

Leveraging VAT for Spatio-Temporal Machine Learning on Environmental Data

Dominik Brandenstein, Henri Dümpelmann, Bernhard Seeger

Data Preparation in VAT

VAT is one of the fundamental services offered in NFDI4Biodiversity. It is a web-based GIS application, powered by Geo Engine, that supports users to visualise, analyse and transform geospatial vector and raster data without requiring deep GIS knowledge. In this use case, VAT generates spatio-temporal training data for a species distribution modelling task. It easily handles the spatio-temporal alignment of different raster layers, while offering multiple curated data products as a solid basis for modelling. Specifically, in this use case, GBIF occurrences of *Arnica montana* are selected and combined with land use classification from the Ökosystematlas, weather data from CHELSA and topographic information. Weather data are also temporally aggregated annually using VAT.

```
import geoengine as ge
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import GridSearchCV, train_test_split
from datetime import datetime

#Initialize VAT
ge.initialize("https://vat.gfbio.org/api")

#Get the training data from VAT
workflow = ge.workflow_by_id("7582cfc3b-3d36-5b86-bb72-e81cef584fae")
time = datetime.strptime('2002-01-01T12:00:00.000Z', "%Y-%m-%dT%H:%M:%S.%f%Z")

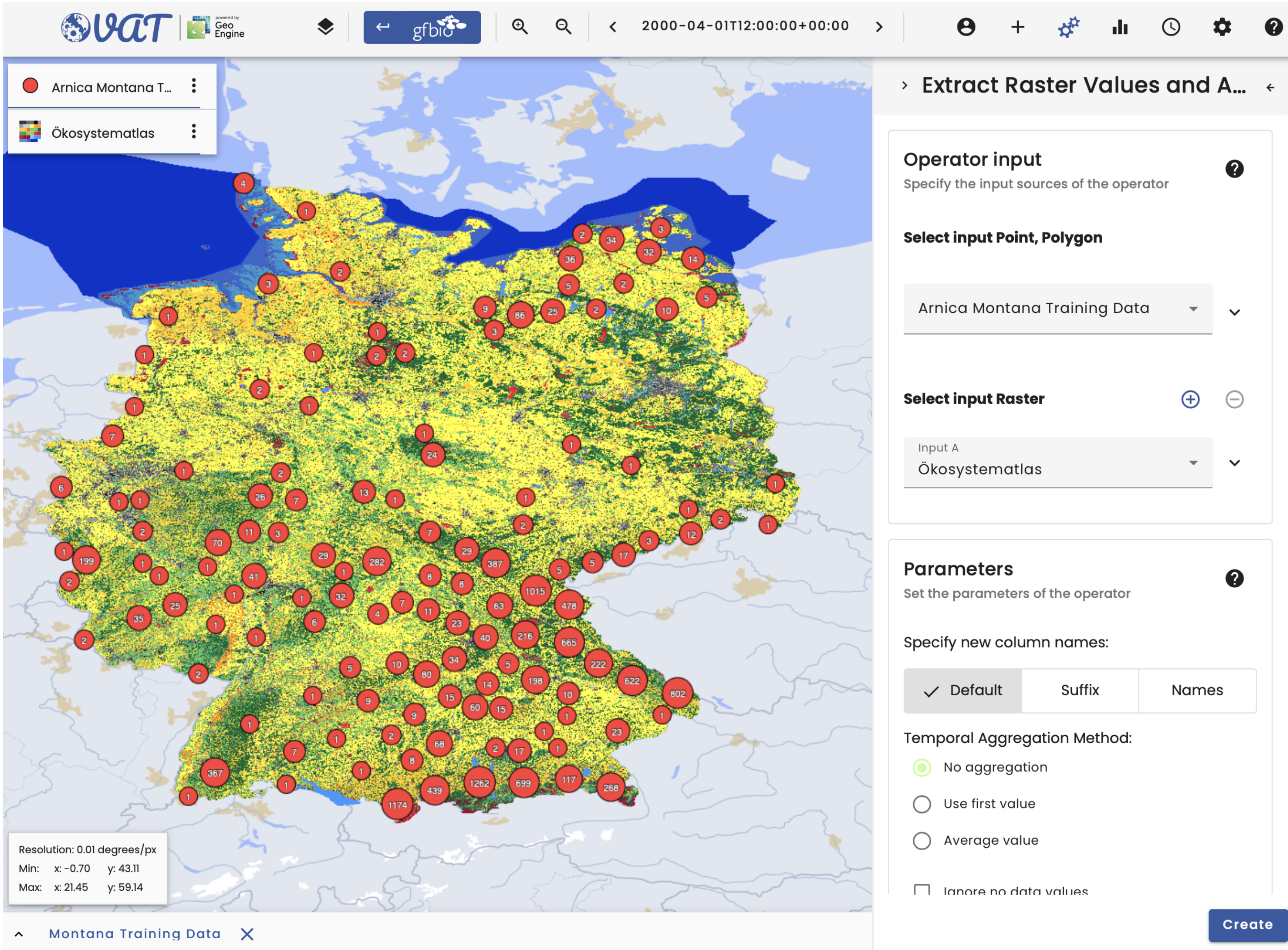
data = workflow.get_dataframe(
    ge.QueryRectangle(
        ge.BoundingBox2D(5.852490, 47.271121, 15.022059, 55.065334),
        ge.TimeInterval(time, time),
        resolution=ge.SpatialResolution(0.01, 0.01),
        srs="EPSG:4326"
    )
)

#Create species frequency as target
training_data = data.round(3)
training_data = training_data.groupby(
    ['Mean Air Temperature',
     'Mean Climate Moisture Index',
     'Precipitation',
     'SRTM',
     'Ökosystematlas']).size().reset_index(name='counts')

#Train test split
X = training_data[['Mean Air Temperature',
                    'Mean Climate Moisture Index',
                    'Precipitation',
                    'SRTM',
                    'Ökosystematlas']]
y = training_data['counts']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

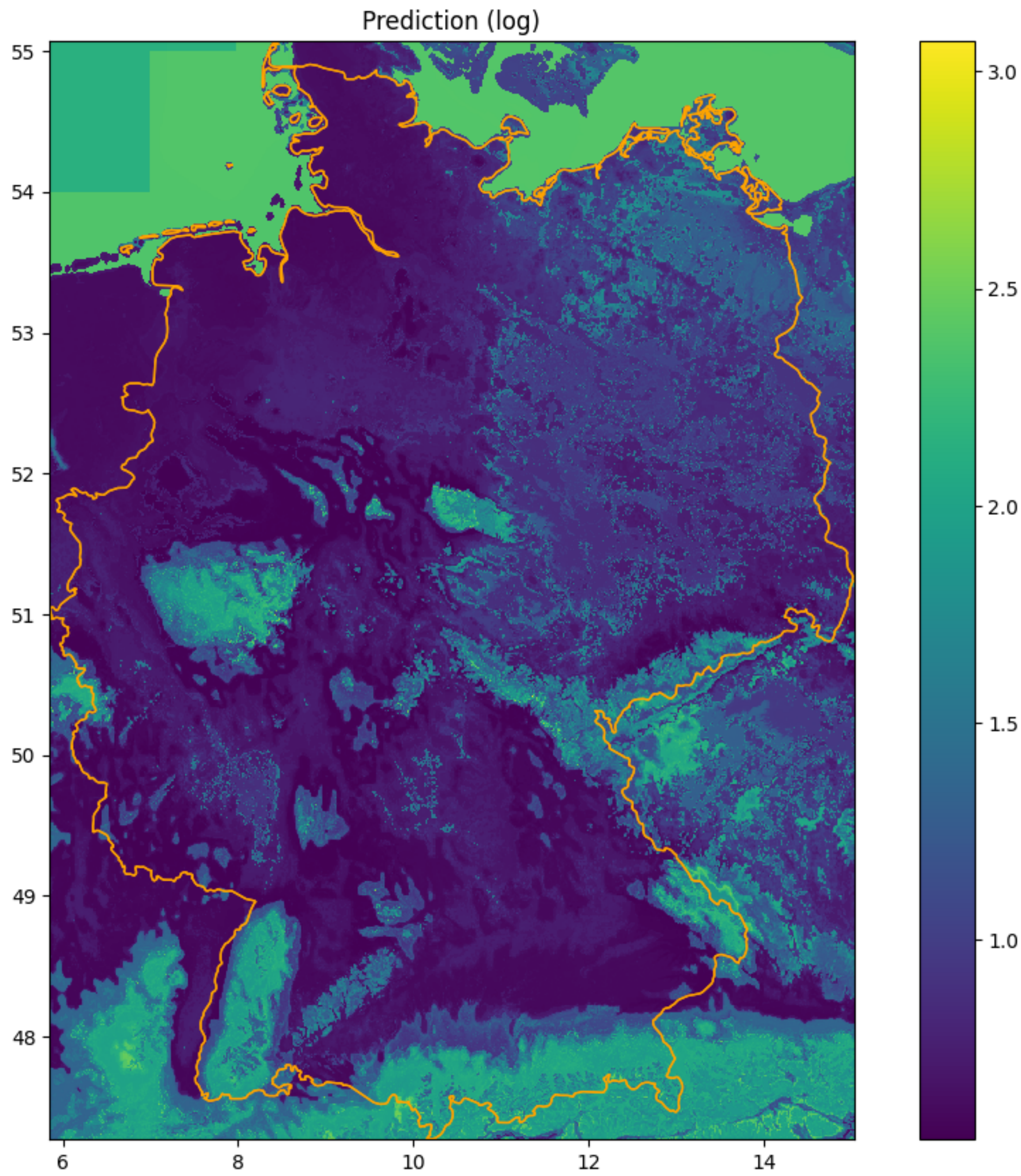
#Train GridSearchCV RandomForestRegressor model
param_grid = {
    'n_estimators': [200, 400, 600, 800, 1000],
    'max_depth': [5, 10, 15, 20],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 4]
}

rf = RandomForestRegressor()
grid_search = GridSearchCV(rf,
                            param_grid,
                            cv=5,
                            scoring='neg_mean_squared_error',
                            verbose=2)
grid_search.fit(X_train, y_train)
```



Machine Learning in Jupyter Notebooks

In addition to the visual interface of a web GIS, VAT provides a Python package for interacting with VAT programmatically, e.g. in Jupyter Notebooks. Users can seamlessly exchange their workflows between the web GIS and the Python package using their unique identifiers. For example, training data generated in the web GIS is loaded into a Geopandas DataFrame using its workflow identifier. The frequency of species occurrences is then grouped by the environmental predictors, which is then used as the target variable. In this way, training data is easily assembled and prepared to train a RandomForest regressor model for species distribution modelling. The advantage of VAT workflows is their non-static nature, i.e. they adapt to changes in the underlying data sources and workflow parameters. This makes it an ideal tool for data scientists who appreciate the flexibility to iteratively develop workflows step by step.

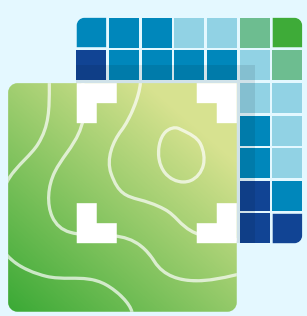


Contribution

VAT empowers users to generate spatio-temporal training data for machine learning applications. In addition, VAT supports several other spatio-temporal use cases. Have a look at the VAT documentation (<https://docs.vat.gfbio.org>), which is frequently updated with the latest features and use cases. If you see potential for your own projects, feel free to email us (seeger@mathematik.uni-marburg.de) or visit us at the poster session. We are always interested in your feedback and ideas on how we can improve the service, not just for you, but for all members of the NFDI community.

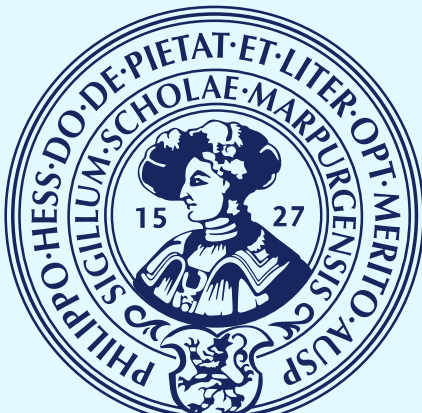
References

- Authman, C., Beilschmidt, C., Dröhner, J., Mattig, M., & Seeger, B. (2015). VAT: a system for visualizing, analyzing and transforming data in science. Datenbank-Spektrum, 15, 175-184.
- Beilschmidt, C., Dröhner, J., Mattig, M., & Seeger, B. (2023). Geo Engine: Workflow-driven Geospatial Portals for Data Science. Datenbank-Spektrum, 1-9.



powered by
Geo Engine

Philipps



Universität
Marburg



NFDI 4
BIODIVERSITY