Completeness Estimation of OpenStreetMap POI Data Using Machine Learning Approaches

Graduate

Dominic Monzón

Marco Crisafulli

Examiner
Prof. Stefan F. Keller

Co-Advisor
Claude Eisenhut, Eisenhut Informatik AG, Burgdorf, BE

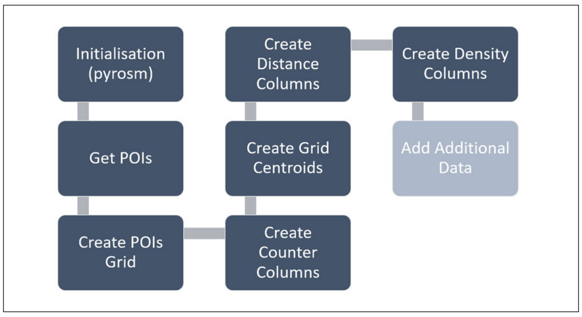
Subject Area
Miscellaneous, Software, Internet Technologies and Applications

Introduction: As OpenStreetMap (OSM) gains traction and is considered a viable alternative to service providers like Google Maps, the question of the quality of the provided data becomes increasingly important. A key factor for the quality of geographical data is completeness of entities that are included or omitted in a dataset. And currently there is no general solution to determine it. The vision of this project is to lay the groundwork for an approach with an open-source tool that can be used by the community and by users to check desired areas for completeness.

Approach / Technology: This work aims to estimate intrinsically - i.e., without comparing to a 'golden dataset' - the number of Points of Interest (POIs) in a defined area. These values compared to the number of existing POIs act as an indicator for completeness. The nature of the problem and the size of available data is predestined for machine learning (ML) methods. An initial model was trained based on high resolution imagery (orthophotos). It showed that there are relationships which can be detected by ML algorithms. Thus, a model was trained using only intrinsic data provided by OSM. Under the assumption that the training and validation areas are completely mapped, the implemented model performs well enough to show a trend where entities are missing.

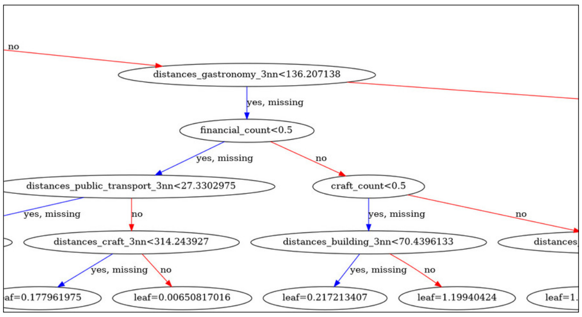
Conclusion: The results are visualized in a color-coded grid showing the areas which are predicted to either be complete, improvable or incomplete. As it is trained on data in Swiss cities it works best for urban areas in Switzerland and neighboring countries because of the geographic and demographic similarities. By use of re-training the model it is possible to predict other areas. One drawback of the intrinsic approach is that a certain amount of existing data is needed to make a prediction. Further, the quality of the prediction itself can only be measured on the assumption that the training and validation areas are well mapped. In conclusion, we provide a model which estimates the completeness of an area and indicates if further investigation is needed.

Data-processing pipeline [Python, pyrosm, GeoPandas]
Own presentment

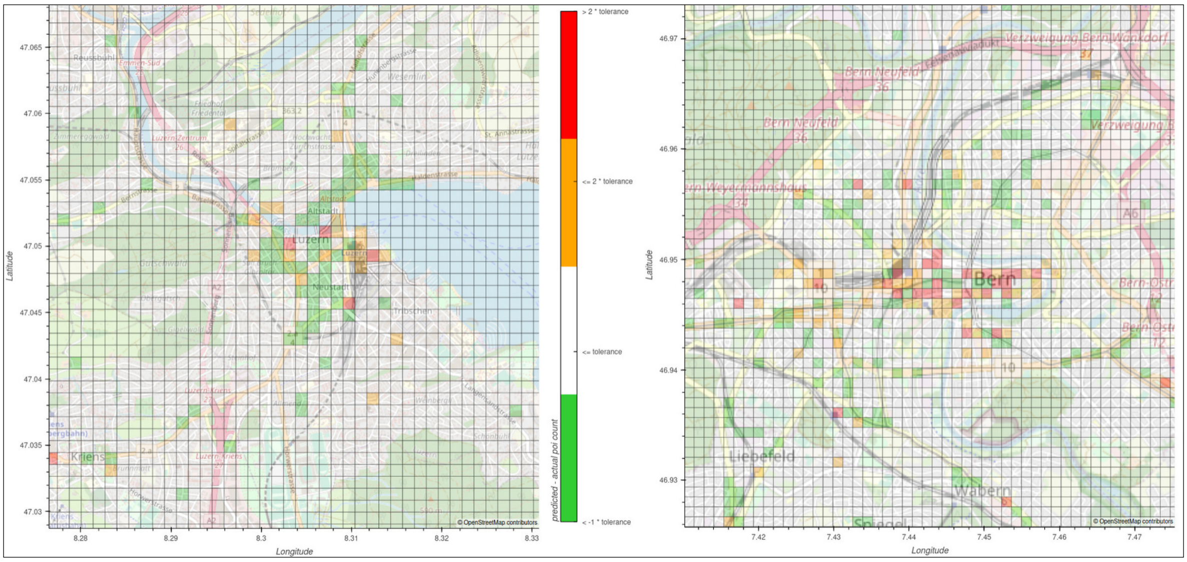


Initialisation (pyrosm)
Get POIs
Create POIs Grid
Create Counter Columns
Create Grid Centroids
Create Distance Columns
Create Density Columns
Add Additional Data

Excerpt of a Decision Tree using the Machine Learning library XGBoost
Own presentment



Prediction for Luzern/Bern in a Hectare Grid (green/white=complete, orange=improvable, red=incomplete) [fast.ai, hvPlot]
Own presentment



OST