ISSN 2007-9737

Agent-Based Modeling for Evaluation of Transportation Mode Selection in the State of Guanajuato, Mexico

David Salas-Rodríguez[1], Luis Arturo Rivas-Tovar[2]

[1] Instituto Tepeyac,
Mexico

[2] Instituto Politécnico Nacional,
ESCA STO,
Mexico

{davino66, larivas33}@hotmail.com

Abstract. One of the negative consequences of the industrialization of Mexico favored by the North American Free Trade Agreement (NAFTA), is the emergence of huge industrial corridors associated with the demand for mobility by commuters who move to their workplace. The demand produces mobility patterns that have a serious impact on air pollution in five cities in the state of Guanajuato that, despite being medium in size, outnumber Mexico City in pollution. The objective of this work is to model a data-driven agent based on the beliefs-desires-intentions model, to predict the selection of transport modes using a J48 decision tree algorithm that was designed from data from the 2015 national census (INEGI). The method is mode based I agent programmed in Net logo. The results show that: it is possible to predict the demand of transport considering the: gender, level of education, transfer times and age in the five cities of Guanajuato, in a horizon of three years. With changes in public policies related to mobility and changes in transportation patterns, air pollution would be reduced. The proposed model could be used to support public policies that improve mobility and positively impact air quality in five cities in the state of Guanajuato.

Keywords. Data-driven, agent-simulation, J48, kappa-index, MCCI, Air pollution, Guanajuato Mexico-

1 Introduction

The North American Free Trade Agreement (NAFTA) started on January 1, 1994, as a consequence, Mexican exports grew steadily. In 22 years, the Mexican economy has been transformed: 82% of its foreign sales are manufactured products. in 1993 oil exports represented 10% of the Gross Domestic Product (GDP) but 40% of government incomes, in 2018 was only 6% and 8 % government incomes At present, Mexico's export profile is made up of aerospace parts and equipment, optical equipment, machinery, electrical and electronics, and author and auto parts.

In 2022, Mexico is seventh producer worldwide. 25 years after the signing of the NAFTA by Mexico, 89.7% of Mexico's exports were manufactured. Mexico went from being an exporting country of oil and raw materials to being an industrialized country.

One of the most negative externalities of NAFTA has been the pollution or the air of its communities as a result of its successful industrialization. Mexicali, Toluca, Ecatepec, Tlalnepantla, Netzahualcoyotl, Salamanca, Leon, Celaya, Irapuato, Mexico City and Monterrey are the most polluted cities in Mexico with PM2.5 particles, which present a medium and long-term health risk.

They exceeded the limit of PM2.5, which is from 0 to 10 μg / m³ of the WHO [37]. The first nine cities plus Monterrey are home to the industrial plant that was created after the industrial boom that transformed the country after 1994.

While the air pollution in Mexico City is explained by its 5.5 million cars, the others cities of this sad and black ranking, are medium, its pollution is associated with its accelerated process of industrialization as a result of NAFTA, now a days T- MEC.

1.1 State of Art in Data-Driven Agent-based Modelling

Agent-based modelling (ABM) aims to simulate human behavior in systems using a programming language. Fundamental to this methodology is the abstraction of human behavior with a prediction function based on quantitative and/or qualitative data, and such approaches are often considered most effective when they are kept as simple as possible (principle by Terano) [34].

One example research that used agent-based modelling was the simulation of pedestrian evacuation in the event of seismic risk at an urban scale in a study that modelled human behavior based on observations from video recordings of real events [5].

Ng, Eheart, Cai, & Braden used agent-based modelling with a Bayesian inference to analyses decision-making among farmers regarding water quality impacts at a watershed scale in carbon emission markets fielding a second-generation biofuel crop [23].

Azar & Menassa developed a model behavior occupant in commercial buildings and their impact on energy use combined a quantitative method of measuring energy use with qualitative techniques for identifying occupant behavior [2].

Klein, Kwak, Kavulya, Jazizadeh, & Becerik-Gerber proposed other model of building occupant behavior, combined sensor data and electronic building controls to reduce energy use using a multi-objective Markov decision problem to determine occupant [14].

Finally, Arel, Liu, Urbanik, & Kohls [1] designed a model of traffic signal control, governed by an autonomous intelligent agent was modelled using neural networks.

Dynamics in ABM have been presented in two dimensions: at microscopic level, which the agent level is represented by the evolution of its dynamic attributes (variables and models) caused by its interaction with other agents or the environment.

At macroscopic level is a consequence of the microscopic dynamics and its mathematical interactions agent representation is not a trivial issue and the model can only address the interaction as the functional relationship between states and parameters [25].

In recent research based on agents, this microscopic dynamic is modelled by deterministic empirical models [26]; theoretical models as objective functions and evolutionary algorithms [9, 19, 20], theoretical empirical such as interviews and algorithms [17]; logit regression [41] or novel theoretical approaches to limited rationality such as Frank-Wolfe's linearization method with the Generalized Benders' Decomposition method [21].

A novel approach is the data-driven agent model [13] that lies between the extremes of totally empirical and totally abstract models; it consists of empirical models elaborated from the data for the dynamics at the micro level using big data and the KDD process (Knowledge Discovery on Database) [8].

Recent research uses big data and machine learning for agent models such as neural networks [6, 40] and complex networks [35].

The aim of this paper is describe beliefs-desires-intentions (BDI) using a data-driven agent-based model to predict the behavioral changes of commuters who must select a means of transportation and thereby affect the dynamic demands on transportation infrastructure using novel data-driven approach with census data and J48 algorithm machine learning tool proposed by Witten & Frank [39].

This is the first work related to five cities of Guanajuato State in Mexico; in order to understand dynamics at the micro level in the selection of the means of transport and at the macro level as the emergent mobility behavior in the five cities.

1.2 Five Cities on the State of Guanajuato

Guanajuato state is located in the center of Mexico with an area of 30,460 km^2, which represents 1.6% of the national territory.

The state comprises 46 municipalities with a total population of 6,166,934 inhabitants, Guanajuato ranks 6th nationally for its number of inhabitants [10], Guanajuato more polluted cities are Leon, Silao, Salamanca, Irapuato and Celaya. Together forms one of the more important industrial cluster in America.

Leon. Located in the East of the State with a population of 1 578 626 inhabitants [11], it is the

largest metropolis in the state; in the last decade it has shown strong growth.

With data from the intercensal count [10] of the inhabitants who move from home to work, the use of public transport predominates with 37%, followed by the use of private cars with 30% and the use of labor transport labor by 7% (labor transport is the means of transportation that companies provide their employees using leased buses).

It is the municipality in the State that has the most extensive bike path network with a total of 108 kms. [18].

In Leon, there is a unique integrated transport system known as Optibus; this is formed by a combination of a bus transport subsystem and a bus rapid transit subsystem. Silao de la Victoria.

Located in the East of the State bordering León and Irapuato and with a population of 189,567, it is rapidly industrializing due to having one of the first automotive plants in the State.

It has an index of registered motor vehicles of 250 (This index is calculated as the number of motor vehicles registered in circulation divided by the estimated total number of population multiplied by 1000. Less is better, the 2015 State average is 270 [12], 20% of commuters use private vehicles, 19% use public transport, and 27% use labor transport.

Irapuato. Located in the south-central part of the State, there is a mixture of agricultural, commercial and industrial activity, with a population of 574 344 and an index of registered motor vehicles of 307, the use of private vehicles predominates with 35%, public transport with 22% followed by 13% of labor transport.

Salamanca. Located in the central zone of the State, there is a mixture *of* agricultural and industrial activity, with the large PEMEX refinery Antonio M. Amor. It has a population of 273 271 inhabitants and an index of registered motor vehicles of 352. Commuters*' use* of private vehicles predominates with 34%, followed by public transport with 22% and 11% for labor transport.

Celaya. Located in the southwestern area of the state, its main economic activity is trade and services followed by industry. It has 494 304 inhabitants, being the third largest metropolis in the state. Its index of registered motor vehicles is 338

Fig. 1. Map of Guanajuato State

and the use of the private vehicle predominates with 34%, followed by public transport with 30% and only with 5% for labor transport.

Figure 1 shows a map of the state, and Table 1 summarizes the population density and the number of observations used for this study [30].

2 Method

We define "agent" as a real entity that has the ability to assimilate information from its environment (input), reason (logical processes), and respond to the environmental input with a decision (output) that results in a behavior.

In operative terms an agent is a resident and commuter of the five Guanajuato cities who moves from home to work.

Consider now all the inhabitants of a city who need to move from their homes to their workplaces. Each of these commuter's reasons is based on a series of personal attributes to arrive at a decision regarding which means of transport to use, resulting in a series of behaviors that will contribute to patterns that arise with respect to the mobility of that city's residents.

Thus, a single agent exhibits individual behaviors that will cause the emergence of patterns in a system of multiple agents.

Agent-based modelling is a technique arising from systems engineering that allows modelling complex systems formed by categorized agents who are related to each other and whose interactions cause the emergence of observable patterns.

The methodology is used to simulate emerging behaviors resulting from the dynamics of socio-ecological systems [32].

The foundational elements for modelling the relationships and interactions among the agents in the system are: the data (agent attributes) and the functions that will determine the changes in those attributes in each iteration in the simulation process.

The feedback from each iteration determines the dynamics of the modelled system, starting from a given initial state [33].

The BDI agent-based model [24] includes (1) a belief system that represents the agent's values or knowledge of the environment as attributes the agent develops over time, (2) a system of desires that represents the agent's established goals related to those beliefs (here considered to be the desire to get to and from work), and (3) a system of intentions that represents the actions aimed at achieving the desired objective (here considered to be the intention to use a specific transport mode to move to and from work).

2.1 Algorithm and Language for Data-Driven ABM

Weka is a software platform developed by the machine learning group at the University of Waikato. It includes a collection of machine learning algorithms for data mining models [36].

Net Logo is a multi-agent programmable modelling environment developed by Uri Wilensky at the Centre for Connected.

Learning and Computer-Based Modelling, Northwestern University [38]. This environment allow the programming of models based on machine learning algorithms through its control structures.

The Net Logo environment enables agent-based modelling by facilitating the programming of agent intentions and their execution in a very simple way, based on code defined by the user and data that inform the creation of the intentions list.

Further details on the manner in which a BDI model is implemented in Net Logo, can be found in Sakellariou [29].

[1] INEGI is the national institute of statistics geography and informatics.

Table 1. Summary of data sample size

Metropolitan area	Population	INEGI[1] commuter agents for ABM (sample size)
Leon	1 578 626	37 400
Silao de la Victoria	189 567	6 692
Irapuato	574 344	11 405
Salamanca	273 271	7 430
Celaya	494 304	9 997
Total	3 110 112	72 924

A classifier algorithm is needed to provide an abstract data-driven model of the reasoning process based on categorical and numerical attributes.

The J48 algorithm allows quantitative and qualitative predictor variables to be used to construct a classifier tree to predict the dependent variable.

Unlike other classification algorithms such as the a priori method (also provided in Weka), in which only binary variables can be introduced to produce association rules in the form $p \rightarrow q$, the J48 classifier was programmed using if ... then... else control structures as the transition function used to determine the agent's selection of a mode of transportation.

The research used an official data sample source that included 72 924 surveys of the inhabitants of Guanajuato state's five mega-cities [10].

To model the agents, the method of Drogoul, Vamderue and Meurisse [7] is adopted for: a) abstract transportation mode selection based on the J48 decision tree, b) formal transportation mode selection, and c) the application of the model in the programming language used for the simulation. Figure 2 shows the methodology for the agent-based model.

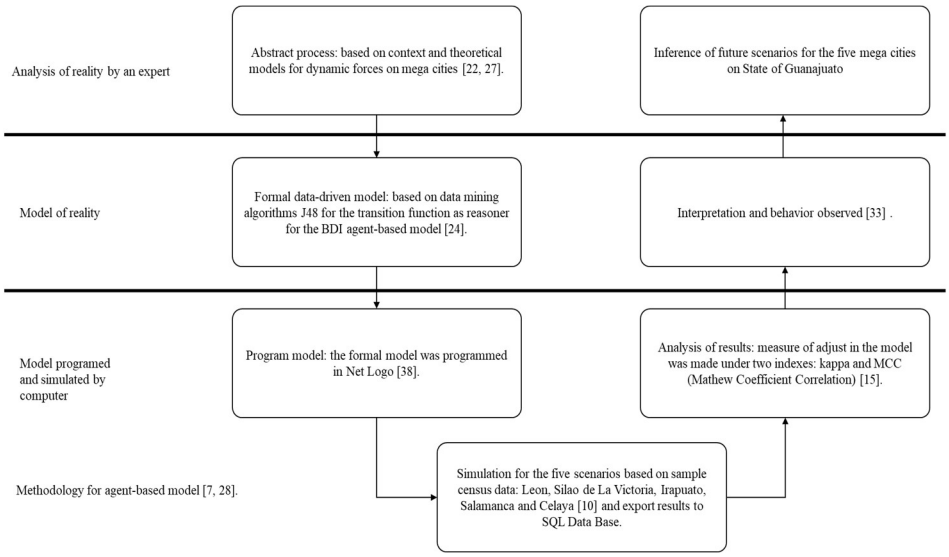

Fig. 3. Methodology for transport demand by an agent-based data-driven modelling

2.2 Scenarios for the ABM for Commuter Patterns

The initial scenarios were configured with agents whose maximum life span was 90 years, modelled independently by city, and the simulations were run to model a horizon of three years, in one-year intervals. The commuter is represented as an independent agent in the model who lives in one of the five cities.

The *initial state* of simulation are 72 924 agents modelled as the commuter types with the attributes of gender, academic level completed, age, means of transport with time of transfer for each, and the city where the agent lives.

The overall choice for transport was: Private vehicle 22 542; Bus, taxi or collective 22 225; Walk 10 835; Bicycle 9 429; Work´s transport 5 694; Not Specified 2 116 and BRT or light rail 83.

The means of transport is identified in the model as the initial means of transport for each agent and is used for comparison with the simulated means of transport at the end of the simulation. Of these agents, those that express the use of more than one means of mobility were 1,951 for two options, and 114 for three options.

The agent-based model simulates the means of transport for each independent agent based on a J38 decision tree modelled from census data. Table 2 shows a summary of the 75 070 observations of the census data used as a set for machine learning training of agents' formal predictive function using WEKA.

At each iteration (tick) of the agent-based model in Net Logo, the reasoned agent function (formal predictive function) uses as scenario input parameter: academic grade, gender, age and city and predicts the mean of transport (as output).

At the *final state* of simulation, a file is generated with, for each agent, its initial means of transport as well as those predicted by the model (can be more than one) and the accumulated transfer time for each one.

Each iteration of the program corresponds to one hour of the day, using a NetLogo extension that establishes date/time utilities and discrete event scheduling to simulate the dynamics in transportation demand [31].

The time that each agent takes from leaving home until arriving at the workplace is established randomly, considering the normal distribution specified in the census data.

Table 2. Data modelled travelers in the five cities studied

	Celaya		Irapuato		Leon		Salamanca		Silao	
Mean of transport	male	female	male	female	male	female	male	female	male	female
Not Specified	142	95	179	117	360	262	149	103	102	52
Bicycle	1 422	137	1 489	70	4 164	156	1 372	66	869	46
Walk	747	654	985	720	3 625	2 917	522	399	794	598
*** Bus, taxi or collective**	1 413	1 566	1 052	1 431	7 624	6 255	674	923	734	543
BRT or light rail	0	0	0	0	58	63	0	0	0	0
Other	58	30	110	21	245	48	42	6	35	7
Work´s transport	407	133	1 110	415	881	248	645	181	1 203	648
*** Private vehicle**	2 349	1 172	2 767	1 258	7 751	3 705	1 724	864	995	363
Totals	6 538	3 787	7 692	4 032	24 708	13 654	5 128	2 542	4 732	2 257
									Total of commuters in census dataset	75 070

* Predominant means of transport in the five cities

The time spent at work is determined randomly, with start times between 7:00 and 9:00 in the morning for work days from Monday to Saturday, and work schedules assigned randomly to between five and ten hours.

2.3 Validation for the Selection of Different Values for Parameters in Data-Driven ABM

The data-driven BDI agent-based model applied in this study is based on INEGI 2015 survey-derived demographic characteristics: gender, academic level completed, city, time of transfer and age. In a section of each survey, the inhabitant expressed: the city where he lives, gender, age, academic level completed, time of transfer to his place of work and up to three Decision.

The selection of the variables of the census questionnaire was carried out by calculating the information gain, selecting the next: Academic level with gain of 0.1752, Gender with 0.0734, City with 0.0637 and Age with 0.0373.

Trees are algorithms used by Maimon and Rokach [16] to predict an output variable based on predicates, which in this case are qualitative and quantitative agent attributes.

The most well-known algorithm of this type is C4.5, which in its most recent version is referred to as J48, implemented in the Weka data mining tool [54].

This tool is efficient and capable of handling large training sets. The precision of this classifier tree model is measured with the kappa (κ) statistic like Landis and Koch [15].

Table 3. Confusion matrix of the J48 decision tree results as output of WEKA for initial scenario

Bus, taxi, or collective	Private vehicle	Bicycle	Walk	Work's transport	Other	Not specified	BRT or light rail	<-- classified as[2]
14 491	4 884	1 843	351	646	0	0	0	Bus, taxi or collective
5 993	14 184	2 034	178	559	0	0	0	Private vehicle
4 064	1 319	**3 789**	233	386	0	0	0	Bicycle *
7 253	1 805	1 804	518	581	0	0	0	Walk
1 529	1 709	1 037	234	1 362	0	0	0	Work´s transport
276	132	154	14	26	0	0	0	Other
664	507	269	41	80	0	0	0	Not Specified
85	33	3	0	0	0	0	0	BRT or light rail
TP Bicycle: 3 789								
TN Bicycle: 4 925								
FP Bicycle: 426								
FN Bicycle: 619								

This index is used to evaluate inter-observer agreement for categorical data, similar to the ANOVA applied to quantitative data, values less than zero are poor, values between 0.21 and 0.4 are fair, between 0.6 and 0.8 is desirable, values greater than 0.81 are almost perfect adjustments.

In this research, this classifier corresponds to the prediction of a commuter's selection for his transportation mode. The Matthews correlation coefficient (MCC) [4] is used for classifications between classes (predicates) in unbalanced data (data with disproportionate frequencies).

An unbalanced data is a data set where the values are grouped by categories and the frequencies are not proportional.

The MCC provides a measure of classification performance in a set of categorical data and can be seen as a discretization of the Pearson correlation for binary variables.

The coefficient is always between -1 and +1, a value of -1 indicates total disagreement with the classification and 1 a perfect classification; value of 0 is for completely random classification (prediction) [3, 4].

For this research, this measure is applied for the measure classification performance in the formal model and ABM model. One of the central aspects of our model is that the algorithm assumes that agents have different transport preferences depending on gender, age, city and educational level.

Thus, women and young people studying at university will be modelled as using public transport while a middle-aged man with university education will be modelled as using private transported means of transport used for that purpose.

The agent decision that the program simulates is the selection of an intended transportation mode for commuting purposes. In the program, the agent's age is a variable that advances, possibly resulting in the selection of a different transportation mode.

[2] The figure 4 shows the simulation of transportation choices in the city of Silao de la Victoria. Which was chosen to represent an intermediate case of the 5 cities. Due to space problems, the other 4 are not shown, however, the agent model developed predicts the behavior of the inhabitants of the other four cities.

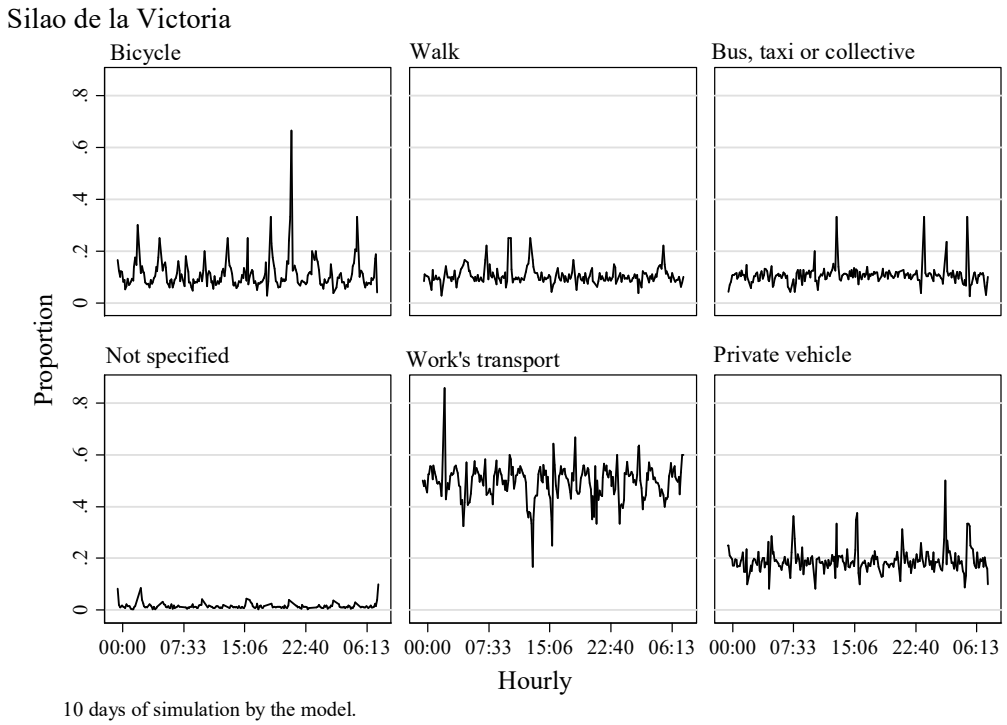

Fig. 4. Diurnal oscillatory behavior results for Silao de la Victoria

For each agent, the model stores as beliefs a variable representing the cumulative total hours of use of each transportation mode in each iteration of the simulation.

The manner in which the states of agents who the simulation failed to categorize was determined as follows. Each time the model evolved, the program counted an hour for the category each agent selected as a belief.

For example, during the simulation, an agent who first walks and then uses a bicycle will have accumulated the first hours for walking and the next for bicycling.

In addition, the initial and final categories were stored as variables (InitialID and id-group, respectively), and an algorithm was applied to compare these values such that the counted beliefs in each category corresponded with the total accumulated hours.

Errors were identified when the initial or final categories hour accumulator is equal to 0 on results. Two states of the model were considered: the initial state q_i corresponding to initial data (Tables 2 and 3) and the final state corresponding to the data exported from the model (Table 4).

The final state consists in a dataset who stores the agent's attributes and the hour counter by each category on mobility that evolve in the simulation.

3 Results

In state of the art review, we could not find an investigation that allows us to compare the findings with previous works, however behavior observed in the hours of use by transportation mode indicated an oscillatory system over a day.

The classifier (or prediction) algorithm produces four outcomes: (1) True Positive (TP), which corresponds to the cases that are correctly classified in the predicted commuter means of transport; for example, it is a commuter that in the observations shows that it uses a bicycle and the algorithm predicts it as such, (2) True Negative

Table 4. Observed and predicted results

	Initial State	Final State	Percent change	Sensitivity (TP ratio)		Percent change	MCC		Percent change
	q_i	q_f	∂	q_i	q_f	∂	q_i	q_f	∂
Walk	10 835	1 224	-89%	0.043	0.064	49%	0.068	0.137	101%
Bicycle	9 429	9 770	4%	0.387	0.41	6%	0.265	0.309	17%
BRT or light rail	83	0	-100%	0	0	0%	0	0	0
Bus taxi or collective	22 225	32 194	45%	0.652	0.677	4%	0.253	0.29	15%
Work's transport	5 694	3 568	-37%	0.232	0.278	20%	0.249	0.304	22%
Private vehicle	22 542	26 078	16%	0.618	0.661	7%	0.411	0.46	12%
Not specified	2 116	41	-98%	0	0.02	0%	0	0.137	14%
Frequency observed in demand for mobility in final state data									
One option for mobility	70 859	69 591	-2%						
Two options	1 951	3 184	63%						
Three options	114	100	-12%						
Statistics									
Po	0.458	0.495	8%						
Pe	0.262	0.268	2%						
Kappa (κ)	0.265	0.311	17%						

*q_i indicates the initial state and q_f indicates the final state of simulation

(TN), which corresponds to the sum of cases which are correctly classified as are not for the predicted category.

For example, they are all the commuters that in the observations show that they do not use a bicycle and the algorithm predicts it as such.

False Positive (FP) are the sum of cases which are classified in a predicted category but not correspond to it, as an example are all commuters that in the observations present that they do not use bicycles and the algorithm predicts that this is the medium they use and False Negative (FN) are the sum of cases which are classified as are not for the predicted category but correspond to it. As an example are all the commuters that in the observations show that they use a bicycle and the algorithm predicts another means of transport.

These values are obtained from confusion matrix as shown in table 3.

For example, the formal model predicts the transportation mode selection for commuters that live in the five cities summarized in the Table 2.

3.1 Predicted Commuter Transportation Mode Selection Patterns to Simulate the Evolution of Transport Activity

Each leaf of decision tree corresponds to a predicted transportation mode selection modelled as control structure in Net Logo program language.

For example, if the agent has an academic grade equal to primary and gender is equal to women, and lives in (Celaya, Silao de la Victoria, Leon-Salamanca, Irapuato)[3] then use bus, taxi or collective; this predicted result corresponds to one leaf of 101 possible predicted results by the model for the five cities.

The 75 070 instances correspond to commuters in census dataset used to build the tree model in Weka, distributed in the categories of transport: predominates Bus, taxi or collective 22 215 and Private vehicle 22 948; Not Specified 1 561; Bicycle 9 791; Walk 11 961; Work´s transport 5 871; BRT or light rail 121 and Other 602.

However, a little subset of 38 of commuters from 121 have academic grade of secondary, 14 women with an age mean of 29 years and standard deviation of 11 and 24 men with an age mean of 31 years and standard deviation of 10; this subset commuter description is a brief example as each one is modelled independently.

The proposed data-driven model generates a confusion matrix describing the mismatch between modelled results and survey answers. In Table 3, a Commuter classified as a walker, green and red color are correctly (True predictions) predicted, blue and grey are errors (False predictions).

In the matrix, the rows correspond to the survey answer, as an example, the bicycle means of transport is used, the sum of the row corresponds to 9 791 commuters who in the survey used this means of transport; the columns correspond to the value predicted by the algorithm, means that only 3 789 were correctly predicted, the rest of the columns were erroneous predictions: 4 064 were predicted as Bus taxi or collective, 1 319 as Private vehicles, 233 as Walk and 386 as Work's transport.

4 Discussion

Figure 4 shows ten days average simulation results for one of the five cities Silao de la Victoria[4]. With the behavior of each agent in the use of the means of transport predicted by the model and the time of use modelled in a random way, it is possible to observe the dynamic behavior of the transport demand. Demand for each category of transport is shown as the proportion of commuters in that category who are in transit.

As we can see, the use of work's transport predominates and the lower peak close to 15:00 hours explain that about 20% of commuters who use this means of transport are traveling at that time.

This sample city was selected to show the behavior in the dynamics of transport demand exhibit reality congruences as smallest of the five studied cities where industrial activity predominates and this category of transportation is used to move its employees. The BDI agent-based model showed an improvement in the kappa and MMC indices compared to table 3 and table 4 of the formal model.

This improvement, although it is a fair value in relation to Kappa index showed in Table 4, means that the agent model predicts as final state the means of transport better than the formal model J48 as initial state.

The improvement in the MMC means that the prediction of the means of transport in each agent shows an improvement compared to the initial state, observing that the model predicts each category in a moderate way as Private vehicle with 0.46.

Table 4 summarizes the results, where q_i represents the initial state and q_f represents the final state; ∂ indicates the percent change after the simulation. The sensitivity of the J48 algorithm, indicated by the true positive (TP) ratio, is also shown. MCC was calculated using model result values.

The commuter demand for mobility predicted by the agent-based model indicates changes in the demand for transport over time. As the result of the use of a data-driven agent-based model, the prediction of the transportation mode varies as a function of the attributes of the agent.

If the age changes during the simulation, for example in the case of a 17 years old Irapuato commuter (agent modelled), at the time of

[3] The model generates predictions and a total of 101 sheets describing the future possible states of travellers

exceeding 17 years it changes from walking to the use of bus taxi or collective simulating the evolution in transport dynamics.

The observed dynamics show that commuter preferences converge towards a single transportation mode over time, with some commuters selecting two options and a minimum of commuters selecting three options.

This change in agent preferences was identified by the reasoning function modelled in the program.

The transportation demand over the three-year horizon indicated an increase in the use of taxi, bus or collective of 45% and an increase of 16% in the use of a private vehicle (automobile, van or motorcycle), which are the two main commuter transportation modes.

In the case of the city of Leon, there is an integrated transport system known as Optibus.

This is formed by a combination of a bus transport subsystem and a bus rapid transit subsystem, however commuters perceive it as transportation by bus more than a real BRT as see in table 4 because the model predicted 0 out of 83 observed.

5 Conclusion

The results show that the patterns of selection of mode of transport in the five cities with the highest air pollution in the state of Guanajuato are predicted.

The model based on can predict the decisions of selecting the mode of transport using variables age, gender, academic level and city of residence. The evolution of transport demand in five cities (macro level) indicated an increase in the use of two main modes of transport: bus and Optibus (in the case of León), taxi or collective by 45% and private vehicle (automobile), truck or motorcycle)) by 16%.

These results indicate that the citizens of the cities studied prefer to use private transport since the supply of public transport is of poor quality. However, if the supply of public transport improved citizens would dispense with their cars.

Research shows that models based on data-based agents can help us understand the evolution of a system in which a large number of people make independent decisions.

Likewise, our study shows that it is possible to construct a population-based model based on demographic surveys to predict transport variables considering: gender, level of education, transfer times and age in the five largest cities of Guanajuato, it is also able to predict the preferred means of transportation for the inhabitants of the five largest cities of Guanajuato.

Finally, it is possible to create a data-based algorithm that can simulate day changes in the choice of transport mode at the micro level. These results are similar to the findings of Witten and Frank [39].

The kappa index and the MCC used to measure the adjustments in the model observed for the algorithm based on data, in the final model (predicted scenario) and its comparison, allow observing that the adjustment values vary between regular and moderate.

The proposed model has the feasibility of being applied not only in other Mexican megalopolises with more than 1 million inhabitants the size of Leon, but also in medium - sized cities that have serious pollution problems, favoring the redesign of the transportation systems that are used in Mexico.

They find concessions to powerful interest groups that have prevented the orderly planning of a fundamental public service such as transportation and the mobility of citizens, which defines the quality of life in a city in such a relevant way.

References

1. **Arel, I., Liu, C., Urbanik, T., Kohls, A. G. (2010).** Reinforcement learning-based multi-agent system for network traffic signal control. IET Intelligent Transport Systems, Vol. 4 No. 2, pp. 128–135. DOI: 10.1049/iet-its.2009.0070.
2. **Azar, E., Menassa, C. C. (2012).** Agent-based modeling of occupants and their impact on energy use in commercial buildings. Journal of Computing in Civil Engineering, Vol. 26, No. 4, pp. 506–518. DOI: 10.1061/(ASCE)CP.1943-5487.0000158.
3. **Baldi, P., Brunak, S., Chuvin, Y., Andersen, C. A., Nielsen, H. (2000).** Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics, Vol.

6, No. 5, pp. 412–424. DOI: 10.1093/bioinformatics/16.5.412.

4. **Boughorbel, S., Jarray, F., El-Anbari, M. (2017).** Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. PloS one, Vol. 12, No. 6. DOI: 10.1371/journal.pone.0177678.
5. **D'Orazio, M., Spalazzi, L., Quagliarini, E., Bernardini, G. (2014).** Agent-based model for earthquake pedestrians' evacuation in urban outdoor scenarios: Behavioural patterns definition and evacuation paths choice. Safety science, Vol. 62, pp. 450–465. DOI: 10.1016/j.ssci.2013.09.014.
6. **Drchal, J., Čertický, M., Jakob, M. (2019).** Data-driven activity scheduler for agent-based mobility models. Transportation Research Part C, Vol. 98, pp. 370-390. DOI: 10.1016/j.trc.2018.12.002.
7. **Drogoul, A., Vamderue, D., Meurisse, T. (2002).** Multi agent based simulation: Where are the agents? Lecture notes in computer science, Vol. 2581. DOI: 10.1007/3-540-36483-8_1.
8. **Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. (1996).** From data mining to knowledge discovery in databases. AI Magazine, Vol. 17, No. 13, pp. 37–54. DOI: 10.1609/aimag.v17i3.1230.
9. **Hernández, C. A., Castilla, G., López, A., Mancilla, J. E. (2016).** A multi-objective algorithm NSGA-II for programming of lamination steps in hot steel. Research in Computing Science, Vol. 120, pp. 65–80.
10. **INEGI. (2020).** Inhabitants number. https://www.cuentame.inegi.org.mx/monografias/informacion/gto/poblacion/default.aspx.
11. **INEGI. (2018).** Information by entity. http://cuentame.inegi.org.mx/monografias/informacion/gto/default.aspx?tema=me&e=11.
12. **IPLANEG. (2015).** Index of registered motor vehicles in circulation, 2015.
13. **Kavak, H., Padilla, J. J., Lynch, C. J., Diallo, S. Y. (2018).** Big data, agents, and machine learning: towards a data-driven agent-based modeling approach. In: Proceedings of the Annual Simulation Symposium, Baltimore: Society for Computer Simulation International, pp. 12.
14. **Klein, L., Kwak, J. Y., Kavulya, G., Jazizadeh, F., Becerik-Gerber, B. (2012).** Coordinating occupant behavior for building energy and comfort management using multi-agent systems. Automation in Construction, Vol. 22, pp. 525–536. DOI: 10.1016/j.autcon.2011.11.012.
15. **Landis, J. R., Koch, G. G. (1977).** The measurement of observer agreement for categorical data. Biometrics, Vol. 33, No. 1, pp. 159–174. DOI: 10.2307/2529310
16. **Maimon, O., Rokach, L. (2010).** Data mining and knowledge discovery handbook. New York: Springer. DOI: 10.1007/978-0-387-09823-4.
17. **Martínez, L. M., Correia, G. H., Moura, F., Mendes Lopes, M. (2017).** Insights into carsharing demand dynamics: Outputs of an agent-based model application to Lisbon, Portugal. International Journal of Sustainable Transportation, Vol. 11, No. 2, pp. 148159. DOI: 10.1080/15568318.2016.1226997.
18. **IMPLAN. (2016).** Cycleways master plan. https://implan.gob.mx/pdf/estudios/movilidad/plan-maestro-de-ciclovias-2016.pdf.
19. **Mogale, D. G., Kumar, S. K., Tiwari, M. K. (2016).** Two stage Indian food grain supply chain network transportation-allocation model. IFAC-PapersOnLine, Vol. 49, No. 12, pp. 49–12. DOI: 10.1016/j.ifacol.2016.07.838.
20. **Mogale, D. G., Kumar, S. K., Márquez, F. P., Tiwari, M. K. (2017).** Bulk wheat transportation and storage problem of public distribution system. Computers & Industrial Engineering, Vol. 104, pp. 80–97. DOI: 10.1016/j.cie.2016.12.027.
21. **Mogale, D., Lahoti, G., Jha, S., Shukla, M., Kamath, N., Tiwari, M. (2018).** Dual market facility network design under bounded rationality. Algorithms, Vol. 11, No. 4, pp. 54–74. DOI:10.3390/ a11040054.
22. **Molina, L. T., Molina, M. J. (2002).** Air quality in the Mexico megacity. An integrated assesment. Kluwer Academic Publishers. Vol. 2, DOI: 10.1007/978-94-010-0454-1.

23. **Ng, T. L., Eheart, J. W., Cai, X., Braden, J. B. (2012).** An agent-based model of farmer decision-making and water quality impacts at the watershed scale under markets for carbon allowances and a second-generation biofuel crop. Water Resources Research, Vol. 47, No. 9, DOI: 10.1029/2011WR 010399.
24. **Norling, E., Sonenberg, L., Rönnquist, R. (2000).** Enhancing multi-Agent based simulation with human-like decision making strategies. Multi-Agent-Based Simulation, pp. 214–228. https://doi.org. 10.1007/3-540-445 61-7_16.
25. **Pereda, M., Zamarreño, J. (2015).** Agent based model: an aproach from system engineering. Ibero-American Magazine of Automatics and Industrial Computing, Vol. 12, No. 3, pp. 304–312. DOI: 10.1016/j.riai. 2015.02.007.
26. **Pijoan, A., Kamara-Esteban, O., Alonso-Vicario, A., Borges, C. (2018).** Transport choice modeling for the evaluation of new transport policies. Sustainability, Vol. 10, No. 4, pp. 1230–1252. DOI: 10.3390/su10041230.
27. **Pinhas, A., Shvainshtein, O., Kishcha, P. (2012).** AOD Trends over megacities based on space monitoring using MODIS and MISR. American Journal of Climate Change, Vol. 1, No. 3, pp. 17–131. DOI: 10.4236/ajcc.20 12.13010.
28. **Rivas-Tovar, L. A. (2017).** Preparation of thesis: structure and metodology. Trillas.
29. **Sakellariou, I. (2010).** Agents with beliefs and intentions in Netlogo. http://users.uom. gr/~iliass/projects/NetLogo/AgentsWithBeliefs AndIntentionsInNetLogo.pdf.
30. **Salas-Rodríguez, D., Rivas, L. A. (2017).** Air pollution in five cities in Guanajuato state (México). Conference on Complex Systems.
31. **Sheppard, C. (2018).** A NetLogo extension that brings date/time utilities and discrete event scheduling to NetLogo. https://github.com/ colin sheppard.
32. **Smajgla, A., Brow, D. G., Valbuena, D., Huigene, M. (2011).** Empirical characterisation of agent behaviours in socio-ecological systems. Environmental Modelling & Software, Vol. 27, No. 7, pp. 837–844. DOI: 10.1016/j.envsoft.2011. 02.011.
33. **Sterman, J. D. (2000).** Business dynamics systems thinking and modeling for a complex world. McGraw-Hill.
34. **Terano, T. (2008).** Beyond the KISS principle for agent-based social simulation. Journal of Socio-Informatics, Vol. 1, No. 1, pp. 175–187
35. **Tomasello, M. V., Vaccario, G., Schweitzer, F. (2017).** Data-driven modeling of collaboration networks: a cross-domain analysis. EPJ Data Science, No. 22. DOI: 10.1 140/epjds/s13688-017-0117-5.
36. **University of Waikato. (2022).** Downloading and installing Weka.https://www.cs.waikato .ac.nz/~ml/weka/downloading.html.
37. **WHO. (2017).** WHO Air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide. World update 2005. Summary of risk assessment. Suiza: OMS. http://apps. who.int/iris/bitstream/10665/69478/1/WHO_S DE_PHE_OEH_06.02_spa.pdf.
38. **Wilensky, U. (2017).** Net Logo. https://ccl.northwestern.edu/netlogo/.
39. **Witten, I., Frank, E. (2005).** Data mining practical machine learning tools and techniques. San Francisco: Elsevier. Vol. 31, No. 1.
40. **Zhang, Y., Grignard, A., Lyons, K., Aubuchon, A., Larson, K. (2018).** Real-time machine learning prediction of an agent-based model for urban decision-making. Proceedings of the 17th international conference on autonomous agents and multiAgent systems Stockholm: International Foundation for Autonomous Agents and Multiagent Systems. pp. 2171–2173.
41. **Zhao, C., Li, S., Wang, W., Li, X. D. (2018).** Advanced parking space management strategy design: An agent-based simulation optimization approach. Transportation Research Record, Vol. 2672, No. 8. DOI: 10.1177/0361198118758671.

Article received on 15/06/2021; accepted on 18/08/2022.
Corresponding author is David Salas-Rodríguez.