Data Requirements

DAIA

Date: 18-09-2023  
Team: Fire&AIce  
Members:

Stoyan Grozdev  
 Timon van de Wal  
 Ahmed Mahadi  
 Michał Raczkowski

Laurenz van de Poll

Timo Maas



# Table of Contents

[Table of Contents 1](#_Toc1271342570)

[Identify Data Types 3](#_Toc953007425)

[1. Numerical Data 4](#_Toc1880263222)

[2. Categorical Data 4](#_Toc158831424)

[3. Time Series Data 4](#_Toc1562768191)

[List Data Elements 4](#_Toc924003913)

[Numerical Data: 5](#_Toc322245455)

[Categorical Data: 5](#_Toc1414796852)

[Time Series Data: 5](#_Toc1531039845)

[Historical Data: 6](#_Toc1253719635)

[Frequency: 7](#_Toc858322484)

[Weather Data: 7](#_Toc1840660077)

[Training and Testing: 7](#_Toc1917185996)

[Data Augmentation: 7](#_Toc2038065468)

[Scalability: 7](#_Toc1082067655)

[Contextual factors 7](#_Toc2130254413)

[Contextual Factors Influencing Traffic in Eindhoven: 8](#_Toc1929969189)

[Define Data Quality Standards 8](#_Toc75703317)

[Consider Ethical and Legal Aspects 8](#_Toc1210207632)

[Finish Documenting Data Requirements 9](#_Toc2020135904)

Introduction

The primary objective of our project is to address the challenges faced by the professionals in JUGO who frequently travel between client locations and their work office. In today's fast-paced world, efficient time management is crucial. Unpredictable traffic conditions can lead to delays, missed appointments, and wasted time. Our aim is to mitigate these challenges by providing a software solution that offers predictive insights into the best times to travel. Our team is committed to developing a user-friendly software program tailored to the needs of professionals who are always on the move. The software will be designed with an intuitive interface, ensuring that even those who aren't tech-savvy can benefit from its features.

# Identify Data Types

## Numerical Data

Numerical data on traffic speed, route distance, and weather conditions are key for predictive models that guide travel decisions. These models categorize delays and consider seasonal variations and alternative routes, offering tailored and practical travel advice.

## Categorical Data

In our project focused on predicting traffic problems, Categorical Data is crucial. We use it to label key variables like weather conditions (sunny, rainy, etc.) and road statuses (clear, under construction, etc.). This helps us understand how different conditions affect traffic. For example, we can easily identify if rainy weather often leads to more accidents on certain roads. By including these categorized variables in our predictive models, we can make more accurate traffic forecasts. Therefore, Categorical Data plays a key role in helping us achieve the project's goals.

## Time Series Data

Weather measurements are time series data recorded over time. They are useful for predicting travel times because they can provide insight into road conditions, safety, traffic flow, and optimize routes and schedules for transportation.

# List Data Elements

## Numerical Data:

* Data Type: Numerical
* Units: Minutes
* Range: 0 to 180
* Example: Delay Duration (Numerical)
  + Units: Minutes
  + Range: 0 to 180 minutes
  + Description: Duration of delay experienced on the route.
  + Source: Rijksoverheid

## Categorical Data:

* Data Type: Categorical
* Categories: "Terrible", "Bad", "Neutral", "Good"
* Example: Delay Category (Categorical)
  + Categories: "Terrible", "Bad", "Neutral", "Good"
  + Description: Categorization of delay based on its duration.
  + Source: Derived from Delay Duration

## Time Series Data:

* Data Type: Time Series
* Time Interval: Daily
* Example: Historical Delay Data (Time Series)
  + Time Interval: Daily
  + Description: Past data on delays on the specific highway route.
  + Source: Rijksoverheid
* Example: Number of Cars (Time Series)
  + Time Interval: Daily
  + Description: Past data on the number of cars on the specific highway route.
  + Units: Numerical
  + Source: Rijksoverheid
* Example: Temperature Readings (Time Series)
  + Time Interval: Daily
  + Units: Celsius
  + Source: KNMI
* Example: Rainfall Readings (Time Series)
  + Time Interval: Daily
  + Units: Millimeters
  + Source: KNMI
* Example: Wind Speeds (Time Series)
  + Time Interval: Daily
  + Units: km/h
  + Source: KNMI
* Example: Visibility (Time Series)
  + Time Interval: Daily
  + Units: Meters
  + Source: KNMI
  + Determine Data Volume

## Historical Data:

Aim for at least 3-6 months of daily data for varied traffic and weather conditions. A full year is ideal for comprehensive coverage.

## Frequency:

Begin with daily data collections, with aspirations to access more frequent data in the future for refined predictions.

## Weather Data:

Ensure consistency in the duration of weather data (e.g., temperature, precipitation) with traffic data.

## Training and Testing:

When modeling, allocate approximately 70% of the data for training, 15% for validation, and 15% for testing.

## Data Augmentation:

Consider methods to maximize data utility, especially when starting with limited datasets.

## Scalability:

Plan for scalable storage solutions as data accumulates over time.

# Contextual factors

Traffic Influences:

* Peak hours leading to congestion.
* Accidents causing unexpected delays.
* Road constructions.
* Public transport failures: more people rely on personal vehicles.

Weather Conditions:

* Heavy rain affecting road grip and visibility.
* Snow leading to slippery conditions and potential road closures.
* Fog affecting visibility and safe driving speed.

## Contextual Factors Influencing Traffic in Eindhoven:

Major Events:

* Dutch Design Week: Held annually in October, this is the largest design event in Northern Europe. It attracts many visitors, potentially leading to increased traffic.
* GLOW Eindhoven: A light art festival held annually in November. It attracts many visitors, potentially causing traffic congestion in certain areas.
* Eindhoven Marathon: Typically held in October, this event can lead to road closures and diversions, affecting traffic flow.
* Philips Stadion Matches: Whenever PSV Eindhoven has a home match, there can be increased traffic around the stadium.

Other Factors:

* University and School Calendars: The start and end of terms, as well as exam periods at institutions like Eindhoven University of Technology, can influence traffic patterns.
* Major Conferences and Trade Shows: Depending on the year and schedule, major conferences and trade shows can attract a significant number of visitors.

# Define Data Quality Standards

Accuracy: Data entries on delays should be accurate to within 5 minutes of the actual delay time.

Completeness: No more than 2% missing data.

Consistency: Consistent units and format across all data sources.

# Consider Ethical and Legal Aspects

Only use open data sources to ensure no infringement on copyrights or proprietary data.

Ensure that no individual user data is stored or processed, adhering to privacy norms and standards like GDPR.

# Finish Documenting Data Requirements

By using the data specifications we've mentioned, we can build a machine learning system to predict when there won't be traffic delays on our routes between the office and client sites. This will help us meet our stakeholders' expectations and ensure smoother travel.