TOAR-classifier v2: A data-driven classification tool for global air quality stations

Ramiyou Karim Mache[1,*], Sabine Schröder[1], Michael Langguth[1,*], Ankit Patnala[1], and Martin G. Schultz[1,2]

[1]Jülich Supercomputing Centre, Forschungszentrum Jülich, 52425 Jülich, Germany
[2]Department of Mathematics and Computer Science, University of Cologne, Cologne, Germany
[*]no longer at institute

Correspondence: Ramiyou Karim Mache (k.mache@fz-juelich.de)

Abstract.

Accurate characterization of station locations is crucial for reliable air quality assessments such as the Tropospheric Ozone Assessment Report (TOAR). ~~This study introduces~~ While urban and rural areas are relatively well-defined, the boundaries and identity of suburban areas remain ambiguous, overlapping with both urban and rural zones and varying due to cultural and social factors. This study investigates a machine learning approach to classify ~~23,974~~ 24,348 stations in the unique global TOAR database as urban, suburban, or rural. We tested ~~several~~ two different methods: unsupervised K-means clustering with three clusters, and an ensemble of supervised learning classifiers including random forest, CatBoost, and LightGBM. We ~~further enhanced the supervised learning performance by integrating~~ integrate these classifiers into a robust voting model, leveraging their collective predictive power. To address the inherent ambiguity of suburban areas, we implement ~~an~~ a grid-search adjusted threshold probability technique. Our models, trained on the TOAR station metadata, are evaluated on 1,~~000~~ 979 unseen data points. K-means clustering achieves ~~70.03% and 71.53~~71.88 % and 87.67% accuracy for urban and rural areas respectively, but only ~~26.36~~15.84% for suburban zones. ~~Supervised~~ The supervised classifiers surpass this performance, reaching over 84% accuracy for urban and rural categories, and ~~62-65~~66% - 72% for suburban areas. The adjusted threshold technique significantly enhances overall model accuracy, particularly for suburban classification. The good separation of our model is confirmed through evaluation with NOx and PM2.5 concentration measurements, which were not included in the training data. Furthermore, manual inspection of ~~25 individual~~ 30 randomly selected sites with Google maps reveals that our method provides a better label for the station type than the labels that were reported by data providers and used in the model evaluation. The objective station classification proposed in this paper therefore provides a robust foundation for ~~type-specific~~ type-of-area-specific air quality assessments in TOAR and elsewhere.

1 Introduction

Ozone in the troposphere plays a crucial role in human and environmental health, (Post et al., 2012; Griffiths et al., 2021). As a significant atmospheric pollutant and greenhouse gas, ozone profoundly impacts air quality and contributes to the dynamics of climate change (Madronich et al., 2023; Orru et al., 2013). Accurate ozone monitoring data is essential for shaping public

health policies and ecological regulations. Modern data infrastructures can be used to provide atmospheric scientists with the necessary metrics to quantify ozone's impact on climate, human health, and vegetation (Gaudel et al., 2018; Fleming et al.; Mills et al., 2018; Teakles et al., 2017; Cooper et al., 2014; Monks et al., 2015; Schultz et al., 2015). In 2021, the International Global Atmospheric Chemistry project (IGAC) launched the second phase of the Tropospheric Ozone Assessment Report (TOAR-II) to undertake a comprehensive review of the global distribution and trends of tropospheric ozone. A key accomplishment of TOAR-II is the development of a new terabyte-scale relational database of surface ozone observations and related variables. This database includes hourly measurement data and enriched metadata from 1970 to 2023, collating information from over 20,000 measurement sites worldwide through collaboration among multiple data centers and individual researchers (Schröder et al., manuscript in preparation). The new TOAR-II database replaces and extends the first TOAR database that has been described in Schultz et al. (2017a). Ozone levels exhibit significant regional variations and distinct patterns across different pollution environments. For example, urban environments with large ozone precursor emissions can exhibit "zero ozone" (i.e., ozone at sub-nmol fractions) situations and very large variability, while concentrations in rural areas tend to be smoother (Zhou et al., 2022; Schultz et al., 2017b). To accurately assess the ozone situation at individual locations and interpret ozone trends across the globe, it is therefore important to characterize measurement sites in a globally consistent and objective manner. While many measurement networks provide information about the station location or "type of station" and "type of station area", this metadata information is inconsistent between regions and error-prone as it involves some subjective judgement (c.f., Tapia et al., 2016).

In the first assessment of TOAR Schultz et al. (2017a) pioneered a new way to classify stations in a globally uniform way based on a set of Earth Observation (EO) datasets that have been processed at the station locations. This method used manually selected threshold values of station altitude, population density, nighttime light intensity, NO_2 column density, and NOx emissions to characterize stations as urban or rural. While this approach provided a useful distinction between "clearly urban" and "clearly rural" sites, it fell short of classifying all sites ~~(~~as almost half of the stations remained unclassified~~) and~~ , see Figure 1. Furthermore, the method was criticized for lack of an objective definition of the threshold values. Especially, the boundaries and characteristics of suburban areas remain ambiguous. Suburban zones, typically located at the periphery of cities and noted for lower density and residential land use, often overlap with both urban and rural regions. Their definition is shaped by cultural, social, and psychological factors, resulting in varied interpretations and a lack of universal consensus, (Hesse and Siedentop, 2018; Airgood-Obrycki and Rieger, 2019). This emphasizes the need for an objective and automated station classification method. In this study, we propose a new machine learning (ML) approach to develop a more advanced and unbiased classifier using similar objective metadata from the TOAR-II database. Our primary objective is to create a machine learning model that classifies stations in the TOAR database as urban, suburban, or rural. We implement and compare two methodologies: unsupervised learning using K-means clustering, and supervised learning classifiers such as random forest, CatBoosting, and LightGBM. In supervised learning, a subset of station characteristics is known and used to train the classifiers which are then used to predict the class of ~~unclassified station locations~~unlabeled stations. The supervised models are evaluated individually and after applying a robust voting method. Furthermore, an adjusted threshold technique is applied to enhance the ~~separation~~ identification of suburban stations. The ~~next section presents~~ remainder of this paper is organized as follows: Section

2 describes the data and ~~methodology, followed by Section 3, which focuses on~~ methods used, Section 3 presents the results and discussion~~. A~~ , and a general conclusion wraps up the paper.

2 Data and methods

This section provides an overview of our data sources~~and methodology~~, machine methods used, and evaluation metrics. We begin by introducing the TOAR-II database and the station metadata that are used as inputs of the ML models. We then detail our data preparation process, including preprocessing, feature engineering, and feature selection, all critical steps in preparing data for ML models. Finally, we present a concise summary of the ML models employed in this study and the evaluation metrics used to assess their performance.

2.1 TOAR-II database and station metadata

Developed in the context of TOAR phase II, the TOAR-II database stands as one of the world's largest collections of near-surface ozone measurements and related information. The database can be accessed through web services which provide a comprehensive suite of ozone-related data products including standard statistics, health and vegetation impact metrics, and trend information (https://toar-data.fz-juelich.de/). The TOAR-II database includes extensive information describing the locations of air quality measurement stations based on pollution-relevant properties. These properties are extracted from EO data and stored as station metadata in the database. This metadata offers contextual information about the measurement site, enabling station location characterization. Table 1 below summarizes all metadata used in this study including references to the data origin.

Some metadata, such as station coordinates, are provided by many air quality agencies and scientific institutions that contribute ~~data to TOAR~~to the TOAR database. The other metadata elements listed in Table 1 stem from ~~the following~~ EO datasets, which were downloaded from the respective provider sites. A special web service called Geospatial point extraction and aggregation service (GeoPEAS) has been developed to compute the aggregate information from the original gridded products. ~~Information~~ More information about the EO datasets used in GeoPEAS can be found on https://toar-data.fz-juelich.de/api/v2/#stat After the metadata extraction with GeoPEAS, all metadata are available as lists of key-value pairs with the keys corresponding to the variable names in Table 1. For further processing, this data was collected into one table with the keys as data columns and the individual stations as rows.

2.2 Data ~~pre-processing~~ preprocessing and feature selection

The first step in our data ~~prepossessing~~ preprocessing pipeline consists of cleaning the dataset. This ~~involved~~ involves removing all duplicate data points, replacing values of -999.0 with NaN to denote missing values, and eliminating rows where all metadata information is missing. To ensure consistency, we filtered out rows with inconsistent values, such as negative population density or negative maximum stable lights. In total, this ~~eliminates 1,596~~ step eliminates 211 stations out of ~~23,974~~ 24,348 stations. For handling missing altitude data, we fill these with mean_topography_srtm_alt_90m_year1994. Other numerical missing values

Table 1. Station metadata in the TOAR database used in this work.

Variable name	Description	Type
lon, lat	longitude, latitude represent the geographical coordinates of station. We did not use these coordinates as predictors in the machine learning model	Numeric
area_code	Unique code of the station in TOAR database	String
altitude	altitude of the station location in meter (m)	Numeric
mean_topography_srtm_alt_90m_year1994 mean_topography_srtm_alt_1km_year1994	mean value within 90m and 1km of relative altitude of the year 1994. data source: NASA Shuttle Radar Topographic Mission (SRTM) (Jarvis et al., 2008)	Numeric
max_topography_srtm_relative_alt_5km_year1994 min_topography_srtm_relative_alt_5km_year1994 stddev_topography_srtm_relative_alt_5km_year1994	maximum, minimum, and standard deviation of the relative altitude within a radius of 5km around the station in 1994 data source: NASA Shuttle Radar Topographic Mission (SRTM) (Jarvis et al., 2008)	Numeric
climatic_zone_year2016	climate zone of the year 2016. Provides information about climatic conditions at a location including whether it tends to be hot or cold, humid or dry, or exhibits a tropical climate data source: University of East Anglia Climatic Research Unit (Harris and Jones, 2017)	Category
mean_stable_nightlights_1km_year2013 mean_stable_nightlights_5km_year2013 max_stable_nightlights_25km_year2013 max_stable_nightlights_25km_year1992	average and maximum nighttime light value of the years 1992, and 2013 in 1km, 5km, and 25km around the station location. The values in this data set represent a brightness index ranging from 0 to 63. data source: NOAA National Centers for Environmental Information (NCEI) (Kroehl, 1982)	Numeric
mean_population_density_250m_year2015 mean_population_density_5km_year2015 max_population_density_25km_year2015 mean_population_density_250m_year1990 mean_population_density_5km_year1990 max_population_density_25km_year1990	Average and maximum population density of the years 1990, and 2015 in 250m, 5km, and 25km radius around the station location data source: The European Commission, Joint Research Centre, (Florczyk et al., 2019)	Numeric
mean_nox_emissions_10km_year2015 mean_nox_emissions_10km_year2000	Average annual NOx emission of the years 2000 and 2015 in a 10km radius around the station location data source: Copernicus Atmosphere Monitoring Service (Granier et al., 2019)	Numeric
timezone	Geographical area, such as Africa, America, Europe, etc.	Category
type_of_area (target)	Characterization of station location (urban, suburban, rural, or unknown) reported by the data providers of the TOAR database. This variable is not used in K-means clustering, and the known part, i.e stations labelled as urban, suburban, rural, are employed to train supervised classifiers.	Category

are estimated using the regression iterative imputer, Rubinsteyn and Feldman (2016), and categorical missing values fill with most frequent instance. Categorical variables are encoded using OneHotEncoder from scikit-learn, Pedregosa et al. (2011b). ~~Additionally, we applied a Box-Cox transformation to the NOx emissions data to normalize its distribution. We then~~

For K-means clustering, numerical features were scaled using standard scaling, and outliers were handled with IQR-based clipping within the range [Q1 - 1.5·IQR, Q3 + 1.5·IQR], where IQR (Interquartile Range) is the difference between the third quartile (Q3) and the first quartile (Q1). These preprocessing steps are essential to mitigate the algorithm's sensitivity to feature scale and outliers. For supervised learning algorithms, we applied a robust scaler, Pedregosa et al. (2011b) to the entire dataset, which proved more effective for this task compared to alternative scaling methods such as standard or min-max scaling. This scaling approach helps mitigate the impact of outliers and ensures consistent feature ranges. In the feature selection process, we prioritized variables containing the most recent available information, ensuring our model utilizes the most up-to-date data for

classification. Notably, geographical coordinates (longitude-latitude) and station codes are excluded from the machine learning models to prevent overfitting to geolocation.

~~After preprocessing~~ Following preprocessing steps, the final dataset comprises 24,125 stations. Of these, 13,~~the dataset consists of 22,378 rows.~~ 225 are labeled as urban, suburban, or rural, while the remaining 10,900 are unlabeled. The distribution of the processed data is presented in Figure 1.

For the K-means clustering analysis, we allocated 22,~~378~~ 111 samples for model training and reserved 15% of the labeled data (1,~~000 samples for testing purposes. The training of the supervised models had to be done with a smaller dataset, because only 12,408~~ 979 samples) for testing. The supervised models were trained on a smaller subset because only 13,225 stations were explicitly ~~categorized~~ classified as urban, suburban, or rural by the TOAR data providers. The remaining ~~9,970~~ 10,900 stations lacked this specific classification, being reported as 'unknown' or without a designated category. ~~This emphasizes the need for an objective station classification method as described in this paper. For the supervised models, we used~~ We used 85% of the labeled data (11,~~408 samples for training and 1,000 for testing, as visualized in (Figure 2~~211 samples) for training while the remaining 15% is allocated for testing. The distribution of training and testing datasets is presented in Figure 1. As shown, the training dataset is imbalanced, with a higher number of urban stations compared to suburban and rural ones. However, the class imbalance is not severe. We address it by applying the Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2002) to the training data before fitting the supervised classifiers. It is important to note that SMOTE was applied exclusively to the training set and not to the test data. The trained classifier is then ~~employed~~ used to predict the characteristics of ~~all 22,378 stations~~ unlabeled stations as illustrated in Figure 2.

To ensure the reliability of our machine learning approach, we manually selected ~~33~~ 35 stations with clear decision ~~boundaries~~ boundaries–that is, stations that are easy to classify as urban, suburban or rural, and excluded them from the training dataset. These stations were explicitly reported by data providers as urban, suburban, or rural, and we used Google Maps (Mehta et al., 2019) to manually verify and label them. During this process, we observed discrepancies between the labels provided by the data providers and those derived from our manual Google Maps analysis. Our models will first be evaluated on these ~~33~~ 35 stations, with accuracy calculated both against the labels reported by the data providers and against the manually verified labels from Google Maps (referred to as hand-labeled data).

2.3 Machine learning algorithms and evaluation methods

This section is devoted to a concise overview of the machine learning techniques employed in this study. Additionally, we describe the evaluation metrics used to quantify the effectiveness and accuracy of our models.

2.3.1 Machine learning algorithms

- **K-means clustering** is a widely-used unsupervised machine learning technique that aims to partition data into k distinct groups called clusters Bahmani et al. (2012); Sinaga and Yang (2020); Pelleg et al. (2000). Each data point is assigned to the cluster with the nearest centroid. The algorithm seeks to minimize the within-cluster sum of squares, which measures the squared distances between data points and their respective centroids. One key requirement of K-means is specifying

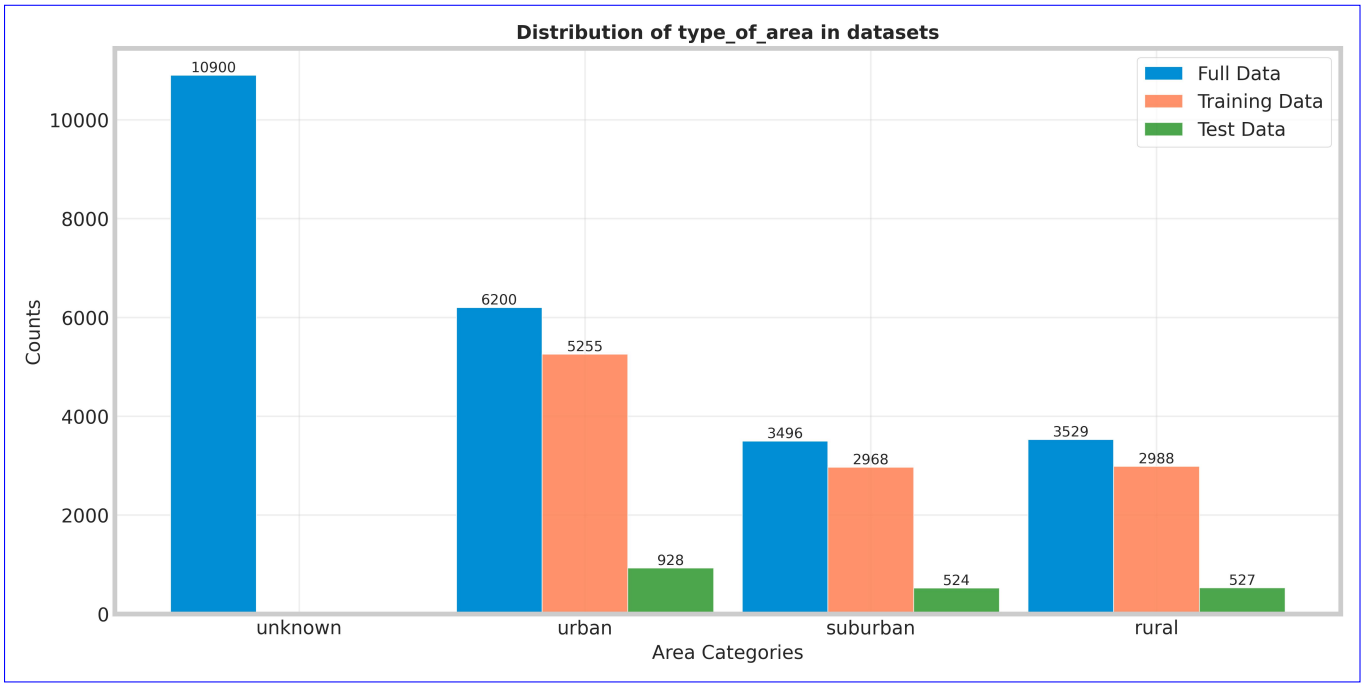


Figure 1. Distribution of unlabeled, training, and test data used for the supervised ML models as described in the text

the number of clusters beforehand. We employed the heuristic elbow method to determine the appropriate number of clusters for our task. The elbow method is a heuristic technique used to determine the optimal number of clusters in K-means clustering. It works by plotting the within-cluster sum of squares (WCSS) as a function of the number of clusters and identifying the "elbow" point on the curve, which correspond to the optimal number of clusters (Ketchen and Shook, 1996). In our K-means clustering application, we determine that three clusters are optimal (see Figure 6(a)), which aligns well with our objective of categorizing the stations into three groups: urban, suburban, and rural. However, it is important to note that visual inspection of correlation plots between individual metadata values Figure 2 reveals rather fuzzy boundaries between clusters. This observation aligns with our expectations, given the diverse nature of urban, suburban, and rural locations across different countries, which can vary significantly in terms of industrial development, population density, and degree of urbanization in different countries (Zhang et al., 2024).

- **Random Forest classifier** is a widely-used machine learning algorithm for classification tasks. As an ensemble learning method, it constructs multiple decision trees through bagging during training and outputs the class that is predicted by the majority vote of the individual trees (Breiman, 2001). Known for its robustness, it naturally resists overfitting through random feature selection and typically requires minimal tuning compared to other algorithms. In our implementation, we train the Random Forest classifier with 500 estimators, employing entropy as the optimization criterion. These specific

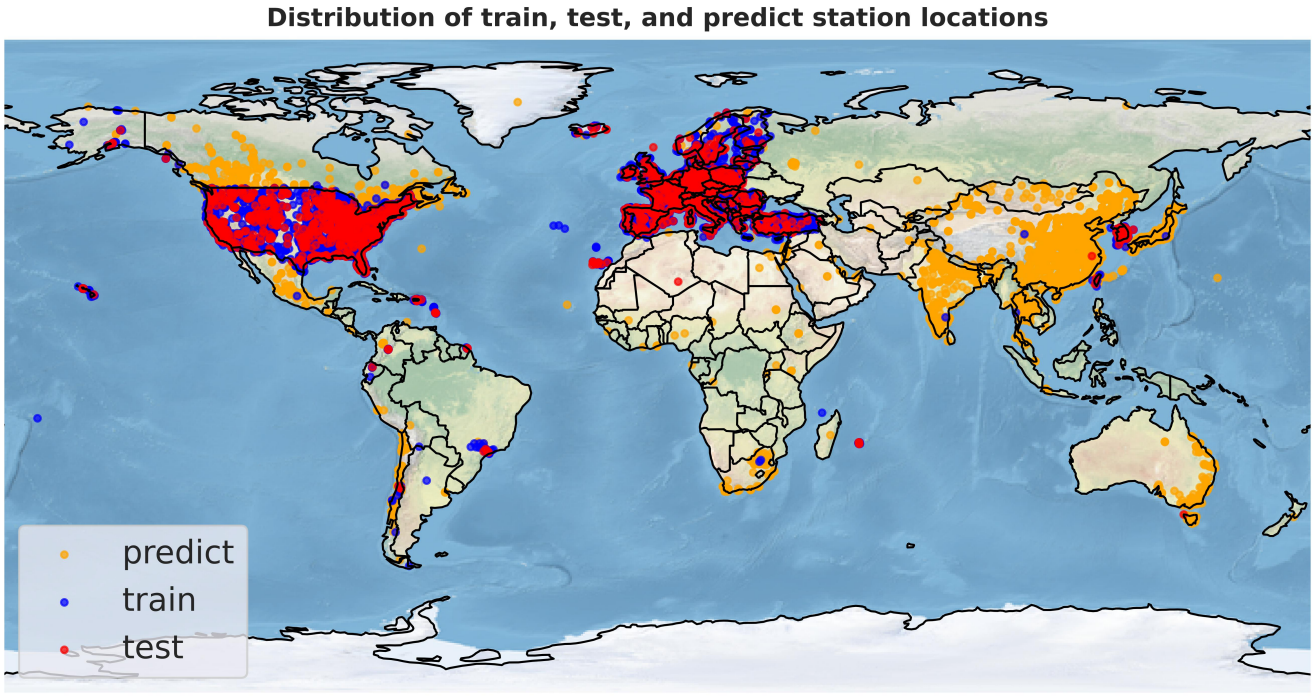


Figure 2. Distribution of unlabeled, training, and test data used for the supervised ML models as described in the text

hyperparameters were determined through a grid search. We utilize the RandomForestClassifier from the scikit-learn library (Pedregosa et al., 2011a).

- **The LightGBM (LGBM) classifier** is a supervised machine learning algorithm that utilizes gradient boosting techniques and tree-based learning. It employs histogram-based algorithms and leaf-wise tree growth strategies, which contribute to accelerated training speeds and reduced memory consumption. LightGBM is particularly well-suited for handling large-scale datasets. Its lightweight architecture and optimized algorithm make it a popular choice for tasks requiring both speed and accuracy in prediction (Ke et al., 2017). In our implementation, we train LightGBM with 500 estimators using Python's open-source library 'lightgbm' (Van Rossum et al., 2007).

- **The CatBoost classifier** is a machine learning algorithm that uses gradient boosting on decision trees, specifically designed to handle categorical features seamlessly Prokhorenkova et al. (2018). CatBoost stands for "Categorical Boosting" and automatically handles categorical variables without requiring manual prepossessing. It uses symmetric trees and ordered boosting to prevent overfitting, and often outperforms other methods on datasets with categorical data. This makes it an attractive option for datasets containing both numerical and categorical variables. In our implementation, we employ the open-source CatBoost library and configure the model with 500 estimators to balance performance and computational efficiency.

- **The Voting Classifier** is an ensemble meta-estimator that combines predictions from multiple base models to enhance overall accuracy and robustness. It functions by either majority vote ('hard') or averaging predicted probabilities ('soft'), effectively balancing the weaknesses of individual classifiers. This approach often yields superior generalization and reduced overfitting. Our implementation uses a soft voting strategy to aggregate predictions from a Random Forest, LightGBM, and CatBoost classifier via the scikit-learn library. (Pedregosa et al., 2011a)

2.3.2 Leveraging Model Uncertainty to Enhance Suburban Classification Accuracy

Considering the inherent subjectivity in defining suburban areas, we refined our prediction methodology as follows: For any given station, if the model's highest probability for either urban or rural classification falls below a threshold~~(we set this threshold to 50%)~~, and if the second-highest probability corresponds to suburban classification, we interpret this as the model's uncertainty in categorizing the area between rural and suburban, or urban and suburban. In such cases, we classify the station as ~~as suburban.~~ suburban. We applied the grid-search strategy in the range [0.35, 0.85] minimizing macro-F1, to find optimal threshold for each supervised algorithms. We obtain the thresholds of 0.5214, 0.4357, 0.5459, 0.4846 for random forest, CatBoost, LightGBM, and voting respectively. This approach acknowledges the model's indecision and leverages it to better capture the nuanced nature of suburban environments.

2.3.3 Evaluation

To evaluate our model's accuracy, we use a separate test dataset of 1,~~000~~ 979 samples, shown as red dots in Figure 2. The test dataset consists of samples that were selected with stratification to ensure it reflects the distribution of real data, and intentionally excluded from both the training phase and the hyperparameter tuning process. This approach ensures that the evaluation metrics provide an unbiased assessment of the model's ability to generalize to new, unseen data, challenging the model in the real-world application scenario. We employed the following evaluation metrics to measure the performance of our machine learning model on this test dataset. ~~We use the following three metrics for the evaluation of our results:~~

- **Accuracy** measures the ability of the machine learning model to accurately predict the outcome for the given input data. It is measured as the proportion of correct predictions to the total number of predictions made by the model, and given by the following formula:

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions} \times 100$$

$$\text{Accuracy} = \frac{\#\text{Correct Predictions}}{\#\text{Total Predictions}} \times 100$$

Here and in the following formulas, # stands for "Number of"

- **Per-class Precision:** For a given class c, precision quantifies the fraction of correctly predicted instances of class c among all instances that the model predicted as class c:

$$\text{Precision(c)} = \frac{\#\text{True Positives}(c)}{\#\text{True Positives}(c) + \#\text{False Positives}(c)} \times 100$$

Here, True Positives (TP) are the instances that actually belong to class c and were correctly predicted as such, while False Positives (FP) are the instances that do not belong to class c but were incorrectly predicted as instance of class c. Precision reflects how reliable predictions of class c are: high precision indicates that when the model predicts c, it is usually correct (Powers, 2020; Brodersen et al., 2010).

- **Per-class Recall (Sensitivity, True Positive Rate):** For a given class c, recall is defined as the proportion of true positive predictions for class c among all actual instances belonging to class c

$$\text{Recall}(c) = \frac{\#\text{True Positives}(c)}{\#\text{True Positives}(c) \ + \ \#\text{False Negatives}(c)} \times 100$$

Here, False Negatives are the instances of class c that the model incorrectly assigned to another class. Recall characterizes the ability of the model to retrieve all relevant instances of class c; high recall indicates that the model rarely overlooks samples from this class (Powers, 2020).

- **Per-class F1 Score:** The F1 score for class c is defined as the harmonic mean of precision and recall:

$$\text{F1}(c) = \frac{2 \cdot \text{Precision}(c) \cdot \text{Recall}(c)}{\text{Precision}(c) \ + \ \text{Recall}(c)} \times 100$$

This metric provides a single, balanced measure of a model's ability to achieve both high precision and high recall for class c, penalizing extreme values in either criterion. A high F1 score indicates that the classifier is effective both in accurately identifying instances of class c as well as capturing the majority of actual class c samples (Powers, 2020). (recall).

- **Macro-F1:** The macro-averaged F1 score is computed as the arithmetic mean of the per-class F1 scores, assigning equal weight to each class irrespective of its prevalence in the dataset:

$$\text{Macro-F1} = \frac{1}{N} \sum_{c=1}^{N} \text{F1}(c) \times 100$$

where N denotes the total number of classes. This metric evaluates overall model performance by averaging across all classes and penalizes poor classification performance on minority classes, as each class contributes equally to the final score (Powers, 2020). Macro-F1 is widely used in multi-class classification evaluation for its insensitivity to class imbalance (Brodersen et al., 2010).

- **Balanced Accuracy:** Balanced accuracy is defined as the average of the per-class recall values:

$$\text{Balanced Accuracy} = \frac{1}{N}\sum_{c=1}^{N}\text{Recall}(c) \times 100$$

where N represents the total number of classes. Unlike standard accuracy, balanced accuracy accounts for class imbalance by assigning equal weight to each class regardless of sample frequency. This metric provides a more equitable evaluation in imbalanced classification scenarios, mitigating the bias introduced when a dominant class disproportionately influences the overall accuracy score Brodersen et al. (2010); Sensoressa (2025).

- **Adjusted Rand Index(ARI)** quantifies the similarity between the true cluster assignments and those predicted by the model. It operates by considering all possible pairs of samples and counting how many pairs are assigned to the same or different clusters in both the predicted and true clusters, Chacón and Rastrojo (2023); Chekir et al. (2017). The ARI score ranges from -1.0 to 1.0. A score approaching 1 indicates strong concordance between the true labels and the model's predictions, indicating that many sample pairs are clustered similarly in both clusters. A score near 0 suggests the clustering is comparable to random assignment. A negative score suggests that the predicted clusters frequently disagree with the true clusters, potentially performing worse than random assignment. This implies that sample pairs are often grouped differently in the predicted clusters compared to the true clusters.

- **Normalized Mutual Information (NMI)** measures the mutual information between the true clusters of the samples and the clusters assigned by K-means, normalized by the average entropy of the two label sets. It ranges from 0 to 1, where a score close to 1 indicates strong agreement between the true clusters and the K-means clusters. A score of 0 indicates no mutual information between clusters (Kvålseth, 2017).

To visualize the K-means clusters, we employed Principal Component Analysis (PCA), a dimensionality reduction technique that projects data onto orthogonal axes of maximum variance, enabling the representation of high-dimensional data in two or three dimensions. (Abdi and Williams, 2010).

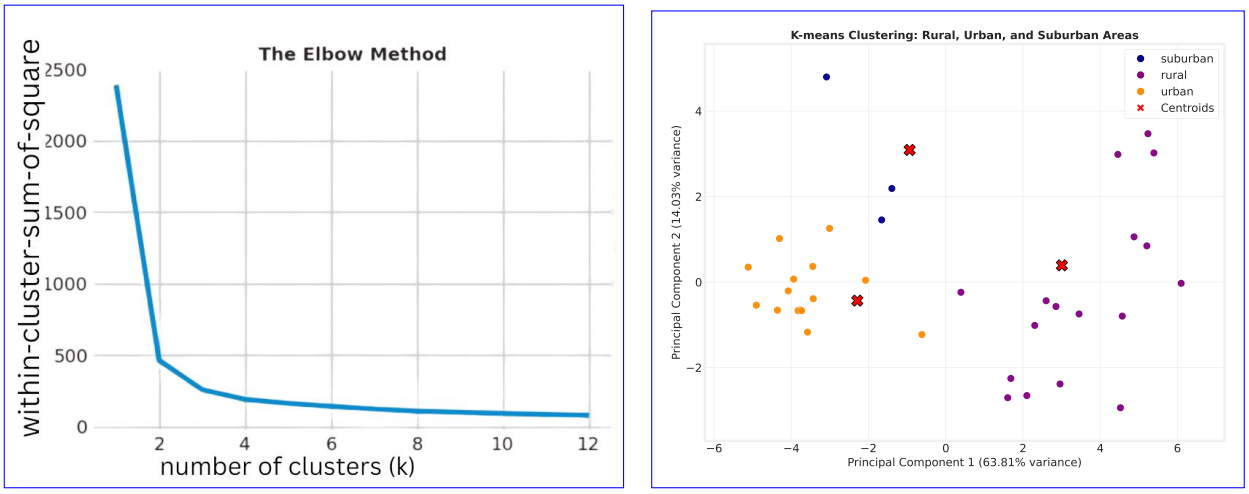
3 Results and discussion

In this section, we present and ~~discuss~~ analyze the results of the ~~different~~ various machine learning models ~~that were used for~~ applied to the TOAR station classification task. ~~In the first subsection , we describe and analyze the results from the unsupervised k-means clustering, and in the second subsection , we discuss~~ The first subsection details the outcomes of the unsupervised K-means clustering. The second subsection presents and analyzes the results from the three supervised methods. Finally, the last subsection discusses the overall performance and comparative insights of the different approaches.

3.1 Results for K-means clustering

Figure 2 shows the elbow plot, a heuristic technique used to determine the optimal number of clusters for K-means clustering. As the gradient of classification accuracy flattens at 3 to 4 clusters, these values for K represent the optimal choices. This result is very encouraging since we want to distinguish 3 different types of stations.

As a first analysis of the K-means clustering, Figure 3(b) shows the K-means predictions evaluated on ~~33~~ 35 manually selected and labeled stations, with clear decision boundary from different categories for sanity check of our method. ~~Table ??~~ Table 2 presents the accuracy of K-means predictions on these manually labeled stations from the test set, comparing them with the characterization report from the TOAR database.



(a) (b)

Figure 3. (a) Elbow method to determine the optimal number of clusters for K-means. (b) Different clusters for selected hand-labeled stations, the red points represent the centroid of different clusters.~~We use Principal Component Analysis (PCA) to project data in 2 dimension for visualization.~~

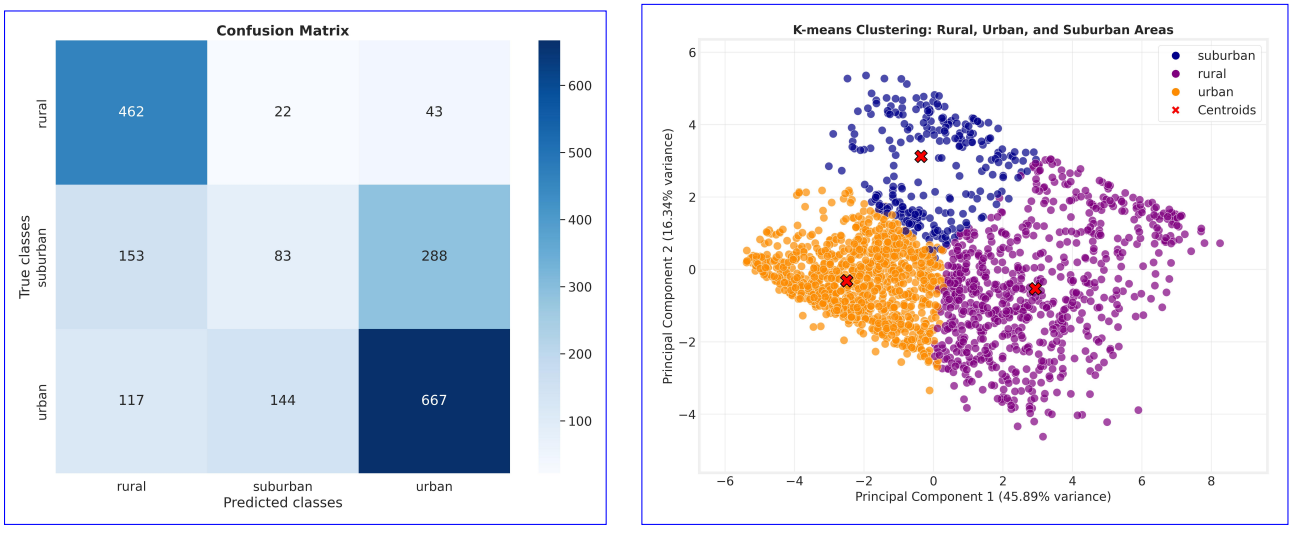
Table 2. Global evaluation of K-means clustering: Accuracy, ~~ARI~~Balanced accuracy, ARI-score, ~~NMI of the K-means method evaluated on 33 hand-labeled stations~~ and NMI-score on ~~1,000~~ unseen ~~stations labelled~~ test set labeled by the TOAR data providers, as well as on the 35 Manually selected with hand-labeled classifications.

Dataset	Accuracy	~~ARI~~ Balanced Acc.	~~NMI~~ ARI-score	NMI-score
35 Manually selected with hand-labels	~~93.94~~77.14%	~~0.77~~ 58.06%	~~0.79~~0.55	0.51
35 Manually selected with TOAR labels	~~87.88~~77.14%	~~0.66~~ 58.82%	~~0.69~~ 0.55	0.546
Unseen test set (1,979 stations)	61.24%	58.46%	0.25	0.22

Table 3. Per-class evaluation of K-means clustering: Precision, recall, and F1-score on 1,979 unseen test set labeled by the TOAR data providers.

Class	~~65.90~~ Precision	Recall	F1-score	Support
Urban	66.83%	~~0.30~~ 71.88%	~~0.31~~69.26%	928
Suburban	33.33%	15.84%	21.47%	524
Rural	63.11%	87.67%	73.39%	527

Figure 4(a) presents the confusion matrix computed from the K-means prediction and the classes from the TOAR database as ground truth, based on 1,~~000~~ 979 unseen test stations (see section 2). This confusion matrix highlights the high accuracy of K-means in classifying urban and rural stations (~~70.53% and 71.03~~87.67% and 71.88% , respectively, see Table ~~??~~3, Recall's column). However, ~~there are also many instances where reported urban and rural stations~~ many instances exist where stations reported as urban or rural are classified as suburban, and vice versa. This misclassification ~~can be attributed~~ is primarily due to the subjective nature of defining suburban areas, which often lie at the interface between rural and urban regions ~~, or which feature~~ or exhibit a mix of ~~urban and suburban or rural and suburban~~ urban-suburban or rural-suburban characteristics. Additionally, Figure 4(b) visualizes the clusters defined by K-means for the 1,~~000~~ 979 unseen test stations. This visual representation clearly illustrates the fuzzy boundaries between the clusters and the noticeable spacing among the three centroids (depicted as red ~~points).~~ cross). ~~Detailed accuracy of K-means prediction evaluated on 1,000 unseen test~~



(a) (b)

Figure 4. (a) Confusion matrix for K-means clustering, evaluated on 1,~~000~~ 979 unseen labelled by the TOAR data providers. (b) Cluster visualization of the previously unseen test data.~~To better visualize the cluster separation, the data is projected into two dimensions using Principal Component Analysis (PCA).~~

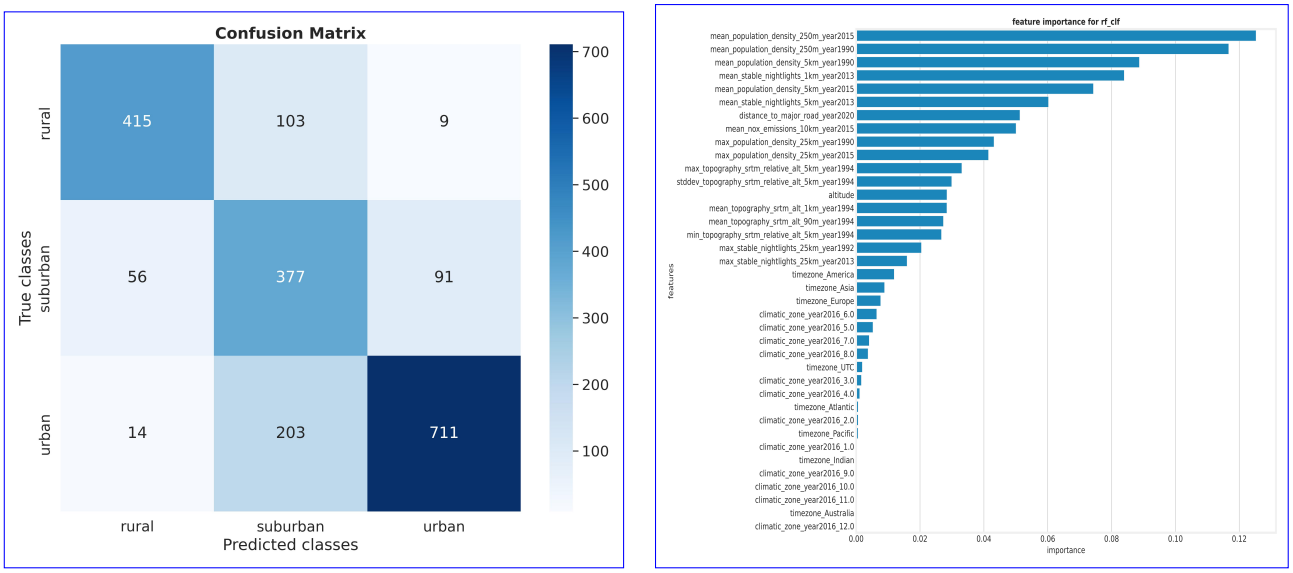
~~data points, for different classes. K-means Global Accuracy 65.90% Accuracy for urban 70.53% Accuracy for rural 71.03% Accuracy for suburban 26.36%~~

3.2 Results for supervised classifiers

Here, we evaluate the results from the three supervised machine learning classifiers, Random Forest, LightGBM, and CatBoost. Furthermore, the results from the three models were subjected to a robust voting classifier to maximize the classification accuracy. We observed that the results of all algorithms are quite similar. The models demonstrated exceptional accuracy (> 83%), high precision (> 85%), and high F1-score (> 81 %) in predicting urban and rural areas. However, all models struggled with the suburban class, yielding accuracies slightly above ~~50~~60% (Table ~~??~~5). As discussed above, the main reason for this lower accuracy can be attributed to the inherent subjective nature of defining this category. To address this issue, we implemented a strategy that capitalizes on model uncertainty. By adjusting the prediction probability threshold, as detailed in Section 2, we significantly enhanced the accuracy of suburban area classifications as shown in Table ~~??.~~ 7. Figure 5(a) presents the confusion matrix for the Random Forest classifier and Figure 5(b) visualizes the feature importance. The ~~accuracy~~ global evaluation, i.e accuracy, balance accuracy, macro-f1 score, and weighted-f1 score of different classifiers ~~is~~ are reported in Table ~~??~~ 4 and Table ~~??~~6, which show the results before and after the probability threshold adjustment, respectively. While the overall accuracy remains relatively similar before and after the adjustment, the probability threshold ~~modification~~ adjustment significantly enhances the prediction accuracy for suburban stations, increasing it from a range of ~~51.74% –54.86% to 58.24~~60.85% - 65.72% to 66.22%–~~62.07%~~71.95%, and also slightly increase the F1-score for the suburban, these can be seen in details in 5 and 7 presenting the per-class evaluation before and after applying the probability threshold adjustment. We also note a slight drop in accuracy when classifying urban and rural stations. However, classification performance remains high across all classifiers, with accuracy values exceeding 80%.

Additionally, we conducted tests on our machine learning models using the manually labeled stations, similar to those used for K-means evaluation. In this test, we found that the classifiers predict the label report on TOAR by data provider for ~~33~~ 35 manually selected stations with 100% accuracy and achieve an 87.88% accuracy for the manual classified stations.

The implementation of the ~~voting procedure and~~ adjusted probability threshold yielded notable improvements in our classification model. While the enhancements for urban and rural station predictions were modest, the impact on suburban area classification was substantial. This is particularly significant given the inherent challenges in accurately identifying suburban zones. When compared to the unsupervised K-means clustering method, our supervised approaches demonstrated superior performance across all categories. The contrast was especially pronounced in the classification of suburban areas, where K-means exhibited a markedly low accuracy of just ~~26.36%~~15.84% , low precision of 33.33% and low F1-score of 21.47%, see Table 3. In contrast, our supervised methods achieved significantly higher accuracy rates, higher precision, and higher F1-score underscoring their effectiveness in navigating the complexities of urban-suburban-rural distinctions.



(a) (b)

Figure 5. (a) confusion matrix for random forest classifier, evaluated on 1,~~000~~ 979 test data points. (b) the feature importance for random, measuring the contribution of each variable in the classification process

Table 4. Global performance evaluation of models. Reported values are Accuracy, Balanced accuracy, Macro-F1 score, and Weighted-F1 score of random forest, LGBM, CatBoost, and voting classifiers before probability threshold adjustment, evaluated on 1,~~000~~ 979 test stations.

Model	~~Random forest~~ Accuracy	~~CatBoost~~ Balanced Acc.	~~LGBM~~ Macro-F1	~~Voting~~ Weighted-F1
Random Forest	~~79.01%~~ 76.25%	~~77.50~~75.39%	~~76.20~~75.07%	~~78.70~~76.53%
CatBoost	~~91.03%~~ 76.25%	~~88.28~~75.70%	~~88.05~~75.27%	~~90.80~~76.66%
LightGBM	~~86.64~~76.81%	~~84.12%~~ 75.54%	~~83.03~~75.42%	~~84.84%~~ 76.93%
Voting	~~53.47~~76.40%	~~54.86%~~ 75.56%	~~51.74~~75.25%	~~54.51~~76.71%

3.3 Discussion

The supervised machine learning approach demonstrates remarkable performance, achieving prediction accuracies > 84% for urban and rural stations when applied to previously unseen test data. This can already be used to accurately predict "urban" and "rural" labels. While the model's performance in identifying suburban areas initially showed slightly lower accuracy, this challenge was effectively addressed through the adjusted probability threshold.

Despite the promising results, the classification results are far from perfect. This can be partially attributed to inherent inaccuracies within the dataset itself. To investigate this issue, we conducted a detailed review and manual inspection of the ~~25 worst~~ 30 randomly selected misclassifications. For these stations, which are listed in Table 8, we visually inspected the areas around the stations on Google Maps Mehta et al. (2019), using a zoom level of 11 or greater. While ~~8~~ 6 of these

Table 5. Per-class performance evaluation of models. Reported values are Precision, Recall, and F1 score, of random forest, LGBM, CatBoost, and voting classifiers before probability threshold adjustment, evaluated on 1,979 test stations.

(a) Random Forest

Class	Precision	Recall	F1-score	Support
Urban	85.02%	79.53%	82.18%	928
Suburban	57.74%	63.36%	60.42%	524
Rural	81.90%	83.30%	82.60%	527

(b) CatBoost

Class	Precision	Recall	F1-score	Support
Urban	86.04%	78.34%	82.01%	928
Suburban	57.00%	65.27%	60.85%	524
Rural	82.40%	83.49%	82.94%	527

(c) LightGBM

Class	Precision	Recall	F1-score	Support
Urban	83.85%	81.68%	82.75%	928
Suburban	59.16%	61.64%	60.37%	524
Rural	82.99%	83.30%	83.14%	527

(d) Voting

Class	Precision	Recall	F1-score	Support
Urban	85.24%	79.63%	82.34%	928
Suburban	57.59%	63.74%	60.51%	524
Rural	82.52%	83.30%	82.91%	527

Table 6. Global performance evaluation of models. Reported values are Accuracy, Balanced accuracy, Macro-F1 score, and Weighted-F1 score of random forest, LGBM, CatBoost, and voting classifiers after applying probability threshold adjustment, evaluated on 1,~~000~~ 979 test stations.

Model	~~Random forest~~ Accuracy	~~CatBoost~~ Balanced Acc.	~~LGBM~~ Macro-F1	~~Voting~~ Weighted-F1
Random Forest	~~76.90%~~ 75.95%	~~77.10~~75.77%	~~77.50~~75.42%	~~78.20~~76.73%
CatBoost	~~82.93%~~ 76.30%	~~83.15~~75.85%	~~85.56~~75.40%	~~84.90~~76.74%
LightGBM	~~80.85~~76.96%	~~81.21%~~ 76.20%	~~82.27~~76.04%	~~82.27%~~ 77.36%
Voting	~~62.07~~76.50%	~~62.07%~~ 76.10%	~~58.24~~75.79%	~~62.07~~77.07%

cases revealed wrong classifications by our best ML model, the model's classification is actually more accurate than the label that was reported by the data providers in ~~15~~ 16 cases. In the remaining ~~two~~ 8 cases, neither the reported nor the ML model derived label was correct. In one of these cases, both the reported and ML based label was urban, while the station site is apparently located in a rural area. In the other case, visual inspection would place the station in the suburban class, while the reported category is urban and the ML model classifies the station as rural. However, for stations misclassified by our machine learning model, we observed ambiguous features that could misguide our model. For instance, surrounding neighborhoods of areas reported as urban but classified as rural by our model exhibited rural characteristics, such as lower population density (which is one of the important features used in the training dataset, see Figure 5(b)) and more green space. This result, while initially counter-intuitive, can be explained by the strong robustness of tree-based ensemble methods to noisy datasets. Algorithms such as Random Forest, CatBoost, and LightGBM are specifically designed to mitigate the effects of

Table 7. Per-class performance evaluation of models. Reported values are Precision, Recall, and F1 score, of random forest, LGBM, CatBoost, and voting classifiers after applying probability threshold adjustment, evaluated on 1,979 test stations.

(a) Random Forest

Class	Precision	Recall	F1-score	Support
Urban	87.67%	76.62%	81.77%	928
Suburban	55.20%	71.95%	62.47%	524
Rural	85.57%	78.75%	82.02%	527

(b) CatBoost

Class	Precision	Recall	F1-score	Support
Urban	86.29%	78.02%	81.95%	928
Suburban	56.98%	66.22%	61.25%	524
Rural	82.67%	83.30%	82.99%	527

(c) LightGBM

Class	Precision	Recall	F1-score	Support
Urban	85.27%	79.85%	82.47%	928
Suburban	57.90%	66.41%	61.87%	524
Rural	85.27%	82.35%	83.78%	527

(d) Voting

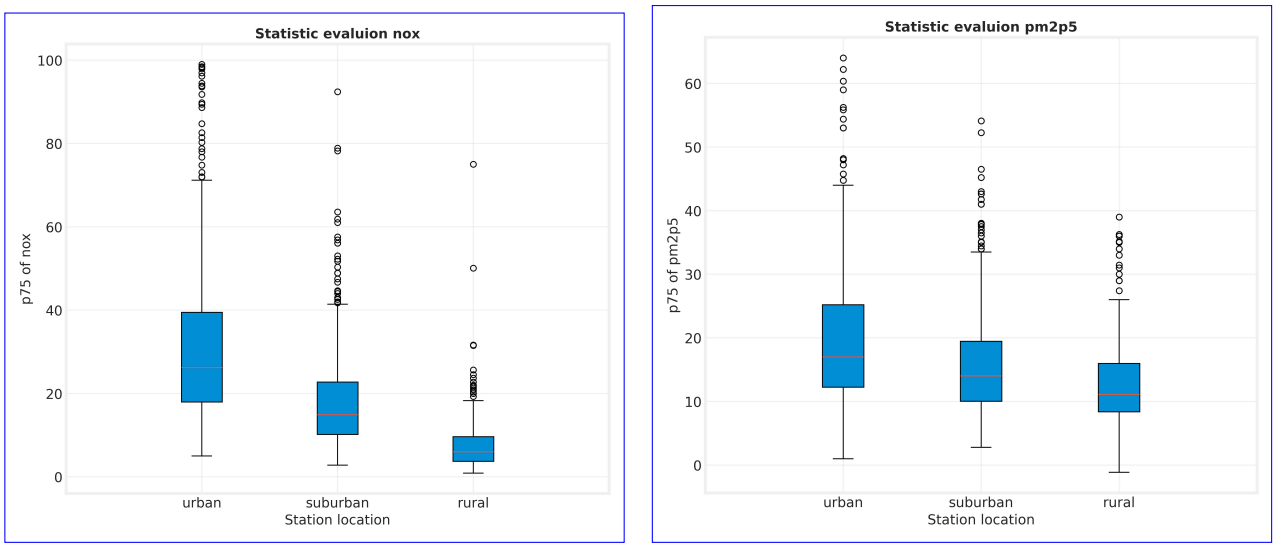
Class	Precision	Recall	F1-score	Support
Urban	86.55%	78.34%	82.24%	928
Suburban	56.80%	68.51%	62.11%	524
Rural	85.01%	81.78%	83.37%	527

noise and outliers through randomization and averaging techniques (Biau and Scornet, 2016). However, these 30 data points are insufficient to draw definitive conclusions.

To further lend confidence to our results, we evaluated the 75-percentile statistics of the primary air pollutant concentrations NO_x and $PM_{2.5}$ from the TOAR database. We chose the 75-th percentile, because urban areas typically exhibit fresh pollution with many high concentration events. This percentile captures such characteristics while being more robust than either the maximum value or a higher percentile. While data on these species is incomplete, there are sufficient measurements from several regions to yield a meaningful statistic. Figure 6 shows box and whisker plots of the 75-percentiles of NO_x and $PM_2.5$ concentrations aggregated for the year 2015 for the three classes. As expected, urban stations typically show substantially higher concentration levels compared to suburban sites, while rural stations show the lowest concentrations.

Table 8. Closer analysis of some misclassified station locations

heightlatitude	longitude	Station code	type of area TOAR	type of area ML	True type of area (Google Maps)	Winner
59.12331433.859662	11.391136-118.200707	rural 06-037-4008	suburban	urban	urban	ML
34.98567044.470336	-84.375193-71.180077	urban 33-007-0015	rural urban	suburban	rural	None
34.71010053.341875	128.587500-6.214075	IE004AP	urban	rural	rural	ML
37.36050042.876469	128.125600-73.071215	50-003-0001	urban	rural	rural	ML
41.28553344.307000	-93.583983-86.242649	26-101-0922	urban	rural	rural	ML
44.78561652.132417	-69.885058-0.300306	GB0954A	urban	rural suburban	rural suburban	ML
35.10263844.835700	-85.162194-108.386000	urban 56-003-0003	rural	urban suburban	rural	TOAR
45.54222235.273460	9.516389-89.961217	urban 47-157-0046	suburban	suburban rural	ML urban	None
39.72045140.734449	-94.872693-75.312389	42-095-1000	urban	suburban	urban suburban	TOAR ML
48.69111345.394410	21.286388-93.885254	urban 27-141-0012	urban	rural	suburban	None
35.25950236.923100	-120.644720-2.463220	04024001	urban	suburban	suburban	ML
37.04911544.390480	-122.0199628.201500	urban IT1233A	rural	suburban	suburban	ML
41.89581248.762780	-87.607683-122.440280	urban 53-073-0015	suburban	urban	TOAR rural	None
41.19358750.575364	1.2367038.492018	rural DEHE095	suburban	rural urban	TOAR urban	ML
34.13171430.958515	-109.282309-88.028332	rural 01-097-0028	suburban	rural	TOAR rural	ML
33.06115031.813370	-112.052040-106.464520	rural 48-141-0693	urban	suburban	suburan suburban	ML
49.07655037.049260	8.406660-86.214870	rural 21-227-0009	suburban	suburban rural	rural	ML
36.18621043.629605	-5.380810-72.309499	33-009-0010	urban	rural	rural	ML
40.262540	-89.230923	17-107-0001	urban	suburban	rural	None
33.089772	-87.459733	01-125-0010	suburban	rural	rural	ML
45.64731044.003853	13.854970-92.414896	27-109-0016	rural	suburban	rural	TOAR
39.652473	-104.925926	08-031-0823	urban	suburban	suburban	ML
43.04526947.689728	-70.71395822.458500	RO0183A	suburban	rural urban	suburban	TOAR
44.01305643.144070	12.420000-2.963370	01036004	suburban	rural urban	rural suburban	ML TOAR
43.18660032.891056	-8.471600-111.570503	04-021-3011	suburban	rural	suburban	TOAR
36.58702740.246528	-89.546742-77.186750	42-041-0101	rural	suburban	rural	TOAR
60.294268	5.324619	NO0121A	suburban	urban	rural	ML None
40.07778057.039915	-6.147220-135.272042	02-220-0007	suburban	rural	rural	ML
35.49871162.486778	-83.31024217.324437	SE0012A	urban	suburban	rural	None
38.583226	-77.121900	11-001-0025	urban	rural	suburban	TOAR None



(a) (b)

Figure 6. Evaluation of the supervised classification with independent data: (a) Box and whisker plot of the 75 percentile of the NOx. (b) Box and whisker plot 75 percentile of the PM2.5

4 Conclusion

We investigated the use of machine learning models to objectively characterize station locations for global air quality data analysis. Specifically, we wanted to improve the station classification in the TOAR-I database that was described by ~~Schultz et al. (2017b)~~ Schultz et al. (2017a) and base it on an objective algorithm. As a side-effect we can now explicitly label stations as suburban that were falling between the urban and rural categories in the TOAR-I classification scheme. Our proposed models demonstrate excellent prediction capabilities for urban and rural areas. With the help of an adjusted probability threshold technique, we also obtain meaningful results on the suburban category, inasmuch this category can be described objectively at all. We noticed a limitation for evaluating the accuracy of our method due to obvious misclassification of stations in official databases. As discussed in ~~(Schultz et al., 2017b)~~(Schultz et al., 2017a), such errors can be introduced for various reasons. In some cases, we speculate that these misclassifications actually reflect true landcover changes (e.g., urban development), which have not been updated in the station metadata at the data providers' sites. Manual inspection of ~~worst~~ random test samples with disagreements revealed that the ML classifier was ~~correct in the majority of cases, where there was disagreement between~~ more often correct than the reported station type~~and our ML-derived one~~.

There is still room for improvement of the methods described here. On the one hand, a larger manual labelling effort using high-resolution EO data, could reduce the number of wrong target labels and reduce the noise in the training data. On the other hand, it may also be possible to employ modern ML methods with spatial context (e.g., (Szwarcman et al., 2024)) on such high-resolution EO data directly as a specialized land cover classification task. Nevertheless, the new TOAR station classifiers

developed in this study provide a clear improvement over the previous method and can be employed in the TOAR-II ozone data analyses that will be reported in the forthcoming assessment papers.

Code data availability

All code accompanying this paper is available in our GitLab repository link(Mache et al., 2025) at the following link: https://doi.org/10.5281/zenodo.15411286. This repository also contains a copy of the data that is used in this sudy as csv files. The data can also be obtained directly from the TOAR-II database.

Author contribution

RKM, SS, and MGS designed the study based on previous work by MGS and SS. RKM, AP, and ML developed the methodology, RKM implemented the methods and evaluated the results. RKM wrote the major part of the text with contributions from all authors. MGS conducted a final review prior to submission.

Competing interests

The authors declare that no competing interests exist.

Acknowledgements

The authors are grateful to the EU for funding the IntelliAQ project under grant ERC-AdG-787576. This allowed the buildup of the TOAR-II database. We also greatly appreciate the effort from hundreds of people around the world who established and operate air quality stations, process the data and make the data available to the TOAR initiative. Sebastian Hickman deserves gratitude for his initial analysis of NOx and PM2.5 data in the TOAR-II database and helpful discussions.

References

Abdi, H. and Williams, L. J.: Principal component analysis, Wiley interdisciplinary reviews: computational statistics, 2, 433–459, 2010.

Airgood-Obrycki, W. and Rieger, S.: Defining suburbs: How definitions shape the suburban landscape, Joint Center for Housing Studies of Harvard University, 2019.

Bahmani, B., Moseley, B., Vattani, A., Kumar, R., and Vassilvitskii, S.: Scalable k-means++, arXiv preprint arXiv:1203.6402, 2012.

Biau, G. and Scornet, E.: A random forest guided tour, Test, 25, 197–227, 2016.

Breiman, L.: Random forests, Machine learning, 45, 5–32, 2001.

Brodersen, K. H., Ong, C. S., Stephan, K. E., and Buhmann, J. M.: The balanced accuracy and its posterior distribution, in: 2010 20th international conference on pattern recognition, pp. 3121–3124, IEEE, 2010.

Chacón, J. E. and Rastrojo, A. I.: Minimum adjusted Rand index for two clusterings of a given size, Advances in Data Analysis and Classification, 17, 125–133, 2023.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P.: SMOTE: synthetic minority over-sampling technique, Journal of artificial intelligence research, 16, 321–357, 2002.

Chekir, A., Hassas, S., Descoteaux, M., Côté, M., Garyfallidis, E., and Oulebsir-Boumghar, F.: 3D-SSF: A bio-inspired approach for dynamic multi-subject clustering of white matter tracts, Computers in Biology and Medicine, 83, 10–21, 2017.

Cooper, O. R., Parrish, D., Ziemke, J., Balashov, N., Cupeiro, M., Galbally, I., Gilge, S., Horowitz, L., Jensen, N., Lamarque, J.-F., et al.: Global distribution and trends of tropospheric ozone: An observation-based review, Elementa, 2, 000 029, 2014.

Fleming, E., Payne, J., Sweet, W., Craghan, M., Haines, J., Hart, J., et al.: Chapter 8: Coastal Effects. Impacts, risks, and adaptation in the United States: The Fourth National Climate Assessment, Volume II [Internet]. US Global Change Research Program; 2018.

Florczyk, A. J., Corbane, C., Ehrlich, D., Freire, S., Kemper, T., Maffenini, L., Melchiorri, M., Pesaresi, M., Politis, P., Schiavina, M., et al.: GHSL data package 2019, Luxembourg, eur, 29788, 290 498, 2019.

Gaudel, A., Cooper, O. R., Ancellet, G., Barret, B., Boynard, A., Burrows, J. P., Clerbaux, C., Coheur, P.-F., Cuesta, J., Cuevas, E., et al.: Tropospheric Ozone Assessment Report: Present-day distribution and trends of tropospheric ozone relevant to climate and global atmospheric chemistry model evaluation, Elem Sci Anth, 6, 39, 2018.

Granier, C., Darras, S., van Der Gon, H. D., Jana, D., Elguindi, N., Bo, G., Michael, G., Marc, G., Jalkanen, J.-P., Kuenen, J., et al.: The Copernicus atmosphere monitoring service global and regional emissions (April 2019 version), Ph.D. thesis, Copernicus Atmosphere Monitoring Service, 2019.

Griffiths, P. T., Murray, L. T., Zeng, G., Shin, Y. M., Abraham, N. L., Archibald, A. T., Deushi, M., Emmons, L. K., Galbally, I. E., Hassler, B., et al.: Tropospheric ozone in CMIP6 simulations, Atmospheric Chemistry and Physics, 21, 4187–4218, 2021.

Harris, I. and Jones, P.: University of East Anglia Climatic Research Unit 2017 CRU TS4. 00: Climatic Research Unit (CRU) Time-Series (TS) version 4.00 of high-resolution gridded data of month-by-month variation in climate (Jan. 1901-Dec. 2015), Chilton, Oxfordshire: Centre for Environmental Data Analysis)(http://doi. org/10.5285/edf8febfdaad48abb2cbaf7d7e846a86), 2017.

Hesse, M. and Siedentop, S.: Suburbanisation and suburbanisms–Making sense of continental European developments, Raumforschung und Raumordnung| Spatial Research and Planning, 76, 97–108, 2018.

Jarvis, A., Reuter, H. I., Nelson, A., Guevara, E., et al.: Hole-filled SRTM for the globe Version 4, available from the CGIAR-CSI SRTM 90m Database (http://srtm. csi. cgiar. org), 15, 5, 2008.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y.: Lightgbm: A highly efficient gradient boosting decision tree, Advances in neural information processing systems, 30, 2017.

Ketchen, D. J. and Shook, C. L.: The application of cluster analysis in strategic management research: an analysis and critique, Strategic management journal, 17, 441–458, 1996.

Kroehl, H.: National Geophysical and Solar-Terrestrial Data Center, EDIS, NOAA, Boulder, Colorado 80303, Copyright 1982 American Geophysical Union. Figures, tables and short excerpts may be reprinted in scientific books and journals if the source is, p. 98, 1982.

Kvålseth, T. O.: On normalized mutual information: measure derivations and properties, Entropy, 19, 631, 2017.

Mache, R. K., Schröder, S., Langguth, M., Patnala, A., and Schultz, M. G.: TOAR-classifier v2: A data-driven classification tool for global air quality stations, GitLab repository, https://gitlab.jsc.fz-juelich.de/esde/toar-public/ml_toar_station_classification/-/tree/develop?ref_type=heads, correspondence: Ramiyou Karim Mache (k.mache@fz-juelich.de), 2025.

Madronich, S., Sulzberger, B., Longstreth, J., Schikowski, T., Andersen, M. S., Solomon, K., and Wilson, S.: Changes in tropospheric air quality related to the protection of stratospheric ozone in a changing climate, Photochemical & Photobiological Sciences, 22, 1129–1176, 2023.

Mehta, H., Kanani, P., and Lande, P.: Google maps, International Journal of Computer Applications, 178, 41–46, 2019.

Mills, M. M., Brown, Z. W., Laney, S. R., Ortega-Retuerta, E., Lowry, K. E., van Dijken, G. L., and Arrigo, K. R.: Nitrogen limitation of the summer phytoplankton and heterotrophic prokaryote communities in the Chukchi Sea, Frontiers in Marine Science, 5, 362, 2018.

Monks, P. S., Archibald, A., Colette, A., Cooper, O., Coyle, M., Derwent, R., Fowler, D., Granier, C., Law, K. S., Mills, G., et al.: Tropospheric ozone and its precursors from the urban to the global scale from air quality to short-lived climate forcer, Atmospheric chemistry and physics, 15, 8889–8973, 2015.

Orru, H., Andersson, C., Ebi, K. L., Langner, J., Åström, C., and Forsberg, B.: Impact of climate change on ozone-related mortality and morbidity in Europe, European Respiratory Journal, 41, 285–294, 2013.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E.: Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, 12, 2825–2830, 2011a.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in Python, the Journal of machine Learning research, 12, 2825–2830, 2011b.

Pelleg, D., Moore, A., et al.: X-means: Extending K-means with E cient Estimation of the Number of Clusters, in: ICML'00, pp. 727–734, Citeseer, 2000.

Post, E. S., Grambsch, A., Weaver, C., Morefield, P., Huang, J., Leung, L.-Y., Nolte, C. G., Adams, P., Liang, X.-Z., Zhu, J.-H., et al.: Variation in estimated ozone-related health impacts of climate change due to modeling choices and assumptions, Environmental Health Perspectives, 120, 1559–1564, 2012.

Powers, D. M.: Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation, arXiv preprint arXiv:2010.16061, 2020.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A.: CatBoost: unbiased boosting with categorical features, Advances in neural information processing systems, 31, 2018.

Rubinsteyn, A. and Feldman, S.: Fancyimpute: An Imputation Library for Python. 2016, URL https://github. com/iskandr/fancyimpute, 2016.

Schultz, M. G., Akimoto, H., Bottenheim, J., Buchmann, B., Galbally, I. E., Gilge, S., Helmig, D., Koide, H., Lewis, A. C., Novelli, P. C., et al.: The Global Atmosphere Watch reactive gases measurement network, Elementa, 3, 000 067, 2015.

Schultz, M. G., Schröder, S., Lyapina, O., Cooper, O. R., Galbally, I., Petropavlovskikh, I., Von Schneidemesser, E., Tanimoto, H., Elshorbany, Y., Naja, M., et al.: Tropospheric Ozone Assessment Report: Database and metrics data of global surface ozone observations, Elem Sci Anth, 5, 58, 2017a.

Schultz, M. G., Schröder, S., Lyapina, O., Cooper, O. R., Galbally, I., Petropavlovskikh, I., et al.: Tropospheric Ozone Assessment Report: Database and metrics data of global surface ozone observations, Elementa: Science of the Anthropocene, 5, 58, https://doi.org/10.1525/elementa.244, 2017b.

Sensoressa, N. A.: Balanced Accuracy: When Should You Use It?, Neptune.ai, 2025.

Sinaga, K. P. and Yang, M.-S.: Unsupervised K-means clustering algorithm, IEEE access, 8, 80 716–80 727, 2020.

Szwarcman, D., Roy, S., Fraccaro, P., Gíslason, Þ. E., Blumenstiel, B., Ghosal, R., de Oliveira, P. H., Almeida, J. L. d. S., Sedona, R., Kang, Y., et al.: Prithvi-EO-2.0: A Versatile Multi-Temporal Foundation Model for Earth Observation Applications, arXiv preprint arXiv:2412.02732, 2024.

Tapia, O., Escudero, M., Lozano, Á., Anzano, J., and Mantilla, E.: New classification scheme for ozone monitoring stations based on frequency distribution of hourly data, Science of The Total Environment, 544, 1–9, 2016.

Teakles, A. D., So, R., Ainslie, B., Nissen, R., Schiller, C., Vingarzan, R., McKendry, I., Macdonald, A. M., Jaffe, D. A., Bertram, A. K., et al.: Impacts of the July 2012 Siberian fire plume on air quality in the Pacific Northwest, Atmospheric chemistry and physics, 17, 2593–2611, 2017.

Van Rossum, G. et al.: Python programming language., in: USENIX annual technical conference, vol. 41, pp. 1–36, Santa Clara, CA, 2007.

Zhang, L., Sun, Y., Li, C., and Li, B.: Promoting Sustainable Development in Urban–Rural Areas: A New Approach for Evaluating the Policies of Characteristic Towns in China, Buildings, 14, 1085, 2024.

Zhou, M., Li, Y., and Zhang, F.: Spatiotemporal variation in ground level ozone and its driving factors: a comparative study of coastal and inland cities in eastern China, International Journal of Environmental Research and Public Health, 19, 9687, 2022.