DATA SCIENCE PILOT STUDY: IDENTIFYING RISK OF DAM FAILURE USING ARTIFICIAL INTELLIGENCE

ÉTUDE PILOTE SUR LA SCIENCE DES DONNÉES: IDENTIFIER LE RISQUE DE RUPTURE DE BARRAGE AVEC DE L'INTELLIGENCE ARTIFICIELLE

J.G. STENFERT
Advisor, HKV

THE NETHERLANDS

D. HONINGH
Advisor, HKV

THE NETHERLANDS

M. van HOEK
Senior Advisor, HKV

THE NETHERLANDS

B. KOLEN
Senior Advisor, HKV

THE NETHERLANDS

I. van de KERK
Advisor, Rijkswaterstaat, Ministry of Infrastructure and Water Management

THE NETHERLANDS

H. JANSSEN
Senior Advisor, Rijkswaterstaat, Ministry of Infrastructure and Water Management

THE NETHERLANDS

K. MIDDELJANS
Head of department, Rijkswaterstaat, Ministry of Infrastructure and Water Management

THE NETHERLANDS

C.H. OOSTINGA
Head of department, Rijkswaterstaat, Ministry of Infrastructure and Water Management

THE NETHERLANDS

E. BOERMA
Advisor, Rijkswaterstaat, Ministry of Infrastructure and Water Management

THE NETHERLANDS

1. INTRODUCTION

Predicting risk of dam failure is conventionally investigated in a physics-driven way. A pure data-driven approach can provide new insights on estimating the risk of failure, comparing dams all over the world and identifying the vulnerability of dams. More and more techniques become available to investigate large data sets with complex dependencies (Ardabili et al., 2019). Artificial Intelligence (AI) studies are also carried out within ICOLD to investigate new analysis possibilities (Lacasse, 2019). In this study, we continue to explore these possibilities of AI for dams.

This study is an experiment for using Artificial Intelligence in combination with the ICOLD World Register of Dams. This dam register contains information on 58.713 dams regarding design and construction, and is available for scientific work and statistical evaluations. Using unsupervised learning, clusters of similar dams could be formed based on a wide set of parameters. The objective of this experiment is to investigate the usage of data techniques, specifically self-organizing maps (SOM), in combination with the ICOLD database to identify similarities and differences between dams all over the world concerning failure of these dams. This analysis could lead to new insights for research into the physics of dams.

2. MATERIALS AND METHOD

2.1 DATA

The ICOLD World Register of Dams was used as a starting point for this study. This World Register of Dams contains dams from member and non-member countries, after validation by the Committee of the Register. The database is filled with fixed criteria including information about geometry, catchment area, construction, location and purpose.

In the context of the ICOLD Incident database Bulletin 99 update (ICOLD, 2019), a database with comparable characteristics has been developed in which only failed dams are present. This database has been added to the ICOLD database.

The failure of a dam depends on the dam itself, but also on the environment. Therefore, we searched for data that describe the environment of dams and hydrological parameters. We have added global geological parameters and general precipitation information.

2.2 SELF-ORGANIZING MAPS

The use of the Artificial Intelligence approach of self-organizing maps for such an application is new in relation to the analysis of dam failure statistics. Therefore, we introduce the concept of this analysis and present an example case in which self-organizing maps are used for the analysis of an international database with a wide set of parameters.

2.2.1. Concept

The method of self-organizing maps is a machine learning technique in which complex patterns and relationships are investigated, while preserving topological properties. It is an objective and unbiased method to investigate complex relations in large databases with many variables. Self-organizing maps are artificial neural networks capable of clustering multidimensional data. It groups a collection of objects into different groups based on multiple properties per object. The self-organizing map uses competitive learning, unlike many other artificial neural networks that use error-correction learning. Competitive learning is a form of unsupervised machine learning in which output nodes in the network compete, based on the given input. The node that is most activated by the given input during training is considered the winner and moves more towards the given input, while the rest of the nodes remain unchanged. This technique is well applicable to unbalanced data sets where the patterns to be investigated are not known in advance. For an in-depth explanation of the method we refer to Miljkovic (2017).

After training, the self-organizing map returns a two-dimensional representation of a multi-dimensional input space. The so called U-matrix (Unified Distance Matrix) is a commonly used representation of the self-organizing map and visualizes the distance between surrounding data points (Ultsch, 2007). The distance between neurons is the parameter of interest, giving information on the appearance of clusters. In the U-matrix, the dark colored nodes depict closely spaced nodes and blue-reddish colored nodes indicate more widely separated nodes. A cluster of dark colored nodes surrounded by reddish nodes is therefore seen as a very unique cluster. In this cluster the combination of parameters is apparently different compared to other data.

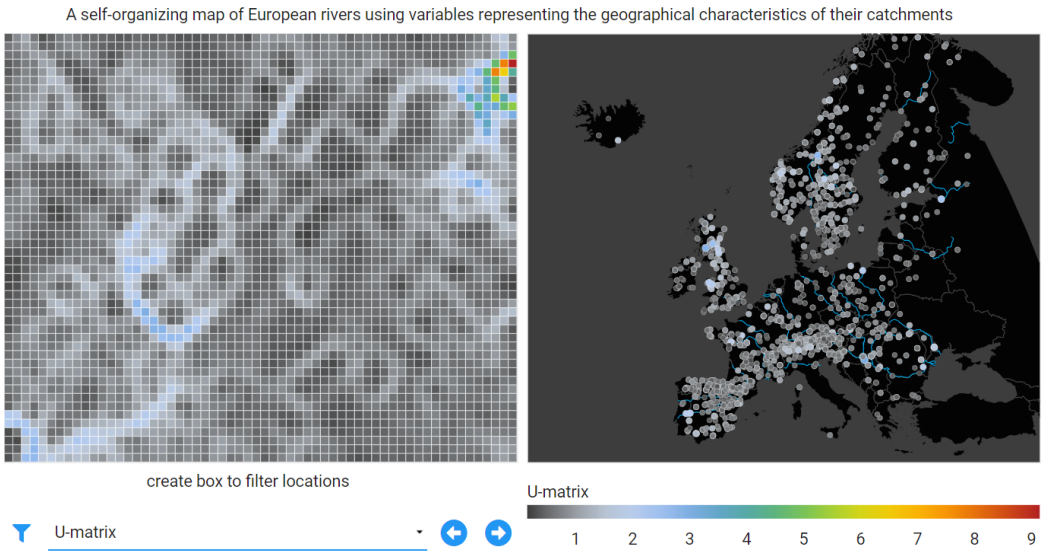

Fig. 1 Example of a U-matrix (left), a 2D-presentation of an input space in which neurons represent underlying data. Distance between neurons is visualized with color. High values (red) correspond to peaks. Large heights (values) mean that there is a large difference in the underlying data for the specific neuron. Low values for the U-matrix mean that neurons are close together.
Exemple de matrice U (à gauche), une présentation 2D d'un espace d'entrée dans lequel les neurones représentent des données sous-jacentes. La distance entre les neurones est visualisée par des couleurs. Les valeurs élevées (rouge) correspondent aux pics. Ces pics signifient qu'il existe une grande différence dans les données sous-jacentes pour le neurone spécifique. Des valeurs faibles pour la matrice U signifient que les neurones sont proches les uns des autres.

2.2.2. Example for European catchments

This example shows the use of a self-organizing map and how the U-matrix can be interpreted. It is in detail available at https://ai.hkvservices.nl/european-catchments.

In this example, characteristic data is used of all catchments in Europe, each describing an aspect of the corresponding catchment area (discharge, information on terrain, land use, etc.). Figure 2 shows the resulting U-matrix (left).

Analysis of the U-matrix patterns itself is very complicated. Linking each neuron in the U-matrix to the corresponding location, not included in training, helps to understand the spatial perspective of the U-matrix. This way of linking the U-matrix to data not already included in the training can be done for every type of

data. The right part of the figure shows geographical regions corresponding to the nodes in the U-matrix.

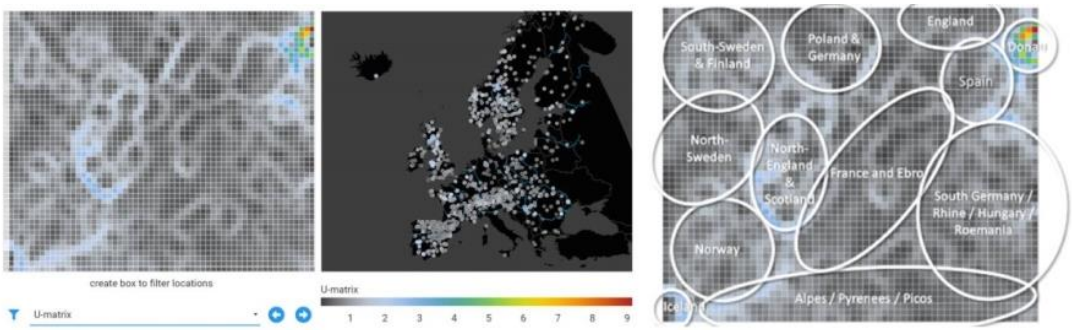

Fig. 2 Left, U-matrix derived from European catchments characteristics; middle, geographical locations of input locations linked to U-matrix node-locations; right, indicated geographical regions in the U-matrix; screenshots from Hoek (2019)
À gauche : matrice en U dérivée des caractéristiques des bassins versants européens; Au centre : emplacements géographiques d'entrée liés aux emplacements des nœuds de la matrice U; à droite : régions géographiques indiquées dans la matrice U; captures d'écran de Hoek (2019)

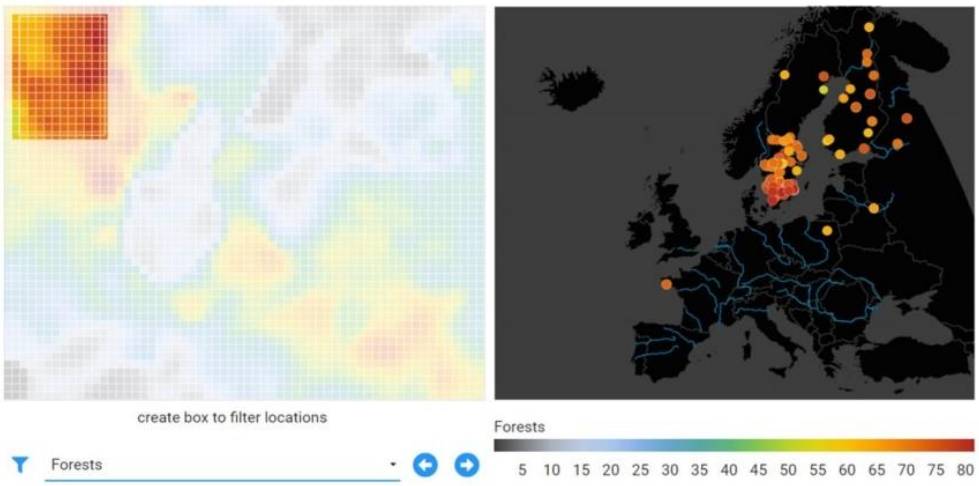

Fig. 3 Selection of 'forest'- parameter connected with (bottom-right) spatial locations (screenshots from Hoek (2019))
Sélection du paramètre «forêt» lié aux emplacements spatiaux (en bas à droite) (captures d'écran de Hoek (2019))

The U-matrix is structured using all input data. It is also possible to view the representation of each individual input variable. As an example, we show the percentage of forest in a catchment area in figure 3. In contrast to the U-matrix, the nodes are now colored by the weight vector. Dark values represent low percentages of forest and blue-reddish values correspond to high percentages of

forest. The selected box in the upper left corner clearly shows a cluster in South-Sweden and Finland with a high presence of forest. This cluster becomes apparent by using all input variables. It is obvious that these two areas have a high presence of forest and therefore resemble each other. However, other forests do not light up in this cluster, as they appear to be different based on other characteristics. For example, a forest area may be at a different level of altitude, receive more or less annual rainfall or have a different type of subsoil.

2.3 DATA PROCESSING AND APPLICATION OF SELF-ORGANIZING MAPS FOR ICOLD DATABASE

To derive a self-organizing map for the ICOLD database the following data processing steps were performed:

2.3.1. Parameter selection and dam selection

To demonstrate the validity of the approach, a selection of parameters was made in such a way that it contained parameters that could be related to dam failure. At same time it was important to include as many dams as possible in the analysis, because this type of analysis demands a fully filled database. The selected parameters relevant for failure were: area of reservoir, catchment area, dam type, length of crest, height, length of reservoir, purposes, reservoir capacity, and year of completion. The area of reservoir, catchment area and reservoir capacity are related to the loading of the dam. The dam type and the crest length is relevant for the sensitivity for failure and year of completion for the aging effect as well as the design and construction practices of that period. Other parameters, like dam height or spillway capacity, would have been interesting to include in the analysis, but these parameters were only available for a small number of dams and were therefore not included. This resulted in a reduction in the number of dams included in the analysis, from 58,000 to 430.

2.3.2. Additional data

Based on the location of these remaining 430 dams, global geological and hydrological information was gathered to enrich the dataset. Because the exact location of the dams is not known in the database, it is not possible to include specific data such as average deformation at the location of the dam in this analysis. Nevertheless, the soil type was included, based on a general geological map (USGS, 1995), and the average rainfall corresponding to the country of the dam (World Bank, 2020), although the latter is a somewhat crude approximation of the rainfall conditions at the dam site.

2.3.3. Preparing for training data

Self-organizing maps only work with quantitative data. However, the database also contains a lot of qualitative data. In preparation, the qualitative data of purposes, geology and dam type is therefore transformed into quantitative data for each variable. Table 1 shows the transformation of qualitative to quantitative data of dam types.

Table 1 Transformation of qualitative data of dam types to numbers of dam types for use of self-organizing maps. Explanations of abbreviations are available at www.icold-cigb.org

Transformation des données qualitatives des types de barrages en nombre de types de barrages pour l'utilisation de cartes auto-organisées. Les explications des abréviations sont disponibles sur www.icold-cigb.org

nr.	dam type	nr.	dam type	nr.	dam type	nr.	dam type
0	TE	5	CB	10	TE/PG	15	TE/VA
1	ER	6	TE/PG/TE	11	TE/BM	16	TE/ER
2	PG	7	PG/TE/ER	12	PG/BM	17	ER/TE
3	VA	8	PG/ER	13	CB/PG	18	XX/VA
4	PG/TE	9	BM	14	CB/TE	19	PG/TE/XX
						20	XX

Information about the specific location in coordinates or city name, and failure status is not included in the dataset for training. After training, data about failure status is added to the dataset to interpret the U-matrix. This makes it possible to investigate the amount of failed and non-failed dams within clusters. After the first training of the self-organizing map, errors were found in the data on the year of completion. Three dams with these data errors have been removed, followed by new training of the self-organizing map. Thus, future users of self-organizing maps should be aware that in practice, this often needs to be an iterative approach.

3. RESULTS

Figure 4 shows the trained self-organizing map, based on the combined dataset, which contains 330 dams that did not fail and 97 failed dams. This can be interpreted in the same way as described in the example case. The complete resulting self-organizing map can be accessed on: http://ai.hkvservices.nl/experiment-dam-risk

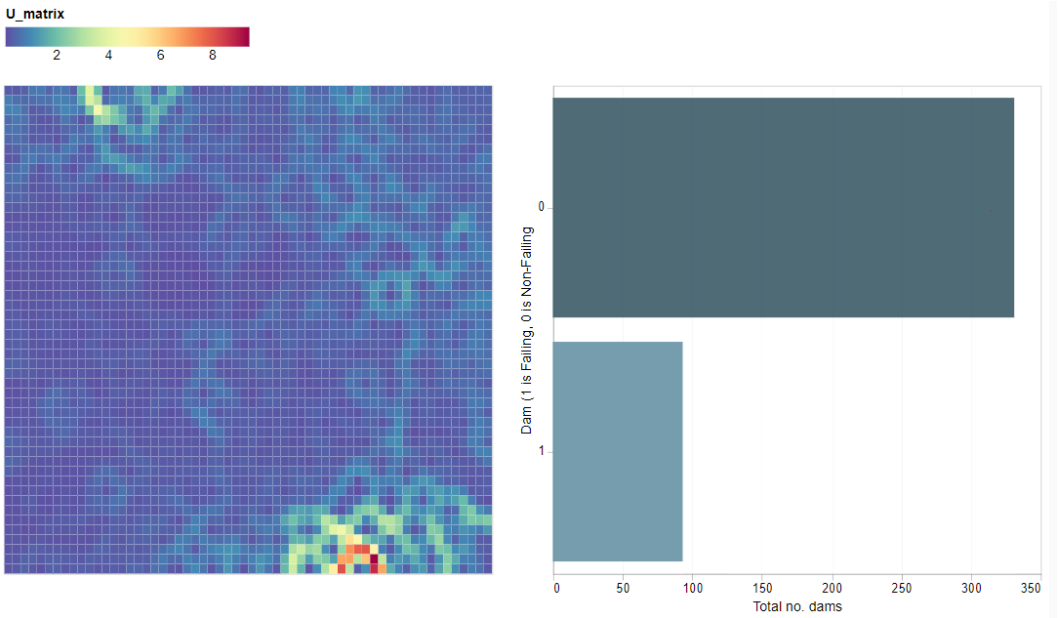

Fig. 4 U-matrix of the ICOLD SOM, with numbers of selected dams in the right figure. In light blue the total number of dams selected and in light red the number of failed dams (dark blue and dark red are the total amount of dams).
Matrice U de la carte auto-organisée des données du CIGB, avec le nombre de barrages sélectionnés dans la figure à droite. En bleu clair, le nombre total de barrages sélectionnés et en rouge clair le nombre de barrages défectueux (bleu foncé et rouge foncé sont la quantité totale de barrages).

The U-matrix is shown in conjunction with a bar chart (right). The bar chart represents the amount of failed and non-failed dams within a selection of the U-matrix, to help investigate clusters. In this figure, the entire U-matrix is selected and shown.

Two cluster regions were found, in the upper left corner and in the lower right corner. Based on all input characteristics, these locations are apparently different compared to their surroundings, because these nodes are enclosed and separated by high (red) values. The center of the U-matrix contains mainly low values, which means that there is relatively little distinction between the different dams based on the input characteristics.

3.1 INTERPRETATION OF CLUSTERS

Figure 5 shows the selection of the clusters at the upper left and the lower right corner of the U-matrix. Based on these selections, it can be seen that these clusters are not formed as a result of failure of these dams, because all dams in the cluster are represented in the bar chart of non-failed dams

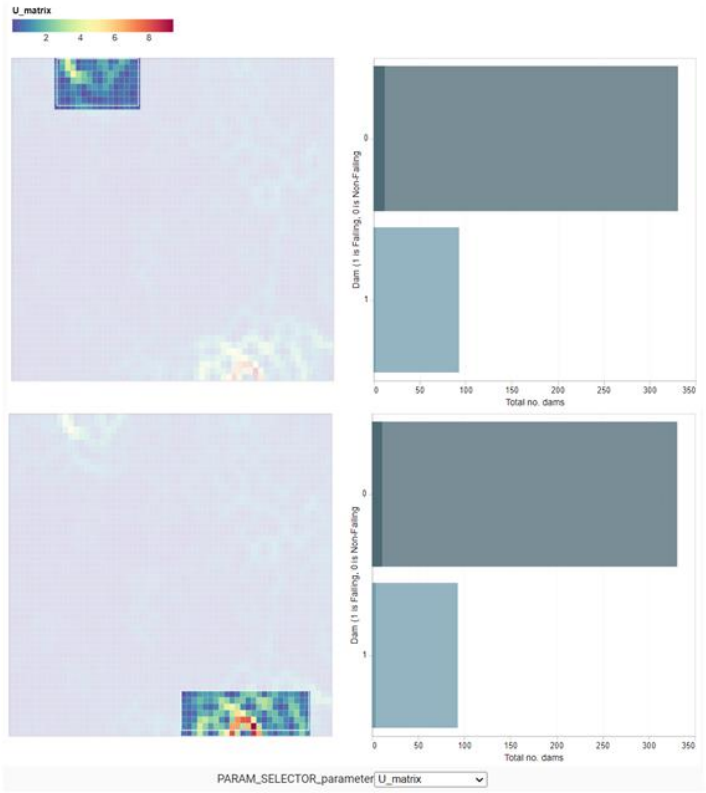

Fig. 5 Selection of formed clusters in U-matrix. Selection is shown left. The amount of failed or non-failed dams within the selection is given in the right part of the figure.

Sélection des clusters formés dans la matrice U. La sélection est affichée à gauche. Le nombre de barrages défaillants ou non défaillants au sein de la sélection est indiqué dans la partie droite de la figure.

With the interactive visualization it is possible to interpret each individual parameter used in this experiment. When a selection is made in the U-matrix it is possible to navigate through parameters to investigate the contribution of

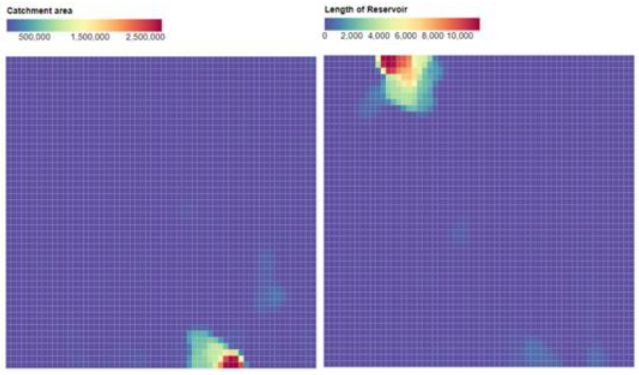

Fig. 6 Individual parameter visualization of catchment area (left) and length of reservoir (right)

Visualisation individuelle des paramètres du bassin versant (à gauche) et de la longueur du réservoir (à droite)

each parameter to the result in the U-matrix. Based on this interactive visualization per parameter, it appears that the cluster in the center below is the result of an extremely large catchment area compared to other dams and the upper left cluster is the result of a large reservoir length. This is shown in Figure 6 in which the data on catchment area is presented at the left and data on the length of reservoir is presented at the right.

3.2 INTERPRETATION OF U-MATRIX

Figure 7 shows a global interpretation of the location of failures in the U-matrix. Approximately 70% of the failed dams are present in the right part of the figure, and the other 30% failed dams are mainly present at the left part of the figure. Although no clear clusters have formed, the failed dams are distributed in some way across the U-matrix, as shown in the figure.

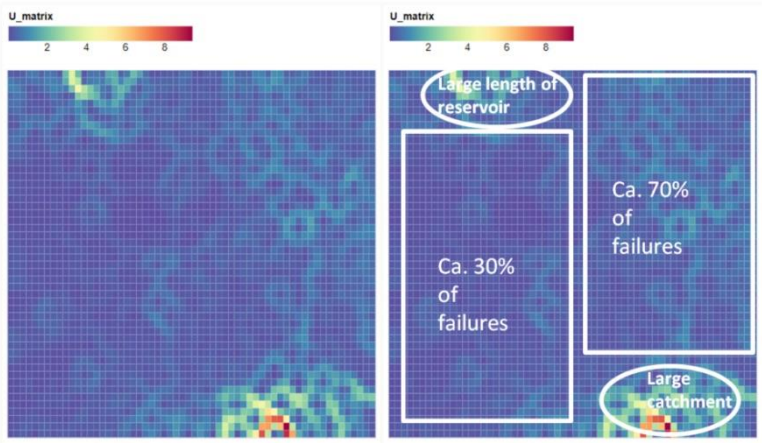

Fig. 7 Regions of U-matrix with failed dams
Régions de matrice en U avec des barrages défaillants

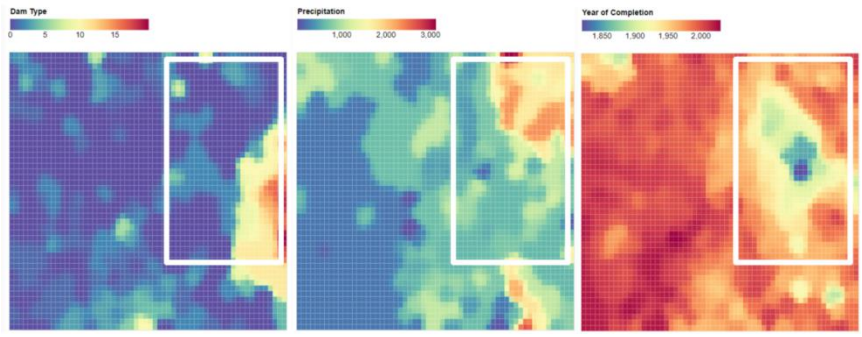

Fig. 8 Individual parameter visualization of dam type (left), precipitation (middle) and year of completion (right).
Visualisation des paramètres individuels du type de barrage (à gauche), des précipitations (au centre) et de l'année d'achèvement (à droite).

Figure 8 shows three individual parameter visualizations of the dam type, amount of annual precipitation and the year of completion. The location of approximately 70% of the failed dams in Figure 7 (in the U-matrix), are in some way related to a certain dam type (see Table 1), relatively high precipitation and older dams (‘low’ year of completion).

4. CONCLUSIONS

This article elaborates on the results of a data-driven, big data experiment with use of a combined dataset of failed and non-failed dams from the ICOLD World Register of Dams and the Dam Failure Database of the ICOLD Dam Safety Committee. The objective of this experiment was to investigate the possibilities for using data science techniques such as self-organizing maps in combination with the ICOLD database to create insights in similarities and differences between dams all over the world concerning failure of these dams for new research on the physics of dams.

Two clusters have been found as a result of data outliers. Dams within these cluster appear to be extremely large compared to other dams. No specific clusters have been found as a result of failed dams. However, the dam type, precipitation and year of completion do show relevance for a large amount of failed dams.

The experiment did not point at a clear and dominant cause for the failure of dams. However, since no specific clusters have been identified in relation to the failing of dams, it can be concluded that in this experiment, the population of failed dams does not seem to be correlated to a single parameter, nor a combination of these.

A limitation of this study was the limited amount of parameters and dams included in the analysis. A trade-off was made between the amount of parameters and the amount of remaining dams. Although unavoidable, the selection of parameters led to a significant decrease in the number of dams included in the analysis. Of the approximately 58,000 dams, 427 dams were suitable for a data-driven analysis. The main reason for this relatively small number of dams included in the analysis is that not all relevant characteristics were available for the dams. It was attempted to include relevant external data , such as soil characteristics and loads. This type of data could only be included on a general level, because a more accurate location of the dams was not available. In many cases it was not possible to determine the exact location of the dam on the basis of the name of the dam and the nearest place. Therefore the addition of this general information on soil characteristics and annual precipitation did not lead to additional insight.

Apart from challenges related to the amount and quality of the data, there is another limitation to this study. Smaller dams are often built and managed

differently, compared to larger dams. In the ICOLD World Register of Dams, there is relatively little information on small dams, which is unfortunate for this type of analysis since small dams are more likely to fail (ICOLD, 2018). Therefore, the question is whether the present type of study will give a full picture of dam failure statistics when only the dams in the ICOLD Dam Register are included in such a study.

5. RECOMMENDATIONS

This study was a first experiment for using self-organizing maps in combination with the World Register of Dams to create insight in the failure of dams, using a data-driven approach instead of a more common physics-driven type of analysis. Of course, this analysis can be improved and expanded on the basis of current knowledge and data. To get more insight, a more complete dataset is needed for a follow-up experiment in order to check whether the present data analysis method will be more successful then.

In addition, it can be very valuable to enrich the dataset with other data sources. In order to do this, the exact location of a dam is important. At the moment, only the dam name and nearest city or village is available. This can provide an indication of the location, but it is not possible to systematically determine the exact location of the dam. When the exact location becomes known, it might become possible to use remote sensing data.

These kinds of studies will not directly lead to statistically sound results. However, the use of self-organizing maps can potentially provide an objective and unbiased view of the dataset, which in turn can lead to new questions about the physics of dam failure. In this way, data-driven analysis of dams could be useful in combination with expert knowledge on dam failure.

REFERENCES

[1] BILJKOVIC, D., Brief Review of Self Organizing Maps, 2017

[2] LACASSE, S., Reliability and risk approach for the design and safety evaluation of dams. *ICOLD-CIGB 2019 Symposium. Ottawa, Canada,* 2019

[3] ULTSCH, A., Emergence in Self Organizing Feature Maps, *Proceedings of the 6th International Workshop on Self-Organizing Maps (WSOM), 2007*

[4] HOEK, M., Understanding Classification of European Catchments using Artificial Intelligence. In *Geophysical Research Abstracts* (Vol. 21), 2019

[5] FAIZOLLAHZADEH ARDABILI, S.; MOSAVI, A.; DEHGHANI, M.; R. VÁRKONYI-KÓCZY, A., Deep Learning and Machine Learning in Hydrological Processes, *Climate Change and Earth Systems: A Systematic Review,* 2019

[6] ICOLD, available online: https://www.icold-cigb.org/GB/icold/organization_mission.asp , 2020

[7] ICOLD, Incident database Bulletin 99 update, Statistical analysis of dam failures. Committee on Dam Safety, International Commission on Large Dams (ICOLD), 2019

[8] WORLD BANK, Data Bank The World Bank. Average precipitation in depth (mm per year) from 1962 till 2014, https://data.worldbank.org/indicator/AG.LND.PRCP.MM , 2020

[9] USGS, General geologic map of the world. https://mrdata.usgs.gov/geology/world/ , 2015

[10] ICOLD, General Report 103 – Small dams and levees, *26th ICOLD Congress, Vienna*, 2018

SUMMARY:

Predicting risk of dam failure is conventionally investigated in a physics-driven way. However, a pure data-driven approach can provide new insights for new questions about estimating risk of failure, comparing dams all over the world and identifying vulnerabilities of dams. This study uses the ICOLD World Register of Dams in combination with the ICOLD Incident database Bulletin 99 update to find and clarify differences between failed and non-failed dams by using Self-Organizing Maps. It is a pilot study in which this relatively new technique is used to do a purely data-driven, big data research. Clusters were found as a result of extremely large dams. No clear patterns were found which could have provided new insights into the failure of dams. A limitation of this study was the selection of parameters. This methodology needs a fully filled database. A trade-off was made between the amount of parameters and the amount of remaining dams in the database. Although unavoidable, the choice of parameters led to significant shrinkage of the usable database. This methodology should be used with a more complete database to gain new insights. Additionally external data sources should be added to enrich the analysis. To achieve this, more information on the exact location of

dams in the database would be required. These kinds of studies will not directly lead to statistically sound results. However, the use of self-organizing maps can provide an objective and unbiased view of the dataset, which in turn can lead to new questions about the physics of dam failure. In this way, this kind of method could be useful in combination with expert knowledge on dam failure.

RESUMÉ:

La prédiction du risque de rupture de barrage est classiquement étudiée d'une manière basée sur la physique. Cependant, une approche purement basée sur les données peut fournir de nouvelles perspectives pour de nouvelles questions sur, l'estimation du risque de défaillance, la comparaison des barrages partout dans le monde et l'identification des vulnérabilités des barrages. Cette étude utilise le Registre mondial des barrages d'ICOLD en combinaison avec la mise à jour Bulletin 99 de la base de données des incidents d'ICOLD pour trouver et clarifier les différences entre les barrages défaillants et non défaillants en utilisant des cartes auto-organisées. Il s'agit d'une étude pilote dans laquelle cette technique relativement nouvelle est utilisée pour effectuer une recherche de méga-données purement axée sur les données. Des grappes ont été trouvées à la suite de barrages extrêmement grands. Aucun modèle clair n'a été trouvé, ce qui a conduit à de nouvelles connaissances sur la défaillance des barrages. Une limite de cette étude était la sélection des paramètres. Cette méthodologie nécessite une base de données entièrement remplie. Un compromis a été fait entre la quantité de paramètres et la quantité de barrages restants dans la base de données. Bien qu'inévitable, le choix des paramètres a conduit à une réduction significative de la base de données utilisable. Cette méthodologie doit être utilisée avec une base de données plus complète pour obtenir de nouvelles informations. De plus, des sources de données externes doivent être ajoutées pour enrichir l'analyse. Pour y parvenir, on a besoin d’information plus spécifique sur la localisation des barrages qui sont inclus dans la base de données. Le présent type d'études ne mènera pas directement à des résultats statistiquement solides. Cependant, l'utilisation des ‘self-organizing maps’ (des cartes auto-organisées) peut fournir une vue objective et impartiale de l'ensemble de données, qui à son tour peut conduire à de nouvelles questions sur la physique de la rupture de barrage. De cette manière, ce type de méthode pourrait être utile en combinaison avec des connaissances d'experts sur la rupture des barrages.

Keywords: Dam Failure, Risk Analysis, Safety, Safety of Dams

Mots-clés: Rupture de barrage, Analyse des risques, Sécurité, Sécurité des barrages