WHERE DO HUMANS BUILD LEVEES? A CASE STUDY ON THE CONTIGUOUS UNITED STATES

M. Ikegawa[1], T. Hascoet[1], V. Pellet[2], M. Watanabe[3], X. Zhou[3], Y. Tanaka[3], T. Takiguchi[1], D. Yamazaki[3]

[1]Graduate School of System Informatics, Kobe University, 1-1 Rokkodaicho, Nada-ku, Kobe, Japan
[2]LERMA, Observatoire de Paris, Paris, France
[3]Institute of Industrial Science, The University of Tokyo, 4-6-1, Komaba, Meguro-ku, Tokyo, Japan

ABSTRACT

Understanding where and why human build levees offers several values: From a hydrological perspective, integration of levees to global flood models has been shown to improve their accuracy. From an Economic intelligence perspective, levee locations provide precious insights into past and future urban developments. However, very little data exists on the location of levees at a global scale, which hinders our ability to reach a global understanding of this question. One rare exception is the National Levee Database (NLD) dataset provided by the U.S. Army Corps of Engineers (USACE). In this study, we hypothesize that levees are built at locations where human activity and flood risk coexist, and develop predictive models that output the probability of levee existence at the hydrological catchment level. Quantitative analysis of these models using the NLD dataset allows to validate our hypothesis, with several important nuances, which we discuss at length.

Index Terms— levee detection, hydrology

1. INTRODUCTION

Current hydrological models simulate surface water storage and floodplain of rivers [1] based on the natural land topology as provided by Digital Elevation Models (DEMs) [2]. However, DEMs usually do not include man-made river flow control equipments such as levees and dams. Very recently, efforts have been put in accounting for such facilities into the global model [3]. Integration of these facilities is expected to improve global hydrological modeling, with important consequences for flood risk assessment and climate studies [4]. In addition, understanding the link between the hydrological model's predictions and levee location will open doors in forecasting future needed facilities building for better mitigation of climate change. Unfortunately, very little data is available on the location of levees. In this paper, we thus propose a preliminary study on using machine learning to infer levee locations. While our ultimate goal is to detect levees globally, this preliminary study focuses on the contiguous U.S. due to the limited data availability on levee sites outside the U.S.

Levees are built to protect human activities from flood damages. So we formulate the hypothesis that levee existence can be predicted as place where human activity and flood risk overlap. We collect global data from both remote sensing and hydrological model simulations to quantify human activity and flood risk, and use this data as input to a predictive model that quantifies the probability of levee existence spatially. The National Levee Database (NLD) dataset provides us with levee locations on the contiguous U.S. We use a subset of this dataset to train our predictive models and evaluate the accuracy of this predictor on a held out test set. We would consider a high accuracy of this predictive model as a validation of our hypothesis. In a series of experiment, we show the following:

Hypothesis validation. Explicitly modeling our hypothesis through feature engineering improves model accuracy, which allow us to validate our hypothesis, to some degree.

Generalization consideration. However, studying generalization across the U.S. reveals that accuracy significantly drops due to regional biases. This suggests that generalizing our approach globally will not be trivial, and may require additional collection of levee information globally.

Hydrological consideration. Careful analysis of the model error shows that our analysis would benefit from more structured hydrological modeling.

2. DATA AND PROBLEM DEFINITION

2.1. Output Data

As ground-truth levee data, we use the data provided by Tanaka et al [3]. In this work, levee locations provided by the National Levee Database (NLD) [5] were first projected onto 10 km resolution scale hydrological catchments derived from the river topography data MERIT Hydro [6], and then projected back to a regular grid. The levee representation we use is thus a binary variable defined on a regular grid across the U.S, in which one pixel represents one hydrological catchment, as illustrated in Figure 1.

2.2. Input Data

To validate our hypothesis, we assemble a dataset aimed to quantify the spatial distribution of flood risk and human activity, as shown in Figure 1. This summarizes all the inputs before the pre-processing used in the predictive model along with the output.

Flood Risk representation: We use the SimFlood [7] and Global Surface Water Occurrence (GSWO) [8] datasets to quantify flood risk. SimFlood provides an estimation of flood risks based on hydrological model simulations. It represents the average depth (in meters) of simulated floods characterized by a return period of 20, 50, and 100 years. GSWO is a dataset summarizing observed floods from remote sensing. In this paper, we use data on the frequency of flooding over the past 30 years. Both datasets are defined at the 90 m resolution.

Human Activity representation: To quantify human activity we used the following variables: GDP ([9]), land use (GFSAD [10] and LCCS (Landcover) [11]), and population (LandScan) [12]. GFSAD is a project providing high-resolution global agricultural land data and its water use to contribute to global food security in the 21st century. This dataset consists of three classes: cropland, watered area, and non-cropland, with a resolution of 90 m. Landcover is a map representing the land cover of the entire North American continent based on satellite imagery. It is classified into 22 classes such as Tree, Grassland, Snow, etc. using the land cover classification system and has a resolution of about 300 m. The GDP map is estimated by Taguchi et al. [9] and has a 1km resolution, the same as LandScan.

The final input features are obtained by applying some pre-processing to these datasets. First, these datasets can be divided into two categories: numerical values and categorical values. For example, GFSAD represents the categorical variables cropland, watered area, and non-cropland. We encode such categorical data into one-hot encoding representations. We aggregate the 22 classes of the Landcover dataset into three categories: cropland, urban, and others. Moreover, GSWO, which is expressed in the range of 0-100%, is difficult to handle as it is, so it is treated as categorical data in five classes: 0, 0-10, 10-90, 90-100, and 100%. Second, since the resolution varies from dataset to dataset, and the resolution of the model output and the corresponding levee dataset is the coarsest, all input features need to be aligned with the resolution of the levee data. Therefore, we average all features over multiple pixels and downsample them to 10 km per catchment.

2.3. Problem Definition

Given the data described in the previous section, we can formulate the problem of levee detection as a per-pixel binary classification problem at the 10km resolution in which for each pixel the input $X \in R^{14}$ aggregates the data presented in Section 2.2, and the label $Y \in \{0, 1\}$ represents wether or not this pixel (catchment) is protected by levees. Our predictor F thus follows the following functional definition.

$$F(X) = P(levee|X), \quad F : R^{14} \longmapsto [0, 1]$$
$$\text{where } P(levee) \text{ is the probability of levee existence} \tag{1}$$

We split our dataset into a training (78% of the dataset catchments) and test (22%) subset. The predictor model is first trained to predict the levee variable on the training subset, and we evaluate its generalization performance in on the held-out test subset. We consider different train/test split strategies as discussed in Section 3.3.

3. ANALYSIS AND RESULTS

This chapter first describes the metrics that support our analysis and then provides a summary with the following structure based on three axis: In a first step, we start by assessing the validity of our hypothesis. To do so, we evaluate both different machine learning models and different feature engineering strategies on the task of levee detection. Our experiments quantitatively suggest that our hypothesis clearly works. In a second step, we investigate the impact of regional biases on our model. We do so by injecting regional biases to the dataset through different training and test split strategies. The visualization and quantitative metrics described in 3.1 show that the regional influence on the model's generalization is significant. In a third step, we analyze the performance of our model at regional scale around the Mississippi delta. Our analysis reveals patterns of errors that suggest that additional hydrological modeling might be needed to accurately predict the existence of levees. Together, these experiments suggest that a different probabilistic modeling approach is needed to achieve automatic detection of levees at a global scale. Our results also provide prior probabilistic models and generalization performance analysis that can further improve those probabilistic models' performance.

3.1. Evaluation Metric

Our dataset presents several difficulties for machine learning approaches: First it is heavily unbalanced: Of the dataset only 0.604% catchments are levees (positive), 99.396% are negative. Furthermore, it is a relatively small dataset with only 1057 positives. Having a large enough test population while maintaining a high enough number of training sample is difficult. Lastly, the dataset contains some false negatives: i.e. catchments that are in reality protected by levees but for which the levee information has not been shared.

Because of these difficulties we have found traditional evaluation metrics to give unstable results. However, we found that evaluating the calibration of the model provided us with a principled methodology to handle the above problems [13, 14], and resulted in stable evaluation. We thus present

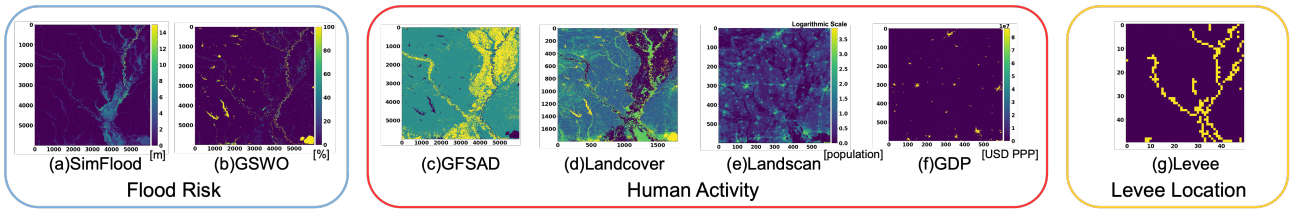

Fig. 1. Illustration of the data used in our experiments. Human activity and Flood risk variables are used as input to the model, and levee locations are used as ground-truth.

our results in terms of Expected Calibration Error (ECE), and visualize calibration curves to illustrate our conclusions. For lack of space, we do not present the details nor the motivation behind this metric here, but redirect interested readers to [13, 14].

3.2. Models and Feature Engineering

We use Logistic Regression (LR) as a baseline model. In addition, we evaluate more powerful non-linear models, across different families of Machine Learning models: We evaluate classical neural networks such as Multi-Layers Perceptron (MLP) and tree-based models such as Random Forest (RF), as well as two heavier boosting based models: XGBoost (XgB) and LightGBM (Lgbm). Finally, our initial hypothesis is that levees should be where human activity and flood risk overlap. Thus, we consider the following featuring engineering strategy that explicitly represent this hypothesis: We use as input features the cartesian product of flood risk variables with human activity variables. We hypothesize that if our assumption is correct, both feature engineering and stronger models should improve the baseline accuracy.

Table.1 shows the results of this experiment. Feature engineering has improved the ECE of almost all the models due to ECE reduction. Furthermore, the combination of nonlinear models and feature engineering scores considerably better than baseline, except for Lgbm. Therefore, our hypothesis allow for a adequate modelling of levee existence.

Table 1. ECE[%] of each model with two patterns of features, original (OR) and feature engineering (FE).

	Models				
Features	LR	RF	Lgbm	XgB	MLP
OR	0.16	0.26	0.22	0.05	0.21
FE	0.15	0.13	0.16	0.06	0.13

3.3. Generalization Study

The finality of our endeavor is to build a global system for levee detection. However, at the time being, we only have access to U.S. data through the NLD dataset. We are thus interested in the ability of our model to generalize spatially to different locations. As a preliminary study of the generalization ability of our model, we start by studying the impact of regional biases within the U.S. To do so, we devise two experiment setups:

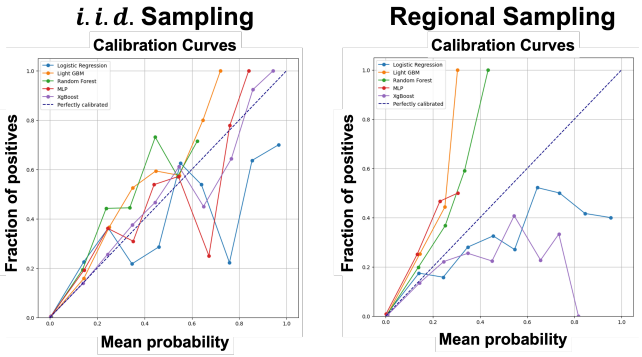

Fig. 2. Calibrated plots of *i.i.d.* sampling (left) and Regional sampling (right). Each color represents a different model and the dotted line, the ideal case. The output probability of the model is binned into 10 bins at 0.1 intervals, and the x-axis represents the mean probability of each bin.

- ***i.i.d.* sampling**: In this setup, we perform *i.i.d.* (independent and identically distributed) sampling of the training and test set across the whole U.S.
- **Regional sampling**: In this setup, we split training and test sets regionally (i.e. we use North U.S. data as training but the whole Mississippi delta as test data).

The more models trained on *i.i.d.* sampling outperform models on the regional sampling setup, the more impactful regional biases. Figure 2 shows that regional sampling is clearly more miscalibrated than *i.i.d.* sampling. MLP, RF, and Lgbm underestimate the probability of levees for many catchments with regional sampling. The graphs of the three models are located above the ideal line, and they only output probability values below 0.5. In contrast, the LR and XgB graphs are below the ideal line, overestimating the probability. These results show that the model's generalization is greatly affected by regional differences, which introduces systematic biases.

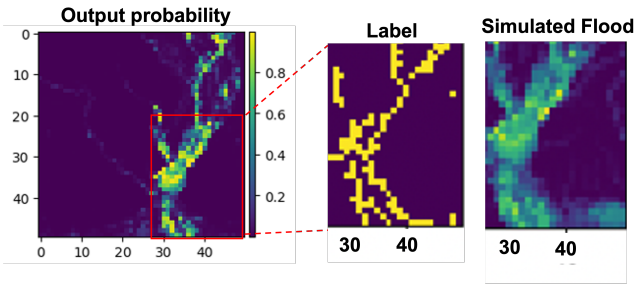

Fig. 3. (left) Probability distribution around the Mississippi River using LR. High values, which means the probability of the levee existence is high, are yellow, and low values are dark blue. (center) The actual levee location used as Label. (right) The flood simulation (SimFlood) as described in 2.2.

3.4. Error Analysis

We then visualize model outputs on a local patch around the Mississippi delta and highlight structural sources of errors. This analysis provides clues on what future work should focus to improve the predictive model. Focusing on the high probability catchments and labels, we see that the position of the levees are captured to some extent (Figure 3 left, center). We observe many catchments where levee existence are over-predicted. These catchments correspond to the cropland and wetland in the Mississippi River floodplain. These catchments are actually protected by the levees along the Mississippi River mainstem. However, because mainstem and floodplains are treated as independent catchments at 10km resolution, floodplain catchments do not have local levees in binary levee existence data. Consideration of the relationship among multiple catchments is needed in order to represent this phenomena. This phenomenon provides an obvious explanation for the error in our model : The predictive models estimate a higher probability for areas protected by levees rather than where they exist.

4. CONCLUSION

Inferring the location of levees is a key challenge for current global hydrological models. In this context, we propose a preliminary analysis leveraging machine learning to infer levee locations on the contiguous U.S. Our analysis reveals that quantifications of human activity and flood risk provide us with a good prior to infer levee locations. In this process, we have solved a number of technical difficulties including instability of standard evaluation approaches, which lead us to consider model calibration as a metric. Generalizing our approach globally will need to address generalization challenges due to regional biases and may require further hydrological modeling.

5. REFERENCES

[1] Yamazaki, Dai, et al. "A physically based description of floodplain inundation dynamics in a global river routing model." Water Resources Research 47.4 (2011).

[2] Yamazaki, Dai, et al. "A high-accuracy map of global terrain elevations." Geophysical Research Letters 44.11 (2017): 5844-5853.

[3] Tanaka, Yoshiaki, and Dai Yamazaki. "The automatic extraction of physical flood protection parameters for global river models." Journal of Japan Society of Civil Engineers, Ser. B1 (Hydraulic Engineering) 75.2 (2019): I_1099-I_1104.

[4] Tanoue, Masahiro, et al. "Residual flood damage under intensive adaptation." Nature Climate Change 11.10 (2021): 823-826.

[5] https://levees.sec.usace.army.mil/#/

[6] Yamazaki, Dai, et al. "MERIT Hydro: A high-resolution global hydrography map based on latest topography dataset." Water Resources Research 55.6 (2019): 5053-5073.

[7] Zhou, Xudong, et al. "The uncertainty of flood frequency analyses in hydrodynamic model simulations." Natural Hazards and Earth System Sciences 21.3 (2021): 1071-1085.

[8] Pekel, Jean-François, et al. "High-resolution mapping of global surface water and its long-term changes." Nature 540.7633 (2016): 418-422.

[9] Taguchi, Ryo, Tanoue, Masahiro and Yukiko Hirabayashi. "Development of a method for estimating business interruption losses from flooding globally." Abstracts of the Journal of Japan Society of Hydrology and Water Resources 31 (2018): 276.

[10] https://www.usgs.gov/centers/wgsc/science/global-food-security-support-analysis-data-30-m-gfsad

[11] http://www.cec.org/north-american-environmental-atlas/land-cover-2010-landsat-30m/

[12] https://landscan.ornl.gov

[13] He, Xinran, et al. "Practical lessons from predicting clicks on ads at facebook." Proceedings of the Eighth International Workshop on Data Mining for Online Advertising. 2014.

[14] Guo, Chuan, et al. "On calibration of modern neural networks." International Conference on Machine Learning. PMLR, 2017.