PLOS ONE

RESEARCH ARTICLE

Advanced susceptibility analysis of ground deformation disasters using large language models and machine learning: A Hangzhou City case study

Bofan Yu[1,2,3], Huaixue Xing[1,2]*, Weiya Ge[1,2]*, Liling Zhou[4], Jiaxing Yan[3], Yun-an Li[3]

1 Nanjing Center, China Geological Survey, Nanjing, The People's Republic of China, 2 The Ministry of Natural Resources' Urban Underground Space Exploration and Evaluation Technology Innovation Center, Qingdao, The People's Republic of China, 3 Zhejiang Institute of Geosciences, Observation and Research Station of Zhejiang Coastal Urban Geological Security, Ministry of Natural Resources, Hangzhou, The People's Republic of China, 4 The Institute of Geological Survey of China University of Geosciences (Wuhan), China University of Geosciences (Wuhan), The People's Republic of China

* 57670204@qq.com, xinghx@mail.cgs.gov.cn (HX); 391743801@qq.com (WG)

Check for updates

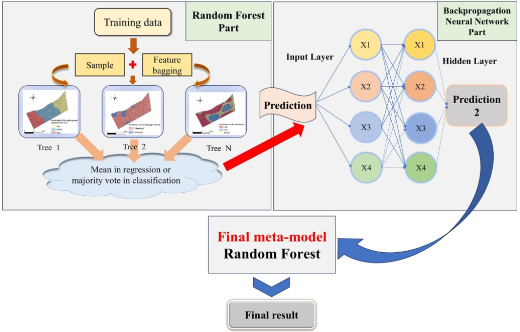
ANALYSIS OF REPRODUCIBILITY OF STUDY

Roman Kyrychenko

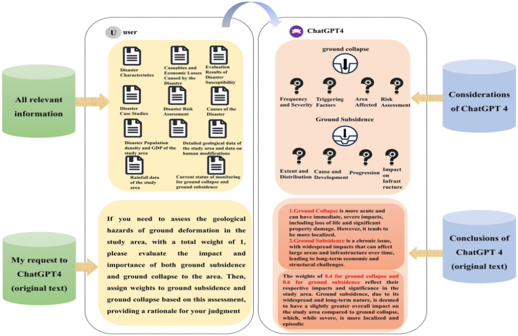
HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

WHAT AUTHORS DID

- Stack model: random forest + Feed forward neural network



- ChatGPT-based decision on weights and its comparison to expert weights



DATA AUTHORS USED

Name of Data	Size	Accessibility
Data source files of LLM and code files	1 long prompt	Public
ArcGIS data	Not shared	Third-party rights
Data processing	27,898 data points	Used Public

WHAT AUTHORS SHARED

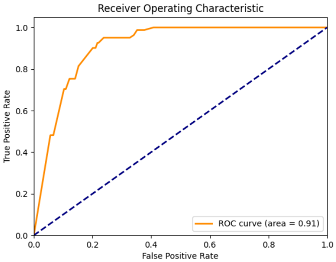
- Prompt for ChatGPT in pdf.
- Excel dataset with 27,898 data points for ground subsidence susceptibility assessment.
- Code for training and testing (without train/test split) in docx format.

WHAT AUTHORS ARE PROUD ABOUT

- Integrated data-driven models into urban ground collapse and subsidence evaluation.
- Used RF-BP neural network coupling model, achieving a 7% increase in AUC value.
- Employed ChatGPT-4 for weight determination, validated by geological experts.
- ChatGPT-4's weights differed by only 3% from expert judgments.
- Conducted comprehensive susceptibility assessment using ChatGPT-4's results.

REPRODUCIBILITY

- It is possible to reproduce results!
- Authors did not share train/test split code, but provided a good description, allowing reproduction from the description.
- Authors achieved 89% ROC-AUC score for ground collapse binary classification; I achieved 91%.
- I obtained the same weights for ground collapse versus subsidence (weight ratio of 0.4:0.6).

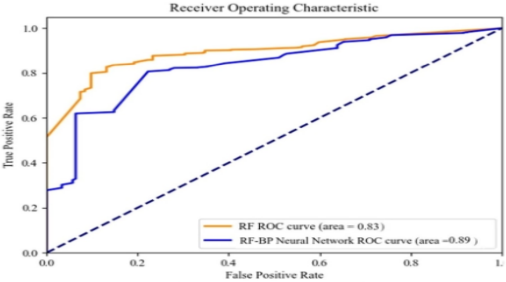


BUT...

Reproducibility $\rightarrow$more transparency.

I noted the following issues:

1. Factual error on the graphs:



Ground collapse

2. Inconsistency between text and graph:

The weights of 0.4 for ground collapse and 0.6 for ground subsidence reflect their respective impacts and significance in the study area. Ground subsidence, due to its widespread and long-term nature, is deemed to have a slightly greater overall impact on the study area compared to ground collapse, which, while severe, is more localized and episodic

In the comprehensive assessment of ground deformation hazards, with a weighting of 0.4 for ground subsidence and 0.6 for ground collapse, we categorized the risk zones as follows: low

SUSPICIOUS METHODOLOGICAL CHOICES

- Strange choice of file formats (docx for Python script, Excel instead of `.csv`).
- Default arguments for Random Forest and Neural Network, no hyperparameter tuning.
- Undersampling with Random Forest instead of using class weights.
- No data scaling, crucial for Neural Networks with wide value ranges (e.g., 1-178111).
- LLM assessment based on a single ChatGPT run, unspecified version and parameters.

MY ATTEMPTS TO SOLVE THESE ISSUES

- Added hyperparameter tuning.
- Scaled input values.
- Changed selection of train values to include all of them and added weight strategy for Random Forest.
- Ran multiple experiments with different GPT API models and various temperature values.

SUSPICION IS GROWING

1. During hyperparameter tuning, I noted that increasing parameters (making the model greedy) leads to better AUC on test data. Usually, greedy models lead to overfitting and failed test results.
2. LLM models give me 0.4:0.6 weights even on gibberish input.

MORE DETAILED LOOK AT DATA

1. During hyperparameter tuning, I noted that increasing parameters (making the model greedy) leads to better AUC on test data. Usually, greedy models lead to overfitting and failed test results.
2. LLM models give me 0.4:0.6 weights even on gibberish input.

MORE DETAILED LOOK AT DATA

Data check:

- Noted a lot of similar rows in the data file.
- Checked for duplicates and found:
  - 27,898 data rows $\rightarrow$625 unique rows.
  - 296 positive classes $\rightarrow$33 positive examples.
- Authors did not mention the problem of duplicated data in the text.
- Many duplicates are simultaneously in train (70% of all rows) and test (30%).

Conclusion: To get better results, the model needs to memorize data points.

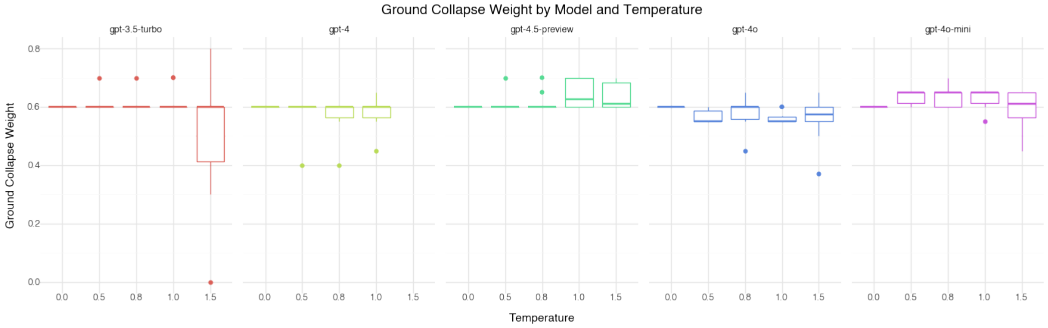
MY ALTERNATIVE AGGREGATION MODEL

- For this dataset, complex modeling is unnecessary due to extensive data duplication.
- A simple dictionary-based approach suffices: match test cases with the training data.
- If a test case isn't found in the training data, predict 0, the more frequent class.

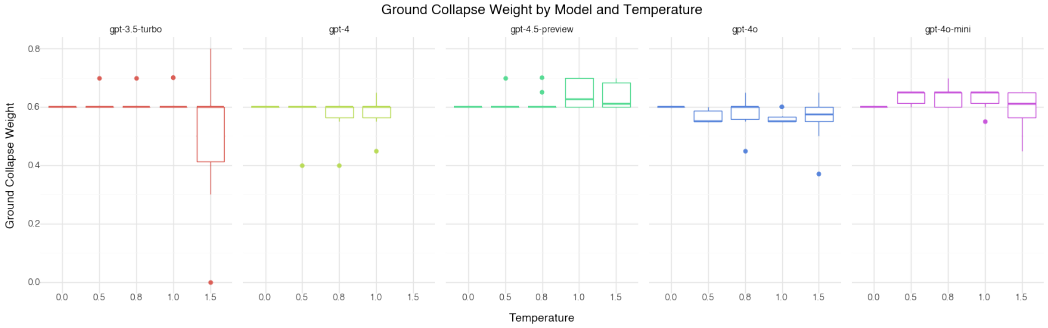
Results: This approach achieves an AUC of 91%, surpassing the authors' model stack by 2%.

...ch a straightforward method (like dictionary-based text analysis) might

HOW RESULTS LOOKS LIKE IF WE DO THIS STUDY CORRECTLY

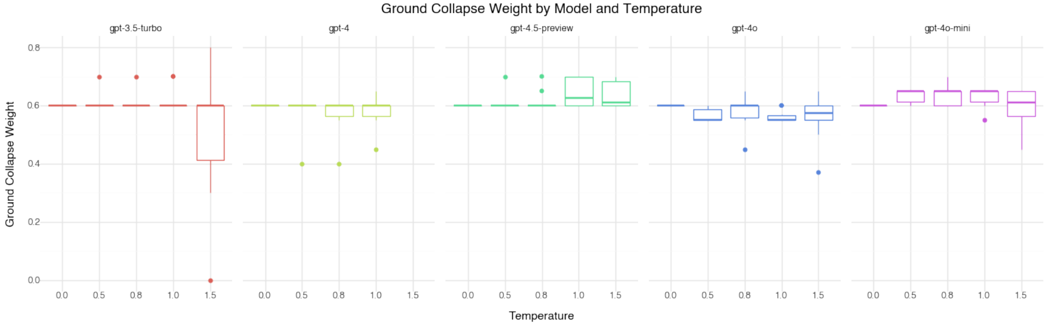


COMPARISON OF DIFFERENT MODELS



LLM RUNS

Weight 0.4 for Ground Collapse is rather outlier.



HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

LLM RUNS

Weight 0.4 for Ground Collapse is rather outlier.

model	ground$_c$*ollapse*				ground$_s$*ubsidence*			
	mean	std	min	max	mean	std	min	max
gpt-3.5-turbo	0.59	0.13	0.00	0.80	0.41	0.13	0.20	1.00
gpt-4	0.58	0.06	0.40	0.65	0.42	0.06	0.35	0.60
gpt-4.5-preview	0.63	0.04	0.60	0.70	0.37	0.04	0.30	0.40
gpt-4o	0.57	0.05	0.37	0.65	0.43	0.05	0.35	0.63
gpt-4o-mini	0.62	0.04	0.45	0.70	0.38	0.04	0.30	0.55