CHEMICAL ENGINEERING TRANSACTIONS

VOL. 91, 2022

A publication of
AIDIC
The Italian Association
of Chemical Engineering
Online at www.cetjournal.it

Guest Editors: Valerio Cozzani, Bruno Fabiano, Genserik Reniers
Copyright © 2022, AIDIC Servizi S.r.l.
ISBN 978-88-95608-89-1; **ISSN** 2283-9216

DOI: 10.3303/CET2291067

A Machine Learning Approach to Analyze Natural Hazards Accidents Scenarios

Antonio Javier Nakhal Akel*[a], Janna S. Hovstad[b], Mathilde S. Ruth[b], Stefano Parmeggiani[c], Riccardo Patriarca[a], Nicola Paltrinieri[b]

[a] Department of Mechanical and Aerospace Engineering, Sapienza University, Rome, Italy
[b] Department of Mechanical and Industrial Engineering, Norwegian University of Science and Technology, Trondheim, Norway
[c] ISPIRA ETS, Rimini, Italy
antonio.nakhal@uniroma1.it

Climate change has contributed to an increasing frequency and severity of natural hazards accidents over recent years, and the increasing trend is expected to continue and escalate. Globally, demographics are changing and urbanization, population growth and increasing coastal populations make societies more exposed and vulnerable to extreme weather events. As a consequence, attention towards natural disasters is increasing along with the interest in approaches to manage emerging risks. Some industries have been experiencing major losses to hazards, while others might be hit harder in the future.

Current research shows that there is a need to further investigate underlying reasons for variations in disaster timing, impacts, and outcomes, as well as mitigation strategies. The purpose of this research is to enhance the understanding of natural disaster mortality and unravel underlying causes and influential factors that can inform decision-making and be relevant for risk reduction efforts. This is achieved by analyzing natural hazards accidents data and using data science techniques to define data clusters and delve into the related factors affecting mortality. The climate-driven, natural disaster events from the International Disaster (EM-DAT) database have been thoroughly explored and visualized to obtain an overview of the current natural disaster situation. More specifically, this manuscript concerns the development of clustering algorithms and analytics to map fatalities and economic damage. The results of the analysis showed the extent to which climate change has a significant effect on resulting fatalities and economic losses from natural hazards accident scenarios. Besides the achieved results of this work, it is acknowledged how further studies should try to dynamically represent vulnerability as well as improve the quality and selection of integrated features to improve the representation of industrial aspects.

1. Introduction

Data analysis of natural hazards accidents can aid risk management by shedding light on disaster characteristics, challenges, differences amongst regions, and similar events. Natural hazard management denotes the systematic actions focused on reducing the negative effects of disasters (Department of Regional Development and Environment Executive Secretariat for Economic and Social Affairs, 1991). Mitigation measures contribute to natural hazard management by minimizing, monitoring, and reducing the probability of severe consequences, the corresponding avoidable impacts, and the unfortunate outcomes of natural hazards (Sarkar and Maiti, 2020). The risk for individuals inflicted by natural hazard disasters differs based on societal vulnerability and exposure, and environmental conditions (ISDR, 2009). Climate change has forced more than 20 million people to move from their homes each year (Masika, 2013). The development level of a country might affect the consequences of a natural disaster. It is often remarked how those living in poverty are hardest hit despite being the least responsible for climate change.

The increasing frequency of natural hazards led to greater attention worldwide devoted to mapping and reducing natural risks (Cruz et al., 2006), unraveling and explaining potential impacts on societies. Vulnerability in this

Paper Received: 1 March 2022; Revised: 24 March 2022; Accepted: 11 May 2022
Please cite this article as: Nakhal Akel A.J., Hovstad J., Ruth M., Parmeggiani S., Patriarca R., Paltrinieri N., 2022, A Machine Learning Approach to Analyze Natural Hazards Accidents Scenarios, Chemical Engineering Transactions, 91, 397-402 DOI:10.3303/CET2291067

context can be a risk factor, but also an outcome: disaster exposure may lead to poverty causing damage to assets and livelihoods (Suarez-Paba and Cruz, 2022). Besides, larger natural disasters often cause extensive property damages and a high number of fatalities. Research has shown that natural disaster-related damages and mortality have increased in the past decades (Jacobsson et al., 2009; Masika, 2013).

Research is needed to develop systematic approaches on disaster causes and impacts to improve responses, and anticipation capacity and design risk prevention and mitigating interventions prior to, or following major natural hazards verification. The International Disaster Database (EM-DAT) developed by the Centre for Research on the Epidemiology of Disasters (CRED) gathers data on natural disasters and maps them into different classification categories, impacts, and causes. This paper aims to study these climate-driven accidents in terms of societal impact, both on populations and properties, as they can be of relevance for industrial systems as well. The manuscript fully relies on EM-DAT and proposes a Machine Learning (ML) algorithm to investigate potential clusters of countries that show commonalities and subsequently can drive to common natural risk management mitigations. The focus of this manuscript spans from natural hazards accidents to technological accidents in order to ensure a wider perspective on all societal impacts.

2. Materials and Methods

2.1 Exploring the database

The EM-DAT database was created following the 1980's investigation by the Centre for Research on the Epidemiology of Disasters (CRED). The study was carried out to serve the purposes of humanitarian action at national and international levels. The initiative aimed to rationalize decision-making for disaster preparedness, as well as provide an objective base to assess vulnerability and set priorities. The database is compiled from various sources, including United Nations agencies, non-governmental organizations, insurance companies, research institutes, and press agencies (e.g.), United Nations Department of Humanitarian Affairs (UN-DHA), European Union Humanitarian Office (ECHO), International Federation of the Red Cross and Red Crescent, the Office of Foreign Disaster Assistance (OFDA-USAID), International Committee of the Red Cross and Red Croissant (ICRCRC, Switzerland), International Decade for Natural Disaster Reduction (IDNDR) (Center for research on the Epidemiology of Disasters, 2021). Currently, EM-DAT collects more than 25000 disasters between 1900 - 2020. All the events in the EM-DAT database fulfill one or more of these entry criteria (Center for research on the Epidemiology of Disasters, 2021):

- Kill (10 or more deaths)
- Affect (100 or more people affected, injuries or homeless)
- Declaration/Appeal (declaration by the country of a state of emergency and/or appeal for international assistance)

The 25000 incidents worldwide involve 189 countries, distributed as follows:

- About 15000 accidents are related to natural impacts (e.g., drought, extreme temperature, flood, landslide, storm, wildfire, etc.),
- About 10000 accidents refer to technological impacts (i.e., industrial, transport, and miscellaneous impacts).

The database incorporates 43 parameters (e.g., location, date, damage, fatalities, disaster type, origin, reconstruction cost, insured damage, appeal, impacts) to fully details the characteristics of the accident and allow the accident identification and analysis (Center for research on the Epidemiology of Disasters, 2021).

2.2 Data clustering through Machine Learning

Machine learning (ML) is known for providing meaning to raw data and solving practical problems in a reliable and efficient way. These problems require machine assistance since the amount of data and the complexity of the statistical patterns imply that humans would not be able to solve them via traditional techniques (Burkov, 2019). ML algorithms learn from examples and are thereby trained to find patterns that can help make decisions and predictions based on new, unseen information (Sharda et al., 2019). A ML pipeline includes training, test, and validation processes. One example of ML refers to clustering. This latter is used to uncover meaningful groups within a dataset based on underlying patterns or structures. Clustering is commonly used for dimensionality reduction and the most common methods are density-based, hierarchical, partition-based, and grid-based methods. This descriptive data mining technique is unsupervised since there are no target values to predict (Murtagh and Contreras, 2012). The clustering algorithm relies on a distance matrix that is created by computing the distance between every pair of data points. For this reason, a clustering algorithm requires standardized, numerical input.

K-Means Clustering

K-Means is one of the most frequently used and effective clustering algorithms, as proved by results obtained in several diverse application contexts (Zhang et al., 2017). K-Means is a general-purpose clustering method preferred for data where a flat geometry. The algorithm tries to group data by minimizing the within-cluster-sum-of-squares which represents the distance between each data point and the cluster centroid (Chen et al., 2005). The most common metric to compute distances in K-Means is the Euclidean distance, as it is flexible to accommodate different operational situations. Another characteristic of the algorithm is that it requires an explicit specification of the resulting number of clusters. The algorithm will always converge, but it is vulnerable to local minima. This will depend on how centroids are initialized. By running the algorithm with a specified number of clusters k, k random samples from the dataset are allocated as cluster centroids. The main steps of the K-Means clustering algorithm are:

- Initialization: the step to choose k initial centroids
- Looping: the iterative step to stabilize centroids, until a certain threshold is reached, or a certain number of iterations has been run. This loop requires two sub-steps:
    - Assigning samples to their nearest centroid based on a selected distance measure.
    - Compute the mean of the assigned samples and create a new centroid.

K-Means with Euclidean distance has been used to map countries' clusters as they appear in the EM-DAT database.

3. Results

The clustering algorithm allowed splitting the 189 countries involved in natural hazard accidents into 40 clusters of varied sizes. The algorithm runs on a set of selected features presented considered relevant for the scope of the analysis: Disaster ID; Country; Location; Year; Disaster group; Disaster subgroup; Disaster type; Event name; Total death; Total damages. The chosen algorithm relies on a distance matrix created computing the distance between every pair of data points. The algorithm has been performed to group data minimizing the within-cluster sum of squares, which represents the distance between each data point and the cluster centroid. Clusters must be validated to check the logical cohesion between the clustered items and to compare the separation among them. A useful metric for validating the significance of clusters is the silhouette, whose scores represent the distance from one sample to the samples in the neighboring clusters (Kingrani et al., 2017). Silhouette coefficients range between -1 and 1 where values close to 1 indicate high compactness within the cluster, which in turn implies longer distances among the sample and the neighboring clusters. Silhouette scores close to 0 indicate overlapping clusters, while negative values indicate a possible misplacement of the sample (Milligan and Cooper, 1985). When examining the obtained results, for demonstration purposes, this manuscript details only the two clusters presenting the higher cumulative number of fatalities. On this basis, cluster 12 (Poland, Germany, Japan, Vietnam, Bangladesh, and South Korea) and cluster 32 (Pakistan, Afghanistan, Iran, Nepal, Sri Lanka, Turkey, Romania, Algeria, and Yemen) being identified have been selected for further explorative statistics. Their average silhouette score was respectively 0,21 and 0,31. Only one element in cluster 12 showed a negative silhouette score (i.e., Poland, -0,04), and it has been manually removed from the following analysis. Details on individual silhouette scores can be retrieved in Table 1.

Table 1. Items in the two clusters were selected for demonstrative purposes, ordered by silhouette score.

Country	Silhouette Score	Cluster	Inclusion
Japan	0,3720	Cluster 12	Included
France	0,2832	Cluster 12	Included
South Korea	0,2367	Cluster 12	Included
Bangladesh	0,2164	Cluster 12	Included
Germany	0,1039	Cluster 12	Included
Vietnam	0,0588	Cluster 12	Included
Poland	- 0,0404	Cluster 12	Excluded
Afghanistan	0,4990	Cluster 32	Included
Nepal	0,4953	Cluster 32	Included
Turkey	0,4266	Cluster 32	Included
Iran	0,4196	Cluster 32	Included
Pakistan	0,4061	Cluster 32	Included
Sri Lanka	0,1781	Cluster 32	Included
Alegria	0,1735	Cluster 32	Included
Romania	0,1510	Cluster 32	Included
Yemen	0,0346	Cluster 32	Included

Table 2 proposes a country classification by a number of deaths and economic damage for the countries being previously selected. It is possible to observe that Bangladesh, Japan, and France account for 98,32% of total deaths count in their cluster, with Bangladesh presenting 89,80%. On the other hand, Japan, Germany, and France represent the 90,93% (75,57%; 8,85% and; 6,50% respectively) of the economic damage in their cluster. Similarly, for cluster 32 the countries: Pakistan, Iran, and Turkey account for 79,06% of total deaths, with however a flatter distribution than the one in cluster 12, i.e. Pakistan 32,14%; Iran 29,04%; and Turkey 17,87%. The same three countries, Pakistan (24,16%), Iran (24,07%), Turkey (23,22%), plus Algeria (10,10%) account for 81,56% of the economic damage in their cluster. Overall, cluster 12 involves 2.769.968 reported deaths, 82,64% more than cluster 32. Likewise, cluster 12 has 82,33% economic damage losses reported more than cluster 32.

Table 2. Country classification by death and economic damage (ordered by number of deaths).

Country	Total deaths	Total economic damage [$]
Bangladesh	3.010.075	21.893.565
Japan	250.305	534.091.500
Pakistan	186.943	30.157.109
Iran	168.942	30.049.696
Turkey	103.996	28.986.670
Sri Lanka	41.046	4.475.364
France	35.177	45.956.100
Vietnam	29.545	23.404.066
Nepal	27.224	6.836.415
Afghanistan	26.890	603.320
Germany	13.775	62.575.505
Algeria	13.382	12.614.846
South Korea	12.734	18.748.034
Yemen	7.539	4.894.400
Romania	5.681	6.199.920

Table 3 proposes a country classification in terms of deaths and economic damage by impact types and technological impacts, this latter divided into sub-types (industrial, transport, miscellaneous). Besides, the analysis has been separated into two parts. In terms of Natural impacts, Bangladesh has the higher number of fatalities in the reports with 2.993.988 deaths, followed by Japan with 239.374 deaths. Moreover, Japan has the higher economic damage 533.908.500 $, followed by Germany with 61.978.605 $. Similarly, about Technological impacts: for accidents related to industries and transports, Bangladesh has the higher number of reported fatalities (1.809, and 13.261 deaths respectively), followed by Germany with 1.650 deaths (industrial impacts) and Pakistan 6.184 deaths (transport impacts). Likewise, Algeria has the higher economic damage reported in the industrial impacts (800.000 $). Nevertheless, South Korea presents 38.400 $ related to transport issues.

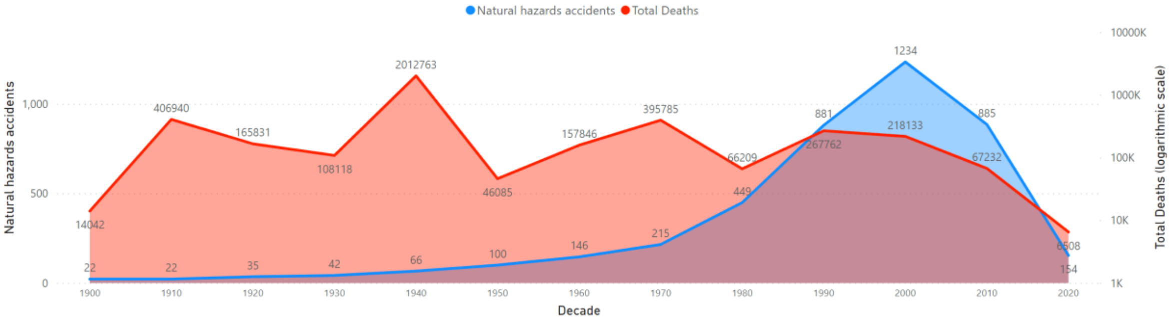

Figure 1. Trend over time of the number of natural hazards accidents (blue line) and deaths, the sum of the values for the two clusters being analyzed (cluster 12, cluster 32).

Errore. L'origine riferimento non è stata trovata. represents a combined line chart describing the count of natural hazard accidents reported and the death losses (logarithmic scale to facilitate comparison) over time. The figure shows a peak in the decades of 1940s, mainly due to a Drought disaster that occurred in Bangladesh accounting for 1.900.000 deaths approximately. Besides, it is possible to observe how the behavior over the decades in terms of count natural hazard accidents is increasing, and the death losses in the last two decades are decreasing.

Table 3. Country classification for detailed impact, with a focus on technological aspects (industrial, transport, and miscellaneous). Background highlights maximum value per category.

Country	Natural Impact		Technological Impact					
			Industrial		Transport		Miscellaneous	
	Deaths	E. Dmg. [$]	Deaths	E. Dmg. [$]	Deaths	E. Dmg. [$]	Deaths	E. Dmg. [$]
Afghanistan	25.197	603.320	280	*No data*	1.184	*No data*	229	*No data*
Algeria	11.874	11.814.846	27	800.000	1.280	*No data*	201	*No data*
Bangladesh	2.993.988	21.893.565	1.809	*No data*	13.261	*No data*	1.017	*No data*
France	28.864	45.892.700	1.223	36.800	3.571	*No data*	1.519	26.600
Germany	10.419	61.978.605	1.650	226.300	1.618	*No data*	88	370.600
Iran	163.347	29.899.696	196	*No data*	4.682	*No data*	717	150.000
Japan	239.374	533.908.500	41	160.500	4.150	16.500	5.840	6.000
Nepal	24.871	6.835.155	No data	*No data*	2.159	*No data*	194	1.260
Pakistan	178.840	29.955.969	861	179.080	6.184	*No data*	1.058	22.060
Romania	4.939	6.199.920	60	*No data*	587	*No data*	95	*No data*
S. Korea	9.111	18.516.257	303	167.300	1.658	38.400	1.662	26.077
Sri Lanka	40.057	4.475.364	25	*No data*	871	*No data*	93	*No data*
Turkey	97.086	28.708.670	1.239	*No data*	3.084	*No data*	2.587	278.000
Vietnam	27.240	23.399.566	762	2.000	1.354	*No data*	189	2.500
Yemen	5.623	4.894.400	129	*No data*	1.647	*No data*	140	*No data*

Figure 2 represents a combined line chart describing the count of natural hazard accidents and the economic damage (logarithmic scale to facilitate comparison) over time. It is possible to notice the economic losses reported for the clusters being analyzed show an exponential trend over time, differently from deaths losses. Moreover, it is possible to observe how the society and countries has been enhance the safety constrains in the industrial processes affected by the natural hazards accidents to reduce the occupational losses, instead, economic damage related by them has increased over the year. These results could be a signal of the increased quality of reporting over years as well as potential improvements in safety management and risk assessment of natural hazards. More specifically, a larger number of events with smaller losses in EM-DAT may be used for analytics and support strategic decision-making also in relation and comparison with world risk indexes (Aleksandrova et al., 2021)

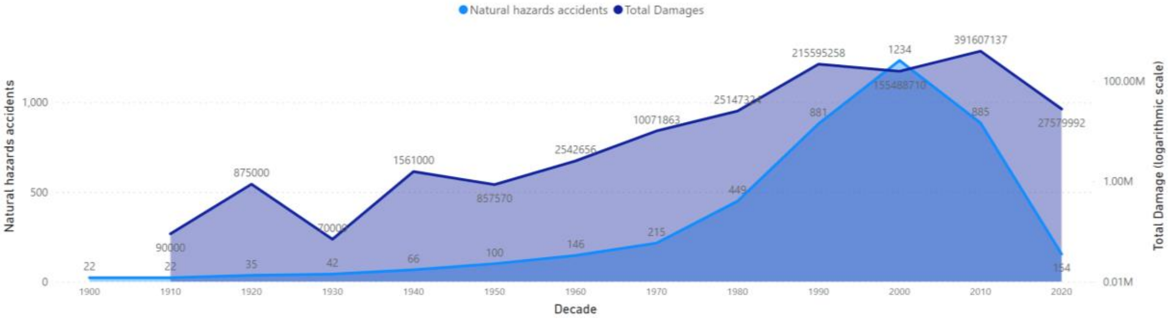

Figure 2. Trend over time of the number of natural hazards accidents (blue line) and economic damage (dark blue line) for the two clusters being analysed (cluster 12, cluster 32).

4. Conclusions

The overall goal of the analysis presented in this paper is to demonstrate the possible usage of data about natural disasters and their implications for societal safety and industrial management. The work presents methodological results obtained from clustering algorithms and analytics referred to fatalities and economic damage. The purpose of the clustering was to find a relevant group of countries that could facilitate inter-country learning opportunities and create actionable insights. The defined similarity is based on a set of general exposure and sensitivity features. Furthermore, the purpose of the proposed analytics may be helpful for policymaker to facilitate comparative analysis of fatality patterns and external factors that affect mortality subsequent to natural and technological events. Overall, the extension of these results to the entire EM-DAT database proves how natural hazards generate more impactful consequences than technological disasters. While from a societal point of view, this result shall motivate the need to invest in both protective and preventive mitigating measures, the extent of natural hazards should also push for specific interest on industrial safety, especially to prevent disastrous cascading consequences.

These early results require further refinement and improvements to further shape future risk learning processes. In this regard, they also constitute the basis for potential additional analyses, (e.g.) using generative model, anomaly detection, as for promising research in this area (Nakhal A. et al., 2021; Patriarca et al., 2022). The analytics may also be used in larger Business Intelligence (BI) solutions to support a multi-variate dynamic analysis both descriptive and predictive (Nakhal A. et al., 2021) if incorporating other ML solutions. A joint BI-ML development may indeed be a crucial instrument to support decision-makers at having a comprehensive understanding of natural hazards to shape risk prevention and mitigation programs.

References

Aleksandrova, M., Balasko, S., Kaltenborn, M., Malerba, D., Mucke, P., Neuschafter, O., Radtke, K., Prutz, R., Strupat, C., Weller, D., Wiebe, N., 2021. The World Risk Index 2021, World Risk Report 2021 F.

Burkov, A., 2019. Machine Learning Engineering. ISSN 2502-3632 ISSN 2356-0304 J. Online Int. Nas. Vol. 7 No.1, Januari – Juni 2019 Univ. 17 Agustus 1945 Jakarta 53 9 , 1689–1699.

Center for research on the Epidemiology of Disasters, C., 2021. The international Disaster Database [WWW Document]. URL https://www.emdat.be/ (accessed 12.2.21).

Chen, T.S., Tsai, T.H., Chen, Y.T., Lin, C.C., Chen, R.C., Li, S.Y., Chen, H.Y., 2005. A combined K-means and hierarchical clustering method for improving the clustering efficiency of microarray. Proc. 2005 Int. Symp. Intell. Signal Process. Commun. Syst. ISPACS 2005 2005, 405–408. doi:10.1109/ispacs.2005.1595432

Cruz, A.M., Steinberg, L.J., Vetere-Arellano, A.L., 2006. Emerging issues for natech disaster risk management in Europe. J. Risk Res. 9 5 , 483–501. doi:10.1080/13669870600717657

Department of Regional Development and Environment Executive Secretariat for Economic and Social Affairs, O., 1991. Chapter 2 - Natural Hazard Risk Reduction in roject Formaulation and Evaluation [WWW Document]. URL https://www.oas.org/dsd/publications/Unit/oea66e/ch02.htm#chapter 2 natural hazard risk reduction in project formulation and evaluation (accessed 12.13.21).

Jacobsson, A., Sales, J., Mushtaq, F., 2009. A sequential method to identify underlying causes from industrial accidents reported to the MARS database. J. Loss Prev. Process Ind. 22 2 , 197–203. doi:10.1016/j.jlp.2008.12.009

Kingrani, S.K., Levene, M., Zhang, D., 2017. Estimating the number of clusters using diversity. Artif. Intell. Res. 7 1 , 15. doi:10.5430/air.v7n1p15

Masika, R., 2013. Gender, Development and Climate Change. Oxfam GB, Oxford.

Milligan, G.W., Cooper, M.C., 1985. An examination of procedures for determining the number of clusters in a data set. Psychometrika 50 2 , 159–179. doi:10.1007/BF02294245

Murtagh, F., Contreras, P., 2012. Algorithms for hierarchical clustering: an overview. WIREs Data Min. Knowl. Discov. 2 1 , 86–97. doi:10.1002/widm.53

Nakhal A., A.J., Patriarca, R., Di Gravio, G., Antonioni, G., Paltrinieri, N., 2021. Business intelligence for the analysis of industrial accidents based on MHIDAS database. Chem. Eng. Trans. 86, 229–234. doi:10.3303/CET2186039

Patriarca, R., Di Gravio, G., Cioponea, R., Licu, A., 2022. Democratizing business intelligence and machine learning for air traffic management safety. Saf. Sci. 146 August 2021 , 105530. doi:10.1016/j.ssci.2021.105530

Sarkar, S., Maiti, J., 2020. Machine learning in occupational accident analysis: A review using science mapping approach with citation network analysis. Saf. Sci. 131, 104900. doi:10.1016/j.ssci.2020.104900

Sharda, R., Delen, D., Efraim, T., 2019. Analytics, Data Science, & Artificial Intelligence: Systems for Decision Support, Eleventh e. ed. Pearson, Hoboken, NJ.

Suarez-Paba, M.C., Cruz, A.M., 2022. A paradigm shift in Natech risk management: Development of a rating system framework for evaluating the performance of industry. J. Loss Prev. Process Ind. 74. doi:10.1016/J.JLP.2021.104615

Zhang, Y., Mańdziuk, J., Quek, C.H., Goh, B.W., 2017. Curvature-based method for determining the number of clusters. Inf. Sci. (Ny). 415–416, 414–428. doi:10.1016/j.ins.2017.05.024