A Machine Learning-Driven Approach for Automated Landfill Site Selection: An Experimental Study on Marmara Region, Türkiye

Atakan Bilgili [1], Tümay Arda [1], Batuhan Kilic [1,2], Melis Uzar [1]

[1] YTU, Dept. of Geomatic Engineering, Civil Engineering Faculty, 34220 Esenler Istanbul, Türkiye – (atakanb, tarda, batuhank, auzar)@yildiz.edu.tr
[2] Firat University, Dept. of Civil Engineering, Engineering Faculty, 23119 Elazig, Türkiye – batuhankilic@firat.edu.tr

Keywords: Municipal Solid Waste Disposal, Landfill, Site Selection, Machine Learning.

Abstract

This study introduces a novel machine learning (ML)-based framework for automated landfill site selection, applied to Türkiye's Marmara Region, a vital area experiencing rapid urbanization and industrial growth. Traditional methods, often reliant on subjective expert opinions and constrained by data complexity, are reimagined using state-of-the-art ML techniques, including Logistic Regression (LR), Random Forest (RF), and Extreme Gradient Boosting (XGBoost). Eighteen critical criteria—spanning hydrogeological, environmental, and infrastructural factors—were integrated into the framework. XGBoost achieved superior performance, with an accuracy of 0.8671, significantly outperforming LR and RF. Interpretability was enhanced using Shapley Additive Explanations (SHAP), which identified land use/land cover, distance to airports, and distance to industrial areas as the most influential factors. The resulting high-precision landfill suitability maps (LSMs) provide decision-makers with a reliable tool for selecting optimal landfill sites. This framework not only advances the technical rigor of landfill site selection but also supports sustainable waste management by addressing environmental, economic, and public health considerations. The study exemplifies the transformative potential of ML in tackling complex geospatial challenges, setting a precedent for integrating artificial intelligence into environmental planning and policy-making.

1. Introduction

In today's era, urbanization, industrialization, population growth, and technological advancements have significantly increased waste generation. Specifically, the replacement of manual labor with technology in agricultural activities has accelerated urban sprawl, leading to the emergence of various waste types. The rapid consumption of resources and shifts in consumption habits have resulted in an increase in waste types that need to be managed, posing serious threats to both environmental and human health (Rahimi et al., 2020). Furthermore, the uncontrolled disposal of waste has led to the contamination of surface and groundwater sources, as well as risks such as fires, landslides, and explosions (Şimşek and Alp 2022). To promote sustainable and integrated waste management, countries implement various policies and measures, such as recycling, reuse, waste reduction, thermal treatment, and landfilling (Tercan et al., 2020). In terms of costeffectiveness, landfills are the oldest and most commonly used approach among all waste management methods (Kuhaneswaran et al., 2024).

The selection of landfill sites involves the evaluation of engineering, technical, and economic protocols along with public health and environmental conditions. Moreover, multiple alternative criteria should be considered, such as land use, topography and soil characteristics, distance to various artificial structures, and hydrogeological features, including geological, groundwater, and surface water resources (Bilgilioglu et al., 2022). As a result, selecting a suitable landfill site becomes a complex process, increasing uncertainties and making it challenging for decision-makers to make sound decisions.

In the past decades, various methods such as diagramming, Geographic Information Systems (GIS), grey system theory, multi-criteria decision-making (MCDM) approaches, and GISMCDM integration have been employed to identify the most suitable locations for landfills (Rezaeisabzevar et al., 2020). However, these methods face several challenges: (1) relying on expert opinions, potentially introducing biases; (2) lacking the capacity to process large and complex datasets; (3) not flexible enough to accommodate changes and updates in data; and (4) encountering generalization issues when assessing conditions with similar characteristics. A potential approach to examining the suitability analysis to overcome the problems introduced by these methods involves the use of artificial intelligence techniques which mimic the cognitive decision-making abilities of humans.

1.1 Research Motivation and Objectives

To our knowledge, there is a lack of research specifically examining the application of machine learning (ML) approaches in the process of selecting suitable locations for landfill sites. Therefore, the primary objective of this study is to develop a pioneering machine learning-based framework to assist in the automated landfill site selection process to overcome the issues introduced by the traditional methods. The sub-aims of this study listed as follows:

- To assess the efficacy of both traditional and state-ofthe-art ML algorithms for identify suitable sites for landfill siting.
- To improve the accuracy and precision of landfill suitability maps (LSMs) by utilizing ML algorithms.
- To interpret the both local and global effects of the site selection criteria on landfill site selection via explainable artificial intelligence (XAI) methods.

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVIII-M-6-2025-73-2025 | © Author(s) 2025. CC BY 4.0 License.

2. Study Area and Data

The Marmara region, situated in northwestern Türkiye, was chosen as the experimental study area. This region covers an area of approximately 72,600 km² and is located between 25° 30′ and 31° E longitude and 39° to 42° 30′ N latitude. The underlying reason for selecting this region for evaluating landfill suitability is that it comprises 30.6% of Türkiye’s population and more than 50% of its economic activities, while also facing intense pollution due to rapid industrialization and population growth in parallel with recent economic and social development. Within the borders of the Marmara region, there are 11 provincial municipalities: Istanbul, Bursa, Kocaeli, Balıkesir, Tekirdağ, Sakarya, Çanakkale, Edirne, Kırklareli, Yalova, and Bilecik. All provincial municipalities offer disposal services by collecting solid waste from settlements and managing it in sanitary landfills. Figure 1 demonstrates an illustration of the designated study area.

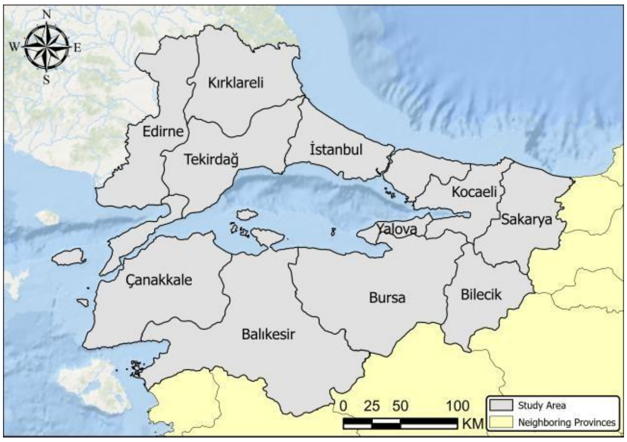

Figure 1. Study area.

2.1 Landfill Inventory

The existing landfill sites in the study area are crucial for establishing LSMs. As the ML models learn from the already established landfill sites to evaluate suitability of potential sites. In this way, a suitability map over a greater area can be generated utilizing a finite number of landfill sites. In Türkiye, there is no such database that holds the records of existing landfill sites. However, the environment status reports established by Ministry of Environment, Urbanization and Climate Change of Türkiye (Environmental Status Reports, 2024) contain necessary information to find locations of existing landfill sites.

The landfill sites were manually digitized according to these status reports into vector (polygon) format using ESRI ArcGIS Pro software, and they were labelled as suitable (labelled as 1). Conversely, ML models require data on unsuitable areas (labelled as 0) for landfill siting to effectively differentiate between suitable and unsuitable locations. To avoid bias, unsuitable land locations were randomly selected across the study area, ensuring an equal number of suitable and unsuitable sites.

2.2 Site Selection Criteria

Identifying potential candidate sites is a critical prerequisite for landfill siting. This process requires evaluating various assessment and restriction factors that enhance, reduce, or constrain the suitability of a candidate site. In this work, eighteen site selection criteria, namely annual rainfall, annual temperature, distance to airports, distance to educational facilities, distance to faults, distance to healthcare organizations, distance to industrial areas, distance to main roads, distance to protected zones, distance to railways, distance to settlements, distance to water bodies, drainage density, elevation, geology, groundwater level, land use/land cover (LULC), and slope were used considering the relevant literature. The format and sources of the criteria were summarized in Table 1.

Criterion	Data Format	Resolution	Source
Annual Rainfall	Raster	1 km	Fick and Hijmans (2017)
Annual Temperature	Raster	1 km	Karger et al. (2017)
Distance to Airports	Vector	-	Overpass Turbo (2024)
Distance to Educational Facilities	Vector	-	Overpass Turbo (2024)
Distance to Faults	Vector	-	Emre et al. (2013)
Distance to Healthcare Organizations	Vector	-	Overpass Turbo (2024)
Distance to Industrial Areas	Raster	100 m	CORINE Land Cover (2018)
Distance to Main Roads	Vector	-	Overpass Turbo (2024)
Distance to Protected Areas	Vector	-	Overpass Turbo (2024)
Distance to Railways	Vector	-	Overpass Turbo (2024)
Distance to Settlements	Vector	-	CORINE Land Cover
Distance to Water Bodies	Raster	100 m	CORINE Land Cover (2018)
Drainage Density	Raster	12 m	Zink et al. (2017)
Elevation	Raster	12 m	Zink et al. (2017)
Geology	Raster	5 km	Asch (2005)
Groundwater Level	Raster	1 km	Verkaik et al. (2022)
Land Use/Land Cover	Raster	100 m	CORINE Land Cover (2018)
Slope	Raster	12 m	Zink et al. (2017)

Table 1. Formats, resolutions, and sources of the landfill site selection criteria.

After the initial processing, each criterion was upsampled or downsampled according to their original resolution into raster format with a spatial resolution of 30 meters. Then, the pixel values for each criterion were then combined with the landfill inventory to generate a data frame suitable for use in machine learning models.

3. Methodology

3.1 Outline of the Proposed Approach

The approach proposed in this work includes five main phases: (1) data collection of 18 criteria and existing landfill sites to serve as features and ground truth data respectively, (2) feature selection which involves correlation and multicollinearity tests to check for relationships between the features and Sequential Feature Selection (SFS) method to check whether there is an irrelevant feature contained in the dataset, (3) model training for three ML algorithms such as Logistic Regression- LR, Random Forest- RF (Breiman, 2001), and Extreme Gradient Boosting- XGBoost (Chen and Guestrin, 2016), and model evaluation through various performance metrics, (4) interpretation of the ML algorithms through Shapley Additive Explanations (SHAP) (Shapley, 1953) method, and (5) generation of the LSMs.

3.2 Feature Investigation

The reliability of LSMs generated by ML models largely depends on accurately identifying the site selection criteria that influence the automatic selection of suitable locations for landfills. The relevant criteria should be identified before the initial model training in order to minimize the hindering effects of those features (i.e., site selection criteria). Feature inspection methods such as Pearson correlation coefficient and multicollinearity tests, and feature selection techniques (e.g., sequential feature selection (SFS)) could help solve this issue.

A multicollinearity test identifies the presence of interrelated measures within the data frame. Multicollinearity was assessed using the variance inflation factor (VIF) (Eq. (1)) and tolerance (TOL) metrics. Specifically, if the VIF value exceeds 10 or the TOL value is below 0.1, it indicates multicollinearity. If multicollinearity was detected, we utilized the Pearson correlation coefficient (Eq. (2)) as a check for the multicollinearity test. Two features are generally considered correlated if the Pearson correlation coefficient is greater than 0.7 or less than -0.7. If two features exceed the threshold levels from either multicollinearity test or Pearson correlation, they are considered correlated, and one of them was removed from the landfill data frame.

$$VIF_i = \frac{1}{1-R_i^2} \tag{1}$$

where VIF_i refers to Variance Inflation Factor for the i-th site selection criterion, and R_i^2 is the coefficient of determination of a regression model.

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{2}$$

where r is the Pearson correlation coefficient, x_i, y_i are the i-th samples for a criterion pair, and, $\bar{x}$, $\bar{y}$ are the means of a site selection criterion pair.

Sequential Feature Selection (SFS) is a wrapper-based feature selection method that can be performed either forward or backward. In backward elimination (which is utilized in this study), the process begins with all features in the dataset, and at each step, the least important feature is removed iteratively until removing any more features negatively impacts model performance. Hence, the best subset available features are obtained.

3.3 Machine Learning Models

Logistic Regression (LR) is a well-established statistical method primarily used for binary classification tasks. The algorithm predicts the probability that an input sample belongs to a particular class label by applying an S-shaped sigmoid function to a linear combination of input features. The output of the algorithm is a probability value between 0 and 1, which can later be used to assign class labels to the input according to a given threshold.

Random Forest (RF) is an ensemble learning method that can be used for both classification and regression tasks. The algorithm constructs several decision trees during training process, where each tree is trained on a random subset of the data and features. The output of the model is determined by aggregating the predictions of all individual trees in a process called majority voting (Breiman, 2001). It is a very popular algorithm that is used for variety of applications as it provides high accuracy along with resistance to the overfitting problem.

Extreme Gradient Boosting (XGBoost) is an ensemble learning method that is built on gradient boosting method. It builds a series of decision trees sequentially, where each new tree corrects the errors made by the previous one. The algorithm uses gradient boosting technique to optimize the model by minimizing the loss function through each iteration. Different from its predecessor gradient boosting machines, it uses second order approximation of the loss function which improves accuracy and efficiency. The output is determined by combining the predictions of all trees, typically through weighted voting. XGBoost is known for its high performance, scalability, and ability to handle complex learning tasks effectively.

After the initial model training, the hyperparameters parameters of each algorithm was tuned through grid search cross validation method to generate final classifiers, and the performance of the classifiers were evaluated through several metrics such as overall accuracy, precision, recall, and F1 score. The outputs of each algorithm were also compared via Cochran’s Q and pairwise McNemar’s tests in order to check for statistically significant differences.

3.4 Interpretation of the machine learning models

In machine learning-based pipelines, the interpretation of models is achieved through the evaluation of feature importances, such as permutation feature importance. However, these methods only provide a general understanding of how a feature contributes to the model's prediction. On the other hand, SHapley Additive exPlanations (SHAP) (Shapley, 1953), a state-of-the-art explainable artificial intelligence (XAI) method, has the capability to provide both global and local insights regarding how a feature or sample affects the model's behaviour.

4. Results

Eliminating irrelevant features from the dataset is a crucial step for machine learning-based site selection pipelines, as they may hinder the model performance significantly. In order the address those multicollinearity tests, Pearson correlation test, and a sequential backward feature selection were performed. The VIF and TOL values for each criterion are given in Table 2. According to the results of the multicollinearity test, the highest VIF was 15.713, while the least TOL value was 0.063 for elevation criterion, as presented in Table 2. The results of the Pearson correlation tests are presented in Figure 2 as a correlation matrix. Based on these results, some of the correlation coefficients exceed the threshold value of 0.7. The highest coefficient value, computed at -0.91, is between the annual temperature and the elevation. Considering both multicollinearity and correlation tests, it can be inferred that elevation and annual temperature are interrelated. As one increases, the other decreases (i.e., negative correlation). Hence, due to its lower original resolution compared to elevation, the annual temperature criterion was removed from the data frame.

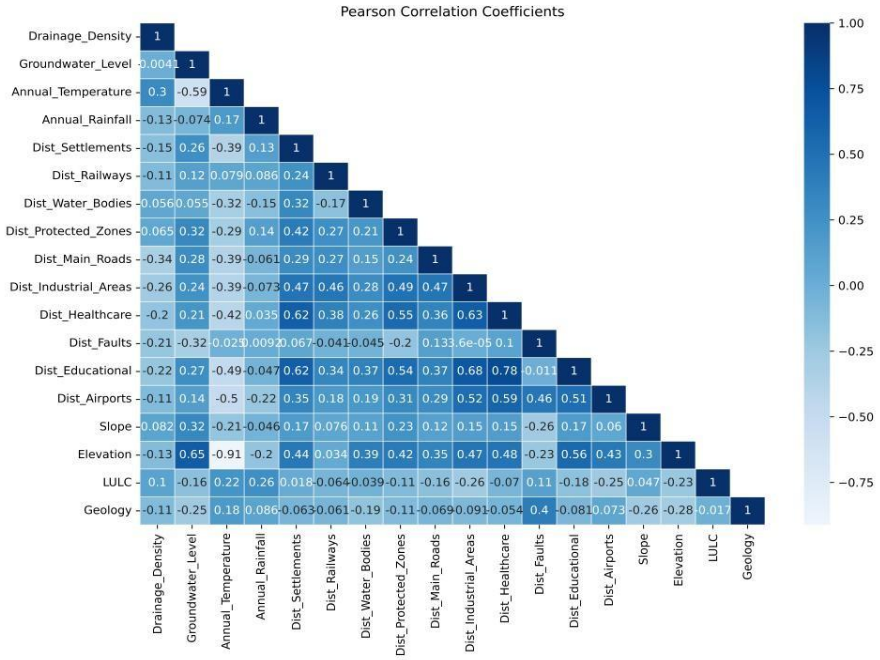

Figure 2. Pairwise correlation matrix.

Criterion	VIF	TOL
Annual Rainfall	1.369	0.730
Annual Temperature	14.728	0.068
Distance to Airports	3.301	0.303
Distance to Educational Facilities	3.726	0.268
Distance to Faults	3.052	0.328
Distance to Healthcare Organizations	3.620	0.276
Distance to Industrial Areas	2.805	0.356
Distance to Main Roads	1.566	0.639
Distance to Protected Zones	1.939	0.516
Distance to Railways	1.875	0.533
Distance to Settlements	2.032	0.492
Distance to Water Bodies	1.628	0.614
Drainage Density	1.845	0.542
Elevation	15.713	0.064
Geology	1.455	0.687
Groundwater Level	2.307	0.433
Land Use/Land Cover	1.372	0.729
Slope	1.241	0.806

Table 2. VIF and TOL values of each site selection criterion.

Furthermore, according to the results of sequential backward selection the 12 criterion namely drainage density, distance to settlements, distance to water bodies, distance to protected zones, distance to main roads, distance to industrial areas, distance to faults, distance to educational facilities, distance to airports, land use/land cover, and geology were kept in landfill data frame while annual rainfall, distance to healthcare organizations, distance to railways, distance to settlements, and groundwater level were eliminated.

A performance evaluation was conducted using several evaluation metrics. The confusion matrices of each ML model are shown in Figure 3. The evaluation metrics accuracy, recall, precision, and F1 score computed through these confusion matrices are presented in Table 3. For overall accuracy, the XGBoost model outperformed other models (0.8671). It was followed by RF (0.8513), and LR (0.8302).

Moreover, Cochran’s Q and pairwise McNemar’s tests were conducted to investigate whether there are statistically significant differences among ML models. According to the results of Cochran’s Q test, there was a statistically significant difference between ML models ($\chi 2(2) = 208.871$, $p < .000$) as it exceeded the threshold value of 12.592 at the 95% confidence interval.

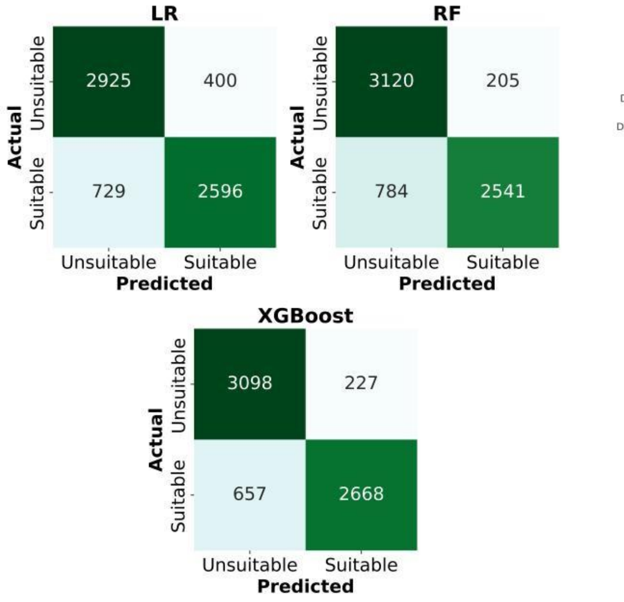

Figure 3. Confusion matrices of the ML algorithms used in the study.

Algorithm	Accuracy	Recall	Precision	F1 Score
LR	0.8302	0.7808	0.8665	0.8214
RF	0.8513	0.7642	0.9253	0.8371
XGBoost	0.8671	0.8024	0.9216	0.8579

Table 3. Evaluation metrics of the ML algorithms used in the study.

The results of pairwise McNemar's tests are given in Table 4. When the Table 4 is examined, it can be clearly seen that the output of all ML models is statistically significant. The greatest significant difference was between XGBoost and LR (173.574), and the least significant difference was between XGBoost and RF (43.092).

	LR	RF	XGBoost
LR	-	**70.515**	**173.574**
RF		-	**43.092**
XGBoost			-

Table 4. Results of the pairwise McNemar's tests.

It is essential to explain how an input site sample influences the predictions of machine learning models to derive meaningful insights for the optimal site selection of landfills. The SHAP summary plots provides both sample-wise and criterion-wise explanations insights for ML models. Figure 4 shows the bee-swarm plots of three ML models used in the study. According to the SHAP plot of XGBoost, LULC, distance to airports, distance to industrial areas, and distance to protected zones are the top contributing criteria, while distance to water bodies, distance to main roads, geology, and distance to settlements were the least contributing criteria. Similar trends can be observed for RF and LR, except for LULC criterion. According to the SHAP summary plots of ML models, closer distances to airports and industrial areas and further distances from faults are more favourable for landfill sites.

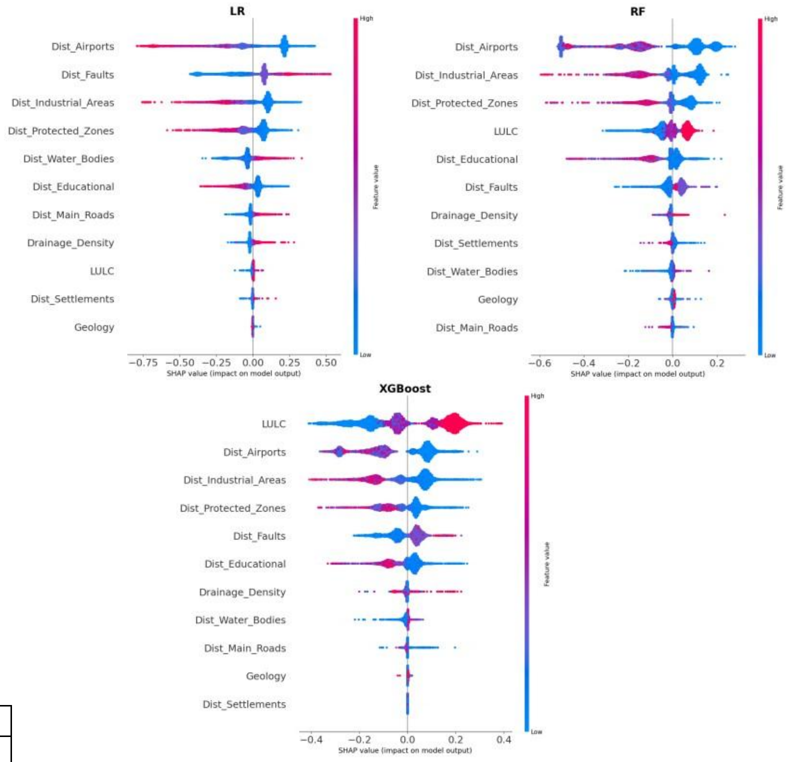

Figure 4. SHAP summary plots of the ML algorithms used in the study.

5. Conclusion

To promote sustainable and integrated waste management, countries implement various policies and measures, such as recycling, reuse, waste reduction, thermal treatment, and landfilling. Landfills are the oldest and most commonly used approach among all waste management methods as they are efficient to retain a sustainable waste management process, and promote cost-effectiveness. One of the most crucial aspects of sustainable waste management through landfills are the choosing optimal locations for landfills as stakeholders should consider several engineering, technical, and economic protocols along with public health and environmental conditions.

In this study, we developed a pioneering machine learningbased framework to assist in the automated landfill site selection process, contributing to the transition toward sustainable futures through effective waste management.

The main findings of our proposed approach are summarized as follows:

- According to results of multicollinearity tests and SFS, the twelve criteria were identified for training ML algorithms.
- The performance of XGBoost is superior compared to the RF and LR, and statistically significant differences found between XGBoost-RF, XGBoost-LR, and RF-LR.
- LULC, distance to airports, distance to industrial areas, and distance to protected zones are top contributing criteria to ML model predictions.
- The generated LSMs are convenient to use by decision-makers for selecting optimal landfill locations.

- Numerous potential sites for landfill siting have been identified, which are not yet established.

The preliminary results suggest that our method is feasible for automatically selecting suitable landfill locations, promoting a sustainable and integrated waste management process.

References

Asch, K., 2005. IGME 5000: 1: 5 million international geological map of Europe and adjacent areas. *BGR (Hannover)*.

Bilgilioglu, S.S., Gezgin, C., Orhan, O., Karakus, P., 2022. A GIS-based multi-criteria decision-making method for the selection of potential municipal solid waste disposal sites in Mersin, Turkey. *Environmental Science and Pollution Research*, 29, 5313–5329.

Breiman, L., 2001. Random forests. *Machine learning*, 45, 532.

Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794.

CORINE Land Cover, 2018 (vector/raster 100 m), Europe, 6-yearly (2024, November 1). https://land.copernicus.eu/en/products/corine-landcover/clc2018.

Emre, Ö., Duman, T.Y., Özalp, S., Elmacı, H., Olgun, Ş., Şaroglu, F., 2013. Açıklamalı Türkiye Diri Fay Haritası. Ölçek 1: 1.250. 000. Maden Tetkik ve Arama Genel Müdürlügü.

Environmental Status Reports of Türkiye Ministry of Environment, Urbanization and Climate Change, 2024. https://ced.csb.gov.tr/il-cevre-durum-raporlarii-82671.

Fick, S.E., Hijmans, R.J., 2017. WorldClim 2: New 1km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, 37(12), 4302-4315.

Karger, D., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R.W., Zimmermann, N.E., Linder, H.P., Kessler, M., 2017. Climatologies at high resolution for the earth’s land surface areas. *Scientific Data,* 4, 170122.

Kuhaneswaran, B., Chamanee, G., Kumara, B.T.G.S., 2024. A comprehensive review on the integration of geographic information systems and artificial intelligence for landfill site selection: A systematic mapping perspective. *Waste Management & Research*. doi.org/10.1177/0734242X241237.

Overpass Turbo-Web based data mining tool for OpenStreetMap, 2024. https://overpassturbo.eu/.

Rahimi, S., Hafezalkotob, A., Monavari, S.M., Hafezalkotob, A., Rahimi, R., 2020. Sustainable landfill site selection for municipal solid waste based on a hybrid decision-making approach: Fuzzy group BWM-MULTIMOORA-GIS. *Journal of Cleaner Production*, 248, 119186.

Rezaeisabzevar, Y., Bazargan, A., Zohourian, B., 2020. Landfill site selection using multi criteria decision making: Influential factors for comparing locations. *Journal of Environmental Sciences*, 93, 170-184.

Shapley, L.S, 1953. Stochastic Games*. *Proceedings of the National Academy of Sciences*, 39, 1095–1100.

Şimşek, K., Alp, S., 2022. Evaluation of landfill site selection by combining fuzzy tools in GIS-based multi-criteria decision analysis: a case study in Diyarbakır, Turkey. *Sustainability*, 14(16), 9810.

Tercan, E., Dereli, M.A., Tapkın, S., 2020. A GIS-based multi-criteria evaluation for MSW landfill site selection in Antalya, Burdur, Isparta planning zone in Turkey. *Environmental Earth Sciences*, 79(10), 246.

Verkaik, J., Sutanudjaja, E.H., Oude Essink, G.H.P., Lin, H. X., Bierkens, M.F.P. (2022). GLOBGM v1.0: a parallel implementation of a 30 arcsec PCR-GLOBWB-MODFLOW global-scale groundwater model. *Geoscientific Model Development*, 17(1), 275-300.

Zink, M. et al., 2017. "The global TanDEM-X DEM — A unique data set". *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, Fort Worth, TX, USA, 906-909.