29 ΜΑΡΤΙΟΥ 2023

REPORT NO. 4

OFFLINE PART OF THE PROJECT

EFSTRATIOS KARKANIS
P19064

Table of Contents

Introduction

In this report, I am going to explain what data we can use and why. Also, I provide a detailed step by step process of how we can build the time series data in the offline part of the project.

About the data

To begin with, since this diploma thesis is structured in top of the Strict Path Queries (SPQ) paper, I strongly recommend using the data that was being used in the SQP analysis. Also, there is a backup of the database, so **we have already the data**. The reasons that I believe this are the following:

- We have already the data cleaned and organized in tables of a PostgreSQL database. So, no further analysis should be considered in this step.

- The source code of the SPQ implementation needs specific tables from the backup data, in order to perform correctly. In other words, the code is adaptable only to the data used for the SPQ analysis. If we are going to choose another dataset of raw data, the whole code of the SPQ implementation should be changed. Also, I am open to any opinion about the data that we can use, but I prefer to keep the SQP data.

The data that was used in the SPQ analysis contain 50 trajectories of vehicles that are moving in the **roads of Attica** (Athens, Euboea Island, and some other islands of Cyclades).

The time intervals (days) that these trajectories were recorded are separated in the following time intervals:

1. 02/09/2002 – 16/09/2002 (consecutive days)
2. 07/08/2002 – 10/08/2002 (consecutive days)
3. 12/08/2002 – 14/08/2002 (consecutive days)

4. 19/08/2002 – 23/08/2002 (consecutive days)
5. 26/08/2002 – 31/08/2002 (consecutive days)

Build of time series data

In this part of the report, I will explain how we can build the time series data using the SQP queries. This final dataset will contain timestamps across the columns and paths across its rows. Each of these paths perform a single route of the road network. The more paths we use, the better cover of the road network we have.

Important: a route contains many consecutive road segments of the existing road network, according to the SPQ paper.

Using SQP queries for each path every constant time interval (ex. each 1 hour or each 16 minutes), we can fill the final dataset. However, because of the high time complexity of the SPQ implementation, this whole process will last much time. After that, we can clean the final dataset (for example there might be many cells that have null values).

Important: SQP queries for a specific path and time interval return the number of trajectories (number of vehicles) that passed from this specific path within the given time interval.

In order to create the time series dataset, I can use python and Jupiter notebook. By this way, I can connect to the database and send SPQ queries to the database.

This is the end of the offline part, after that, we proceed to time series forecasting algorithms, to **forecast** the traffic flow of each path that belongs to the time series dataset.