Mapping building coordinates to building polygons using a Random Forest classifier

The Buildings and Dwellings Register (=GWR), led by Statistics Austria, contains vast amounts of information on buildings, including building coordinates (=BC). BCs are required to be within the respective building's polygon (=BP). However, these BPs are not part of the GWR. Roughly 600,000 (20%) of BCs are not within their BPs. The project aims to establish a match between BCs and BPs using a Random Forest (=RF) classifier.

In the initial phase, only residential BCs with one coordinate per property were retained, resulting in a reduction of misplaced BCs from 600,000 to 192,000. Two additional datasets - the BP and property polygon (=PP) layers - were required, and were spatially joined in a 1:N relation. Consequently, the BC was linked to each BP within a PP.

Several features were defined to describe the relationship between the BC and BPs. A binary target was assigned to roughly 7000 (2%) randomly selected observations. Validation plots ensured that the distribution of the training data (2%) and feature data (98%) matched. The training data was then randomly split into 75% used for training the model and 25% for its validation. An analysis of the feature importance and hyperparameter tuning was conducted to improve the model's accuracy. The best-performing RF model had an accuracy of 94.9%, an F1-Score of 0.928, and an AUC-ROC of 0.944. Significant improvements to the GWR's data quality could be achieved by implementing a RF classifier, that facilitated the automatic assignment of approximately 125,000 misplaced building coordinates.