Smart Drivers: Simulating the Benefits of Giving Twitter Information about Traffic Status

Ana L. C. Bazzan[1*], Pedro G. Araújo[1], Cristiano Galafassi[1], Anderson R. Tavares[1], Alessandro Dalla Vecchia[1], Antônio Rodrigo D. de Vit[2], Glaucio R. Vivian[2]

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul
Caixa Postal 15064 – 91501-970 Porto Alegre, RS

[2]Universidade Federal de Santa Maria, campus de Frederico Westphalen, RS, Brasil

Abstract. *One of the main pillars to reach smart cities is a smart transportation system, both public and private. Our long term goal is to develop an agent-based infrastructure that can be used for investigations that are key to evaluate the effects of concepts related to smart transportation systems. However, such infrastructure demands an amount of data that few traffic authorities can afford to have. An alternative to installing sensors is to use human and their mobile devices as sensors. However, this poses challenges for the gathering and management of such data. In this paper we propose a methodology to deal with this problem. It aims at capturing and treating traffic data (mainly streets and links statuses) that appear in social networks, microblogs, etc. Specifically, we illustrate the approach with data that appears in the blog "Trânsito" that is managed by the daily paper OESP, online edition. With the implemented prototype, we have simulated thousands of agents that can do en-route adjustments on their routes based on updated knowledge of the traffic status. We were able to derive conclusions that would not be possible if only macroscopic simulation methods were used, as for instance the extent of improvement in the travel time of drivers that receive information.*

1. Introduction

Due to changes in demographical patterns around the world, it is expected that the urban population is going to increase drastically. In particular, the increase in the number of mega-cities has strong consequences to traffic and transportation.

Contrarily to data networks, for example, which can be monitored and controlled to a given extent, it is barely impossible to efficiently control vehicular traffic because drivers and users of the transportation network in general act as “intelligent” and autonomous decision-makers, seeking to maximize their own utility in a bounded rational way. This may cause serious issues as the so-called price of anarchy
[Koutsoupias and Papadimitriou 1999] reveals.

Information systems have a key role to allow citizens to better plan their trips and activities. In order to provide the necessary information to the user of the transportation network, traffic simulation is key in the ITS effort. All simulation paradigms depend on traffic data, but the microscopic one is perhaps the most data intensive. Collecting and

*corresponding author: bazzan@inf.ufrgs.br

using data for microscopic simulation has been a problem, especially in countries such as Brazil, where the investment necessary to collect that kind of data has not been done so far, except in a handful of big cities. This way alternative ways to gather data and information about traffic status is necessary in order to run a microscopic simulation.

In this paper we aim at using data from social networks (in particular data from microblogs like Twitter) to feed a microscopic simulation of traffic. Our methodology is to extract data from the web about current status of key links in a traffic network, transform status into numerical values and use them as modified costs in a shortest path algorithm, in order to compute routes for each trip that is generated in the network.

The challenges are manifold. For example, not only sources are mostly not yet geo-referenced and not completely credible, but also data is poorly structured.

We illustrate the use of our methodology with a case from the city of São Paulo and data provided by the website `http://blogs.estadao.com.br/transito`, a service that summarizes traffic status from Twitter users (and other sources). Besides, we use real trip data of São Paulo, provided by an origin-destination survey (which includes all trips in the metropolitan area, available at `http://www.metro.sp.gov.br`). For the simulation, we use SUMO, a microscopic simulator available for download at . We remark that any microscopic simulator could be employed, as [Passos et al. 2011] has shown in a comparison study.

In order to generate the initial routes for each driver when the network is empty SUMO's default route computation algorithm considers only the length of each link of the network. This way, routes generated by SUMO ignore the time that is necessary to travel a link under different traffic conditions. For comparison, we also use SUMO's DUA (acronym for dynamic user assignment, an iterated procedure that seeks the user equilibrium).

Our preliminary results show that if drivers follow SUMO's default route generation algorithm, this results in congestion in several links. When we incorporate information from social network, i.e., when the actual status of the link is considered, the route that are generated tend to disperse the total demand more evenly in the network, as it is the case in the real world. The rest of this paper is organized as follows: In the next section we give a brief background about traffic simulation as well as pointers to related work. Methods are detailed in Section 3, while the settings and results achieved so far are discussed in Section 4. Conclusions and the future work are then presented in Section 5.

2. Background and Related Work

Smart transportation systems are likely to depend more and more on a good modelling of the transportation system at hand because only this way can the effect of new technologies be tested and evaluated. Regarding modelling of traffic, while a macroscopic model is mainly concerned with the movement of platoons of vehicles, that means, with the aggregate level, in the microscopic modelling one may go to the individual level. Each road user can be described as detailed as desired (given computational restrictions), thus permitting the model of travelers' behaviours. Travel and/or route choices may be considered, and this is a key issue in simulating traffic, since those choices are becoming increasingly more complex. Among the microscopic modeling forms, the agent-based technique is a promising one.

Modeling of traffic scenarios using agent-based modeling techniques is not new. However, few works address agent-based assignment of demand. Examples appear in: [Balmer et al. 2004], [Chmura and Pitz 2007], and [Bazzan et al. 2011]. In all these works the aim is to investigate what happens when the driver chooses a route using information (its own experience or available from the environment or navigation devices). The common line is that information changes the decision pattern regarding route selection. Thus, information is a key commodity in traffic simulation. In this paper we discuss how to get it from the Internet and incorporate it within the simulation for route assignment purposes.

We remark that we deal with vehicular traffic, i.e., the focus is not on public transportation. For works in this area see, e.g., [Jaques et al. 2012, de Lima et al. 2012]. Also, readers interested in the issue of agent-based modeling and simulation are referred to [Bazzan and Klügl 2013].

We also note that in the last years there has been an increase in the number of applications that run on smart phones and navigation devices, which are intended to help the commuter/driver. However, to the best of our knowledge, there are just initial attempts to deal with the mass of data that is generated in this process. Most of these works are proprietary (e.g., google traffic) but one can guess that users with mobile devices are being tracked in order to provide traffic information (speed, location).

3. Methods

As mentioned, the aim of our work is to run microscopic simulations producing results as close to the reality as possible. In order to do this, a lot of data is necessary, as for instance real-time data from traffic status at key links of the network, as well as an estimate of the trips that occur in a given traffic network. Although the latter is normally available (as it refers to historical data), real time data is very hard to get. Most cities, especially in emerging economies, do not have the necessary infrastructure to monitor and measure traffic status. One exemption is the city of São Paulo, where the CET (the local traffic engineering public company) monitors around 800 kms. of arterials deemed important in the city traffic. However, other initiatives also exist: MapLink, Google Traffic, Waze, etc. Increasingly, the social network is playing a role in filling the gap that exists regarding traffic information for the end user. Mining this information is however a challenge. Not only sources are scattered and not completely credible, but also data is poorly structured.

In this section we report the progress achieved so far regarding collecting and using information about traffic status from social networks. We start giving some details about SUMO (Section 3.1) and how it does the assignment of trips. We then describe our methods to collect and process data from the Internet (Section 3.3), as well as how to generate synthetic data that emulates the tweets that are then processed (Section 3.4). The routing methods as well as the driver model is given in Section 3.5.

3.1. Microscopic Simulation using SUMO

SUMO [Behrisch et al. 2011] is based on a microscopic simulation model called car-following. In this model, a vehicle's operational behavior regarding acceleration or braking is influenced by its leading vehicle. The simulation is continuous in space and discrete in time, so that we have the precise physical location of the vehicles in the road network.

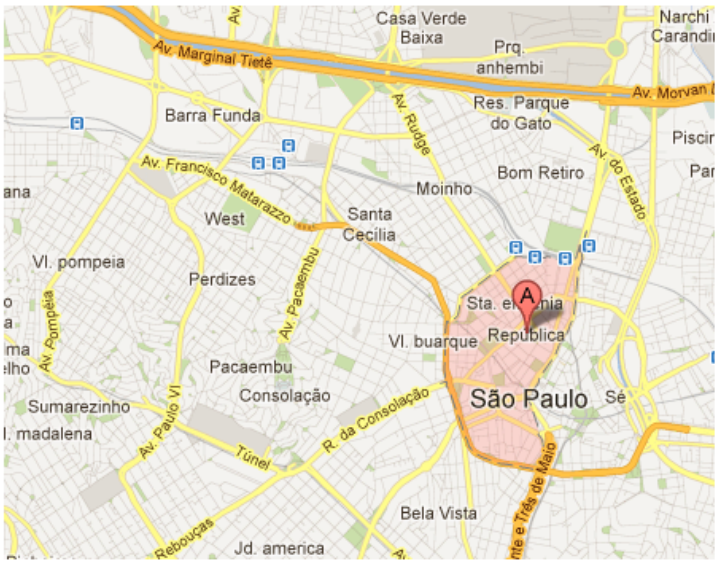

Figure 1. Map of the center-western of the city of São Paulo showing the district of República, that is marked in Fig. 2 by letter A (source: Google Maps)

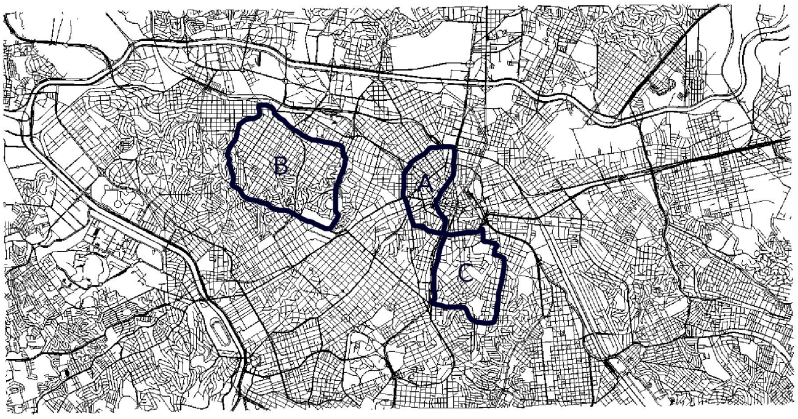

Figure 2. Map of the city of São Paulo showing three of the districts that appear in the OD matrix (downloaded from OSM and annotated)

In our experiments, we simulate a commuting scenario. Given an origin-destination matrix (OD) matrix, routes are generated for each trip. Then, each iteration simulates a fixed period of a working day: the same drivers will travel from the same origins to the same destinations. The simulation procedure is as follows: during an iteration, at each simulation time step, the vehicles are moved according to the car-following rules; drivers update their knowledge bases when they leave or enter a link.

3.2. Scenario

To illustrate our methods, we give the example of the city of São Paulo, Brazil. The map of the city (excluding the metropolitan area) was imported from Open Street Maps (`www.osm.org`), henceforth OSM. It is shown in Figure 2.

Specifically, the extracted area ranges from coordinates -23.5001 (N), -23.5999 (S), -46.7556 (W), and -46.543 (E). The map contains 21620 intersections (nodes of the

graph) and 49990 edges. These edges could accommodate around 2 million vehicles (if all lanes were completely occupied).

The OD matrix was generated based on information available at `http://www.metro.sp.gov.br`. Note that this survey includes all trips in the metropolitan area of São Paulo. We have used the data for motorized trips (cars, motorcycles, trucks).

We generate a number of trips that corresponds to around 10% occupancy of the whole network on average. This corresponds to roughly 200.000 trips, which were generated in a period of 3 hours, i.e, 200.000 vehicles must enter the simulation distributed in a 3 hour period. We remark that we do not use only arterials but all streets of the city, thus this occupancy is not as low as it may appear. This number is realistic given that the distribution of trips, not being even, means that some portions of the network are highly overloaded, whereas others have almost no traffic.

These steps seem trivial but are associated with some challenges. First, after the download of the objects from the OSM, some parts rendered disconnected. Thus a posprocessing was required to fix these problems.

Second, the OD matrix has to be generated by hand, a time consuming task. Zones and districts that generate and attract the trips were manually transcribed. For instance, Figure 2 shows three of these districts, from a total of 35 districts we have used:

- A - República;
- B - Perdizes;
- C - Liberdade.

Each district is composed by at least one link.

Finally, a third challenge, as discussed in the next subsection, is the extraction of real time information is not trivial given that the text has no fixed structure, is not geo-referenced, appearing mostly in natural language (Portuguese).

3.3. Information Retrieval from Internet Traffic Forums

In order to collect real-time data from the Internet to feed SUMO, we propose a methodology that, given a source of traffic status information, constantly checks for new information, filter it, and transform it to a given format. We illustrate the current status of the work by means of information that is available at the website of the daily newspaper (online edition) OESP (`http://blogs.estadao.com.br/transito/`), which has a blog where textual entries regarding traffic status in the city of São Paulo appear with a relatively high frequency. It is possible to see that the information becomes available at different times of the day, normally covering all day.

Typical blog entries have the format as follows. These particular two examples refer to June 1st (2012) when the city has reached the historical record of traffic jam (295 Km., out of the around 800 that are monitored by CET), as in `http://blogs.estadao.com.br/transito/2012/06/page/10/`:

8:15PM - Paulista Avenue has 3.2 km of heavy traffic, direction Consolação Avenue, from Oswaldo Cruz Sq. until Augusta Street.

8:19PM - Brasil Avenue has 3.2 km of slow traffic, direction Ibirapuera Avenue, from Venezuela Street until Pedro Álvares Cabral Avenue.

In order to process information like this, a tool was developed to extract specific terms from these blog entries. In this particular case we extract the following attributes: date, time, road, location / traffic direction, and status. For example, from the first blog entry just given, our tool extracts the following values for the given attributes:

1. Date: June, 1st, 2012
2. Time: 8:15PM
3. Traffic direction: "to CONSOLAÇÃO AVENUE"
4. Road/street: "PAULISTA"
5. Status: "HEAVY TRAFFIC"
6. Location: "FROM OSWALDO CRUZ SQUARE UNTIL AUGUSTA STREET"

Values for these attributes were collected during a period of time and stored in a database. This way, we have the most common values occurring for each attribute.

The attribute status is the most important. Given one text value for a status (e.g., heavy traffic in the just mentioned example), we assign a numerical value to it. This value is used to "inflate" the travel time of each edge in the path given by the attributes Road-/Street and Location, thus acting as weight / cost for the shortest path algorithm. So far we are grouping all values that appear in the website into four groups and assigning the following weights for the traffic statuses: 1 if links are under free flow conditions, 1.6 if the status is slow traffic, 12 if it is jammed, and 21 if it is stopped. These values are computed using a volume-delay function (Eq. 1), largely used in transportation engineering [Ortúzar and Willumsen 2001], where t is the travel time per unit of distance on the link (such as min/km), t_{ff} is the travel time per unit of distance under free-flow condition, v is the current volume on the link, C is the link's nominal capacity, and a and b are calibration parameters (here, set to 20 and 5 respectively). Thus, for example, if $\frac{v}{C} = 0.5$, then $t = 1.6 \times t_{ff}$.

$$t = t_{ff} \times [1 + a(\frac{v}{C})^b] \tag{1}$$

A difficulty here is to match the values of the attributes Location, Road/Street and Traffic direction with the objects of the map, given that there is a huge number of such objects in the map of the city of São Paulo. In order to extract these informations from the website of the newspaper, the sequence of steps is:

1. query to the mentioned website, where the information is captured.
2. identification and extraction of attributes from blog entries.
3. generation of a CSV file with all blog entries in segmented components extracted in Step 2.

The extraction tool was developed using Java; Apache Web Server; Apache Tomcat; Ubuntu Linux; and an open source database. Also, we used a lexical analyzer and regular expressions to identify the correct "Location" attribute.

The CSV file is then processed to generate the numerical values that are used as cost in the shortest path algorithm. It is possible to use SUMO's native shortest path algorithm or an implementation of A* that we have added to SUMO. The final step is to run the simulation, considering a given number of trips (user defined), each of which is determined by an OD pair.

3.4. Synthetic Tweets

Given the challenges associated with processing the Twitter information coming from the OESP website, (e.g., linking geo-referenced objects to information in natural language), in order to assess results and benefits, our methodology also includes the generation of synthetic tweets, which are generated during simulation time. This works as follows.

As the simulation proceeds, the average volume at each link is measured. The current volume is then compared with the average volume. If the volume in a link is higher than the average, a tweet about that specific link is generated. For the experiments reported in this paper, tweets were generated if volume is 30% higher than the average volume.

A synthetic tweet has the same information as the ones coming from the OESP blog, with the exception that tweets generated within the simulation have a precisely known location, and therefore are easily associated with the corresponding links.

3.5. Routing and Modeling of Drivers

The motivation for our work is to simulate what happens if routes are calculated by the drivers themselves, when some of them get information (in this case from Twitter). To do this, we simulate the fact that the information described in the previous section is processed and given to a share of drivers (as not all drivers may have access to the information).

Drivers have a knowledge base (KB) where not only their origin and destination is kept, but also the estimated travel time. Moreover, these drivers constitute a heterogeneous population in the sense that they have different information. They are able to recalculate their travel times, compare them to the expected ones and replan their routes during the trip (en-route planning). A new route will be adopted if the following condition is satisfied: expected travel time is higher than travel time in the newly computed route, times a factor DF. DF is the delay factor, a parameter of the model.

Three routing strategies were used. The first is based on SUMO's default route generation for trip assignment: given an OD, for each trip, find a route between the origin and the destination using Dijkstra algorithm, where the cost is the edge length.

The second is our implementation of route computation, based on the A* algorithm, in which cost are the lengths of each edge, multiplied by a constant, depending on the current edge status, as explained before. This way, a completely blocked edge is multiplied by a factor of 20 and so on for less severe congestion statuses.

Finally, the third is SUMO's dynamic user assignment (DUA), an iterative process that seeks the user equilibrium based on Gawron's method [Gawron 1998].

4. Experiments

In this section we report our preliminary results using synthetic tweets. We remark that the only difference is that by using the synthetic data we do not have to make the association between a position in the map and the non-referenced information coming from the OESP tweets. Although this is already implemented and tested, so far we have no results using actual tweets because simulations are very time consuming. However we note that there is basically no difference in the methodology for processing actual or synthetic tweets.

4.1. Metrics

In order to assess the results of the experiments that were performed, we use the following metrics:

1. travel time efficiency: mean travel time when tweets are used divided by the mean travel time when tweets are not used (this means a simulation under the otherwise same conditions is repeated with and without use of tweets).
2. number of waiting vehicles divided by the overall demand: this is an indirect measure of overall congestion of the network; because SUMO cannot insert a driver in its origin if the corresponding link is congested, it will keep such vehicles in a list of waiting vehicles; ideally this would be a low number but in fact many vehicles have to wait to be inserted in the simulation.
3. speed of the link divided by the free flow speed: if there is no congestion, this factor is close to one because drivers can travel at free flow speed; however, if a link is congested at some extent, then its speed is lower and hence the quotient is lower than one.

4.2. Parameters and their values

The first parameter of the model is the delay factor DF, used by the driver to replan its trip. This means that a new route will be adopted if its travel time is higher than the expected travel time in the original route times DF. Values of 1.0 and 1.3 were used for DF. The other important parameter is the number of drivers that have access to the information coming out of the tweets. We have used 10% and 50% because these are two quite different situations. Henceforth we denote this quantity by %ID (percentage of informed drivers).

4.3. Results

Next we discuss our preliminary results, focussing on those for DF set to 1.0. We discuss the results in terms of efficiency and number of waiting vehicles. Speed is basically highly correlated with travel time thus we skip this point here.

Table 1 shows the travel time efficiency in two cases: 10% and 50% of informed drivers (ID). We show results for three populations of drivers, which is something that only microscopic, agent-based simulations can provide since they can deal with heterogeneous populations of agents (in this case, drivers). We start with results related to the overall efficiency, i.e., measured over all drivers. As seen, the travel time is nearly the same when only 10% of drivers are informed, compared with the situation in which no information is given. However, when 50% of the drivers are informed, the efficiency factor is below 1 meaning that overall travel time has decreased in this case.

More interestingly, the increase in efficiency is more significant for the sub population of drivers that receive information. As seen in the third column, these have travel times that are indeed around 20% lower, compared to the situation in which no information is given.

As shown in the last column, the 90% and 50% of *non* informed drivers respectively have an increase in their travel times, when information is given to the corresponding share of drivers. One explanation for this could be that drivers who used to use free

	Efficiency		
%ID	All Drivers	ID	non ID
10%	1.0099	0.8475	1.1602
50%	0.9158	0.8031	1.0704

Table 1. Travel time efficiency (travel time with tweets / travel time no tweets)

portions of the network (i.e., those not subject to tweets as they were not jammed), now have to share their routes with informed drivers who deviate from their original routes and hence compete with "old costumers" of such routes. This result is akin with those found in [Bazzan and Azzi 2012] where travel times were compared for increasing penetrations of use of navigation devices: It was shown that drivers of non-arterial roads had an increase in travel times due to the competition with owners of navigation devices, who would venture using unknown routes, instead of just known arterials.

Next we discuss the number waiting vehicles (normalized by the total of vehicles) along simulation time. The effect of the tweet information can be also seen in the reduction of the number of waiting vehicles. Figure 3 depicts waiting vehicles for the 3 cases: drivers have no information, 10% and 50% of the drivers receive Twitter information.

In Figure 3, it is possible to see that giving information about congestion via tweets also helps to reduce the number of waiting vehicles. Clearly, by receiving information, drivers relocate their trips, the overall distribution is better than when everyone uses its shortest path, and hence, less drivers have to wait and the waiting times are smaller. When no information is given, the peak of waiting vehicles reaches 65% of the total demand at the beginning of the simulation, when many drivers want to start their trips. If information is given about jammed links, some drivers divert, thus leaving more room for those in the waiting queue to enter the simulation sooner.

5. Conclusion

Our long term goal is to develop an agent-based infrastructure (possibly over SUMO) that can be used for investigations that are key to evaluate the effects of concepts related to smart transportation systems. Being microscopic, this infrastructure demands an amount of data that few traffic authorities can afford in countries like Brazil. An alternative to installing sensors is to use human as sensors, as explained. However, this poses challenges for the gathering and management of such data.

In this paper we propose a methodology to deal with this problem. It aims at capturing and treating traffic data (mainly streets and links statuses) that appear in social networks, microblogs, etc. Specifically, we illustrate the approach with data that appears in the blog "Trânsito" that is managed by OESP, online edition.

We have implemented a prototype of such an approach to a given extent. With this prototype, we have simulated thousands of agents that adjust their own route based on knowledge of the traffic network status.

By means of the agent-based approach we were able to derive conclusions that would not be possible if only macroscopic methods were used, as for instance the fact that there are differences in the travel time of drivers that receive information.

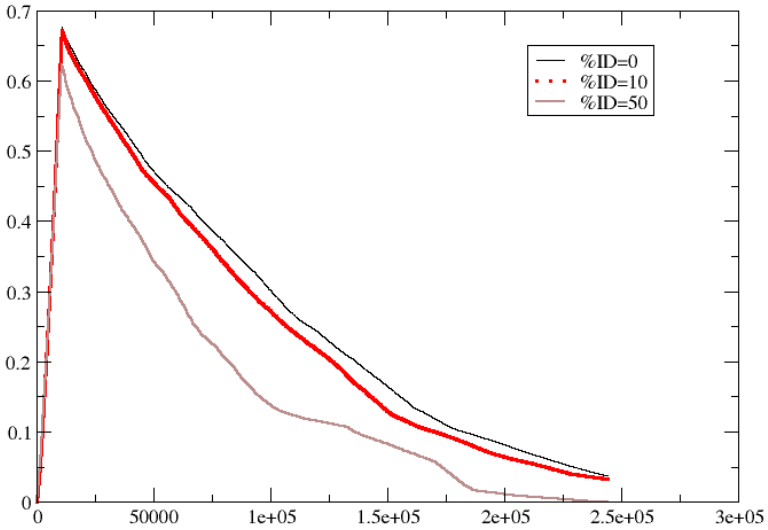

Figure 3. Waiting vehicles (ratio waiting / total vehicles) along simulation time

Work is being done regarding matching the location that appear as part of the information carried by the tweets, with its object in the simulation. Alternatively, we are looking for other sources of information, preferentially geo-referenced.

Another direction of future work is the application of a similar methodology to cases in which data comes directly from sensors installed in the vehicles.

Acknowledgements

We would like to thank the following funding agencies for their partial support to the authors and to the projects: FAPERGS (RS-SOC project), CTIC/RNP (SIMTUR project), and CNPq.

References

Balmer, M., Cetin, N., Nagel, K., and Raney, B. (2004). “Towards truly agent-based traffic and mobility simulations”. In Jennings, N., Sierra, C., Sonenberg, L., and Tambe, M., editors, *Proceedings of the 3rd International Joint Conference on Autonomous Agents and Multi Agent Systems, AAMAS*, volume 1, pages 60–67, New York, USA. New York, IEEE Computer Society.

Bazzan, A. L. C. and Azzi, G. G. (2012). “An investigation on the use of navigation devices in smart transportation systems”. In *SBSI Anais do VIII Simpósio Brasileiro de Sistemas de Informação*, volume 1, pages 156–161, São Paulo/SP, Brasil. SBC.

Bazzan, A. L. C., de B. do Amarante, M., Azzi, G. G., Benavides, A. J., Buriol, L. S., Moura, L., Ritt, M. P., and Sommer, T. (2011). “Extending traffic simulation based on cellular automata: from particles to autonomous agents”. In Burczynski, T., Kolodziej,

J., Byrski, A., and Carvalho, M., editors, *Proc. of the Agent-Based Simulation (ABS / ECMS 2011)*, pages 91–97, Krakow. ECMS.

Bazzan, A. L. C. and Klügl, F. (2013). A review on agent-based technology for traffic and transportation. *The Knowledge Engineering Review*, FirstView:1–29.

Behrisch, M., Bieker, L., Erdmann, J., and Krajzewicz, D. (2011). "SUMO - simulation of urban mobility: An overview". In *SIMUL 2011, The Third International Conference on Advances in System Simulation*, pages 63–68, Barcelona, Spain.

Chmura, T. and Pitz, T. (2007). An extended reinforcement algorithm for estimation of human behavior in congestion games. *Journal of Artificial Societies and Social Simulation*, 10(2).

de Lima, V. G., de M. R. Magalhães, F., de O. Tito, A., dos Santos, R. A., Ristar, A. R. R., dos Santos, L. M., Vieira, V., and Salgado, A. C. (2012). "UbibusRoute: Um sistema de identificação e sugestão de rotas de ônibus baseado em informações de redes sociais". In *VIII Simpósio Brasileiro de Sistemas de Informação, (SBSI 2012)*, São Paulo. Sociedade Brasileira de Computação.

Gawron, C. (1998). *Simulation-based traffic assignment*. PhD thesis, University of Cologne, Cologne, Germany.

Jaques, P., Pasin, M., Chiwiacowsky, L. D., Bazzan, A. L. C., Moraes, R., and Bastos, R. (2012). "Provendo informações para atores do sistema de transporte público: um passo na direção de sistemas inteligentes de transporte". In *XXVI ANPET - Congresso de Pesquisa e Ensino em Transportes*, Joinville, SC. ANPET.

Koutsoupias, E. and Papadimitriou, C. (1999). "Worst-case equilibria". In *Proceedings of the 16th annual conference on Theoretical aspects of computer science (STACS)*, pages 404–413, Berlin, Heidelberg. Springer-Verlag.

Ortúzar, J. and Willumsen, L. G. (2001). *Modelling Transport*. John Wiley & Sons, 3rd edition.

Passos, L. S., Kokkinogenis, Z., and Rossetti, R. J. F. (2011). Towards the next-generation traffic simulation tools: a first appraisal. *3rd Workshop on Intelligent Systems and Applications (WISA), 6th Iberian Conference on Information Systems and Technologies (CISTI'11)*.