Chaos, Solitons and Fractals 90 (2016) 37–45

Contents lists available at ScienceDirect

Chaos, Solitons and Fractals

Nonlinear Science, and Nonequilibrium and Complex Phenomena

journal homepage: www.elsevier.com/locate/chaos

Analyzing human driving data an approach motivated by data science methods

Peter Wagner[a,*], Ronald Nippold[a], Sebastian Gabloner[b], Martin Margreiter[b]

[a] *Institute of Transportation Systems, Deutsches Zentrum für Luft- und Raumfahrt, Germany*
[b] *Engineering Faculty Construction, Geo, Environment, Department of Traffic Techniques, Technical University of Munich, Germany*

ARTICLE INFO

Article history:
Received 30 October 2015
Revised 5 February 2016
Accepted 8 February 2016
Available online 2 March 2016

Keywords:
Car-following
Big data analysis
Maximal information content

ABSTRACT

By analyzing a large data-base of car-driving data in a generic way, a few elementary facts on car-following have been found out. The inferences stem from the application of the mutual information to detect correlations to the data. Arguably, the most interesting fact is that the acceleration of the following vehicle depends mostly on the speed-difference to the lead vehicle. This seems to be a causal relationship, since acceleration follows speed-difference with an average delay of 0.5 s. Furthermore, the car-following process organizes itself in such a manner that there is a strong relation between speed and distance to the vehicle in front. In most cases, this is the dominant relationship in car-following. Additionally, acceleration depends only weakly on distance, which may be surprising and is at odds to a number of simple models that state an exclusive dependency between acceleration and distance.

1. Introduction

Driving behavior models have become important tools in transportation science and engineering applications since their first appearance in the 1950s [21]. By now, a lot of models that seek to describe human driving behavior, i.e. acceleration processes (and lane changing maneuvers) have been introduced, see [6,10,15,16] for albeit incomplete overviews. The complexity of these models ranges from very simple models with few parameters like Newell's lower order model [18] or the cellular automaton model [17] up to multi-regime models like Wiedemann's psycho-physical perception threshold model [26] or the model implemented in the MITSIM-lab [1] open source simulator with lots of thresholds and different conditions describing an assumptive behavior in specific situations.

It is commonly believed, that the behavior of a human driving a vehicle can be described quite generally by the two equations

$$v(t+\Delta t) = v(t) + a(\Delta v, g, v)\Delta t \tag{1}$$

$$x(t+\Delta t) = x(t) + \frac{\Delta t}{2}(v(t+\Delta t) + v(t)) \tag{2}$$

where $(x(t), v(t))$ are the position and speed of the vehicle at time t, $(g, \Delta v)$ are the distance and speed difference to the vehicle in front, and $a(\cdot)$ is the acceleration function. There is no need for Δt to be equi-distant, and in fact the so called action-point models [22,26] claim that a driver changes her course of action only from time to time, based either on external forces or even on no apparent reason which may lead to an exponential distribution of the Δt-values [24].

Since a lot of empirical research has been done on this topic, too, only occasionally approaches have been made to connect empirics and models in a physicist's manner. A course of action that is often pursued in driver modeling

* Corresponding author. Tel.: +49 30 67055 237; fax: +49 30 67055 291.
E-mail addresses: peter.wagner@dlr.de (P. Wagner), ronald.nippold@dlr.de (R. Nippold), sebastian.gabloner@tum.de (S. Gabloner), martin.margreiter@tum.de (M. Margreiter).

http://dx.doi.org/10.1016/j.chaos.2016.02.008

is to do calibration and validation [8] of known models. By their very nature, these exercises often yield reasonable results (of the order of 10–20% root-mean-square error [4]) even for models that are known to be a bad description of driving, like e. g. the simplest cellular automaton models [17] or the optimal velocity model [3].

Here, a different approach is adopted. A large data-set of car driving (collected in the second half of 2012 in the German simTD project) is analyzed with the help of methods from data science [9]. This research will try to look on these data and try to find some general relationships between the measured data. Within this work, the approach is limited to the four variables speed v, acceleration a, distance to the lead vehicle g, and the speed-difference Δv to the lead vehicle. However, in principle it is also possible to include other parameters that might influence driving behavior, such as the time of day when the driver is driving, the street-type, the weather conditions, or the acceleration of the lead vehicle which might be accessible to the driver by observing the lead vehicle's braking lights. Finally, this approach here can find a few general features of the function $a(g, \Delta v, v)$ in Eq. (1) above.

The present paper consists of four parts: A description of the data set under consideration is contained in Section 2. The following Section 3 gives an overview of the statistical measures maximal information content (MIC) and mutual information (M) used in this paper, and states the main results. Finally, the conclusions of this analysis are presented in Section 4.

2. Description of the simTD data

The data to be used here have been recorded during the field test of the simTD project from July 2012 to December 2012. Altogether 120 vehicles where driving around for 98 days. Different drivers were assigned to the vehicles, but this assignment is kept confidential and therefore not part of the data-set. Because of the main goal of the project, to estimate the efficiency of vehicle-to-vehicle and vehicle-to-infrastructure communication, the vehicles drove on predefined routes only, thereby covering an area of roughly 15 × 45 km around the German city of Frankfurt.

All the data were made available by the six participating German car manufacturers via the CAN-bus in a frequency between 200 Hz and 0.5 Hz. To abstract the manufacturer-specific protocols, all the data were extracted from the internal network of the car (the CAN-bus) by a specialized VehicleAPI (VAPI) which was especially developed within this project. Therefore, all data were available in the same generic format for the project. All the signals from the CAN-bus where synchronized using the GPS time.

The data used in this paper have been recorded by four sensors, that were built into the vehicles: a GPS sensor, an acceleration sensor that was aligned with the car's geometry and therefore allowed the measurement of the longitudinal (in driving direction) and lateral acceleration (perpendicular to the driving direction), the distance and velocity difference to the lead vehicle by a radar/lidar sensor, and the speed from the traditional wheel sensors. The acceleration data was noisy, but not in an unreasonable manner, so it was decided to use them unfiltered as they are. This does not rule out, that the internal machinery within the cars itself does some filtering, but from what have been seen by visual inspection, this does not seem very likely.

Within this project, the GPS data have been enhanced into a differential GPS by correction signals received over UMTS and ITS-G5 (802.11p). These data have been matched on an underlying digital road network, but this has not been used in this paper but was used in the communication part of the projects. All the cars were normal cars (with all the sensors) that can be bought in exactly this form, only the data-acquisition had been added within the project.

The data were recorded asynchronously. That means, that the variables to be analyzed here (distance g, speed v, speed-difference Δv, and acceleration a) are recorded in their raw format not at the same time, and the time-difference between subsequent readings even of the same sensor is not guaranteed to be equidistant. The variables are acquired by three different sensors: a radar sensor which measures the distance and speed-difference to the vehicle in front (sometimes it also picks a tree at the border of the road), the vehicle's internal measurement of the speed, and the acceleration which is recorded by a dedicated sensor.

As explained above, the current position of the vehicle is measured by a GPS receiver. Note, that the GPS provides an additional measurement of the speed which has not been utilized here. Also, it could have been used to determine the acceleration of the vehicle, which also has not been done here for the analysis below systematically. A brief view into this, however, reveals that the data obtained from GPS are in good agreement with the recordings to be used in the following.

The typical frequency with which the data are recorded is about 10 Hz. Therefore, it was decided to force the data into a common time-basis by aggregating them to 0.1 s. This worked well, in most cases less than 4 data-points fall into one time-bin, which is then averaged to get the time-series that will be analyzed subsequently.

It is safe to assume that the data are not error-free. From visual inspection of the data, it turns out that there are a lot of points where the data-stream disconnects, i.e. there are gaps in time which are larger than 0.1 or 0.2 s, where no values are recorded. To transform this into a measure of data-quality, the one-step ahead prediction error is used in the following. This error can be computed from estimates of the gap and speed of the subject vehicle for the next step in time, which is given by:

$$\hat{g}_k = g_{k-1} + \Delta v_{k-1}(t_k - t_{k-1}), \tag{3}$$

$$\hat{v}_k = v_{k-1} + a_k(t_k - t_{k-1}). \tag{4}$$

This is compared with the actual measurements at the time-point $k+1$, thereby defining a measure of consistency of the time-series:

$$e_k^{(g)} = \hat{g}_k - g_k, \tag{5}$$

$$e_k^{(v)} = \hat{v}_k - v_k. \tag{6}$$