

統計的手法と機械学習に基づく地すべりおよび土砂崩壊地点の特徴評価

Evaluation of the characteristics at the landslide occurrences based on statistical methods and machine learning

12TM0300 古川 昭太
Shota Furukawa

指導教員 丸山 喜久

SYNOPSIS

This study evaluated the topographic characteristics at the landslide outbreak locations based on the covariance structure analysis and the machine learnings. The support vector machine (SVM) and random forest (RF) were employed in this study to achieve the objective. The estimated results were compared to investigate the accuracy of the detection of landslide occurrences. The SVM and RF gave reasonable results if a technique for oversampling is applied to correct for a bias in the training dataset. Lastly, the discriminated results were projected onto a map to evaluate the possibility of landslide occurrences.

1. はじめに

わが国では、自然災害の発生に伴って多くの地すべりの被害を被ってきた。2004年10月23日に発生した新潟県中越地震では地すべり131ヶ所、崖崩れ115ヶ所、土石流21ヶ所と多くの土砂災害の被害が出た¹⁾。また2011年東北地方太平洋沖地震では北海道と青森を除く東北地方と群馬・栃木・茨城・埼玉・群馬で大規模な範囲での斜面崩壊が確認された²⁾。その他、広島県広島市北部で発生した平成26年8月豪雨³⁾など、豪雨による土砂災害も多く発生している。地震調査委員会では今後30年間でマグニチュード7クラスの地震が首都圏で起こる確率は70%としており、大規模な土砂災害の発生も懸念される。このことから、どのような地形的特徴が地すべりの発生の原因になり得るのかをあらかじめ予測しておき、地すべりの危険性が高い地域を予測することは有意義であると考えられる。

そこで本研究では、まず複数の構成概念間の関係を検討できる統計的手法である共分散構造分析 (Covariance structure analysis)⁴⁾を行い、どの素因が地すべりや土砂崩壊の発生に大きく影響するのかを明らかにする。さらに、比較的少数の説明変数を用いたサポートベクターマシン (Support vector machine)⁵⁾とランダムフォレスト (Random forest)⁶⁾に基づく機械学習を実行し、地すべりや土砂崩壊発生地点を予測するモデルを構築することを目的とする。分析手法ごとの結果と既存の地すべり地形分布図と比較し精度を評価することによって、今後どのような地域で地すべり発生が起こり得るのかを予測することを目指す。

2. 対象地域と使用データ

本研究の対象地域は宮城県仙台市と2016年熊本地震の影響を受けた熊本県とした。使用した主なデータは国土交通省が整備する国土数値情報⁷⁾、微地形区分⁸⁾、防災科研の地すべり地形分布図⁹⁾である。

国土数値情報は標高・傾斜角などの地形データが5次メッ

シュ (250 mメッシュ) ごとに格納されているものを使用した。地すべり地形分布図⁹⁾は、地すべり地形を航空写真から判読したものである。宮城県仙台市の地すべり地形分布を図-1に示す。

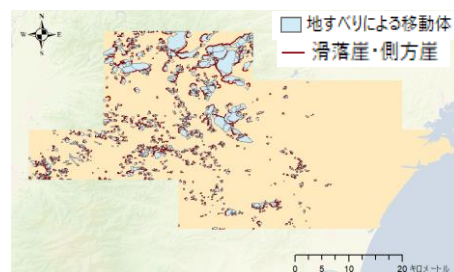


図-1 宮城県仙台市の地すべり地形分布図

3. 共分散構造分析に基づく地すべり発生確率の評価

共分散構造分析では、素因(観測変数)が誘因(潜在変数)の影響を受けるものとし、さらに、それらがいくつかの要因に集約できると仮定する。要因と地すべり発生の確率との相関を評価する。基本パスモデルを図-2に示す。なお、基本パスモデルは既往研究¹⁰⁾を参考に仮定した。

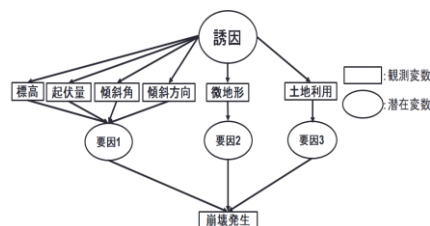


図-2 本研究で仮定した基本パスモデル

メッシュごとの地すべり発生確率をベイジアンモデルとファジーセットモデルの二つの手法により求める¹¹⁾。ベイジアンモデルの発生確率の計算式を式(1)に示す。

$$S_q = \frac{P\{c_{1j}\} \dots P\{c_{mj}\}}{P\{c_{1j}, c_{2j}, \dots, c_{mj}\}} \cdot \frac{P\{T_q\}}{P\{T_q\}} \cdot \frac{P\{T_q|c_{1j}\}}{P\{T_q\}} \dots \frac{P\{T_q|c_{mj}\}}{P\{T_q\}} \quad (1)$$

ここで、式(1)を計算するためには、

$$P\{c_{ij}\} = \frac{N_{ij}}{A} \quad (2)$$

$$P\{c_{1j}, \dots, c_{mj}\} = \frac{N_x}{A} \quad (3)$$

$$P\{T_q\} = \frac{N_0}{A} \quad (4)$$

$$P\{T_q|c_{ij}\} = \frac{N_{d_{ij}}}{N_{ij}} \quad (5)$$

をそれぞれ代入する必要がある。ここで、 N_{ij} は*i*番目の素因におけるカテゴリ*j*のメッシュ数、 A は全メッシュ数、 N_x はメッシュ*q*における*i*~*m*番目の素因に対する全カテゴリにおいて等しい属性のメッシュ数、 N_0 は地すべり地形のメ

ッシュ数、 $N_{d_{ij}}$ は*i*番目の素因におけるカテゴリ*j*の領域に

属する地すべり地形のメッシュ数である。また、ファジーセットモデルは式(6)に帰着する。

$$S_q = \mu_s \{q|c_{1j}, \dots, c_{mj}\} = 1 - \prod_{i=1}^m \left\{ 1 - \frac{N_{d_{ij}}}{N_{ij}} \right\} \quad (6)$$

ベイジアンモデル、ファジーセットモデルの両方で地すべり発生確率を計算し、図-3の基本パスモデルの適合度を算出する。なお、共分散構造分析に用いる観測変数の値は式(5)の条件付き確率とし、平均0、標準偏差1に標準化してある。さらに、基本モデルから観測変数を1または2個減らしたモデルから、最適なパスモデルを探索した。

それぞれのモデルの地すべり地形の的中率を算出し、最も的中率の高いモデルを採用することとする。ここでの的中率とは、実際に地すべり発生危険がある地点を「危険あり」、危険がない地点を「危険なし」とそれぞれ正しく判別したメッシュ数の和を全体のメッシュ数で除したものである。具体的には、横軸にパスモデルをもとに算出した標準化された地すべり発生確率を、縦軸にその度数および累積頻度を示す。さらに、負極側から地すべり地形の累積頻度曲線を、正極側から非地すべり地形の累積頻度曲線を描く(図-4)。つまり、地すべり地形をAグループ、非地すべり地形をBグループとすると、

$$F_A(X) = 1 - \int_{-\infty}^X f_A(x) dx \quad \int_{-\infty}^{+\infty} f_A(x) dx = 1 \quad (7)$$

$$F_B(X) = \int_{-\infty}^X f_B(x) dx \quad \int_{-\infty}^{+\infty} f_B(x) dx = 1 \quad (8)$$

となり、 $F_A(X)$ 、 $F_B(X)$ の交点が判別の閾値となる。

図-5に最もの中率の高かったファジーセットモデルにおける基本モデルから起伏量と土地利用を抜いたモデルのパス図と的中率を示す。パスモデルの相関係数から、地すべりには要因1(標高、起伏量、傾斜角、傾斜方向)が大きく影響していることがわかる。また、要因1内の相関係数を見ると、標高と傾斜角の相関係数が大きいことがわかる。採用したパスモデルに用いられている説明変数を使用して、次章の機械学習を行う。

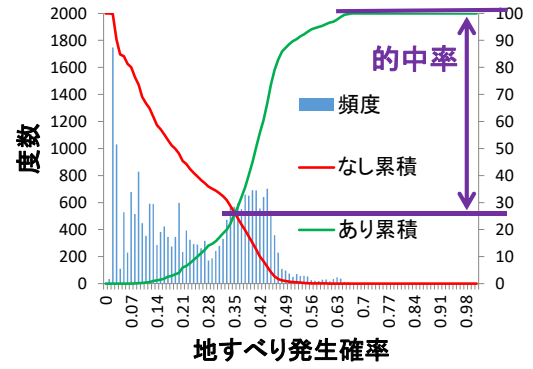


図-4 的中率を求める際のグラフ

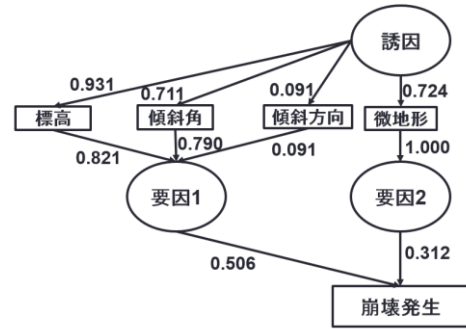


図-5 採用したパスモデル(的中率: 74.00%)

4. 機械学習に基づく地すべり発生箇所の評価

本研究では、共分散構造分析に加えて、機械学習に基づく分析を実施した。ここでは、サポートベクターマシンとランダムフォレストの二手法を利用した。全データの10%を学習データとして地すべり発生箇所の分類器を作成し、学習に使用しなかったデータをテストデータとして分類器の精度評価に利用した。

(1) サポートベクターマシン

サポートベクターマシン(SVM)は、教師あり学習を用いたパターン認識手法の一つである。基本的には2つのクラスを識別するための識別機を構成するための学習手法とされており、認識性の優れた学習モデルの一つといわれている。

SVMによる2クラス分類では、式(9)で表される最適化問題を解く。ここで、 $\mathbf{w}$ は分離超平面の法線ベクトル、 ϕ は入力ベクトル $\mathbf{x}$ を特徴空間 F へ非線形写像する関数(式(10))、 b はスカラー変数を表す¹¹⁾。

$$\begin{aligned} \min_{\mathbf{w}, b, \zeta} & \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i \in [n]} \zeta_i \\ \text{s.t. } & y_i (\mathbf{w}^T \phi(\mathbf{x}_i) + b) \geq 1 - \zeta_i, i \in [n], \zeta_i \geq 0, i \in [n] \end{aligned} \quad (9)$$

$$K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j) \quad (10)$$

さらに、本研究では非線形写像のための関数にRBFカーネルを用いる。

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2) \quad (11)$$

式(11)の γ と式(9)の C は、ハイパーパラメータと呼ばれ分類結果に大きく影響する¹²⁾。

本研究ではグリッドサーチ¹³⁾に基づきハイパーパラメータを設定する。グリッドサーチとは、 C と γ の値の範囲を任意に設定し、交差検証法により、最適な C と γ の値を求める手法である。本研究では、全データの10%を学習データ、 γ を $10^{-5} \sim 10^5$ 、 C を $10^{-2} \sim 10^2$ の範囲に設定して、交差検証法($k=10$)によって最適なパラメータを求めた。

図-6に分析結果と既存の地すべり地形分布図との比較を行い、作成した地すべり地形評価図を示す。また、的中率(地すべりの有無を正しく分類した割合)と、地すべり地形のメッシュだけで算出した的中率を示す。的中率は共分散構造分析に比べ向上したが、地すべり発生地点の抽出精度が悪く、特に実際に地すべりの危険がある地点を危険なしと判定する危険個所の見落としが多くみられた。

(2) ランダムフォレスト

ランダムフォレスト(RF)は、複数の決定木を用いる集団学習法(アンサンブル学習)の一つである。各決定木での予測結果を多数決することにより、結果の取得を行う。決定木学習とは、データの種類に応じて決定木を成長、分類させていく学習手法であり、式(12)に示す情報利得IGが最大となるようにする¹⁴⁾。

$$IG(D_p, f) = I(D_p) - \sum_{j=1}^m \frac{N_j}{N_p} I(D_j) \quad (12)$$

ここで、 D_p は学習データ、 N はノード、 j は注目している

データ、 I は不純度、 m は個々の木を分割するノード数である。不純度はデータに偏りがあるほど、大きな値になる指標であり、エントロピー、ジニ係数、分類誤差などが用いられる。式(13)はエントロピーの式¹⁵⁾を表している。

$$I(N) = -\sum p(k|N) \log p(k|N) \quad (13)$$

ここで $p(k|N)$ は、各ノード N でクラス k を取る確率である。

図-6に地すべり地形評価図および的中率を示す。SVMによる分析と同様に、的中率そのものは向上したものの、地すべり発生地点の抽出の精度が低く、危険個所の見落としが多くみられた。

(3) 不均衡データの整形

SVMとRFの結果、地すべり発生地点の見落としが多かったことが課題として挙げられた。その原因としては、本研究で取り扱っている地すべり地形分布図のデータが、正例(地すべり発生が認められるメッシュ)に対して負例(地すべり発生が認められないメッシュ)の数が極端に多い不均衡データであるため、負例を過剰に抽出しまったことが考えられる。そこで、本研究ではこのような不均衡データにおける誤分類を解決するための手法として、サンプリング法の一つであるSMOTE(Synthetic Minority Over-sampling Technique)¹⁷⁾を適用した。

SMOTEとは、Chawlaらによって提案されたオーバーサンプリング法の一つであり、 k -最近傍法を基にしたアルゴリズムである。任意の正例を指定し、類似度計算により最近傍にある他の正例を特定する。そして、その二点を結ぶ直線上に新たな正例を作成する。この作業を指定の回数繰り返すことで正例の個数を増やし、データ間の偏りを減少

させる¹⁸⁾。式(14)は類似度計算の式である。

$$\text{sim}(m_a, m_i) = \sqrt{\sum_{j=1}^n (v_{a,j} - v_{i,j})^2} \quad (14)$$

ここで、 m は任意の点、 v は点 m における説明変数、 n は説明変数の数を表す。

本研究では、正例と負例の比が1:4なので、データが均衡になるように、正例を4倍にオーバーサンプリングして学習データを作成、分析を行った。図-6に不均衡データ整形後のSVMおよびRFの地すべり地形評価図と的中率を示す。データ整形前と比べて、全体的中率は下がったが、地すべり発生地点の抽出の精度は向上したため、オーバーサンプリングの効果と考えられる。

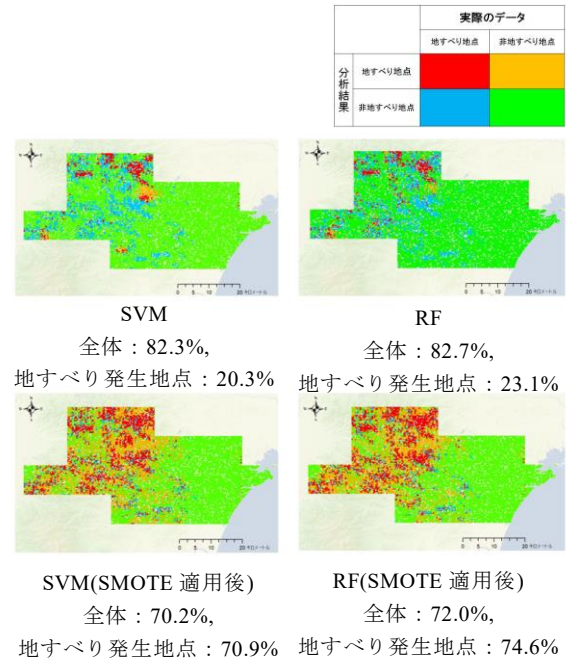


図-6 各手法の地すべり地形評価図と的中率

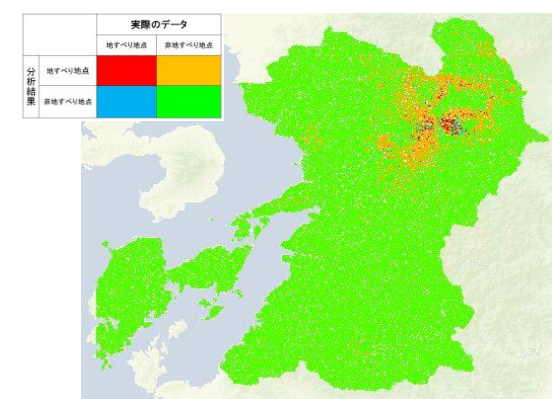
5. 熊本地震による土砂崩壊への適用

前章までの検討では、地すべりを引き起こした誘因を考慮していない。そこで、熊本地震によって発生した土砂崩壊地の地形的特徴を前章と同様にSVM、RFの二手法の機械学習を用いて評価した。本章では前章までの分析で用いた説明変数のほかに、土砂崩壊の誘因として地震動強さを考慮するため、QuiQuake¹⁹⁾の最大速度(PGV)分布を追加した。また、土砂崩壊発生地点のデータには国土地理院の土砂崩壊地分布図²⁰⁾を利用した。これは、国土地理院が熊本地震後に撮影した航空写真から、土砂崩壊地の分布を判読したものである。対象は熊本県内109741メッシュである。

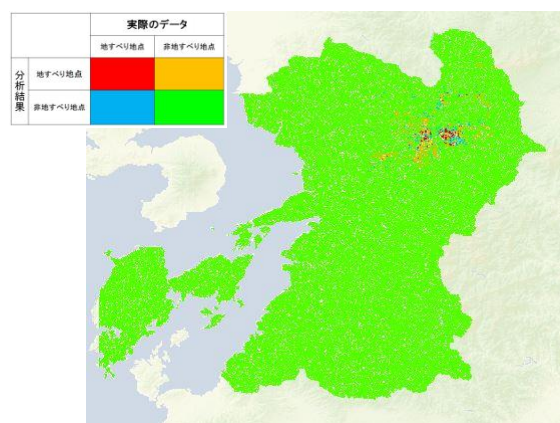
仙台市の分析と同様に、全データの10%を学習データとして分類器を構築し、残りの90%を予測データとして精度を評価した。仙台市のデータと比べより負例の数が多不均衡データであったため、分析結果が負例に引っ張られてしまいどちらの手法でも土砂崩壊地の抽出的中率が0%となった。そこで、この問題を解決するためにSMOTEによるオーバーサンプリングを行い、再度分析を実行した。正例と負例の比率から、正例が160倍となるようにデータをオーバーサンプリングした。

データ整形後の機械学習の分析結果から作成した土砂崩

壊評価図および各手法の的中率を図-7に示す。仙台市の分析結果とは異なり、各手法の的中率の差がみられ、SVMでは土砂崩壊地抽出の精度が、RFでは全体の的中率が高かった。



SVM(SMOTE 適用後)
全体：93.6%，土砂崩壊発生日点：60.0%



RF(SMOTE 適用後)
全体：98.6%，土砂崩壊発生日点：25.4%

図-7 熊本地震後の土砂崩壊評価図

6. まとめ

本研究では、共分散構造分析とサポートベクターマシン、ランダムフォレストの二種類の機械学習手法に基づき、宮城県仙台市との地すべりの危険箇所の特徴を評価した。また、誘因を加えた分析として、熊本県内において熊本地震により発生した土砂崩壊地の特徴を評価した。

機械学習を実行した結果、負例が正例に比べて極端に多い不均衡データであったため、分析による危険箇所の見落としが多くなる問題が起こった。そこで、SMOTEによるオーバーサンプリングを行い、データを整形したところ、機械学習の精度を向上させることができた。機械学習の二手法を比較すると、仙台市の分析においてはSVMとRFの全体精度、地すべり地形抽出精度ともにほぼ同程度であった。熊本県の分析では、土砂崩壊の誘因としてPGVを考慮したことが影響し、機械学習の手法間での精度の違いが見られ、SVMの精度がRFよりも高かった。

参考文献

- 1) 新潟県土木部砂防課：新潟県中越地震と土砂災害，
http://www.pref.niigata.lg.jp/HTML_Article/864/780/chuet_sujisin.pdf
- 2) 国土交通省：災害・防災状況，
http://www.mlit.go.jp/saigai/saigai_110311.html
- 3) 国土交通省砂防部：平成26年8月豪雨による広島県で発生した土砂災害への対応状況，
http://www.mlit.go.jp/river/sabo/H26_hiroshima/141031_hiroshimadosekiryu.pdf
- 4) 豊田秀樹：共分散構造分析[R編]，東京図書，2014.
- 5) 栗田多喜夫：サポートベクターマシン入門，<http://home.hiroshima-u.ac.jp/tkurita/lecture/svm.pdf>
- 6) 波部 斉：ランダムフォレスト，<https://www.slideshare.net/HitoshiHabe/ss-58784421>
- 7) 国土交通省：国土数値情報ダウンロードサービス，
<http://nlftp.mlit.go.jp/ksj/>
- 8) 若松加寿江，松岡昌志，久保純子，長谷川浩一，杉浦正美：日本全国地形・地盤分類メッシュマップの構築，土木学会論文集，No. 759/1-67, 213-232, 2004.
- 9) 防災科学技術研究所：地すべり地形分布図データベース，<http://lsweb1.ess.bosai.go.jp/>（2015年5月28日閲覧）
- 10) 小島尚人，大林成行，青木太：共分散構造分析を導入した斜面崩壊危険箇所評価アルゴリズムの構築，土木学会論文集，No.714/VI-56, pp.79-93,2002.
- 11) 大林成行，小島尚人，Chang-Jo F.Chung：斜面安定性評価モデルの精度比較とその実用化への提案，土木学会論文集，No.630/VI-44, pp.77-89,1999.
- 12) 竹内一郎，鳥山昌幸：サポートベクトルマシン（機械学習プロフェッショナルシリーズ），講談社，2015
- 13) 荒川正幹，宮尾知幸，船津公人：ドラッグライクネスモデルの構築とその可視化，Journal of Computer Aided Chemistry，Vol. 9, pp. 70-80, 2008.
- 14) 波部 斉：ランダムフォレスト，コンピュータビジョンとイメージメディア，pp. 1-8, 2012.
- 15) 山岡 啓介：ランダムフォレスト，映像情報メディア学会誌，Vol. 66, No. 7, pp. 573-575, 2012.
- 16) 小林寛武，戸田航史，亀井靖高，門田暁人，峯恒憲，鶴林尚靖：11種類のfault密度予測モデルの実証的評価，電子情報通信学会論文誌D，Vol. 96, No. 8, pp. 1892-1902, 2013.
- 17) Rpubs：SMOTEで不均衡データの分類，
https://rpubs.com/hoxo_m/54954（2017年7月4日閲覧）
- 18) 亀井靖高，門田暁人，松本健一：Fault-pronenessモデルへのオーバーサンプリング法の適用，
<http://www.empirical.jp/paperdb/papers/archive/102/102.pdf>（2017年7月4日閲覧）
- 19) 産業技術総合研究所：QuiQuake—地振動マップ即時推定システム—，<https://gbank.gsj.jp/QuiQuake/>（2016年10月11日閲覧）
- 20) 国土地理院：土砂崩壊地分布，
<http://www.gsi.go.jp/BOUSAI/H27-kumamoto-earthquake-index.html#cc>