An AI unsupervised clustering of airports - a tool to find suitable humanitarian cooperation for disaster preparedness

MARIA BROWARSKA, Delft University of Technology, Netherlands
KARLA SALDAÑA OCHOA, University of Florida, School of Architecture, College of Design, Construction and Planning, USA

Fig. 1. Satelite images of airports in the database.

In recent years, natural disasters have increased in frequency, causing significant damage to communities and infrastructure worldwide. When a natural disaster strikes, airports in the affected region have to adapt quickly from serving regular passengers to becoming a humanitarian hub handling a massive increase in passengers and cargo. Several countries are particularly vulnerable and prone to such a devastating event. Although existing initiatives aim to raise awareness and improve airport preparedness, authorities are often isolated in their resilience efforts as they tend to act individually, and their response is often bound by local experience. Consequently, this research aims to broaden the field of view from a local to a global one by compiling a database of 971 airports worldwide with corresponding socio-technical characteristics in various data modalities. In addition, through a data science approach, a transformation of the different data modalities was performed to extract numerical feature vectors so that in future studies a correlation between airports can be found, to find similar airports from which different approaches to disaster preparedness and response can be learned.

Additional Key Words and Phrases: airports database, disaster preparedness, AI-based clustering

ACM Reference Format:
Maria Browarska and Karla Saldaña Ochoa. 2021. An AI unsupervised clustering of airports - a tool to find suitable humanitarian cooperation for disaster preparedness. In . ACM, New York, NY, USA, 18 pages. https://doi.org/XXXXXXX.XXXXXXX

Manuscript submitted to ACM

1 INTRODUCTION

When a natural disaster strikes, the nearest airport becomes the critical link for delivering and organizing relief aid while trying to stay efficient in evacuating citizens and receiving emergency personnel [5]. However, the existing infrastructure often cannot handle the sudden spike in the volume of incoming goods [6]. When airports become nonoperational, the only way to receive valuable aid is via road, rail, and water, which is often much less efficient and time-consuming [18].

Even though disasters and humanitarian aid are not the newest challenges, there is still much room for improvement. Airports are set in an environment of technical and operational challenges, laws and regulations, international and regional cooperation of stakeholders from various fields improving humanitarian logistics. To characterize an airport, we need to consider various features that describe their complexity, a) geospatial and airport-specific data: area surrounding, reachability, number of runways, taxiways; b) demographic data: urban indexes, and population around the airport; and c) geographic and urban data: seaport data and built environment information. All of the aforementioned characteristics influence airports' preparedness for a potential disaster and collecting them can help experts better address the problem in a broader scope.

Thus, this research explores how data science could help establish a base for forming collaborations between airports that might face similar challenges in disaster preparedness efforts. The goal is to build a comprehensive database describing airports from the perspective of their disaster preparedness that will help future researchers find similarities between them, based on their intrinsic socio-technical features, so that perhaps an airport in Indonesia could be matched with its sibling airport in the Caribbeans. The research involved several programming operations––starting with collecting data, through data processing, up to experimenting with Self Organising Maps (SOM) algorithm in order to find airports that share features that are relevant in terms of their disaster preparedness. The database can be found in the following repository.

https://gitlab.com/maria.browarska/OSM-SOM

The proposed database of airports and their numerical features are the first step to a process that will conclude creating group-specific policy advice for similar airports. With this article, we want to describe the steps from collection, normalization, and pre-processing of the data to transforming the multimodality of the gathered data to a numerical feature vector that can be used for the grouping of similar airports through Unsupervised Machine Learning algorithms that can cluster similar airports based on similar numerical features. Having a relevant scenario to apply ML that benefits society at large.

2 KNOWLEDGE GAP AND RESEARCH GOAL

In order to define key concepts, narrow down the scope of the research and precisely define the knowledge gap, a literature review was conducted, followed by 5 semi-structured interviews with industry experts.

2.1 Literature review

Most of the reviewed articles focused on a case study as the research approach, often looking at individual airports and assessing historical events. Researchers analysed the behaviour of airports in specific disastrous events, mainly focusing on organisational processes and stakeholders' cooperation [17, 18, 25]. While all the considered features, without a doubt, influence logistical operations, they are also unique for each airport. Hence, it is challenging to draw general

Table 1. Socio-technical features

Structural and capacity features	Accessibility features	Organisational features	Risk related features
Runways and their characteristics	Airport connection	How much staff is available	Risk of occurrence of a natural disaster
Aircraft parking and its characteristic	Geographical surroundings	How well the staff is trained	Regional capacity for handling disasters
Terminals and their characteristics	Alternative airports and seaports	Who owns the airport	What is the airport's main Purpose (civil / military)
Storage facilities both open-air and covered warehouses			Whether the airport was part of any preparedness programs

conclusions that could apply to other airports since their organisational structure may differ, due to international and regional regulations, resources and needs.

Some of the authors pointed out the importance of the geographical location of an airport, structural features as well as reachability [4, 24, 26]. Pandey et al. [15] proved that utilising geo-spacial data is beneficial for airport humanitarian response planning and that airport authorities are interested in tools that can help to plan logistical procedures.

While some of the authors suggested that cooperation between airports that struggle with similar challenges would have a positive outcome [10, 17], none of them explored the possible backbone of such cooperation. That finding, combined with the idea of structural features of airports having an impact on their humanitarian logistical procedures, led to defining the knowledge gap.

The specific methods applied in this research were used in the field of humanitarian aid-related research before, but on a local or national scale, as shown by Saldaña Ochoa & Comes, and Chen [3, 13]. The global approach is a challenge due to the limited availability of reliable data, but if successful, it paves the way for more detailed research on a global scale, which could benefit the less developed countries, that often do not have resources for local advanced research and preparedness strategies.

Until now, the practitioners in the field, such as Get Airports Ready for Disaster (GARD), have used straightforward methods for assessing the vulnerability of airports and had to prepare different strategies for each client. GARD's capacity is minimal, and this research could lead to new ways for authorities to prepare, thanks to establishing collaborations directly with other airports facing similar challenges.

2.2 Research goal

The goal of this research is to (1) better understand the challenges that airports face when a natural disaster strikes and their preparedness activities. This understanding shall then be (2) translated into a list of socio-technical features influencing the level of preparedness and airport capabilities in facing a disaster. The finding of key features is relevant for (3) building a database containing valuable humanitarian aid-related information about several airports worldwide, composed solely from publicly available sources. The focus on publicly available data is conditioned by a large number of airports being analyzed, which makes it impossible to conduct surveys and obtain information directly within the resources and time frame of this research.

3 METHODOLOGY

In order to find specific qualities and features that influence airports' preparedness for a disaster, a thorough understanding of activities and the environment in which they take place is needed. This information was derived from a desk study accompanied by semi-structured interviews (table 3 in the Appendix lists organizations contacted for interviewing) with experts on airports' disaster preparedness and performance, summarized in table 1. The next step was to translate identified challenges influencing the performance of an airport in a post-disaster scenario into socio-technical features to achieve a good starting point for the data mining process.

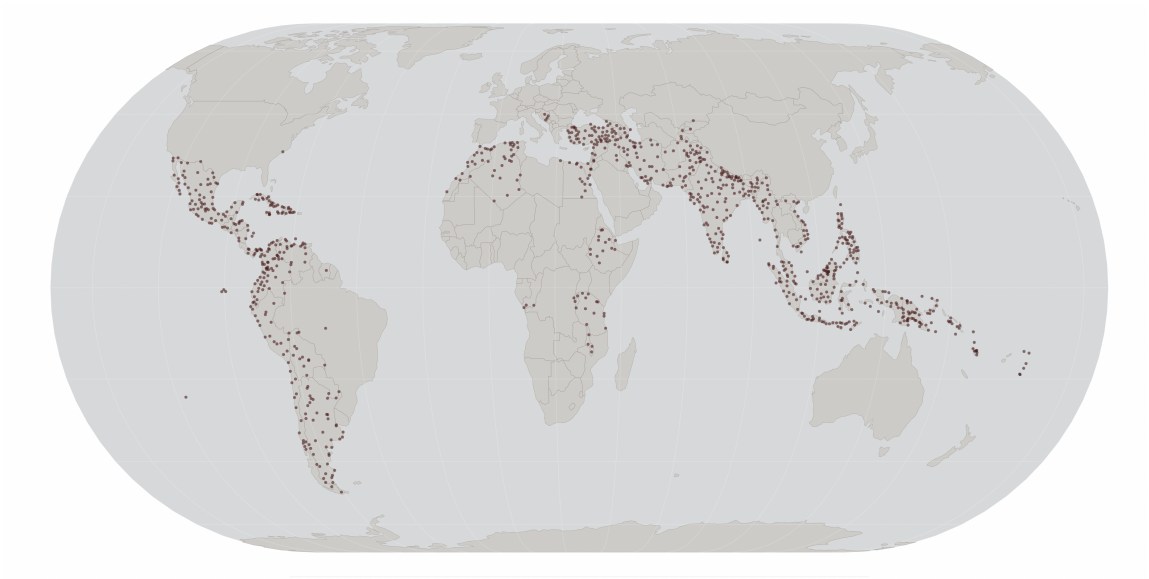

Fig. 2. 971 airports chosen to be analyzed, placed on a world map

The data mining process was composed of two main iterative phases. First, the identified socio-technical features of airports had to be translated into measurable data points –– numerical, categorical, or descriptive. The second phase was retrieving data from publicly available sources, as described in more detail in diagram 6. When building a database from publicly available sources, it is crucial to have a strong understanding of what we want to describe to allow for flexibility and easy replacement or adjustment of originally planned measures.

To start building the database, we choose vulnerable countries and airports using the INFORM Risk Index as qualification criteria for choosing. First, a list of all airports that are located within these countries was exported. Next, the airports.csv file from OurAirports was used to select only airports currently operating, i.e., have scheduled services. An additional criterion was the airport type - heliports, seaplane bases, and closed ones were excluded, while small, medium, and large were chosen (the size of an airport was defined based on the number of scheduled flights as described by OurAirports' data). These operations resulted in forming a list of 971 airports, with their names, coordinates, International Air Transport Association (IATA) codes, and International Civil Aviation Organization (ICAO) codes. This list would form the base for all mass queries applied via APIs to collect data for each airport. Figure 2 presents the 971 airports on a World map.

4 BUILDING THE DATABASE

Data used in this research came from a multiplicity of sources in various data modalities and formats. In order to translate socio-technical into comparable sets of numerical features, a number of conditions need to be taken into account, such as availability of data, methods of measuring and quantifying specific characteristics, their correlations, and level of importance. In order to keep track of changes and make the database easy to navigate, the SQLite database was built with the use of DB Browser software. The OSM queries, the GeoDB - cities API were connected to the database through Python queries, as seen in the attached GitLab repository. To add records and features to the database, outputs

Table 2. Description of the database.

Source	Feature	Type of data	Data handling	Relevance
OurAirports	iata	text	no additional handling needed	airport identification and location
	airport_name	text		
	latitude_deg	numerical		
	longitude_deg	numerical		
	country	text		
	elevation_ft	numerical	empty fields inputed with mean value	
	lighted	categorical	empty fields imputed with '0'	runway description for assessing airport's capacity and accessibility
	max_length_ft	numerical	empty fields inputed with mean value	
	width_ft	numerical	empty fields inputed with mean value	
	airport_type	categorical	text values converted into categorical values '0', '1'	general assessment of the airport traffic size
OSM	seaport_count	numerical	manual verification	identifying potential alternative seaports within 100 km radius
	airport_count	numerical		identifying potential alternative airports within 100 km radius
	build_count	numerical		describing the surrounding within 5 km
	industrial_count	numerical		assessing airport's cargo handling preparedness
	tourism_count	numerical		
	terminal_count	numerical		assessing airport's capacity
	runways_count	numerical		assessing airport's capacity
GeoDB	name_city_n	text	obtaining data about three closest cities	assessing the distance between the airport and potential casualties
	dist_city_n	numerical		
	population_city_n	numerical		assessing the number of potential casualties in the area
Global Airports	aptclass	categorical	text values converted into categorical values '0', '1'	assessing airport's capacity international / domestic
	apttype	categorical		assessing airport's capacity Airport / Airstrip / Airfield
	authority	categorical		assessing airport's organisational structure: civil / military
	humuse	categorical		assessing airport's humanitarian operation preparedness
INFORM Index	natural_dis_risk	numerical	empty fields inputed with mean value	assessing regional disaster risk
	informrisk	numerical		assessing regional disaster preparedness
Logistics Performance Index	lpi_customs	numerical		assessing regional logistical capacity and preparedness
	lpi_infrastructure	numerical		assessing regional logistical capacity and preparedness
GARD	gard	categorical	text values converted into categorical values '0', '1'	assessing airport's humanitarian operation preparedness
Self calculated	airport_area	numerical	calculated based on OSM data	assessing airport's capacity
	population_around	numerical	calculated based on GeoDB data	assessing the number of potential casualties in the area
	iso_country	text	no additional handling needed	identification purposes

from various sources were converted into the .csv format. Results of OpenStreetMap (OSM) and API queries were automatically written into the database directly. A detailed description of each data source and steps taken in the process of extracting data can be found in B.1 and B.2.

4.1 The database

As the plan is to compare airports based on numerical features, each data modality was turned into an *understandable* form for mathematical processing. Depending on the modality of data, various preprocessing methods were applied, based on several scientific sources [7, 9, 20, 21] and can be seen in Appendix C. The final list of all airports and corresponding features were built in the DB Browser and made available through the GitLab depository, both as a .csv file and an SQLite database. Features selected for each airport, together with the corresponding source, preprocessing methods, and a description of their relevance for assessing disaster preparedness, are presented in table 2.

5 UNSUPERVISED MACHINE LEARNING

This section describes the process of applying an unsupervised machine learning algorithm - Self Organising Maps - (SOM) on the data set built in previous steps. Apendix B shows an initial trial with other clustering algorithms and it explains the reason why we selected SOM to proceed with the experiment.

5.1 Self Organising Maps

In order to cluster airports based on their distinctive features relevant for disaster preparedness, an unsupervised machine learning algorithm was applied with the use of SOMPY Python library ([22]). The whole process was thoroughly documented in the attached GitLab repository.

5.1.1 Training. The data set consisting of 971 records with airports and their features was split into two smaller sets - the training set with randomly chosen 70% of all records, and the testing set with the remaining 30% - resulting in the training set with 650 data points and the test set with data points.

After the pre-processing was finished, all records from the training set were transformed into input vectors that can be processed by the SOM. For the first attempt each vector was a series of 36 numerical values, describing all the chosen features for each airport. Within the SOMPY API, each vector was normalised before the training of the SOM.

The training phase was repeated 100 times for various, randomly chosen sizes of the final SOM map, in order to find the best performing one, based on the calculated topographic and quantisation error of each training run. The smaller these values, the better the performance of the feature map ([1]). Once the best performing map was chosen, a visualisation of each feature on a map was performed, as shown in figure 3.

5.1.2 Analysing results. Initially, the clustering was achieved only on a small number of airports. By comparing the result of SOM to the individual representation in figure 3, we discovered that there were multiple dominating features that led the process of clustering. The dominant features were the categorical ones, a problem known as the *the curse of dimensionality* ([23]). Simply put, there were too many 0/1 dominating features that influenced the whole clustering process.

5.1.3 Adjusting input vectors. Given the result of the first attempt at applying the SOM algorithm on the whole data set, a number of attempts at adjusting the input vectors were performed.

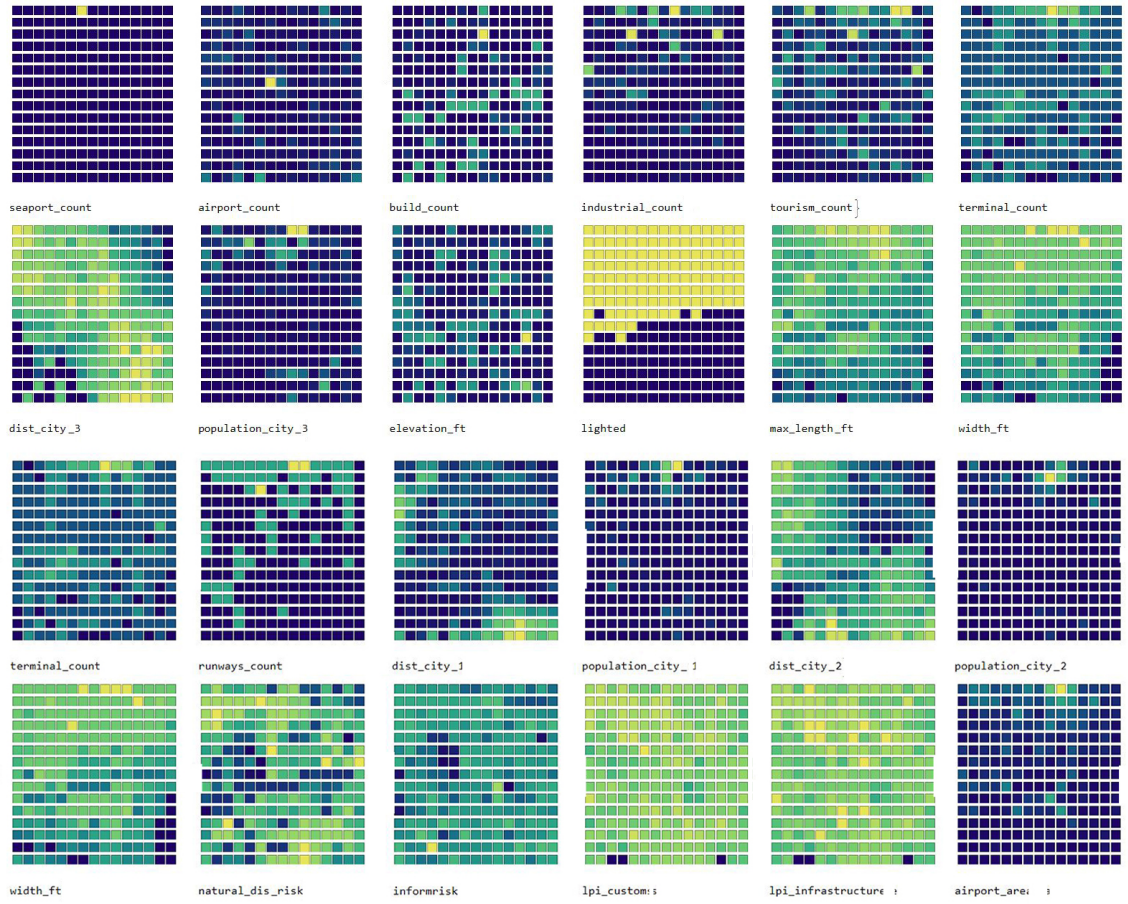

Fig. 3. Visualisations of each feature in the first attempt at SOM. Each airport is placed in one of the cells on the map - the brighter the cell colour, the higher the value of the feature (eg. more terminals) or the value is equal to 1 for categorical features (eg. lighted airstrip - yes). From this figure, we can derive that there is a number of features that may become dominating, due to their distribution - concentration of the bright cells in a small area. It may lead the clustering algorithm to focus on these strong features, which does not necessarily reflect their importance in real life. We can also observe some correlations - large airports have more terminals and runways, which tend to be longer and wider than at medium and small airports. While it is a very straightforward conclusion, it can serve as a verification tool.

First, the most dominating features were excluded from the data set, based on the individual representation of each feature in figure 3. The categorical feature of airport type was changed from the binary representation to a translation of *small, medium, large* into numerical values: 1, 2, 3. While it should not be performed for features describing non-continuous categories, the airport type does in fact sort the airports from the ones with smallest traffic to the largest, therefore it is acceptable to translate it into continuous values.

5.1.4 Adjusted input vectors - results. Again, the remaining features went through all the steps of pre-processing, transformed into input vectors and normalised. The result of running the SOM algorithm on input vectors reduced to 20 features is represented in figure 4.

When analysing the SOM cell by cell, we can observe that cell 15, which consists of 14 records represents airports that have no seaports in their vicinity, have 2-4 alternative airports, have 1-2 terminals, are all of the medium traffic type and have the natural disaster risk between 4.0 and 6.7. For the rest of the features, no dominant value exists, there is a broad representation of each feature.

Another cell - number 46, that consist of 11 records, includes airports with a large number of alternative airports - between 6 and 27, 0-1 terminals, small traffic and higher natural disaster risk then the previous group - between 5.8 and 7.7.

5.1.5 Adjusted input vectors - clustering. The next step would be to cluster individual cells into groups. An example result is shown in figure 5. Applying the K-Means algorithm led to defining key 4 groups of airports.

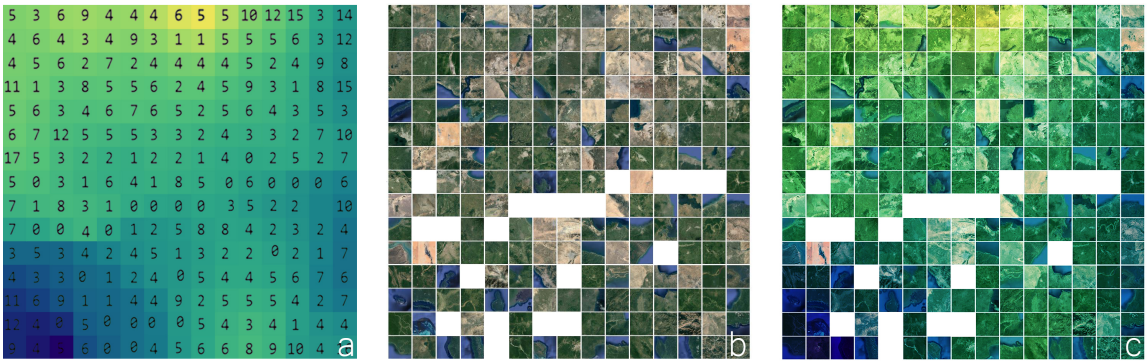

Fig. 4. Results of the adjusted SOM. a) shows a color changing spectrum that visualized the consistency in clustering, b) shows the satellite images of airports that were closer to each best matching unit, c) shows an overlapping of the color spectrum and the satellite images.)

5.1.6 Verification. In order to verify the the way the SOM operates, a vector with verification data and added to the input data. This vector consisted of feature values nearly identical to one of the records in cell 46. The algorithm was run again and the result was positive - the verification vector was added to the cell with other similar ones, proving that the SOM operates correctly.

5.2 Using the SOM map in practice

Regardless of the current performance of the clustering algorithm, or rather, the level of preparedness of the input data - since those two are strongly dependent - we can discuss how the proposed approach could be used in practice.

The attached repository allows for investigating the output map in details. With a result like the one shown in figure **??**, it can be derived which airports were put together in a cell, meaning - which ones were chosen as the similar ones. This is the starting point for determining on what areas these airports could cooperate with one another. Sometimes, the similarity will result from a specific dominating features, with others fairly different, therefore it is important analyse the result before stating which airports are similar, only by looking at their cell membership. On this cell-level analysis we can also find very small groups of 2-3 airports that are grouped together, which could constitute an opportunity for a stronger cooperation. The higher level of similarity can be derived from additional clustering of cells, as presented in figure 5. Here bigger groups of airports are formed - while still different from one another, there will be a bigger

diversity within members of each group. This can form a base for another type of cooperation, with more members who might not be identical, but still have some strong similarities. Here again it is important to analyse what are the common features that mainly influenced the grouping.

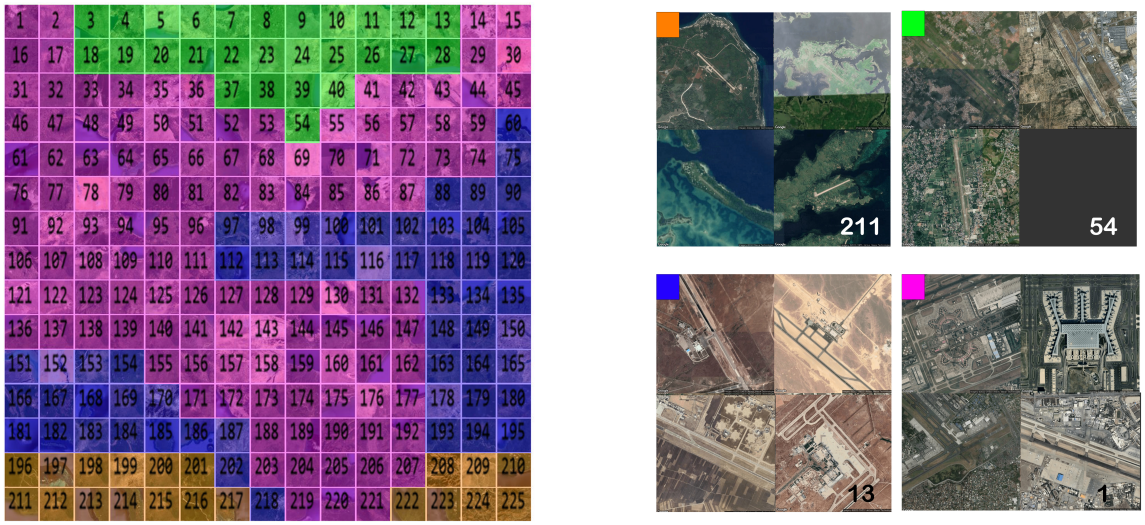

Fig. 5. Result of clustering a SOM. Here, on top of the original SOM clustering and additional K-means clustering is performed. Cells from 5.1.4 were put into larger groups in order to find main 4 types of airports. On the right we can observe how similar the satellite images of airports grouped in specific cells are, as well as how distinctively different each cluster is.

An example described in section 5.1.4, with airports grouped in one cell showing strong similarity in the low number of alternative seaports and airports, and medium traffic type, could be used to form a cooperation focusing on ways of preparing an airport with these specific conditions. Even though the airports themselves can be in distant parts of the world, their preparedness strategies can be similar, given their dominant features. Of course, these are only a couple of areas in which these airports can be seen as similar, and it is important to note the possible organisational and cultural differences. While the *airport authority* feature aims at describing the possible organisational scheme, there still might be more factors at place.

To sum up, the SOM map can be used as a tool to quickly group and find the dominating features of a big group of airports. The better the data describing the grouped institutions, the more accurate the result will be. It is easy to visualise and interpret, and it can be used by humanitarian aid and aviation experts without advanced programming or mathematical skills. A task of grouping such a big number of records while taking into account more than 20 factors would be impossible to perform by hand, therefore it is a great combination of using sophisticated unsupervised machine learning algorithm in an easy to interpret way.

6 LIMITATIONS

The quality data sources used in the research can sometimes be contested, as the level of detail available for various airports and their surroundings was not always equal, which may lead to inaccurate results. This is also a problem with official sources widely used by the humanitarian community, such as the Logistics Capacity Assessment. Interviewees mentioned the importance of access to dynamic data that describes the state of each airport and its surroundings at

a precise moment in time, after a disaster strikes, because the static information gathered in assessments earlier can be inaccurate the moment a disaster strikes. However, interviewees involved in preparedness programs rather than immediate response operations underlined the importance of building comprehensive data sets with static information to assess better what can be done ahead of a tragic event.

Another challenging factor is the accuracy of assumptions made––especially for assessing airport connectivity. As proved by historical disasters, the inability to distribute humanitarian relief from the airport to the population in need can undermine the airport's operations and preparedness. A more sophisticated and accurate way of quantifying the level of connectivity could be used in future research.

7 DISCUSSION AND CONCLUSION

The database built in this research is a valuable resource for future clustering analysis or future research related to airports' preparedness for humanitarian disasters. It can be further analyzed in more detailed research, updated accordingly, and used to assess airports' venerability and preparedness. From the scientific perspective, this research proves that there are now ways of analyzing complex, specific challenges with a global overview based on numerous publicly available data sets. It also shows that scientists need to be very careful when using not precisely scientific sources and that building a specific, tailored database is a lengthy, challenging process. Nevertheless, it can be achieved not only by IT professionals but also by multidisciplinary researchers.

This research provided a valuable framework for approaching complex socio-technical environments of airports and their disaster preparedness, through building a database with relevant features, based on interviews and literature review, using only publicly available data, followed by a comprehensive data selection, collection and pre-processing. The challenges and problems encountered along the way, both solved, and unsolved can form a valuable tool for other professionals and scientists willing to conduct similar research, not only related to the domain of aviation and disaster preparedness.

An additional finding is that we identified the need for a common, reliable database with all relevant information about airports in vulnerable locations. The one designed during this research could form a base for a one built with official data sources that are otherwise unavailable to the public. With that, however, comes the challenge of security; since detailed information about airports can be viewed as sensitive data, therefore access to such a database should be regulated.

7.1 Future research

The ideas for future research can be divided into three sections - (1) related to the data mining and the process of building the database, (2) data pre-processing and applying an unsupervised clustering algorithm and (3) using the results in various ways in order to improve airports' disaster preparedness.

Building a database solely from publicly available sources has some drawbacks, as discussed in section 6, such as limited trustworthiness and inability to retrieve the exact types of information that are needed in order to describe specific features. In the future, it is worth considering building a similar database with direct involvement of the airports that are being described––with the use of surveys and possible involvement of international humanitarian and aviation related organisations such as ACI or OCHA. This would allow for retrieving more specific data, up to date information. Moreover, if regularly updated and maintained, it could become a useful resource for airports that themselves would like to know more about capabilities of alternative ports in the region––not only for research purposes, but for operations

once a disaster strikes and help from neighbouring ports is needed. Other scientists could also use such a database for various additional analyses, saving time for gathering the data and focusing on what can be derived from it.

However, the database that was built in this research is itself a valuable resource for performing other research related to airports' preparedness for humanitarian disasters. With additional iterations of the data pre-processing, there is room for gathering insightful knowledge on similarities between airports, that would form a solid base for establishing cooperations. In order to achieve that, future research should focus on identifying the dominating features and adjusting the algorithm accordingly. This could require more sophisticated methods of data pre-processing and automating the process of analysing results, in order to quickly pick up combinations of features that cannot offer trustworthy results.

Building policy advice based on the database could be achieved by identifying airports that are especially vulnerable, due to to their intrinsic features and capabilities. This process would have to be accompanied by a thorough analysis of historical events that took place at similar airports, and the lessons learned could be used for improving preparedness of those that might face similar challenges in the future, leading to achieving the full potential of this research.

REFERENCES

[1] Le Anh Tu. 2020. Improving Feature Map Quality of SOM Based on Adjusting the Neighborhood Function. *Sustainability in Urban Planning and Design* (2020). https://doi.org/10.5772/intechopen.89233

[2] Jean-François Arvis, Lauri Ojala, Christina Wiederer, Ben Shepherd, Anasuya Raj, Karlygash Dairabayeva, and Tuomas Kiiski. 2018. *Connecting to Compete 2018.* Technical Report. The World Bank. https://doi.org/10.1596/29971

[3] Ning Chen, Lu Chen, Yingchao Ma, and An Chen. 2019. Regional disaster risk assessment of china based on self-organizing map: Clustering, visualization and ranking. *International Journal of Disaster Risk Reduction* 33, October 2018 (2019), 196–206. https://doi.org/10.1016/j.ijdrr.2018.10.005

[4] Sunkyung Choi and Shinya Hanaoka. 2017. Diagramming development for a base camp and staging area in a humanitarian logistics base airport. *Journal of Humanitarian Logistics and Supply Chain Management* 7 (06 2017), 00–00. https://doi.org/10.1108/JHLSCM-12-2016-0044

[5] Deutsche Post DHL Group. 2019. *GoHelp Program - Disaster Preparedness and Response.* Technical Report. https://www.dpdhl.com/en/responsibility/society-and-engagement/disaster-management.html

[6] Deutsche Post DHL Group. 2021. Disaster Preparedness - Get Airports Ready for Disaster. https://www.dpdhl.com/en/sustainability/social-impact-programs/disaster-management/disaster-preparedness.html

[7] Purva Huilgol. 2020. Feature Transformation and Scaling Techniques to Boost Your Model Performance. https://www.analyticsvidhya.com/blog/2020/07/types-of-feature-transformation-and-scaling/

[8] Humanitarian Data Exchange. 2019. Global airports - Humanitarian Data Exchange. https://data.humdata.org/dataset/global-airports

[9] Gota Kikugawa, Yuta Nishimura, Koji Shimoyama, Taku Ohara, Tomonaga Okabe, and Fumio S Ohuchi. 2019. Data analysis of multi-dimensional thermophysical properties of liquid substances based on clustering approach of machine learning. *Chemical Physics Letters* 728 (2019), 109–114. https://doi.org/10.1016/j.cplett.2019.04.075

[10] Jakub Kraus, Vladimír Plos, and Peter Vittek. 2014. The New Approach to Airport Emergency Plans. *International Journal of Aerospace and Mechanical Engineering* 8, 8 (2014), 2406 – 2409. https://publications.waset.org/vol/92

[11] MIT. 2021. Python Wrapper to access the Overpass API. https://github.com/DinoTools/python-overpy

[12] M. Mogley. 2017. GeoDB Cities API Documentation. https://rapidapi.com/wirefreethought/api/geodb-cities

[13] Karla Saldana Ochoa and Tina Comes. 2021. A Machine learning approach for rapid disaster response based on multi-modal data. The case of housing shelter needs. arXiv:2108.00887 [cs.LG]

[14] OurAirports. 2007. About OurAirports. https://ourairports.com/about.html#overview

[15] B.H. Pandey, Carlos Ventura, P. RioFrio, J. Pummell, and S. Dowling. 2014. Development of response plan of airport for mega earthquakes in Nepal. *NCEE 2014 - 10th U.S. National Conference on Earthquake Engineering: Frontiers of Earthquake Engineering* (01 2014). https://doi.org/10.4231/D3TH8BN7T

[16] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[17] Abdussamet Polater. 2018. Managing airports in non-aviation related disasters: A systematic literature review. *International Journal of Disaster Risk Reduction* 31 (2018), 367–380. https://doi.org/10.1016/j.ijdrr.2018.05.026

[18] Abdüssamet Polater. 2020. Airports' role as logistics centers in humanitarian supply chains: A surge capacity management perspective. *Journal of Air Transport Management* 83 (2020), 101765. https://doi.org/10.1016/j.jairtraman.2020.101765

[19] QGIS Development Team. 2009. *QGIS Geographic Information System.* Open Source Geospatial Foundation. http://qgis.org
[20] Jimin Qian, Nam Phuong Nguyen, Yutaka Oya, Gota Kikugawa, Tomonaga Okabe, Yue Huang, and Fumio S Ohuchi. 2019. Introducing self-organized maps (SOM) as a visualization tool for materials research and education. *Results in Materials* 4 (2019), 100020. https://doi.org/10.1016/j.rinma.2019.100020
[21] H Ritter and T Kohonen. 1989. Self-organizing semantic maps. *Biological Cybernetics* 61, 4 (1989), 241–254. https://doi.org/10.1007/BF00203171
[22] Sevamoo. 2018. sevamoo/SOMPY. https://github.com/sevamoo/SOMPY
[23] G. V. Trunk. 1979. A Problem of Dimensionality: A Simple Example. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-1, 3 (1979), 306–307. https://doi.org/10.1109/TPAMI.1979.4766926
[24] Michael Veatch and Jarrod Goentzel. 2018. Feeding the bottleneck: airport congestion during relief operations. *Journal of Humanitarian Logistics and Supply Chain Management* 8, 4 (jan 2018), 430–446. https://doi.org/10.1108/JHLSCM-01-2018-0006
[25] Bartel Walle and Julie Dugdale. 2012. Information management and humanitarian relief coordination: findings from the Haiti earthquake response. *Int. J. of Business Continuity and Risk Management* 3 (01 2012), 278 – 305. https://doi.org/10.1504/IJBCRM.2012.051866
[26] Martijn Warnier, Vincent Alkema, T. Comes, and Bartel Walle. 2020. Humanitarian access, interrupted: dynamic near real-time network analytics and mapping for reaching communities in disaster-affected countries. *OR Spectrum* 42 (09 2020). https://doi.org/10.1007/s00291-020-00582-0
[27] Sanford Weisberg. 2001. Yeo-Johnson Power Transformations. *Department of Applied Statistics, University of Minnesota* 2 (2001), 1–4. http://stat.umn.edu/arc/yjpower.pdf

A PROCESS FLOW

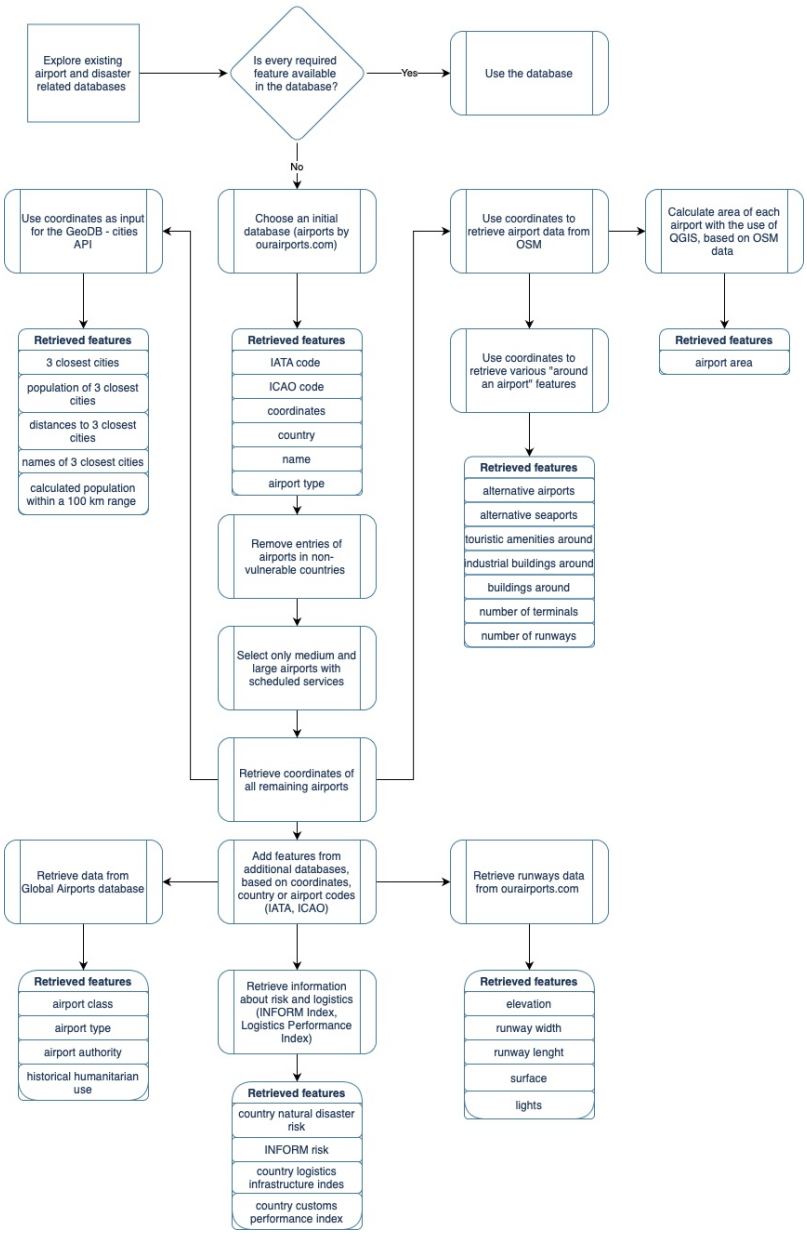

Fig. 6. Process flow of data mining.

Table 3. Affiliation of interviewees

Interviewee	Organisation
Chris Weeks	GARD
Virginie Bohl	OCHA, IMPACCT Working Group
Thomas Romig	ACI

B CLUSTERING COMPARISON

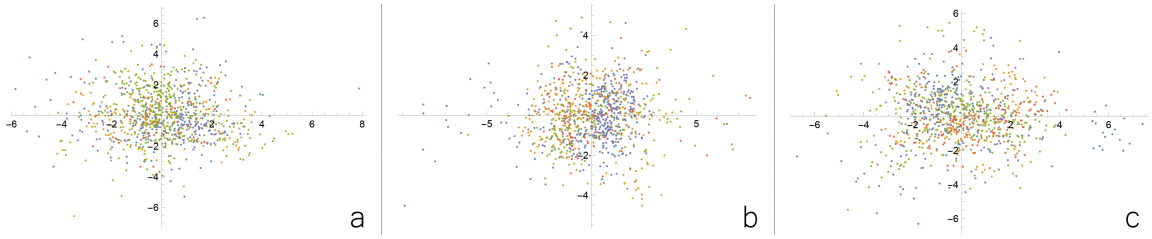

Fig. 7. Clustering comparison. a) clustering using K-Means and an algorithm for dimensionality reduction. b) clustering using dbscan and an algorithm for dimensionality reduction. c) clustering with a spectral clustering algorithm and an algorithm for dimensionality reduction. After trying this method we decided to work with Self Organizing Maps (SOM). The reason why we choose SOM is because we identify that the visualization of the SOM results in a user-friendly interaction and it has a visual output that helps understand the clustering.)

B.1 Data sources

B.1.1 ***OSM.*** In order to extract data from OSM, Overpass turbo was used - a web-based data mining tool, designed to run OSM API queries and present them on a map. Since data needed to be extracted for over 900 airports, multiple scripts were written, with the use of the OverPy API, published under the MIT license [11]. A detailed documentation of the scripts and queries can be found in the attached GitLab repository.

B.1.2 ***OurAirports.*** OurAirports is a free and public service that maintains data about airports around the world. Similarly to OSM, it is run by volunteers - members create records individually - but at the same time much of the information comes from official governmental institutions such as the U.S. Federal Aviation Administration [14]. In addition from exploring an online interactive map-based tool, users can also download daily updated files with data records of all airports that are part of the service. For this research, data set of all airports and runways was used.

B.1.3 ***Global airports.*** The most comprehensive, publicly available, data set aimed at providing information on disaster logistics is called *Global airports* and was published by the Humanitarian Data service [8]. Officially coordinated by the World Food Programme, based on openly available data from sources such as OSM and OurAirports, it also contains inputs from partners though the Logistics Cluster and Logistics Capacity Assessments [8]. Even though the data set is updated, according to a WFP representative interviewed, for many places the data has not been checked since the original upload in 2013. Furthermore, the data set contains fairly basic information on airports. Data points presented in the table are not available for every airport in the set.

B.1.4 ***The Logistics Performance Index.*** The Logistics Performance Index (LPI) provides information on how easy or difficult it is to transport goods in the analysed countries. The World Bank, together with various logistics-related partner organisations conducts the survey every two years [2]. While aimed at assessing the logistical capacity in the context of trade and merchandise, some of the indicators are relevant for humanitarian logistics, such as the ones chosen to be included in this research: the assessment of customs procedures and the assessment of general quality of trade and transport related infrastructure.

B.1.5 ***The INFORM Risk Index.*** Led by the European Commission, INFORM is a global, open-sourced risk index for humanitarian disasters and crises, that describes three dimensions: hazard exposure, vulnerability and lack of coping capacities. In addition to being the qualification criteria for the final airport database, parts of the INFORM Risk index were also used to characterize airports.

B.2 Extracting data

B.2.1 ***Airport surroundings.*** Two strategies in OSM were tested in order to asses the surroundings of each airport. First, the "landuse" tag was explored - all the nodes containing information on the land use within 5km radius from each airport were extracted. However, this led to inconsistent results - visual validation of multiple query outputs was conducted and it led to a conclusion that buildings-related nodes are highly over represented as compared to fields or other unused spaces. Therefore, for many airports, the result only showed a number of buildings within that radius, and no information describing the empty fields that were the true dominant surrounding.
The second strategy, which led to more representative results, was one based on purely the number of nodes with the tag "building". The assumption was that if the buildings are well tagged in OSM, simply the number of those nodes within the radius would describe how densely built the surrounding of the airport is. The lower the number of buildings

around - the more useful space for organising humanitarian aid. A visual validation of multiple records was conducted, with a special focus on the outliers - airports with very low or very high number of buildings around. The surroundings of some remote airports was underrepresented, resulting in 0 buildings reported. While it was not true, the number of buildings was very little and the result was still useful.

B.2.2 ***Alternative airports.*** To find and alternative airport, we focused on the surroundings within a 100km radius. Unlike with choosing airports for the main database, with alternative ones there was no exclusion of those that are smaller or do not have an IATA code. The assumption was that any kind of airport within a close vicinity to the main one might work as a supporting space, even if not for landing the same size of airplanes, but perhaps storage and other humanitarian operations. Since airports are well tagged in OSM, the validation of results was positive - there were no overlooked airports found. However, depending on the quality and density of roads, an airport within 100 km radius might in fact be many hours away, which would not be a useful alternative. In future research it is worth considering finding a more accurate qualifying feature than the radius.

B.2.3 ***Alternative seaports.*** Similarly to alternative airports, alternative seaports were inspected within a radius of 100km. Vast majority of results showed 0 seaports and that was validated thoroughly and resulted to be true. Validation was also conducted for a high number of seaports counted - for some, the counted results was higher than the actual number of ports, because of multiple tags within the same seaport. It did however indicate the size of the seaport - often the nodes were indicating more seaport terminals or storage facilities. Given the small number of records that indicated seaports at all, all results higher than 0 were validated and manually corrected if needed.

B.2.4 ***Tourism vs. industry.*** In order to asses how well an airport is equipped to handle a sudden influx of cargo handling and not only a growth in passenger turnaround, it was decided that it can be assessed by the surrounding of an airport. Based on the insights from the interview with Chris Weeks of GARD, it was determined that airports that are situated in mainly touristic destinations are less likely to have a good capacity for handling cargo. Therefore, for each airport the amount of nodes tagged as "industrial" and "tourism amenities" was calculated. In order to account for over / under representation of certain regions, a ratio of tourism and industry related facilities is calculated - based on the assumption that if the region is under / over represented in OSM, it will happen for both types of amenities.

B.2.5 ***Runways.*** The number of runways was calculated for each airport by counting the number of nodes/ways/relations with a "runway" tag. All outliers were manually validated - those that resulted in 0 runways were corrected since a functioning airport cannot have 0 runways. The same was done for all records that showed more than two runways since it is not very common for airports to have multiple runways, especially in remote places, which happens to be where most of the airports from the database are.

B.2.6 ***Cities and distances.*** In order to asses how distant an airport is from the population it might be serving when a disaster strikes, three closest cities for each record were found, together with the direct distance (not by road) and population of each city. For this purpose, the GeoDB - cities API was used [12]. Based on the coordinates of each airport the three closest cities within 100km, containing population information were chosen. Validation was performed for a number of randomly chosen records and outliers, and manually corrected if needed. The API works with GeoNames and WikiData, which similarly to OSM are considered trustworthy sources, thanks to the user community input and validation scheme.

B.2.7 ***Population.*** Data gathered to describe surrounding cities was used to calculate the general population around each airport - as a summation of population in all three closest cities found by the GeoDB cities API.

B.2.8 ***Airport area.*** In order to assess the storage capacity as well as the area available for setting up a humanitarian hub, the area of each airport was calculated. In OSM, each airport is not only indicated by a single node, but by a relation that indicates its borders. This geodata was exported and analysed with the QGIS software [19]. Thanks to built in features, the area of each airport was calculated. Validation was conducted on a random sample of results and the method proved to be effective.

C DATA PRE-PROCESSING

In order for airports to be comparable for the unsupervised machine learning algorithms, the features that are describing them need to be turned into an *understandable* form for mathematical processing.

In this section, the pre-processing of text, categorical and numerical features is described.

C.1 Empty fields

Due to the fact that various data sources were used, there was a number of empty fields for some features. Depending on the feature, these empty fields were filled either with zeroes or the mean value of all existing records. Missing fields in features describing whether the runway is lighted and whether there was a GARD training conducted before, as it was decided that if there is no information available, it is safer to assume the negative outcome. The elevation, length of the runway, width of the runway and missing INFORM and LPI risks were replaced with the mean values.

C.2 Categorical data

A number of features in the final data set describes each airport as a member of a certain category. For example, the `airport type` feature categorises airports into `small airport, medium airport, large airport`. While it is a clear and understandable distinction for a human eye, the mathematical algorithms require a numerical expression [16]. As proposed in the original publication on Self Organising Maps [21], the categorical feature with three values was transformed into three binary features, with on equal to 1, and all others to 0, for each airport. An example result can be seen in table 4. To achieve that for each categorical feature, the LabelBinarizer function from SciKit [16] was used.

C.3 Numerical data

It is common for many machine learning algorithms to require standardised data inputs, in order to perform well [16]. This also the case with unsupervised learning algorithm used in this research - the SOM. There are various mathematical transformations that can help to achieve a normally distributed data and it is important to choose one that fits the type of data the best. Again, the SciKit documentation, supported by various scientific sources [7, 9, 20] and experiments was used to choose the right approach.

The Yeo-Johnson transform [27] was used to change the distribution of numerical data, since it was one of a few transformations that can be applied on negative and zero values, which the data set contained. The effect of the transformation can be seen in figures 8 and 9. While it was not possible to successfully transform all features, especially the ones consisting of 0/1 values, for most features the improvement is visible.

Table 4. An example of encoding categorical features

	small_airport	medium_airport	large_airport	airport_type
Airport A	1	0	0	small_airport
Airport B	0	0	1	large_airport
Airport C	1	0	0	small_airport
Airport D	0	1	0	medium_airport

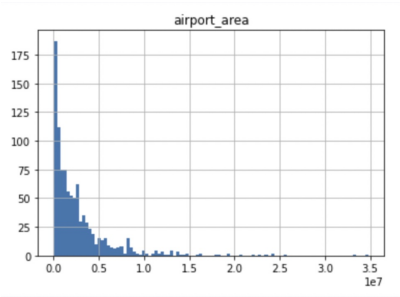

Fig. 8. An example of data distribution before the Yeo-Johnson transform. Most of the data points are concentrated around the lower values. Applying SOM directly on a non-normally distributed data could lead to specific features being over represented, therefore the transformation is needed.

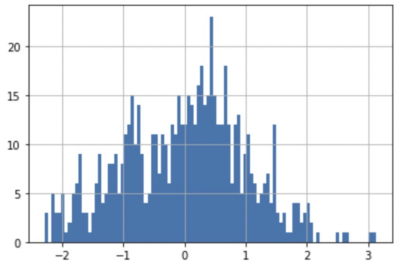

Fig. 9. An example of data distribution after the Yeo-Johnson transform. The range of values has changed, however the relations between specific values are kept and the distribution is now closer to normal.