Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 113 (2017) 502–507

Procedia Computer Science

www.elsevier.com/locate/procedia

The 2nd edition of the International Workshop on Data Mining on IoT Systems (DaMIS)

Prediction of bicycle counter data using regression

Johan Holmgren[a,b,*], Sebastian Aspegren[a], Jonas Dahlström[a]

[a]*Department of Computer Science and Media Technology, Malmö University, Malmö 205 06, Sweden*
[b]*K2 (The Swedish Knowledge Centre for Public Transport)*

Abstract

We present a study, where we used regression in order to predict the number of bicycles registered by a bicycle counter (located in Malmö, Sweden). In particular, we compared two regression problems, differing only in their target variables (one using the absolute number of bicycles as target variable and the other one using the deviation from a long-term trend estimate of the expected number of bicycles as target variable). Our results show that using the trend curve deviation as target variable has potential to improve the prediction accuracy (compared to using the absolute number of bicycles as target variable). The results also show that support vector regression (using 2nd and 3rd degree polynomial kernels) and regression trees perform best for our problem.

Peer-review under responsibility of the Conference Program Chairs.

Keywords: Bicycle counter, regression, trend curve, regression algorithm comparison

1. Introduction

The bicycle has become an important part of urban transport due to its ability to contribute to fast, sustainable, and cost efficient transport. It also contributes to a healthy, active, life style, and the popularity of the bicycle is accentuated by the increase of bicycling that can be observed around the world. Due to the positive effects of bicycling, there is an increasing interest from public authorities to increase the use of the bicycle. However, in order to achieve a modal shift towards bicycling (from motorized transport), it is important to increase the attractiveness of the bicycle. This can be achieved by implementing various types of policy measures, including the construction and improvement of biking infrastructure, such as bicycling lanes and safe parking facilities. Other initiatives include bicycle sharing systems, which are currently being implemented in cities around the world [1,2]. Bicycle sharing systems enable, for example, fast multimodal passenger transport, where public transport and the bicycle can be combined in an efficient way [1]. The recent introduction of electrical bicycles, is another factor that increases the attractiveness of the bicycle [3].

However, in order to build a transport system that encourages bicycling, it is important to fully understand the current bicycle flows, and what factors influence the travelers' choices whether to travel by bicycle, to use some other mode of transport (e.g., car or bus), or to not travel at all. Hence, it is important to collect various types of traffic, transport and bicycle related data, which can be done using Internet of Things (IoT) connected devices, such as bicycle

* Corresponding author. Tel.: +46-40-6657688 ; fax: +46-40-665 76 46.
E-mail address: johan.holmgren@mah.se

1877-0509
Peer-review under responsibility of the Conference Program Chairs.
10.1016/j.procs.2017.08.312

counters and mobile phone applications that enables registering the movement of travellers. Bicycle counters, which are the focus of the current paper, allow to continuously register the bicycles that pass some particular point in a transport network. Due to the possibility to register a large share of the passing bicycles, bicycle counters (typically built using inductive loop detectors) are commonly used to collect bicycle flow data.

In the presented study, we analyzed data collected by a bicycle counter located in Malmö, Sweden. An important purpose of our work was to quantify how various factors, such as day of week, time of year, and weather (temperature and precipitation), is expected to influence the amount of bicycle traffic at a particular point in the traffic network. We studied how it is appropriate to formulate a regression problem that can be used to estimate the number of bicycles registered by a bicycle counter. In particular, we investigated whether the use of a long-term trend estimate of the number of registered bicycles has the potential to improve the regression accuracy. We also compared different regression approaches, in order to identify which approach is most suitable for the considered problem. The current study builds on the Bachelor's thesis of Aspegren and Dahlström[4], who compared a set of regression algorithms regarding their ability to estimate the number of bicycles registered by our bicycle counter. Aspegren and Dahlström limited their analysis to consider only working days, whereas we include all days in the regression problem, explicitly considering day of week, school breaks, national holidays, and bridge days as input features.

Our work aims to provide input for passenger transport analysis models used by city and transport planners, e.g., for assessing the impact of transport policy measures. The relevance in this direction is emphasized by the fact that bicycling is currently being incorporated in passenger transport analysis models.

The current paper is organized in the following way. In the next section we give an account to previous research related to our work. In Section 3, we describe the data processing that we conducted in the beginning of our study. In Section 4, we present our regression modeling, which is followed in Section 5 by our computational results. We finalize the paper in Section 6 with some conclusions and pointers to future work.

2. Related work

The research related to bicycle data analysis has been quite intensive during the recent years. Romanillos et al.[5] provide an overview of big data approaches applied in the bicycling context. A large amount of research concern bicycle sharing systems, where the studied problems include bicycle repositioning[6] and location of base (or docking) stations[7]. Data mining has been applied in the bicycle sharing context, for example, in order to estimate usage patterns[8,9]. Data mining also plays an important role in travel demand estimation (including bicycle demand analysis), which is an integral part of traffic and transport analysis models (both in urban and in regional contexts). Traditionally, travel demand is estimated using travel survey data, often combined with GPS trajectories[10]. Bicycle demand can be further estimated using different types of discrete choice models, which have been used, for example, for bicycle route and destination choice estimations[11]. In addition, there exists research on how various factors, including weather, calendar events, and work related factors, influence the choice whether or not to use the bicycle[12,13].

The current paper focus on regression analysis using bicycle counter data in order to quantify how factors such as weather are expected to influence the amount of bicycling. According to the best of our knowledge, there exist no such previous study, except for the work by Aspegren and Dahlström[4].

3. Data pre-processing

In our study, we considered the time period September 13, 2006 to March 31, 2014, where we used bicycle volume data from a bicycle counter located in the city center of Malmö, weather data (i.e., temperature and precipitation), and information about national holidays and school breaks. We obtained information about school breaks from the web pages of the public schools in Malmö; however, as complete information about school breaks were not publicly available for the considered time period, we made a few assumptions concerning school breaks. In particular, we assumed that the longer school breaks occur during the same weeks each year, which was partially confirmed by the municipality of Malmö. The bicycle counter and weather data sets, which we received from the municipality of Malmö, specify values hourly. However, in the regression problem, where we considered each day as a data point, we aggregated the bicycle counter data for each day, and we used the averages of the temperature and precipitation values for each day. In addition, the bicycle counter and weather data sets had some missing values, which we estimated