Proposal for Semester Project

Patterns & Trends in Environmental Data / Computational Movement Analysis Geo 880

Semester:	FS24
Data:	GPS Data produced with different kinds of movement methods
Title:	Identifying Modes of Transportation from Personal GPS Data in Switzerland
Student 1:	Michael Fehr
Student 2:	Johannes Guler

Abstract

This proposal aims to develop a method to classify different modes of transportation using GPS-derived metrics in Switzerland. By leveraging features like speed, elevation change, stop frequency and similarity measures, the research will evaluate the accuracy of a self-developed model in differentiating transportation modes, including jogging, biking, hiking, and ski touring.

Research Questions

- How accurately can we identify transportation modes based on GPS data features such as speed, elevation change, similarity measures and stop frequency?
- What are the challenges in distinguishing between similar modes (e.g., jogging & biking)?
- How do environmental factors like street data influence the accuracy of transportation mode detection models?

Results / products

- Model Accuracy: Develop a classification model with high accuracy for identifying hiking, jogging, cycling, and ski touring using GPS features and provide a visualization of the results.
- Key Features: Determine the most critical GPS features for differentiating transportation modes.
- Challenges Addressed: Identify challenges in distinguishing similar modes (e.g., jogging & biking) and propose solutions.
- Practical Applications: Highlight potential applications in fitness tracking and urban planning.

Data

- Own collected GPS data with different modes of transportation (hiking, jogging, skitouring, cycling)
- Street Data from SwissTLM3D (streets, hiking paths)

Analytical concepts

Feature Engineering To begin with, feature engineering will play a critical role. Key features such as instantaneous and average speed, elevation change between consecutive GPS points, similarity measures (Edit Distance, Fréchet Distance) between the GPS tracks and already existing street data and stop frequency (identifying stops through minimal movement over defined time intervals) will be calculated. These features will serve as the primary inputs for classification models.

Exploratory Data Analysis (EDA) Next, Exploratory Data Analysis (EDA) will be used to visualize and understand the data. Techniques like histograms, box plots, and scatter plots will help reveal patterns and relationships between different transportation modes. This step is crucial for identifying overlaps and distinct characteristics in the feature distributions.

Model Development For model development, the additional parameters calculated through the feature engineering will be used to statically classify the GPS tracks into different transportation modes.

Model Evaluation Model evaluation will involve assessing performance metrics such as accuracy, precision, recall, and F1-score using the pre-known mode of transportation through the name of the tracks (validation). A confusion matrix will be used to analyze misclassifications, providing insights into the model's strengths and weaknesses.

Conceptual Movement Spaces and Trajectory Modelling The conceptual framework for this study will consider different movement spaces, such as urban vs. rural environments, and respective modelling approaches for trajectories. This includes understanding how different paths (streets, hiking paths) and spatial contexts influence transportation modes. Trajectory modelling will involve tracking and analyzing the paths taken by individuals to identify distinct patterns associated with each mode of transportation.

Spatial Analysis Methods Additional spatial analysis methods will be employed to enhance the model's accuracy. Similarity measures will be crucial to evaluate how additional street data affect the classification accuracy. In order to reduce the computational load of the project, the street data will be cut to the spatial extend of this project to exclude streets, that are not relevant. This will be done using QGIS Open Source Software using the Buffer and Spatial Intersection Tools.

Visualization and Discussion Finally, visualization tools such as confusion matrices and map plots will be used to illustrate the model results, showcasing accuracy and performance comparisons. Feature importance graphs will highlight the critical features influencing mode classification. A thorough discussion will address challenges in distinguishing similar modes (e.g., jogging & biking) and propose practical solutions. The potential real-world applications of the model, such as in fitness tracking and urban planning, will be explored, demonstrating the practical relevance and utility of the research findings.

R concepts

Data Manipulation: dplyr: For data manipulation and transformation (e.g., mutate, filter, summarize, group_by). readr: read and impor the data tidyr: For data tidying and reshaping (e.g., gather, spread, unite, separate).

Data Visualization: ggplot2: For creating comprehensive visualizations such as histograms, box plots, scatter plots, and feature importance graphs.

Model Evaluation: caret: For calculating performance metrics (accuracy, precision, recall, F1-score) and generating confusion matrices.

Feature Engineering and Similarity Measurement: geosphere: For calculating distances and geographic features from GPS data. sf: For handling spatial data and performing spatial operations. lubridate: For managing and manipulating date-time data. SimilarityMeasures: For calculating Edit and Fréchet Distance.

Spatial Analysis Methods: sp: For spatial data manipulation and analysis. rgdal: For interfacing with GDAL (Geospatial Data Abstraction Library) to read and write spatial data formats.

Spatial Data Visualization: leaflet: For creating interactive maps that visualize the paths taken by users and their classification results. tmap: For thematic mapping and advanced spatial visualizations.

Risk analysis

The biggest challenges include data quality issues due to GPS signal loss, difficulty distinguishing similar transportation modes, and the impact of similarity measures on model accuracy. Overfitting and high computational requirements are also concerns. To mitigate these, we will employ data cleaning, and validation techniques. If these issues persist, Plan B involves using alternative data sources, developing simplified models,

collaborating with other research groups, and initially focusing on more distinct transportation modes. This flexible approach ensures robust and actionable results despite potential obstacles.

Questions?

- What would be the most useful similarity measure? Are the two ones proposed suitable?
- Is the scope of this project suitable for this course?
- Are there any additional features or techniques that could improve the differentiation between similar transportation modes?
- What are the best practices for handling missing or inaccurate GPS data?
- Can you recommend specific data cleaning and interpolation techniques for GPS datasets?