響が強い宮城、茨城、福島各県と熊本で因子得点が高く、複数の項目に及んでいる。北海道と新潟では通信施設と農地損壊を起点とする因子の得点が高い。当該期間（2007～2018年）中に集中豪雨があった広島、岡山、和歌山では複数の因子の得点が高く、多様な影響が及んでいることが確認できる。

4. おわりに

以上、NMFを用いて、この10年余りの自然災害による被害の発生パターンの抽出を行い、概略的に傾向を握るために有用であることが確かめられた。

本研究の今後の発展方向として、既存の研究が指摘するように被害の発生の背後にある、ハザード、暴露量、および脆弱性という視点からの分析を行うことが必要である。その際、外部情報を取り入れてパラメータを細かく設定した事前分布を用いることで、外部情報が分析結

果に反映されるようにすることが考えられる。

参考文献

- 1) 阿部雅人・藤野陽三：自然災害リスクの特性に関する統計的分析，土木学会論文集 A, Vol.64, No.4, pp.750-764, 2008.
- 2) 下野勘智・菊本統・伊藤和也・大里重人・稲垣秀輝・日下部治：自然災害に対する全国 47 都道府県のリスク指標の試算と考察，土木学会論文集 F6(安全問題), Vol. 72, No. 1, 1-10, 2016.
- 3) 篠原慶規・小松光：近年の土砂災害による死者・行方不明者数の経年変動，砂防学会誌, 68, 5, pp.3-9, 2016.
- 4) 牛山 素行：日本の風水害人的被害の経年変化に関する基礎的研究，土木学会論文集 B1(水工学) Vol.73, No.4, I_1369-I_1374, 2017.
- 5) 亀岡弘和：非負値行列因子分解，計測と制御，第 51 巻第 9 号，pp.835-844, 2012.
- 6) 元井初音・奥村誠・水谷大二郎：年齢構成の共通性に着目した地域社会の比較分析法，土木計画学研究・講演集, Vol.57, No.17-09, 2018.
- 7) Mikkel N. Schmidt, Ole Winther, and Lars Kai Hansen: Bayesian Non-negative Matrix Factorization, ICA 2009, LNCS 5441, pp.540-547, 2009.

(Received October2, 2020)

(Accepted October2, 2020)

STRUCTURE OF NATURAL DISASTER IMPACT STATISTICS, ANALYZED BY NMF

Makoto OKUMURA and Daijiro MIZUTANI

In recent years, serious natural disasters have occurred almost every year, and it has been advocated to strengthen regional disaster prevention and emergency response systems. Depending on the type of hazard that causes the disaster and the characteristics of the disaster area, the impact on damage items such as the number of casualties, infrastructure damage, house damage, and farmland damage will differ. In order to efficiently promote disaster prevention in the future, it is useful to understand what kind of damage combination is likely to occur in which area and what kind of hazard disaster. In this study, we applied the non-negative matrix factorization (NMF) method to the disaster statistical data by prefecture and natural hazard in Japan over the last ten years.