SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
ENVIRONMENTAL SCIENCES
GEOCHEMISTRY
GEOLOGY

Received
7 July 2014

Accepted
30 October 2014

Published
17 November 2014

Correspondence and requests for materials should be addressed to T.K. (kuwatani@mail.kankyo.tohoku.ac.jp)

Machine-learning techniques for geochemical discrimination of 2011 Tohoku tsunami deposits

Tatsu Kuwatani[1], Kenji Nagata[2], Masato Okada[2,3], Takahiro Watanabe[1], Yasumasa Ogawa[4], Takeshi Komai[1] & Noriyoshi Tsuchiya[1]

[1]Graduate School of Environmental Studies, Tohoku University, Sendai 980-8579, Japan, [2]Graduate School of Frontier Sciences, University of Tokyo, Kashiwa 277-8561, Japan, [3]Brain Science Institute, RIKEN, Wako 351-0198, Japan, [4]Graduate School of Engineering and Resource Science, Akita University, Akita 010-8502, Japan.

Geochemical discrimination has recently been recognised as a potentially useful proxy for identifying tsunami deposits in addition to classical proxies such as sedimentological and micropalaeontological evidence. However, difficulties remain because it is unclear which elements best discriminate between tsunami and non-tsunami deposits. Herein, we propose a mathematical methodology for the geochemical discrimination of tsunami deposits using machine-learning techniques. The proposed method can determine the appropriate combinations of elements and the precise discrimination plane that best discerns tsunami deposits from non-tsunami deposits in high-dimensional compositional space through the use of data sets of bulk composition that have been categorised as tsunami or non-tsunami sediments. We applied this method to the 2011 Tohoku tsunami and to background marine sedimentary rocks. After an exhaustive search of all 262,144 ($=2^{18}$) combinations of the 18 analysed elements, we observed several tens of combinations with discrimination rates higher than 99.0%. The analytical results show that elements such as Ca and several heavy-metal elements are important for discriminating tsunami deposits from marine sedimentary rocks. These elements are considered to reflect the formation mechanism and origin of the tsunami deposits. The proposed methodology has the potential to aid in the identification of past tsunamis by using other tsunami proxies.

A tsunami deposit provides direct evidence of the inundation area of past tsunamis. A large number of publications have described the diagnostic signatures and identification criteria for past tsunamis, including evidence from sedimentology, geomorphology, stratigraphy, archaeology, anthropology, and macro- and micropalaeontology[1–12]. However, it is still difficult to identify tsunami deposits because the criteria thus far determined are neither applicable nor sufficient due to the various origins, mechanisms, and temporal variation of tsunami deposits. Although microfossils are considered one of the most useful signatures of tsunami deposits, not all tsunami deposits contain microfossils[1,13]. In the 2011 Tohoku tsunami, distinctive sand deposits, which are widely used criteria for a tsunami inundation area[2,8,9], extended to only 60% of the total inundation distance in the Tohoku district[14]. This observation suggests that sand is deposited relatively early seaward, and thus, the sand layer can only be used as a lower limit for the inundation area.

The geochemical discrimination of recent and past tsunami deposits is now recognised as a useful and quantitatively tractable proxy, especially when other proxies cannot be used[11,13–16]. Recent studies of modern tsunami sediments have described the geochemical signatures of tsunami sediments for the 2004 Indian Ocean tsunami[10,17–23], the 2009 South Pacific tsunami[10] and the 2011 Tohoku tsunami[14,24,25]. Sediment geochemistry has been successfully used for the identification of past tsunami deposits together with other proxies[11–13,26–30]. In particular, it has been suggested that geochemical indicators can be useful in the identification of tsunami deposits by aiding in the identification of the marine origin of fine-grained sediments beyond the limit of recognisable sand deposition[13,14]. A large amount of high-dimensional data on bulk composition have been acquired during the process of geological sampling by many researchers[14]; however, these data are not always used.

Despite the high potential of geochemistry for identifying past tsunami deposits, the geochemical proxy is still not recognised as one of the standard proxies, which include sedimentological and micropalaeontological evidence, most likely because the geochemical signature depends on both the source and the background material[11,13,14]. In addition, the geochemical signature can be modified after its deposition by various physical and

chemical processes, such as weathering and diagenesis[31]. Thus, universal criteria for a geochemical proxy cannot be as fully established as for other proxies. Instead, with geochemical analyses, it is very important to establish the best discrimination criteria for each situation, taking the particular location and background composition for each tsunami into consideration. Therefore, the development of mathematical methodology is essential in establishing the most appropriate criteria by using the available data sets at the maximum extent.

In this study, we propose a mathematical methodology to establish the criteria for the geochemical discrimination of the 2011 Tohoku tsunami deposits and the background marine sediments using machine-learning techniques. Machine learning is the science of using computers for the automatic detection of patterns in data, and this field has recently developed rapidly in association with the rapid progress of computational capability[32]. Machine-learning techniques enable us to maximise the amount of useful knowledge and information extracted from the available high-dimensional data and to construct models for making predictions for the unknown data. Two powerful techniques are used: one is a support vector machine (SVM) classifier used for determining the appropriate decision plane that categorises samples into tsunami deposits and non-tsunami sediments[33], and the other is a cross validation (CV) technique used for evaluating the discrimination performance of the SVM for each combination of elements used for discrimination[32].

In this study, we focused the geochemical discrimination on discerning the 2011 Tohoku tsunami deposits from the background marine sediments. However, a single use of a geochemical proxy cannot definitively identify the tsunami deposits, as is the case with other proxies[13]. Integration of all the available proxies is required for the precise identification of past tsunami deposits. Moreover, at present, the results of the analysis cannot contribute to the discrimination between tsunami deposits and other marine inundation events such as high-energy storm events. The discrimination between a tsunami event and a storm event is very important because these events have largely different risk profiles, but both create very serious problems[5,7,26,34–37]. However, the subtle differences in the geochemistry of deposits resulting from the physical and chemical processes of these two types of events can potentially be detected by applying the proposed method to discriminate between these two types of high-energy events in a specific area.

Results

We used 129 samples of the deposits from the March 11, 2011, Tohoku tsunami. Samples were taken along the coastline from Kuji City, Iwate Prefecture, via Miyagi Prefecture, to Minami-Soma City, Fukushima Prefecture, from April to August 2011[38] (Fig. 1a). Many tsunami deposits were sampled from a depth of 0.5–5 cm underground; these deposits consisted primarily of mud and sand, ranging from silt to coarse sand. In the stricken region, the tsunami deposits were sampled only from well-preserved locations, such as inundation points and basements in surviving buildings, because much of the sediment was disturbed by the removal of rubble. Although we cannot deny the possibility of a slight modification of the geochemical pattern by diagenesis and/or leaching by rainfall, Chague-Goff and co-workers reported that the geochemical signature was still retained in most sediment samples seven months after the 2011 Tohoku tsunami[14]. The details of the sampling and analytical methods will be discussed in the Methods section.

For the non-tsunami sediments, we used 75 samples of rock and soil, including sand, silt, and the shell-bedding layer, sampled from various Neogene marine sedimentary layers distributed in the inland area and the coastal area of the Tohoku district. These samples were taken from Plio-Pleistocene-age sediments along the Pacific coastline, especially in the Sendai Plane (Tatsunokuchi Formation), which had been deposited in shallow and embayment degree under mostly reduced conditions[39–41]. Siltstone and mudstone were mainly composed of quartz, feldspar, and montmorillonite with additional

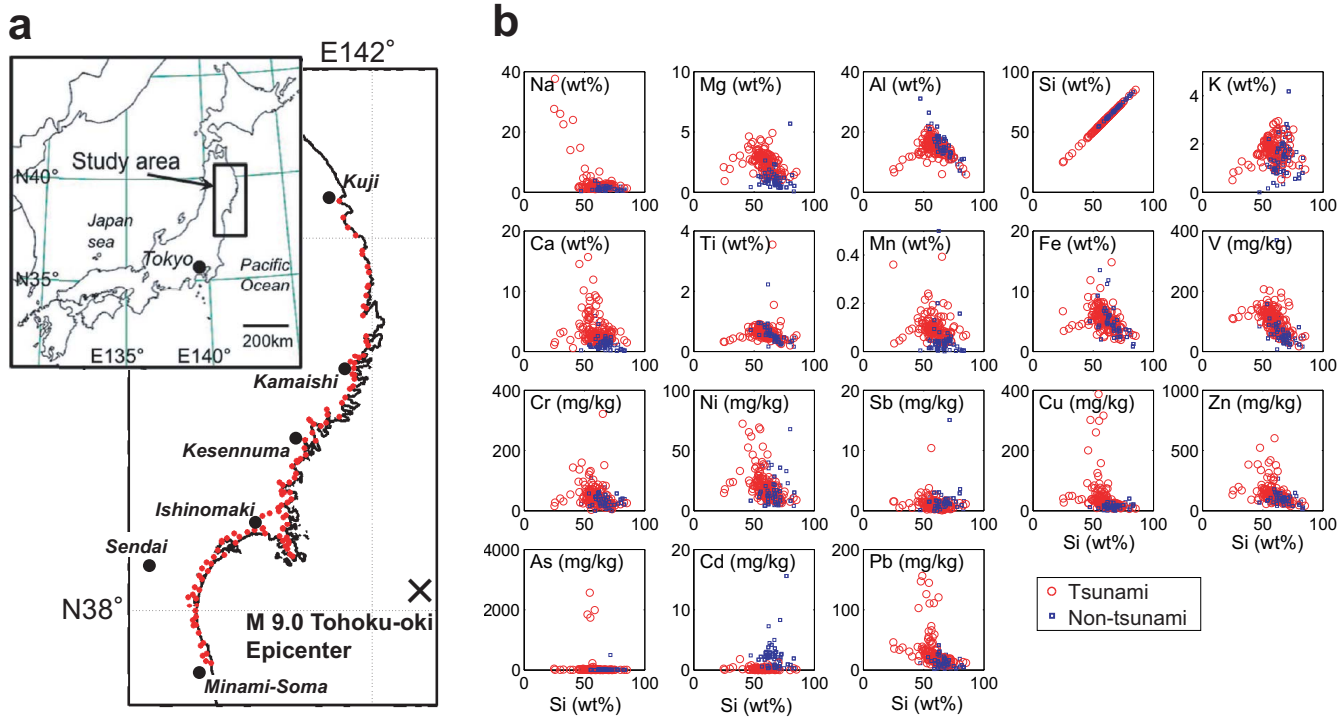

Figure 1 | (a) Sampling locations of 2011 Tohoku tsunami deposits[38]. The locations of the tsunami samples are indicated by red circles. This figure was generated using MATLAB. (b) Scatter diagrams of whole-rock compositions. The horizontal axis is Si, and the vertical axes are the elements named at the top of each diagram. Red circles indicate the tsunami deposits, and blue squares indicate the non-tsunami sediments. The unit wt% indicates weight percent oxides.

arsenopyrite, pyrite, and gypsum as accessory minerals. Several contemporaneous heterotopic sedimentary facies of the Tatsunokuchi Formation are distributed in NE Japan. These marine sediments contain relatively high amounts of arsenic and cadmium, which cause current soil contamination along the Pacific coast and hinterland[42].

In this study, data analyses were conducted using the bulk compositions of the following 18 major elements and heavy-metal elements: Na, Mg, Al, Si, K, Ca, Ti, Mn, Fe, V, Cr, Ni, Sb, Cu, Zn, As, Cd, and Pb. These heavy-metal elements were selected because they are important to environmental pollution and thus are commonly analysed in sediment samples. For some elements (e.g., Fe, Mn, and V), the distribution areas of tsunami and non-tsunami sediments overlap, which indicates that precise discrimination would be difficult using only a single element (Fig. 1b).

For our first task, we considered the multi-dimensional compositional space defined by the bulk compositions and then determined the decision hyperplane, which divides the samples into tsunami deposits and non-tsunami sediments. This type of problem is known as supervised classification in the fields of machine learning and pattern recognition[32]. In supervised classification, the available labelled data (those for which the class is known) are regarded as the training problem of the supervisor and are used to determine the decision hyperplane that best classifies the data. In this study, we used N-dimensional bulk compositional data $\boldsymbol{x}$, labelled as either tsunami or non-tsunami sediment, to determine a linear decision hyperplane in the N-dimensional compositional space: $\boldsymbol{w} \cdot \boldsymbol{x} + b = 0$, where $\boldsymbol{w}$ is the weight vector that determines the slope of the hyperplane and b is the bias, as shown in Fig. 2a. For this classification, we used a powerful supervised-clustering method, the SVM, which was developed during the 1990s[33].

For effective discrimination, we must determine the best elements to use and those that can be ignored; thus, the discrimination performance of the hyperplane obtained by the SVM was evaluated for each combination of elements. It is important to note that the intended aim of the discrimination method is not to analyse the available known data but, rather, for new unknown samples of the 2011 Tohoku tsunami, which have not yet been classified as tsunami or non-tsunami deposits.

We calculated the discrimination rate for unknown samples using CV, which is a versatile and simple evaluation technique for machine learning[32]. In the CV method, the discrimination rate for unknown data is precisely and objectively calculated using only the available data, the labels of which are already-known. The available data set is divided into two subsets as follows: one subset is for the training of the learning model (in this case, the discrimination plane), and the other is for the evaluation of the classification capability for the trained model. In other words, we first analyse one part of the data as if it were unknown data and then evaluate the discrimination rate. When we choose to use or not use all combinations of elements from N elements, the number of subsets is 2^N, including the null set (using no elements). Thus, the analyses were conducted for all 262,144 (2^{18}) combinations of elements.

We first present the results for the 7 major elements (Na, Mg, Al, Si, K, Ca, and Fe), which each constitute greater than 1wt% of the earth's crust (Fig. 3a, c–e). Most combinations of these elements had a discrimination rate of approximately 80–90% (Fig. 3a). The highest discrimination rate was 91.2%, which was achieved for the three-element combination of [Mg, Si, Ca]. Fig. 2a, b shows that many of the tsunami deposits are distributed on the high-Ca side of the decision plane, whereas the non-tsunami sediments are on the low-Ca side. Alternatively, the lowest discrimination rate, 62.9%,

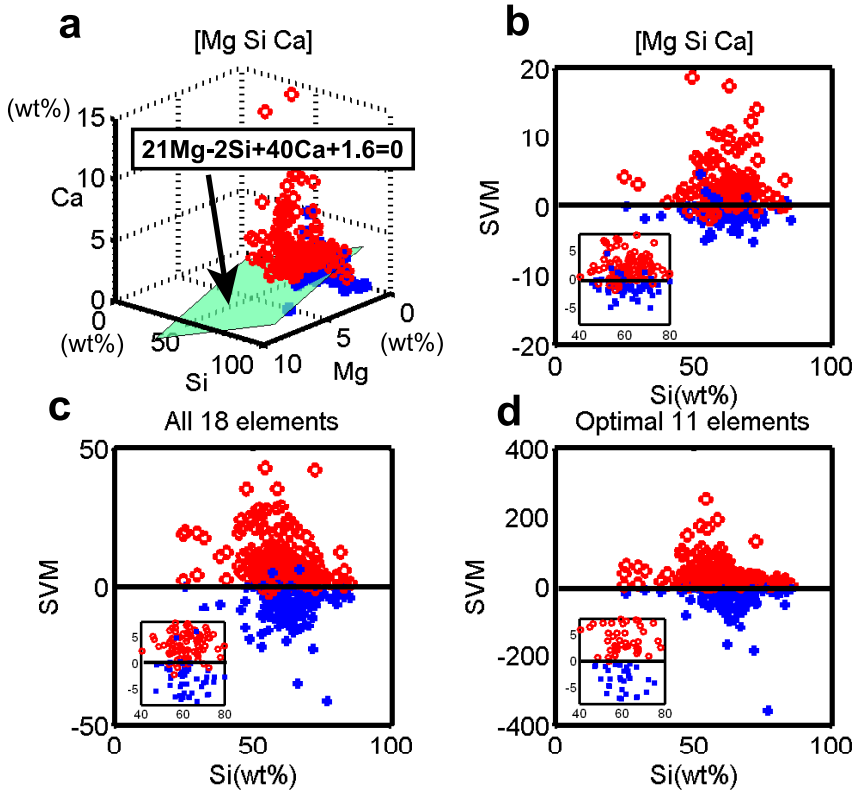

Figure 2 | Determination of the appropriate decision hyperplane. (a) Decision hyperplane in the compositional space for the combination [Mg, Si, Ca] with the highest discrimination rate among the combinations of 7 major elements. (b–d) The scatter diagrams of test data for discrimination using a support vector machine (SVM) for [Mg, Si, Ca] (b), all 18 elements (c), and the optimal 11 elements (d). The discrimination rates are 91.2%, 95.6%, and 100%, respectively. The horizontal axis is Si, and the vertical axis is $\boldsymbol{w} \cdot \boldsymbol{x} + b$ for each combination of elements.

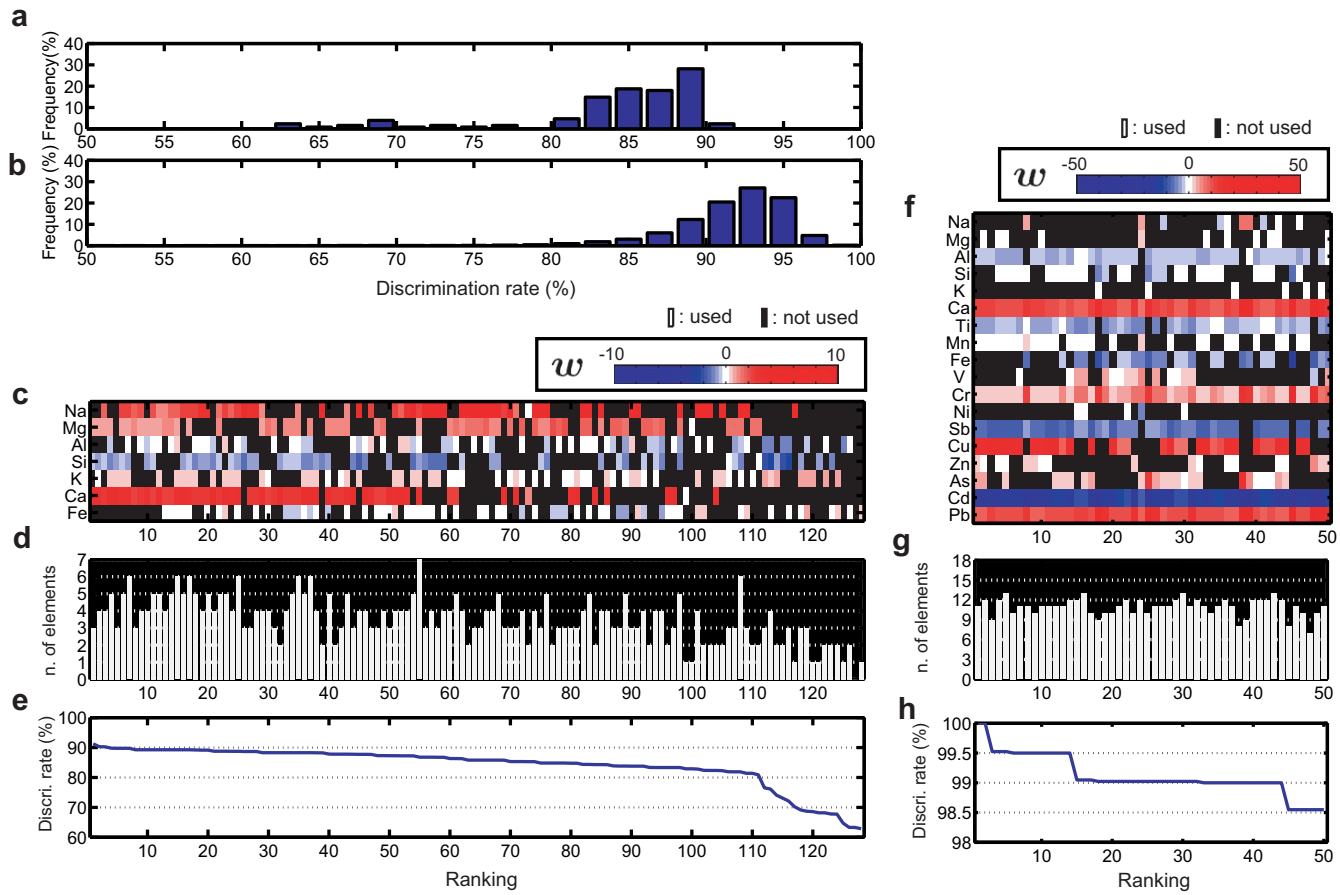

Figure 3 | Results of the cases using (a, c–e) the major 7 elements (128 cases) and (b, f–h) all 18 elements (242,144 cases). (a, b) Histograms of the discrimination rates. (c, f) Combinations of elements used or not used and the slope of the decision plane for each element. The elements not used are shown in black, and the slope of the decision plane $\boldsymbol{w}$ for each element is illustrated by a colour map in which red indicates positive and blue indicates negative. For comparison, the slope of $\boldsymbol{w}$ is obtained for the case in which each composition is linearly transformed with its average and variance set to 0 and 1, respectively. (d, g) Number of elements used. (e, h) Discrimination rates. In (c–e) and (f–h), horizontal axes are the ranking of the discrimination rates in descending order. All 128 cases using the major 7 elements are shown in (c–e), whereas only the top 50 cases with discrimination rates higher than 98.5% are shown in (f–h).

was achieved using only a single element, [Fe]. The discrimination rate of the case using no elements (null set) was 63.3%, which is nearly equal to the proportion of the tsunami samples of the total (129/204 = 63.2%). This result occurs because the decision hyperplane was randomly drawn using no information about the elemental composition. Fig. 3c shows that Na, Mg, and Ca appear frequently in combinations that have high discrimination rates, which indicates that these elements are important in the discrimination of tsunami deposits. Alternatively, Al, Si, K, and Fe are dispersed in higher- to lower-ranking combinations, indicating that they are comparably unimportant elements. Notably, the specific combinations of [Mg, Ca] and [Si, Ca] are important because they frequently appear in combinations that have high discrimination rates.

When using all 18 elements in various combinations, the mode value of the peak performance is approximately 92–94% (Fig. 3b). Overall, the proportions of combinations with high discrimination rates are higher than those that used only the 7 major elements. This tendency for higher discrimination rates when using 18 elements reflects that more information can be derived from using more elements. For example, there were 44 combinations with discrimination rates higher than 99.0%, accounting for 0.017% of all combinations (Fig. 3f–h). The highest discrimination rate was 100.0%, which was achieved by two combinations, using either 11 or 12 elements: [Al, Ca, Ti, Mn, Cr, Sb, Cu, Zn, As, Cd, Pb] and [Mg, Al, Ca, Ti, Mn, Cr, Sb, Cu, Zn, As, Cd, Pb] (Fig. 2d). Alternatively, the lowest discrimination rate was 57.3%, achieved when using the four elements [Al, Ti, Fe, Cr]; there were 76 combinations with a discrimination rate lower than 63.3%. In addition to Ca, several other elements, such as Cr, Cd, and Pb, were important for precise discrimination (Fig. 3f). The combinations with discrimination rates higher than 99.0% included [Al, Ca, Ti, Mn, Cr, Sb, Cu, Cd, Pb] and [Ca, Fe, Cr, Cd, Pb].

Discussion

In general, the discrimination rate increases with the number of elements used for discrimination. This observation is consistent with our intuitive sense that as the amount of known information increases. However, the best discrimination was not attained by using all 18 elements, which was only 95.6%, corresponding to 10,250th place (Fig. 2d). This phenomenon is known as overfitting, which is a very familiar concept in information science for classification and prediction problems. When the number of explanatory parameters is sufficiently greater than the number of samples or when the prediction model is too complex for the data sets, even though a model may classify training data very well, the classification and prediction capability may fail for unknown data sets[32]. Thus, to avoid overfitting, it is important to select the appropriate combinations of elements based on a data-driven approach that formulates a prediction model by utilising the available data sets to the maximum extent.

We also conducted classical multivariate analyses, such as correlation analysis, principal component analysis (PCA) and cluster analysis (Fig. S2–4). We can find that the behaviours of several sets of elements resemble each other using the correlation analysis (Fig. S2). Based on the variance and covariance structure, the PCA is able to extract the important low-dimensional subspace from the high-dimensional geochemical data (Fig. S3). The cluster structure of geochemical data can be extracted by the cluster analysis (Fig. S4). These multivariate analyses are very useful in capturing the statistical structure of geochemical data sets. However, these analyses cannot utilise the label data for tsunami or non-tsunami sediments. For discriminating between tsunami and non-tsunami sediments, the proposed method has a very large advantage over the previous multivariate methods because it utilises labelled (supervised) data. In addition, the proposed method can obtain the most important combination of elements using an exhaustive search, which aids in understanding the geochemical process of tsunami sediments.

A high content of Na, Mg, and Ca is important for a precise discrimination, and these elements were previously identified as important in identifying tsunami deposits[13,14]. This key role is in accordance with the fact that these elements are abundant in sea water and marine sediments derived from bioclasts and shell hash. The slopes of the elements Si and Al, which are the main components in sediment, are either positive or negative, which contributes to the precise discrimination for a wide range of compositions of sand and mud. The slopes of the elements Fe and Ti are negative, which indicates that the low content of these elements may be used as an indicator for tsunami deposits. Similar observations have been reported for the 2011 Tohoku-oki tsunami deposits in the central part of the Sendai Plain by Chagué-Goff et al.[14]. These results are in agreement with the observation that Fe and Ti have lower concentrations in seawater than in freshwater[43] and have been used as indicators for the terrestrial origin of sediments[28].

For the minor elements, the weights of Cr, Cu, and Pb are positive, whereas those of Sb and Cd are negative. Tsuchiya et al. proposed a possible scenario for the 2011 Tohoku tsunami in which several of the heavy-metal elements, including As, originated from marine sediments deposited due to mining in the Tohoku district[38]. A similar process has been reported for the 2004 Indian Ocean tsunami[17]: The high amounts of heavy-metal elements in tsunami sediments most likely originated from marine sediments containing the anthropogenic heavy metal[17,21]. In a future study, we intend to more fully investigate the origin of these tsunami sediments and the mechanisms that produce the concentrated elements.

Recent studies suggest that geochemical proxies have great potential in identifying historical and palaeotsunami deposits for postdepositional processes such as leaching and weathering[14,29]. With our discrimination method, we can consider many elements, which may contribute to a better discrimination. Thus, by applying our method to data from a historical tsunami for which there is concrete sedimentological evidence, we might be able to determine a set of important elements that will be robust to such postdepositional processes. Because chemical composition and its post-depositional changes may vary according to the origin of tsunami sediments and locality, spatial information and several proxies, including sedimentology and palaeontology, are recommended for use in more precise discrimination. Following the 2004 Indian Ocean tsunami and the 2011 Tohoku tsunami, many researchers have investigated past tsunami deposits all over the world, including analyses of the chemical compositions of the deposits. Notably, many samples have been investigated using chemical scanning techniques. This study can contribute to the analysis of these enormous chemical data sets. Additionally, the proposed method can be widely used for high-dimensional geochemical data sets, regardless of the target.

Methods

Chemical analysis. During the chemical analysis preparation, samples were air dried, gravel and organic matter were removed, and the samples were then divided into tsunami sediments with a grain size smaller than 2 mm using a fine (2 mm) sieve. The whole-sample compositions were analysed using an energy dispersive X-ray fluorescence spectrometer (ED-XRF), Epsilon 5, manufactured by PANalytical. More detailed descriptions of these samples and the chemical analysis have been published previously[38,42,44].

Support vector machine (SVM). A two-class classification problem was considered here for the discrimination of tsunami and non-tsunami sediments. We used N-dimensional compositional data as the input feature vectors and output labels as being from tsunami or non-tsunami sediments. This procedure was performed for samples $i = 1, \ldots, M$ as

$$(\boldsymbol{x}_i, t_i) \in \boldsymbol{R}^N \times \{+1; -1\}, \tag{1}$$

where $\boldsymbol{x}_i$ represents the N-dimensional input vector and t_i represents the output label, tsunami $(+1)$ or non-tsunami (-1), for sample i. The SVM determines the decision function $f: \boldsymbol{R}^N \rightarrow \{+1; -1\}$, called the maximal margin hyperplane. In this study, we considered a linear decision function:

$$f(\boldsymbol{x}) = \mathrm{sgn}(\boldsymbol{w} \cdot \boldsymbol{x} + b), \tag{2}$$

where $\mathrm{sgn}(x)$ is the signum function that gives -1 when $x < 0$ and $+1$ when $x > 0$, $\boldsymbol{w}$ is the weight vector which determines the slope of the hyperplane, and b is the bias. By definition, the maximal margin hyperplane maximises the distance between itself and the boundaries of the two objective classes. The boundary of a class is defined by several input vectors $\boldsymbol{x}_i$s, called support vectors. The hyperplane determined by the SVM is not only the farthest from the two training-data classes, but it is also robust for unknown data sets.

Cross-validation (CV). CV is a versatile and easy evaluation technique for determining the generalisation capability of the trained model in the case that we do not have a sufficient amount of available data. The generalisation capability is the prediction capability for unknown data. Because we apply the discrimination function not to available data but to unknown data sets in actual situations, we must obtain the discrimination function with a high generalisation capability. In the CV method, the available data set is divided into two subsets as follows: one subset is for the training of the learning model, and the other is for the evaluation of classification capability for the trained model. In other words, we regard a part of the data sets as unknown data and then evaluate the prediction capability for unknown data.

In this study, we evaluated the decision function obtained by SVM using 10-fold CV. The K-fold CV method divides the available data set into K subsets: $C_1, \ldots, C_K$. The parameters of the decision function, $\boldsymbol{w}$ and b, are trained using all the data sets, which are included in the $K - 1$ subsets, and then the prediction capability is evaluated by the number of misclassifications when using the trained decision function. The trained and evaluated cycles are repeated for all K combinations, and the generalisation capability is calculated as the mean of all K combinations. Therefore, the CV error ($CV\,E$) is calculated as follows:

$$CV\,E = \frac{1}{M} \sum_{k=1}^{K} \sum_{i \in C_k} E(t_i, f_{-k}(\boldsymbol{x}_i)) \tag{3}$$

where $E(t_i, f_{-k}(\boldsymbol{x}_i))$ is the function that gives 0 for true or 1 for false discrimination for sample i for the decision function trained by all subsets other than subset k. Because the $CV\,E$ is divided by the number of all samples M, it is the ratio of the misclassified data to all data. In this study, we defined the discrimination rate (%) as $(1 - CV\,E) \times 100$.

1. Goff, J., Chagué-Goff, C. & Nichol, S. Palaeotsunami deposits: a New Zealand perspective. *Sediment. Geol.* **143**, 1–6 (2001).
2. Cisternas, M. *et al.* Predecessors to the giant 1960 Chile earthquake. *Nature* **437**, 404–407 (2005).
3. Dominey-Howes, D., Humphreys, G. S. & Hesse, P. P. Tsunami and palaeotsunami depositional signatures and their potential value in understanding the late-Holocene tsunami record. *Holocene* **16**, 1095–1107 (2006).
4. Dominey-Howes, D. Geological and historical records of tsunami in Australia. *Marine Geology* **239**, 99–123 (2007).
5. Dawson, A. G. & Stewart, I. Tsunami deposits in the geological record. *Sediment. Geol.* **200**, 166–183 (2007).
6. Kortekaas, S. & Dawson, A. G. Distinguishing tsunami and storm deposits: An example from Martinhal, SW Portugal. *Sediment. Geol.* **200**, 208–221 (2007).
7. Switzer, A. D., Yu, F., Gouramanis, C., Soria, J. L. A. & Pham, D. T. Integrating different records to assess coastal hazards at multicentury timescales. *J. Coastal Res.* **70**, 723–728 (2014).
8. Jankaew, K. *et al.* Medieval forewarning of the 2004 Indian Ocean tsunami in Thailand. *Nature* **455**, 1228–1231 (2008).
9. Monecke, K. *et al.* A 1000-year sediment record of tsunami recurrence in northern Sumatra. *Nature* **455**, 1232–1234 (2008).

10. Chagué-Goff, C., Schneider, J. L., Goff, J. R., Dominey-Howes, D. & Strotz, L. Expanding the proxy toolkit to help identify past events - Lessons from the 2004 Indian Ocean Tsunami and the 2009 South Pacific Tsunami. *Earth-Science Reviews* **107**, 107–122 (2011).
11. Goff, J., Chagué-Goff, C., Nichol, S., Jaffe, B. & Dominey-Howes, D. Progress in palaeotsunami research. *Sediment. Geol.* **243**, 70–88 (2012).
12. Cuven, S. *et al.* High-resolution analysis of a tsunami deposit: Case-study from the 1755 Lisbon tsunami in southwestern Spain. *Marine Geol.* **337**, 98–111 (2013).
13. Chagué-Goff, C. Chemical signatures of palaeotsunamis: A forgotten proxy? *Marine Geol.* **271**, 67–71 (2010).
14. Chagué-Goff, C., Andrew, A., Szczucinski, W., Goff, J. & Nishimura, Y. Geochemical signatures up to the maximum inundation of the 2011 Tohoku-oki tsunami: Implications for the 869 AD Jogan and other palaeotsunamis. *Sediment. Geol.* **282**, 65–77 (2012).
15. Minoura, K. & Nakaya, S. Traces of tsunami preserved in inter-tidal lacustrine and marsh deposits: Some examples from northeast Japan. *J. Geol.* **99**, 98-111 265–287 (1991).
16. Goto, K., Chagué-Goff, C., Goff, J. & Jaffe, B. The future of tsunami research following the 2011 Tohoku-oki event. *Sediment. Geol.* **282**, 1–13 (2012).
17. Ranjan, R. K., Ramanathan, A., Singh, G. & Chidambaram, S. Assessment of metal enrichments in tsunamigenic sediments of Pichavaram mangroves, southeast coast of India; *J. Environ. Monit. Assess.* **147**, 389–411 (2008).
18. Szczucinski, W. The post-depositional changes of the onshore 2004 tsunami deposits on the Andaman Sea coast of Thailand. *Natural Hazards* **60**, 115–133 (2012).
19. Szczucinski, W. *et al.* Contamination of tsunami sediments in a coastal zone inundated by the 26 December 2004 tsunami in Thailand. *Environ. Geol.* **49**, 321–331 (2005).
20. Szczucinski, W. *et al.* Effects of rainy season on mobilization of contaminants from tsunami deposits left in a coastal zone of Thailand by the 26 December 2004 tsunami. *Environ. Geol.* **53**, 253–264 (2007).
21. Srinivasalu, S. *et al.* Evaluation of trace-metal enrichments from the 26 December 2004 tsunami sediments along the Southeast coast of India. *Environ. Geol.* **53**, 1711–1721 (2008).
22. Srinivasalu, S., Jonathan, M. P., Thangadurai, N. & Ram-Mohan, V. A study on pre- and post-tsunami shallow deposits off SE coast of India from the 2004 Indian Ocean tsunami: a geochemical approach. *Natural Hazards* **52**, 391–401 (2010).
23. Veerasingam, S. *et al.* Identification and characterization of tsunami deposits off southeast coast of India from the 2004 Indian Ocean tsunami: Rock magnetic and geochemical approach. *J. Earth Syst. Sci.* **123**, 905–921 (2014).
24. Jagodzinski, R., Sternal, B., Szczucinski, W., Chagué-Goff, C. & Sugawara, D. Heavy minerals in the 2011 Tohoku-oki tsunami deposits - insights into sediment sources and hydrodynamics. *Sediment. Geol.* **282**, 57–64 (2012).
25. Chagué-Goff, C. *et al.* Environmental impact assessment of the 2011 Tohoku-oki tsunami on the Sendai Plain. *Sediment. Geol.* **282**, 175–187 (2012).
26. Goff, J., Mcfadgen, B. G. & Chagué-Goff, C. Sedimentary differences between the 2002 Easter storm and the 15th-century Okoropunga tsunami, southeastern North Island, New Zealand. *Marine Geol.* **204**, 235–250 (2004).
27. Schlichting, R. & Peterson, C. 2006. Mapped overland distance of paleotsunami: high velocity inundation in back-barrier wetlands of the central Cascadia margin, U.S.A. *J. Geol.* **114**, 577–592 (2006).
28. Ramirez-Herrera, M. T. *et al.* Extreme wave deposits on the Pacific coast of Mexico: tsunamis or storms? – A multi-proxy approach. *Geomorphology* **139**, 360–371 (2012).
29. Nichol, S. L. *et al.* Lagoon subsidence and tsunami on the west coast of New Zealand. *Sediment. Geol.* **200**, 248–262 (2007).
30. Sun, L. *et al.* Preliminary evidence for a 1000-year-old tsunami in the South China Sea. *Sci. Rep.* **3**, 1655 (2013).
31. Shanmugam, G. Process-sedimentological challenges in distinguishing paleo-tsunami deposits. *Nat. Hazards* **63**, 5–30 (2012).
32. Bishop, C. M. *Pattern recognition and machine learning* (Springer, New York, 2006).
33. Vapnik, V. *Statistical Learning Theory* (Wiley-Interscience, New York, 1998).
34. Komatsubara, J. *et al.* Historical tsunamis and storms recorded in a coastal lowland, Shizuoka Prefecture, along the Pacific Coast of Japan. *Sedimentol.* **55**, 1703–1716 (2008).
35. Kumar, K. A., Achyuthan, H. & Shankar, N. *Paleo-tsunami and storm surge deposits* (Taylor & Francis Group, London, 2007).
36. Morton, R. A., Gelfenbaum, G. & Jaffe, B. E. Physical criteria for distinguishing sandy tsunami and storm deposits using modern examples. *Sediment. Geol.* **200**, 184–207 (2007).
37. Nanayama, F. *et al.* Sedimentary differences between the 1993 Hokkaido-nansei-oki tsunami and the 1959 Miyakojima typhoon at Taisei southwestern Hokkaido northern Japan. *Sediment. Geol.* **135**, 255–264 (2000).
38. Tsuchiya, N. *et al.* Risk assessments of arsenic in tsunami sediments from Iwate, Miyagi and Fukushima prefectures, northeast Japan, by the 2011 off the Pacific coast of Tohoku earthquake. *J. Geol. Soc. Japan* **118**, 419–430 (2012).
39. Aida, Y. *et al.* in *Japanese geology* (eds Hatayama, Y. et al.) 59–67 (Kyoritsu shuppan, Tokyo, 2005).
40. Takai, F., Matsumoto, T. & Toriyama, R. *Geology of Japan* (University of Tokyo Press, Tokyo, 1963).
41. Geological Survey of Japan, AIST Seamless digital geological map of Japan 1 : 200,000. https://gbank.gsj.jp/seamless/maps.html. (2014), accessed on Jun 6, 2014.
42. Suto, K. *et al.* Effect of weathering on changes of leaching properties and chemical forms of heavy metals in a sedimentary rock of Tatsunokuchi Formation. *J. Japan Soc. Enginee. Geol.* **54**, 181–190 (2010).
43. Wedepohl, K. H. Environmental influences on the chemical composition of shales and clays. *Phys. Chem. Earth* **8**, 305–333 (1971).
44. Matsunami, H. *et al.* Rapid simultaneous multi-element determination of soils and environmental samples with polarizing energy dispersive X-ray fluorescence (EDXRF) spectrometry using pressed powder pellets. *Soil Sci. Plant Nutrition* **56**, 530–540 (2010).

Acknowledgments

This work was supported by a Grant-in-Aid for Scientific Research on Innovative Areas, 'Initiative for high-dimensional data-driven science through deepening sparse modeling', Nos. 25120005 and 25120009.

Author contributions

T.Ku., M.O., T.Ko. and N.T. designed the study. T.W., Y.O. and N.T. collected data sets. T.Ku. and K.N. performed data analysis.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Kuwatani, T. *et al.* Machine-learning techniques for geochemical discrimination of 2011 Tohoku tsunami deposits. *Sci. Rep.* **4**, 7077; DOI:10.1038/srep07077 (2014).