An Algorithmic Approach to Route Planning for Performance-Minded Cyclists: Using Segmentation and Machine Learning to Visualise Training Routes

Jack Jibb
Student ID: 001408490
MSc Computer Science

April 22, 2025

Project Proposal

Supervisor

Dr. Elena Irena Popa

Topic Area

Data Science, Web Development, Segmentation Algorithms

Keywords

Bike Training, GPX, Web Application, Go, React, SQLite, Route Analysis, OpenStreetMap, Leaflet.js, Segmentation Algorithm, Clustering, Protocol Buffers

MSc Modules Contributing to this Project

- Fundamentals of Data Science - Gained a fundamental understanding of data manipulation in Python, which will help with organising large data collections from GPX and OSM sources
- Enterprise Software Engineering - Learned the logistics and administration behind developing a strong software project, including Architecture Diagrams and Agile Methodologies
- Systems Administration and Security - Learned basics of Linux, web stack development, and web security, such as HTTPS, Apache, and the LAMP stack.

1 Motivation

Existing GPS routing apps for cyclists mainly address simple metrics like distance, time, and terrain. They don't have tools for analyzing suitability in terms of effectiveness towards particular training objectives. The tools that do exist don't address dynamic and context-specific factors such as traffic, road environment (e.g., road furniture), or availability of unbroken sections suitable for prolonged effort. Therefore, coaches and athletes have to depend on personal knowledge or manual examination via services such as Google Street View or VeloViewer to gauge training value. This can be a laborious and inexact process.

2 Overview

This thesis puts forward an algorithmic approach to display cycling training appropriateness on a section-by-section basis from GPX route files. The method will be to create a software application that divides a GPX track into spatially and contextually meaningful segments, then score these segments based on the following metrics (note that these metrics are subject to change based on project feedback):

- Safety
- Adventure
- Amenities
- Difficulty
- Technicality
- Interval Suitibility
- Endurance Suitibility

The preliminary principal part of this project is to design a segmentation algorithm that can divide GPX track points into grouped segments, and save these segments to a database. This segmentation process will use data from OpenStreetMap (*OpenStreetMap Wiki* 2024), retrieved through the Overpass API (OpenStreetMap Wiki 2024*a*), and will determine the range of gpx points that contribute to a segment. After segmentation, a scoring system will analyse each segment. This will score segments based on certain performance-critical metrics via a Training Suitability Algorithm, so that users can rapidly gauge the training value of that particular segment, giving them the knowledge to use or omit it from their route.

3 Objectives

Objective 1: Research and Analyse Segmentation Algorithms, Map Data APIs, and Data Sanitation Methods

Activities:

- Review existing algorithms and methodologies for evaluating training suitability of cycle routes.
- Investigate APIs that can gather anonymous training data from external sources.
- Research methods of road data collection from OpenStreetMap and other GIS sources.
- Analyse prior academic work and technical case studies on route scoring and geospatial analysis.

Deliverables:

- Literature Review
- Proof of Concept programs, cleaned data collection, and API test scripts

Objective 2: Plan and Define the Project Architecture and Development Roadmap for the Web Application, and the Algorithms

Activities:

- Create a detailed project plan including a schedule of tasks and technical milestones.
- Identify requirements, architecture designs, and risks.
- Choose technology stack for implementation of project.

Deliverables:

- Requirements Specification
- Evaluation and Testing plan
- Gantt Chart
- Project Timeline
- Technical design document including Architectural Diagrams, data models, and system interfaces.
- Technology stack report

Objective 3: Design and Implement the Web Application

Activities:

- Develop a front end for route plotting and segment score visualisation.
- Develop a back end for data processing, scoring, and segment generation.
- Design and implement a database for storage of GPX files, metrics, and segment data structures.
- Build a simple REST API to link the front end and back end systems.

Deliverables:

- Fully functional web application source code and version history.
- Technical documentation and user guide for web application and API

Objective 4: Develop and Evaluate the Route Segmentation Algorithm

Activities:

- Parse GPX files to extract sequential trackpoints and associated cycling metrics.
- Pull nearby data from Overpass API.
- Group trackpoints into logical sections based on metric and OSM similarity.
- Develop a segmentation algorithm to generate ”Routelets” that compose the larger GPX route.
- Store generated segments into the database for the web application to access.

Deliverables:

- Segmentation Algorithm Source Code
- Sample GPX segment data
- Segment Structure Documentation

Objective 5: Develop and Evaluate the Suitability Score Algorithm

Activities:

- Query OSM API for road and environment data.
- Engineer OSM/GPX parser to characterise an input segment from matching GPS Coordinates.
- Define scoring model to generate Suitability Scores, either via Machine Learning or a heuristic approach.
- Implement and evaluate a scoring algorithm for each defined Suitability Score.

Deliverables:

- Suitability Score Algorithm Documentation
- Scoring System Implementation Code
- Segment-to-Score Dataset

Objective 6: Test, Validate, and Evaluate the Algorithms in Context With the Web Application

Activities:

- Test and validate both algorithms using real-world GPX data and reconnaissance via field testing or Google Street View.
- Perform unit, integration, and usability testing of the application.

Deliverables:

- Evaluation Reports
- Testing logs and scripts
- Final summary of findings and limitations

4 Legal, Social, and Ethical Issues

Legal

- **ODbL OpenStreetMap Licensing:** All data used from OpenStreetMap is licensed under the Open Database License (ODbL), which requires proper attribution and that any derivative databases must also be openly shared under the same license (OpenStreetMap Foundation 2024). This impacts the project if enriched or processed OSM data is redistributed.
- **Attribution Requirements for OSM and Leaflet.js:** OpenStreetMap and Leaflet.js both require clear attribution in visualizations and public-facing applications (*OpenStreetMap Wiki* 2024, Leaflet.js 2024). This must be explicitly included in the final application.
- **Use of Public API and Rate Limiting TOS:** The Overpass API enforces strict rate limits and usage policies to ensure fair access for all users (OpenStreetMap Wiki 2024*b*). Excessive automated requests for segment data could violate these terms and require rate limiting or caching on the backend.
- **Handling of User-Generated Data under Copyright Law:** GPX tracks and annotations created by users may be subject to copyright if they are derived from unique contributions. If user data is uploaded, its handling must comply with copyright and ownership principles.
- **Data Retention and Access Control:** Any user-submitted or processed data stored on the system should be retained only as long as necessary, with adequate access controls and encryption to protect it (European Union 2018).
- **Legal Implications of Data Scraping:** Even open APIs may have terms prohibiting scraping or excessive querying without permission. Bulk downloads or programmatic querying must respect the service's legal boundaries (OpenStreetMap Wiki 2024*a*).
- **Use of Restricted Government or Commercial Map Layers:** Care must be taken not to use government or commercial tiles (e.g., Ordnance Survey, Google Maps) that are not licensed for redistribution or analysis.
- **Licenses and Machine Learning Usage:** Some open data sources restrict use for model training, or require that resulting models not be commercialised unless attribution or relicensing occurs (Wiley & Kumar 2018).

Social

- **Discrimination and Area Labeling:** Labeling a route or area as ”unsafe” based on scoring algorithms may unintentionally stigmatize certain neighbourhoods, particularly lower-income or minority-dense areas.
- **Equity in Route Scoring:** Bias in data or scoring criteria may make it harder for some users (e.g., rural, disabled, or low-income groups) to access safe or suitable routes, reinforcing digital inequality (Haklay 2010).
- **Data Literacy and Accessibility:** Users may misinterpret complex visualizations or scoring metrics without sufficient explanation or interface clarity, reducing the tool's effectiveness and inclusiveness.

- **Data Quality Discrepancies in Low-Income Areas:** OpenStreetMap coverage is uneven and often less complete in economically disadvantaged areas, which could skew scoring or produce biased results (Goodchild 2007).
- **Privacy Concerns in Public GPX Trace Usage:** While GPX files are often shared publicly, repeated use could enable de-anonymization or tracking, especially when traces include home locations (Leszczynski 2015).
- **Unintended Consequences of Route Suggestions:** Suggesting ”adventurous” or high-score routes based on environmental data may inadvertently lead users into potentially unsafe areas, especially where real-time hazards (e.g., wildlife, road works) aren't accounted for.

Ethical

- **Transparency in Score Calculation:** Users should be able to understand how segment scores are calculated and what factors are considered. Algorithmic opacity can erode trust and create misinformation (Floridi & Taddeo 2016).
- **Consent in User Data Usage:** If user-uploaded data (like GPX traces) is used to refine the scoring models, informed consent must be obtained and users should have the option to opt out (European Union 2018).
- **Misrepresentation of Safety:** If users rely heavily on the app's safety score, it may create a false sense of security, especially if real-world dangers (e.g., traffic, poor lighting) aren't well represented (Tufekci 2015).
- **Fairness and Bias in Machine Learning:** Machine learning models may inadvertently inherit biases from training data, leading to unfair or inaccurate scoring. Methods must be tested for bias and explained clearly (Danks & London 2017).
- **Conflicting Metrics and Trade-Offs:** Users may prioritize different route aspects (e.g., safety vs. challenge), so scoring must be transparent about how trade-offs are handled and allow for customization.
- **Anonymisation Before Storage or Transmission:** Any GPX or telemetry data transmitted to a central server must be anonymised and stripped of identifying metadata to protect user privacy .

5 Resources

Programming & Data Resources

- Language for data processing, machine learning, and API calls; Most likely Python. (*Python 3 Documentation* 2024)
- A concurrent easy-to-code programming language will be used for GPX parsing and backend logic; most likely GoLang (*Go Documentation* 2024)
- Leaflet.js potentially used for front end map visualisation (Leaflet.js 2024)
- OpenStreetMap with Overpass API for map data and base maps (OpenStreetMap Wiki 2024*a*)
- Neovim Editor on Arch Linux for development.
- Github for version control and continuous integration.
- Personal GPX activity files (sample data)
- Open repository GPX data (if available)
- Local HTTP server for development with Nginx or Apache

Hardware Resources

- Personal Development Machines
- Dedicated GPU machine for training models
- Internet Access
- Resources for on-site reconnaissance (bike, camera, GPS device).
- GPX-capable GPS Device

Financial Resources

- Cloud Computing Credits
- University-Provided Funding (including access to academic databases, computer labs, software licences, department funding)

Academic Resources

- Papers and information on Geospatial Analysis, Routing Algorithms, and Machine Learning
- OpenStreetMap Overpass API, Leaflet.js, scikit-learn Documentation
- University and Research databases

6 Critical Success Factors

For each objective to be considered successful, the following requirements should be fulfilled:

#	Critical activity / resource	Key risk(s)
1	High-quality, complete OSM & GPX data	Sparse coverage in rural areas; Overpass rate-limiting or downtime
2	Sound research foundation & data-sanitation pipeline	Important literature missed weak algorithm design
3	Clear architecture & realistic roadmap	Scope creep; under-estimated task durations
4	Seamless front-end / back-end / DB integration	API contract drift; latency or data-format mismatch
5	Accurate GPX segmentation algorithm	Over- / under-segmentation producing invalid "routelets"
6	Trustworthy suitability-scoring model	Over-fitting, data bias, poor interpretability for athletes
7	Data-privacy / GDPR compliance	Storage of identifying location data in raw GPX traces
8	Comprehensive testing & validation	Insufficient time for usability tests and field checks
9	Regular supervisor / peer feedback	Drift from academic requirements; late course-corrections
10	Adequate computing resources	Computer unable to handle large Overpass queries or model training

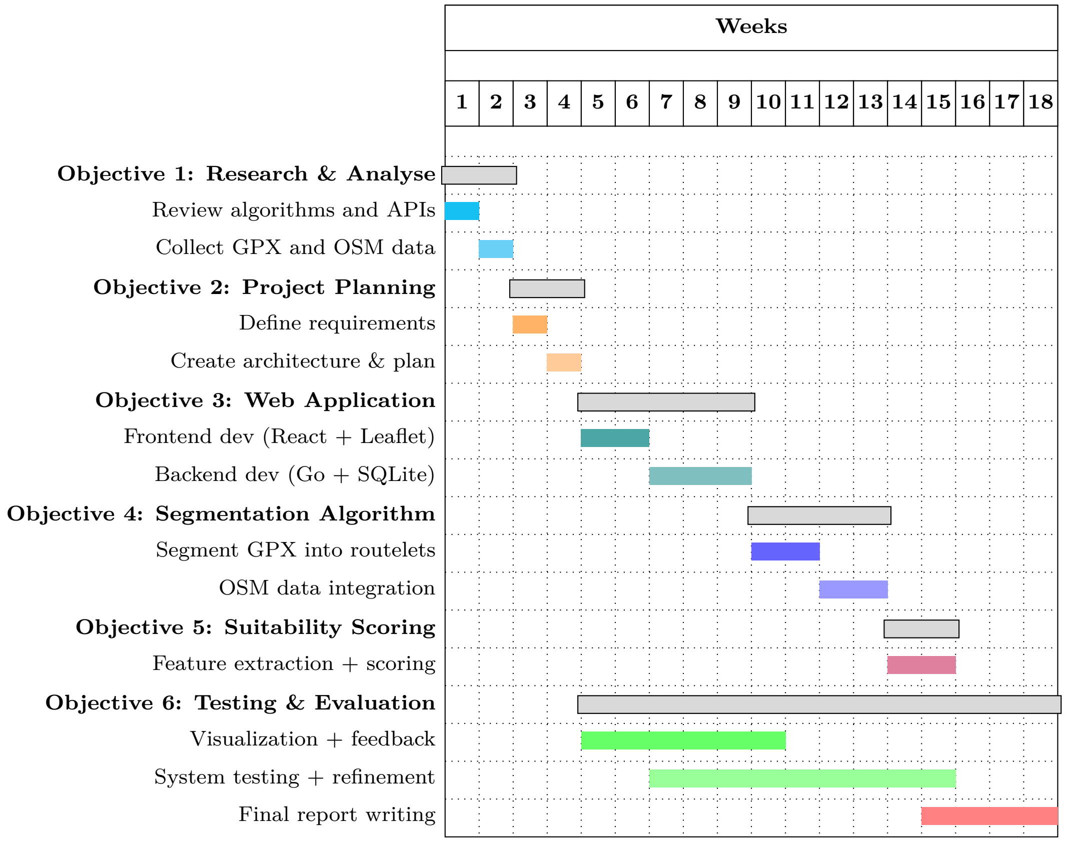
Table 1: Critical activities/resources and associated risks

7 Risk Matrix

Risk	Likelihood	Impact	Mitigation Strategy
Incomplete OSM data in target area	Medium	High	Cache OSM data; allow user corrections or fallback data sources (e.g., MetroExtracts)
Algorithm fails to segment GPX correctly	Medium	High	Use visual validation tools and segment with heuristics as fallback
Overpass API rate-limiting or downtime	High	Medium	Use data caching; throttle requests; prefetch large areas
Inaccurate or misleading scoring model	Medium	High	Start with explainable heuristic scores before switching to ML; gather user feedback
Frontend-backend integration bugs	Medium	Medium	Use Postman tests; define schemas up front; automate integration tests
User GPX data privacy breach	Low	High	Anonymise data on ingest; strip timestamps and identifiable home location
Scope creep or time underestimation	Medium	High	Prioritise MVP; follow weekly planning and retrospectives; maintain issue tracker
Lack of supervisor feedback at key stages	Low	High	Book recurring meetings in advance; prepare concise progress notes
Hardware/API quota limits during testing	Medium	Medium	Test locally when possible; apply for academic cloud credits; limit frequency of external calls

8 Schedule

- **Total time to completion:** 18 Weeks



Deadlines

- **May 5:** Objective 1 Deliverables: Literature Review, and Proof-of-Concept code
- **May 10:** Interim Report
- **May 23:** Objective 2 Deliverables: SRS, Evaluation & Testing Plan, in depth GANTT chart and project timeline, Technical Design document, and Technology Stack report
- **June 2:** Basic front end completed with Documentation
- **June 23:** Back end and API completed with Documentation
- **July 7:** Basic Segmentation Implementation with dummy data
- **July 21:** Segmentation Algorithm Integration with OSM Data
- **August 4:** Suitability Scoring Algorithm complete implementation
- **August 11:** Evaluation Report and Testing Logs
- **August 18:** All Documentation and Code completed
- **September 6:** Final report and Source code submitted

References

Danks, D. & London, A. J. (2017), Algorithmic bias in autonomous systems, *in* 'Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)', pp. 4691–4697.

European Union (2018), 'General data protection regulation (gdpr)', `https://gdpr.eu`. Accessed: 18 April 2025.

Floridi, L. & Taddeo, M. (2016), 'What is data ethics?', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **374**(2083), 20160360.

Go Documentation (2024), `https://go.dev/doc/`. Official documentation for Golang, supporting development of the GPX and backend tooling.

Goodchild, M. F. (2007), 'Citizens as sensors: the world of volunteered geography', *GeoJournal* **69**(4), 211–221.

Haklay, M. (2010), 'How good is volunteered geographical information? a comparative study of openstreetmap and ordnance survey datasets', *Environment and Planning B: Planning and Design* **37**(4), 682–703.

Leaflet.js (2024), 'Leaflet documentation', `https://leafletjs.com/reference.html`. Accessed: 18 April 2025.

Leszczynski, A. (2015), 'Spatial big data and anxieties of control', *Environment and Planning D: Society and Space* **33**(6), 965–984.

OpenStreetMap Foundation (2024), 'Open database license (odbl)', `https://wiki.openstreetmap.org/wiki/Open_Database_License`. Accessed: 18 April 2025.

OpenStreetMap Wiki (2024*a*), 'Overpass api', `https://wiki.openstreetmap.org/wiki/Overpass_API`. Accessed: 18 April 2025.

OpenStreetMap Wiki (2024*b*), 'Overpass api – public usage policy', `https://wiki.openstreetmap.org/wiki/Overpass_API#Public_Usage_Policy`. Accessed: 18 April 2025.

OpenStreetMap Wiki (2024), `https://wiki.openstreetmap.org`. Useful for understanding the structure, tagging schema, and guidelines of OSM data.

Python 3 Documentation (2024), `https://docs.python.org/3/`. Core reference for scripting, data processing, and API communication in Python.

Tufekci, Z. (2015), 'Algorithmic harms beyond facebook and google: Emergent challenges of computational agency', *Colorado Technology Law Journal* **13**, 203–218.

Wiley, B. & Kumar, P. (2018), 'Data licensing and the machine learning ecosystem', `https://datasociety.net`. Data & Society Research Institute. Accessed: 18 April 2025.