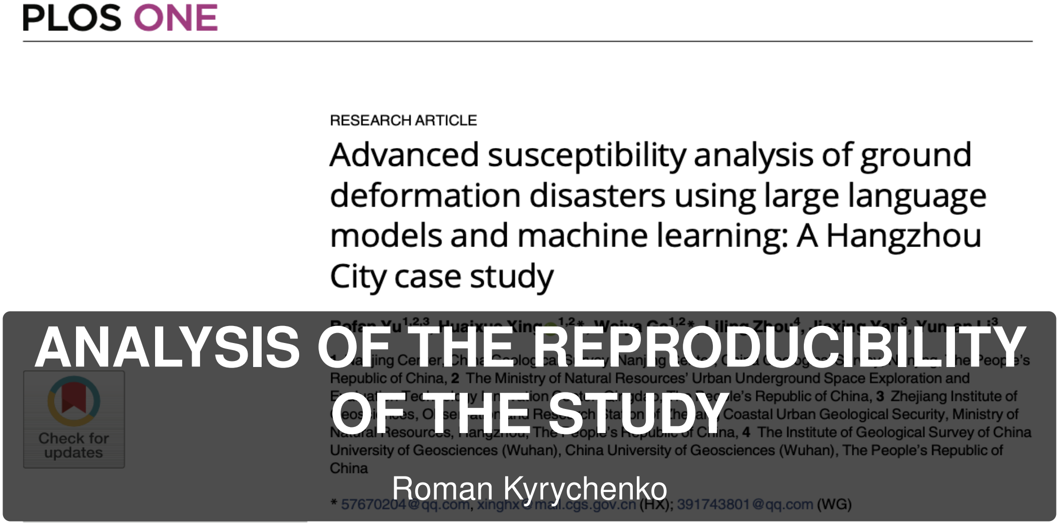
ANALYSIS OF THE REPRODUCIBILITY OF THE STUDY

Roman Kyrychenko



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]

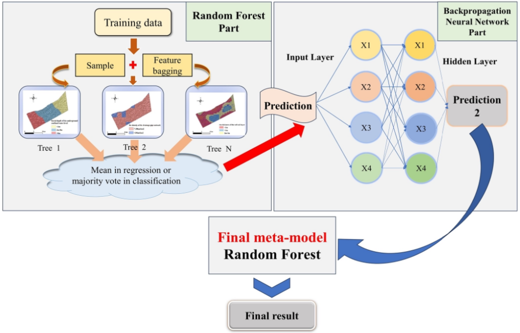
1 Nanjing Center, China Geological Survey, Nanjing Center, China Geological Survey, Nanjing, The People's Republic of China, **2** The Ministry of Natural Resources' Urban Underground Space Exploration and Evaluation Technology Innovation Center, Qingdao, The People's Republic of China, **3** Zhejiang 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)

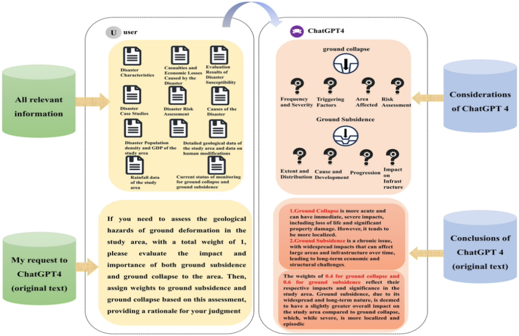
Check for updates

WHAT AUTHORS DID

- Stack model: random forest + Feed forward neural network



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


U user
Disaster Characteristics
Casualties and Economic Losses Caused by the Disaster
Evaluation Results of Disaster Susceptibility
Disaster Case Studies
Disaster Risk Assessment
Causes of the Disaster
Disaster Population density and GDP of the study area
Detailed geological data of the study area and data on human modifications
Rainfall data of the study area
Current status of monitoring for ground collapse and ground subsidence
All relevant information
My request to ChatGPT4 (original text)
ChatGPT4
ground collapse
Frequency and Severity
Triggering Factors
Area Affected
Risk Assessment
Ground Subsidence
Extent and Distribution
Cause and Development
Progression
Impact on Infrastructure
Considerations of ChatGPT 4
Conclusions of ChatGPT 4 (original text)

If you need to assess the geological hazards of ground deformation in the study area, with a total weight of 1, please evaluate the impact and importance of both ground subsidence and ground collapse to the area. Then, assign weights to ground subsidence and ground collapse based on this assessment, providing a rationale for your judgment

1.Ground Collapse is more acute and can have immediate, severe impacts, including loss of life and significant property damage. However, it tends to be more localized.
2.Ground Subsidence is a chronic issue, with widespread impacts that can affect large areas and infrastructure over time, leading to long-term economic and structural challenges.

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

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

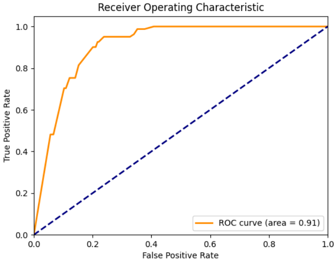
- Prompt for ChatGPT in pdf.
- Excel dataset with 27,898 data points for ground subsidence susceptibility assessment.
- Code for training and testing (without train/test split) in docx format.

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).

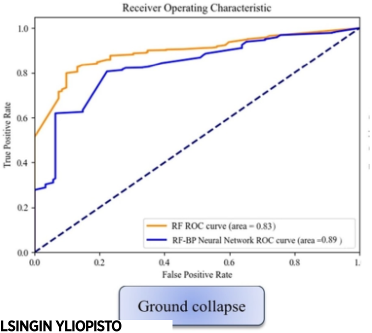


BUT...

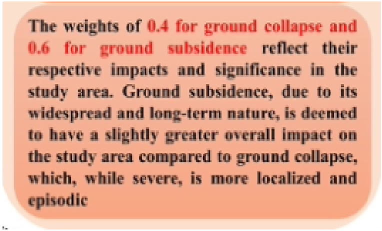
Reproducibility →more transparency.

I noted the following issues:

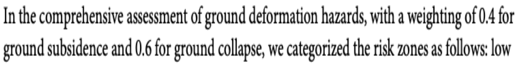
1. Factual error on the graphs:



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

Data check:

- Noted a lot of similar rows in the data file.
- Checked for duplicates and found:
  - 27,898 data rows →625 unique rows.
  - 296 positive classes →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.

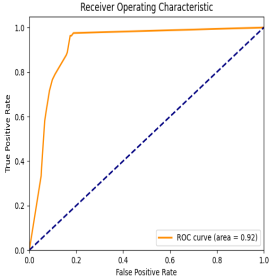
MY ALTERNATIVE AGGREGATION MODEL

- 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.

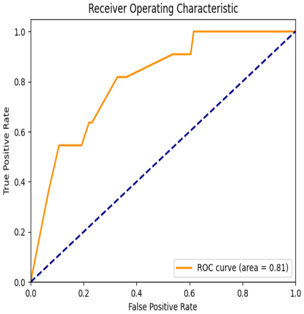
- If a test case isn't found in the training data, predict 0, the more frequent class.

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



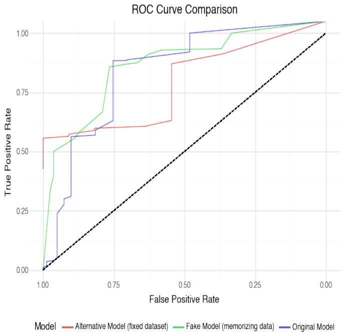
HOW RESULTS LOOKS LIKE IF WE DO THIS STUDY CORRECTLY

We won’t sell it! Or maybe...



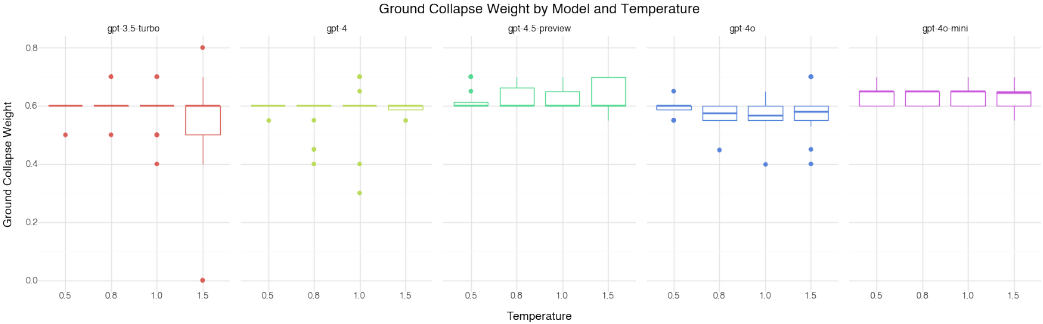
COMPARISON OF DIFFERENT MODELS

Probably with more time for hyperparameter tuning it’s possible to reach the level of broken models.



LLM RUNS

Weight 0.4 for Ground Collapse is rather outlier.



LLM RUNS

Weight 0.4 for Ground Collapse is rather outlier.

model	ground_collapse				ground_subsidence				
	mean	std	min	max	mean	std	min	max	count
gpt-3.5-turbo	0.58	0.11	0.00	0.80	0.39	0.09	0.00	0.60	80
gpt-4	0.59	0.06	0.30	0.70	0.41	0.06	0.30	0.70	64
gpt-4.5-preview	0.63	0.04	0.55	0.70	0.37	0.04	0.30	0.45	80
gpt-4o	0.58	0.05	0.40	0.70	0.42	0.05	0.30	0.60	79
gpt-4o-mini	0.63	0.03	0.55	0.70	0.37	0.03	0.30	0.45	80

CONCLUDING THOUGHTS

- The quality of reviewers' work in 1Q journals seems lacking; they had ample opportunities to detect these issues but did not.
- The study is reproducible, but the practices of sharing code and data could be improved.

REPRODUCIBILITY OF THIS REPRODUCTION

- GitHub repository with data, code for the study and analysis, presentation, and visualizations:
  `https://github.com/RomanKyrychenko/groud_collapse`
- Docker image:
  `https://github.com/RomanKyrychenko/groud_collapse/blob/master/Dockerfile`