Analyzing Patterns of a Bicycle Sharing System for Generating Rental Flow Predictive Models

Johnattan Douglas Ferreira Viana[1], **Oton Crispim Braga**[1], **Lenardo Chaves Maia**[2], **Francisco Milton Mendes Neto**[3]

[1]Programa de Pós-Graduação em Ciência da Computação - PPgCC
UFERSA/UERN, Mossoró, RN, Brazil.

[2]Departamento de Engenharias e Tecnologia - DETEC
Universidade Federal Rural do Semi-Árido, Pau dos Ferros, RN, Brazil

[3]Centro de Ciências Exatas e Naturais - DCEN
Universidade Federal Rural do Semi-Árido, Mossoró, RN, Brazil

{johnattandouglas, otoncbraga}@gmail.com,

{lenardo, miltonmendes}@ufersa.edu.br

Abstract. *Urban mobility has been highlighted as one of the most relevant themes in Smart Cities. Alongside this, following a principle of resource optimization and seeking greater sustainability, Bicycle Sharing Systems (BSSs) have stood out as a resource that can be used to assess urban mobility. The correct analysis of these data and the understanding of the dynamics in these systems can aid in decision making, in addition to optimize the complex urban mobility system. Thus, this work analyzes a BSS dataset, which is enriched for us with meteorological and seasonal information. In order to achieve our results, we recognize cyclist activity patterns related to date and climate information, as well as we identify a set of parameters that influences bicycle rental flow. Finally, we explore the relationship between these parameters and patterns, in order to present predictive regression models for rental flow prediction. In our results, Random Forest algorithm was the best approach for the creation of an effective regression model, explaining 95% of the explanatory variables.*

1. Introduction

Assuming the current growth trend of the metropolises, it is estimated that by 2050, more than 80% of the world population will be concentrated in urban centers [Zheng et al. 2014]. This expansion, which often happens uncontrolled, has been problematic in several aspects (i.e., energy consumption, traffic, safety). For that reason, it requires solutions that are consistent with current mobility challenges, such as the rising number of vehicles, increasingly time-consuming congestion and physical exhaustion of citizens due to chaotic traffic, resulting in loss of productivity.

These problems bring new opportunities for innovative technologies in Smart Cities, a concept focused on the integration of urban infrastructures and services to Information and Communication Technologies (ICTs) [Randhawa and Kumar 2017]. The context information collected by these scenarios has been an advance that has allowed the emergence of several applications that can do the integration of services in order to assist

decision-making in these scenarios. In this context, urban mobility has been highlighted as a challenging theme for public management, as well as one of the most relevant study areas in the ambit of Smart Cities [Georgescu et al. 2015].

In this sense, cycling has stood out as a great alternative to urban mobility problem, since the adoption of the bicycle reduces traffic congestion [Hamilton and Wichman 2018], besides impact positively the health of the population and the environment [Souza and Gomes 2014]. As a result, the number of bicycles in cities has grown [Fishman et al. 2013].

In addition, the use of bicycles has been enhanced with the emergence of Bicycle Sharing Systems (BSSs), an evolution of traditional bicycle rental systems, where the entire process of registering, lending and returning is automatic. These systems have evolved rapidly and are already in their fourth generation [Mátrai and Tóth 2016]. BSSs are shown as a complement to traditional public transport services such as buses and subways [Jäppinen et al. 2013] and, for that reason, the number of cities implementing them is growing [Correa et al. 2010].

BSSs have several stations throughout the city, where users rent a bicycle for a certain period and return it to a station at the end of the journey. The practicality of the systems encourages new cyclists, once it dispenses maintenance and parking concerns, which helps to minimize the use of cars on shorter journeys [Fishman et al. 2013]. In addition to the practical benefits of BSS, the data generated by these systems make them attractive for research, since variables such as travel time, departure time and arrival positions are explicitly recorded. Unlike other transportation services, this feature makes BSS a resource that can be used to assess urban mobility by identifying patterns and monitoring collected data.

How could we explore information from these systems to understand the dynamism of citizens' habits in order to improve urban mobility in Smart Cities? How can managers properly analyze dataset from public and private BSSs in order to provide better decisions in the strategic and operational context of their companies? The correct analysis of these data and the construction of predictive models, such as those that will be shown in this paper, can help answer these questions and assist in decision-making solutions. For this purpose, this work has the following contributions:

- enrichment of a database with context information and making it available for other researches;
- identification of attributes that influence bicycle rentals;
- recognition of patterns in the relationship of identified attributes;
- generation of a predictive model capable of inferring hourly rental flow.

The remainder of this paper is organized as follows: Section 2 shows some related works, which are focused on data from BSSs; In Section 3 the methodology of this work is presented; in Section 4 the results are discussed; finally, Section 5 presents the conclusions obtained in this work, as well as directions for future work.

2. Related Work

This work was inspired by [Fanaee-T and Gama 2013] that proposes an alternative for labeling events using background knowledge. They enriched 2011 records from a BSS

with weather and seasonal information. We replicate this approach with a more recent dataset for another purpose, with the aim of generating predictive models. On the other hand, some researches have already addressed the analysis of BSS data, mainly through a spatial-temporal perspective, in order to support in decision-making [Vogel et al. 2011]. For instance, after processing more than 18 million bike rides in Mexico City from 2010 to 2015, [Moncayo-Martínez and Ramirez-Nafarrate 2016] carried out an analysis of the mobility patterns using clustering to understand users' behavior related to usage in each stations. [Chen and Jakubowicz 2015] presented a model capable of inferring patterns of travel behavior, evaluating real data from a series of stations in Washington D.C.

[Borgnat et al. 2011] and [Kaltenbrunner et al. 2010] used statistical predictive models for different purposes, at Lyon and Barcelona stations, respectively. [Borgnat et al. 2011] used these models to predict the number of rents at a given time. [Kaltenbrunner et al. 2010] also used statistical predictive models to indicate the amount of free bikes to rent at certain stations. However, unlike these approaches, we use contextual information to generate an enriched dataset. We also identify the parameters that influence the quantity of rentals, finally generating a flow inference model.

[Razzaque and Clarke 2015] proposed an intelligent bicycle-sharing scenario based on Internet of Things. In addition, a number of services are proposed for the enhancement of current BSSs. Therefore, it is a model for the next generations of BSS using real-time information. For that, [Razzaque and Clarke 2015] evaluated an existing system, analyzing factors such as climate.

In our work, the analysis is not focused on the geo-referenced parameters of these data, but rather on the understanding of cyclists' temporal activity patterns related to contextual information, such as meteorological and seasonal variables.

3. Methodology

In order to extract useful knowledge from BSS' data, our work followed a methodology inspired by the classic process of Knowledge Discovery in Database (KDD) [Düsing 2000], using Artificial Intelligence techniques to aid in the decision-making process [Wickham and Grolemund 2017]. We scraped some contextual information from Web (i.e., weather) and integrated them to a BSS dataset. The integration step represents data enrichment process, resulting in a enriched dataset. We import, tidy, integrate, transform, visualize and model this dataset [Wickham and Grolemund 2017] to identify patterns and generate predictive rental flow models (Section 4). The methodology adopted is represented by Figure 1.

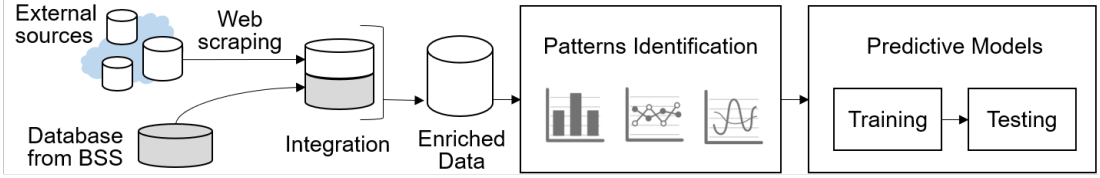


Figure 1. Methodology overview.

In this paper, the case study was applied to Capital BikeShare (CBS), the largest BSS in the United States across 6 jurisdictions: Washington, DC.; Arlington, VA; Alexan-

dria, VA; Montgomery, MD; Prince George's County, MD; and Fairfax County, VA. This system has been running since September 2010 and currently has more than 500 stations and about 4300 bicycles[1]. CBS has two types of users: $registered$, who is annual member, 30-day member or day key member; and $casual$, who does single trips or uses 24-hour pass, 3-day pass or 5-day pass. For practicality, in figures 2, 7 and 8, $registered$ and $member$ refer to the same user type.

CBS conducts studies that aim at continuous improvement such as removing the poorly used stations or adding new ones. For these surveys, stations have a computerized system that stores rental data. These data are publicly available, motivating the analysis and exploitation of these information. The data provided by CBS contains trip duration, start time, end time, info about start station and end station, bike identification and user type [System 2018]. The dataset used in this work represents rental flow from January 1 to December 31, 2017. It has about 3.75 million records, with approximately 35.4% of casual users and 64.6% of registered users.

In order to generate our enriched database using 2017 records, we performed a Web Scraping using a Node.js script, retrieving weather information from Freemeteo Website[2]. For the integration process, it was considered the closest meteorological measurement to beginning of trip. In addition to meteorological data, information about Washington D.C holidays was collected from the Department of Human Resources (DCHR) Website[3].

To extract as much information as possible from the $start\_date$ parameter, it was divided into $hour$, day, $month$, $season$, $weekday$, $workday$ (Monday to Friday, excluding holidays). $start\_station$, $start\_station\_name$, $end\_station$ and $end\_station\_name$ parameters represent where the cyclist's trip begins and ends, containing the numbering and the station name where bike was rented and returned. The $duration$ parameter is the subtraction of $start\_date$ and end_date. These data were enriched by aggregating the lowest distance values between $start\_station$ and $end\_station$ ($m\_distance$, in meters) and the probable duration using this shorter route ($m\_duration$, in minutes), both calculated by Google Maps API[4].

We used MySQL[5] commands to integrate and manipulate these additional information. In this work, we investigate this enriched dataset we generated[6]. Some parameters are listed in Table 1, except variations from $start\_date$ parameter.

In order to facilitate the patterns identification, an hourly grouping of the records was carried out. In sequence, statistical analysis was performed and several descriptive graphics (bars, dispersion and curve) were generated. For this purpose, the $ggplot$ library [Wickham 2016] for R language[7] was used. Predictive models were also generated in the RStudio[8] tool.

[1] www.capitalbikeshare.com
[2] freemeteo.com
[3] https://dchr.dc.gov/
[4] https://cloud.google.com/maps-platform/
[5] https://www.mysql.com/
[6] Available in https://goo.gl/FEFeXf
[7] https://www.r-project.org/
[8] https://www.rstudio.com/

Table 1. Attributes from enriched database.

$bike^{*}$	$start\_longitude^{*}$	$duration^{*}$	$temperature^{**}$
$type^{*}$	$end\_longitude^{*}$	$m\_duration^{\diamond}$	$r\_temperature^{**}$
$start\_station^{*}$	$start\_latitude^{*}$	$m\_distance^{\diamond}$	$wind^{**}$
$end\_station^{*}$	$end\_latitude^{*}$	$holiday^{\bullet}$	$humidity^{**}$
$start\_station\_name^{*}$	$start\_date^{*}$	$weather\_id^{**}$	dew_point^{**}
$end\_station\_name^{*}$	end_date^{*}	$cut\_description^{**}$	$pressure^{**}$

Data origin: CBS*, Google Maps API$^{\diamond}$, DCHR$^{\bullet}$ and Freemeteo**.

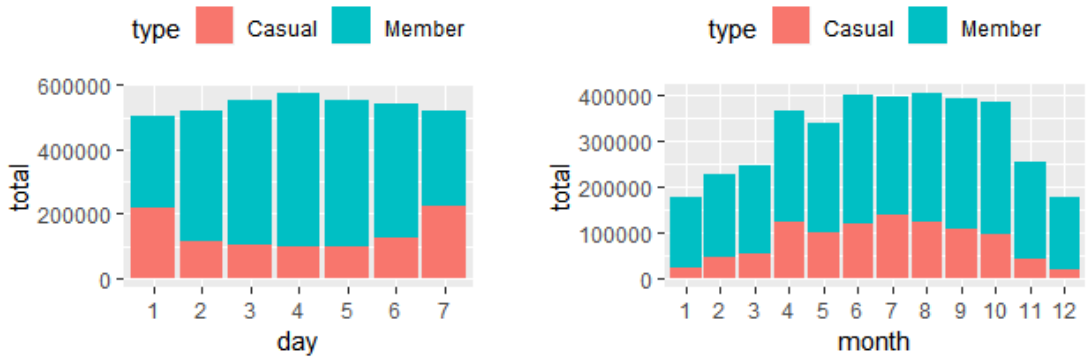
4. Results and Discussions

After data enrichment, the generated dataset was examined to identify some rental flow patterns (Section 4.1) and to create predictive models using algorithms based on Machine Learning (Section 4.2). Besides that,it is provided a comparative evaluation to assess what is most efficient algorithm for this purpose (Section 4.3).

4.1. Pattern Identification

Analyzing the $type$ parameter, it was possible to observe a notable discrepancy in the rental flow of $casual$ and $registered$ users. In total, the number of rents made by $registered$ users (2.775.979) represents almost triple the amount made by $casual$ ones (981.798). It indicates that 64.63% of rides were made by $registered$ users. The average $duration$ of registered rents is approximately 12 minutes, while that of casual ones is approximately 39 minutes. Thus, although the amount of casual rentals in general is lower, they tend to be longer, during on average three times more than registered rents.

We also identified a relationship between the $type$ of rent and weekdays. It has been noticed that casual-type rents occur in greater numbers on Sundays and Saturdays. Among the rents of registered users, the amount is higher in the middle of the week (specifically on Wednesdays). Figure 2(a) represents the number of rents per weekday.



(a) Total for each weekday.

(b) Total for each month.

Figure 2. Total of rents per type user.

Analyzing the amounts of rents per month, it was observed that June, July, August, September and October are the months with the highest amount of rents, as shown in Figure 2(b). January and December represent the months with the lowest amounts of

rents. We presume that this decrease is directly related to the average temperature of these months, which are usually the coldest of the year.

Investigating the rental flow per $season$, it was verified that in the winter occurred 572,440 rents (15,2%), with an average of 266 rents per hour and about 6,360 rents per day. In the spring there were 1,072,608 rents (28,5%), with an average of 487 rents per hour and about 11,658 rents per day. In the summer, 1,197,255 (31.9%) were recorded, with an average of 542 rents per hour and approximately 13,013 rents per day. In the fall, 915,474 (24.4%) rents occurred, averaging 419 per hour and about 10,060 rents per day. Thus, winter is the season with less rents. On the other hand, summer is the season with more rents, justifying the peak of rents in these months (Figure 2(b)).

There is a proportional relationship between the $duration$ and $season$, since the average duration of rents is usually higher in seasons of the year with more rents and lower in seasons of the year with less rents. Hence, rents tend to be longer in the summer. Figure 3 shows that in most months the amount of rents varies during the month. The months with the most continuous flow of rents are June, July, August and September, those that are part of the summer.

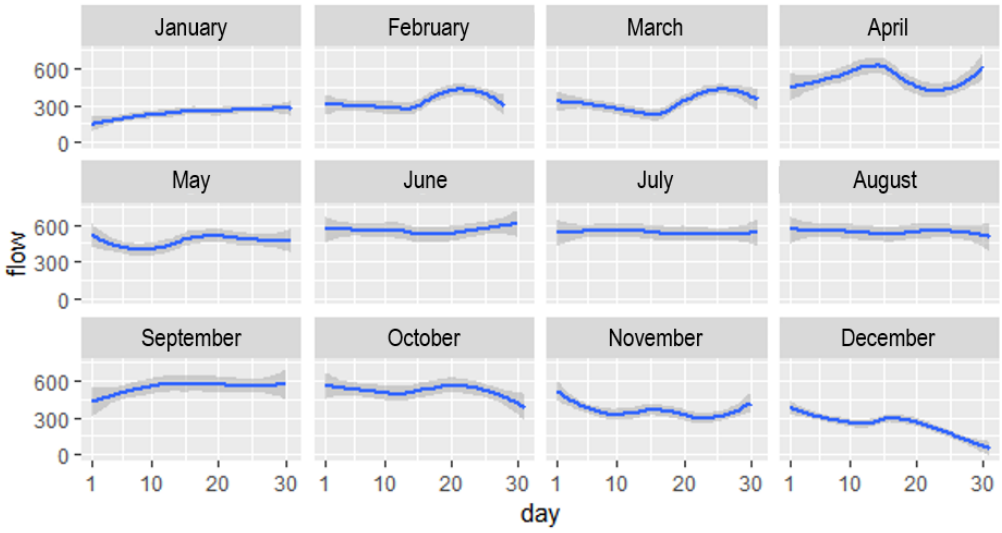


Figure 3. Quantity of rents per day of the month.

In addition, a similarity is seen in the rental flow per hour in each season. Figure 4 presents two peak times around 07:30 and 17:30. In winter, even with the lowest total amount of rents, the rental line flow still resembles the other seasons. These times are usually related to when people go to work and return home. This evidence that, regardless of season, most CBS users use it for the home-work route.

Regarding $hour$, rental flow is higher at times of greater movement in the city. Equivalently, as expected, the rental flow is not high at dawns. It was also noted an hourly patterns in weekends and weekdays (Figure 5). From Monday to Friday, it is clear two rental spikes, which happen around 07:30 and 17:30. On Saturdays and Sundays, there is a continuous increase in rents from 7:30 and a decrease after 15:00.

On holidays, the hourly patterns are similar to those found on weekends. Holidays are also included in Figure 4. However, because they were minor in relation to the working

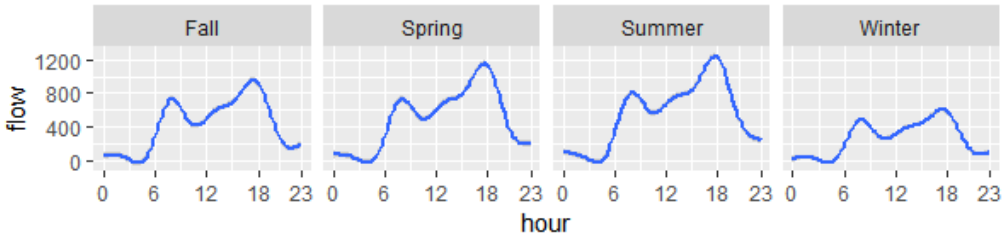


Figure 4. Amount of rents per hour in each season.

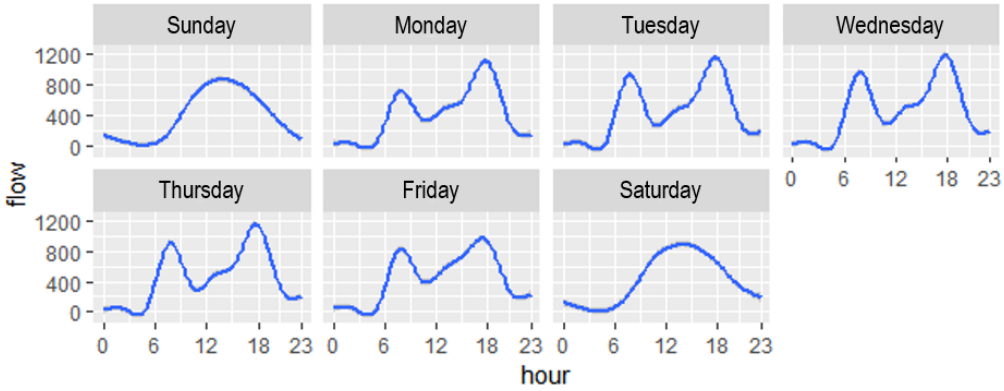


Figure 5. Amount of rents per hour in each weekday.

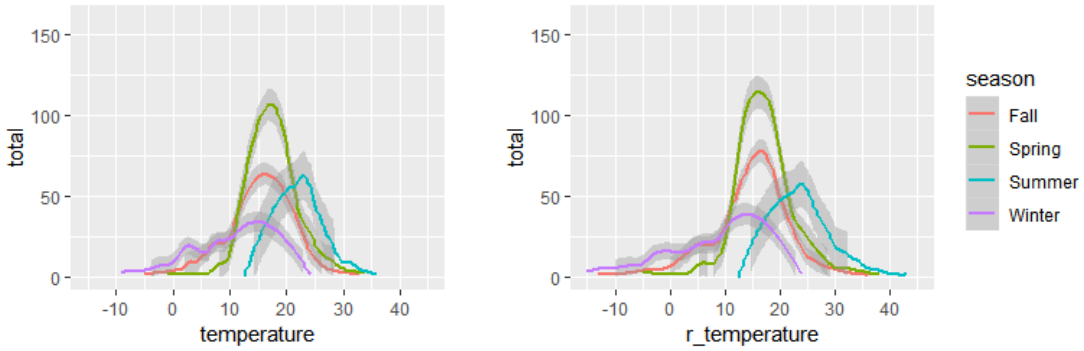
days, they were not enough to influence the curves in Figure 4. This same pattern of peak in rental flows resembles the hourly peak shown in Figure 5. In addition, rental flow has a relation with the climatic variables. Table 2 details the minimum, median, mean and maximum values of the climatic variables in dataset.

Table 2. Statistical summary of climatic variables.

Label	Min	Median	Average	Max
$temperature$	-9	17	15.83	36
$r\_temperature$	-17	17	15.24	43
$wind$	0	13	14.11	63
$humidity$	13	66	65.13	100
dew_point	-19	10	8.60	27
$pressure$	990	1016	1017	1042

The analysis of the $temperature$ (in °C) and the feels like temperature - real temperature ($r\_temperature$, in °C) revealed a great relation of these two variables with the rental flow. As shown in Figure 6, the highest rent flow happens in spring and summer, which are periods of mild temperatures.

The analysis of the mean rental curve in each season shows weather intensity increases or decreases rental flow. In other words, users tend to do less rentals when it is extremely cold or too hot. Figure 6 demonstrates although there is a slight variation between the $temperature$ and $r\_temperature$ parameters. In general, the average rental flow is higher between 15°C and 20°C.



(a) Total rentals for temperature.

(b) Total rentals for real temperature.

Figure 6. Relation of climate condition with rental flow.

In order to understand the influence of holidays on the rents dynamics, a comparison was made between the types of users for each of these days (Figure 7). However, on some holidays, the amount of casual rents approaches or exceeds the amount of rents recorded, unlike the pattern shown in Figure 2. Perhaps this happens because of the increase of visitor flow in the city during those days, considering the tourism potential of those dates. On some holidays, such as Independence Day and Memorial Day, there is a significant increase in rents, especially at stations close to historic or tourist monuments such as the White House and the Park National. For instance, during Independece Day, National Park Service (SPN) strongly encourages visitors of this holiday to use public transport for activities in the city, since public parking is extremely limited and cars are not allowed inside or around National Mall. In addition, as numerous roads are interdicted around George Washington Memorial Parkway, this is believed to be the main reason for the amount of rents on this day.

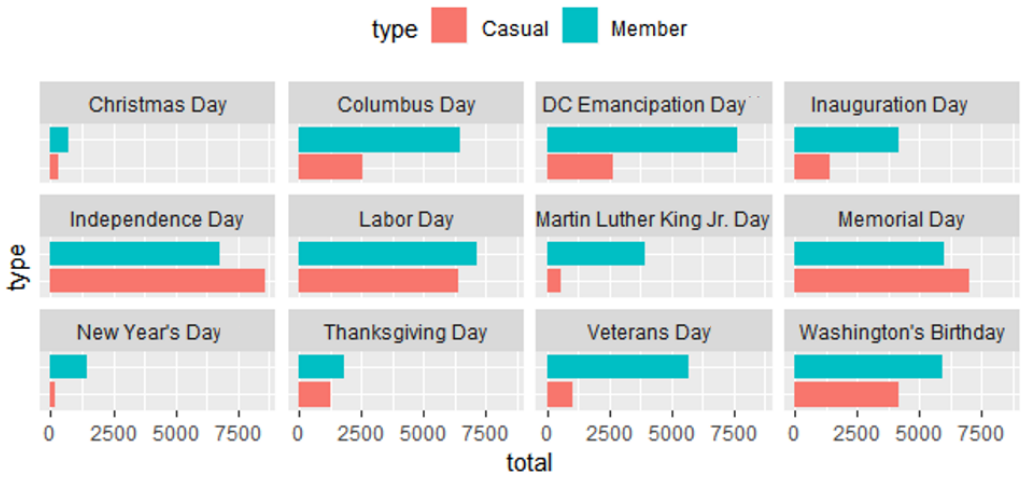


Figure 7. Amount of rent per type on each holiday.

In addition, in some holidays, the amount of rentals decreases rather than increasing, especially at Christmas and New Year. It is believed that in addition to the low temperatures, the type of holiday strongly influences this flow reduction, since, in general, people travel to spend these holidays as a family.

Another pattern found represents the relation between the rent duration ($m\_duration$) and the distance covered ($m\_distance$). In this relation, shown in a scatter plot in Figure 8, it was identified two rental profile groups, one that is usually faster and that travels farther, and another that usually takes longer and travels smaller distances.

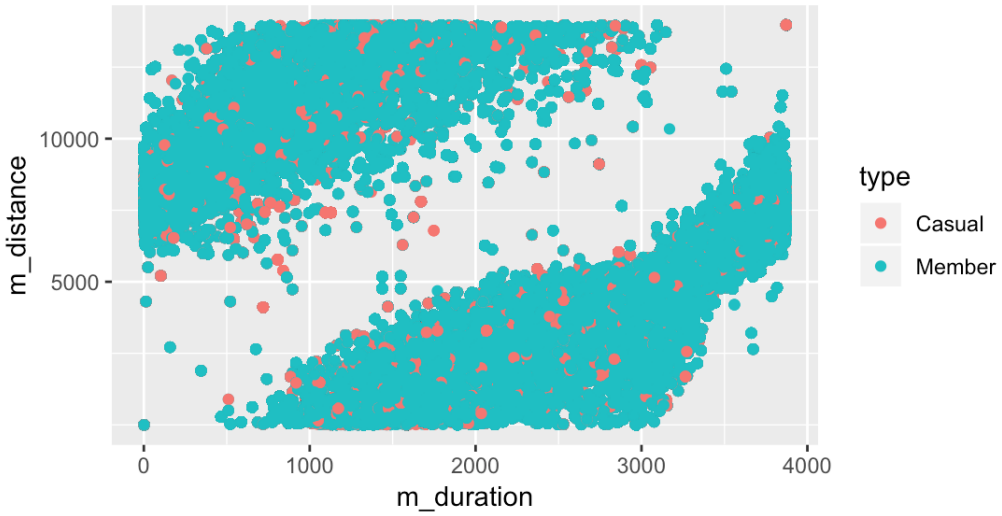


Figure 8. Relationship between distance and duration of rent.

It was believed that these two groups had correlation with $type$. The user who rents a bicycle to go to work, for example, wants to make the journey as fast as possible. On the other hand, it is thought that other users (i.e. tourists), take longer rents, but return the bicycle at nearby stations, and may even return it to the same place as the rent, resulting in shorter distances. However, as Figure 8 evidences, $type$ does not seem to be the variable responsible for differentiating these two groups. In our analyzes, we also tested day, $weekday$ and $season$ variables, however, none of them effectively explains the grouping. When comparing the average rental $m\_duration$ with $m\_distance$ based on $start\_station$ and $end\_station$ (calculated using the coordinates), it was realized that many rents took much more time than expected. This happens mainly at stations near parks, during holidays or weekends. Thus, we believe that these rent profiles are related to the rental purpose, although eventually registered users do leisure trips and casual users do home-work-home trajectory.

4.2. Definition of the Predictive Model

The patterns identified in the data analysis process were used to apply algorithms based on Machine Learning techniques in order to generate an predictive hourly rental flow model. For that, as said previously, it was done the hourly grouping and total rent in each hour was saved in a qtd variable. In total, the new dataset[9] after grouping contains 8737 records. To generate the predictive model, we selected 13 variables divided into two sets: $season$, $month$, day, $weekday$, $hour$, $workday$, $holiday$, that derive from $date$; and $temperature$, $r\_temperature$, $wind$, $humidity$, dew_point, $pressure$, that derive from the weather. The parameters with average hourly measurements were calculated for the grouping of climatic values. Section 4.1 showed some analysis of these variables.

[9] Available in https://goo.gl/FEFeXf

In this case, we search for the number of rents (qnt), a discrete value. Thus, three Machine Learning algorithms were tested and evaluated:

Linear Regression is a statistical predictive model that treats scattered data in a linear way. The algorithm generates a better fit line in the scatter plot, which represents the relation of the explanatory variables to the dependent variable. Using the best fit line it is possible to predict a value for a dependent variable given a new input [Long et al. 1993]. Equation 1 represents the best fit line.

$$Y_i = \alpha + \beta\, X_i + \epsilon_i \tag{1}$$

Y_i is the dependent variable; α is a constant, which represents the intercept of the line with the vertical axis; β is another constant, which represents the line slope (angular coefficient); X_i is the explanatory variable (independent) that represents the explanatory factor; and, finally, ϵ_i is a random variable that represents the possible measurement errors.

Decision Tree is an algorithm based on gaining information, which is calculated using categorization techniques (i.e. Gini, Chi-square, entropy). The gain is expressed by Equation 2, which follows:

$$G(A) = I(p,n) \sum_{v}^{i=1} \frac{p_i + n_i}{p+n} I(p_i; n_i) \tag{2}$$

where

$$I(p,n) = -\frac{p}{p+n} log_2 \frac{p}{p+n} - \frac{n}{p+n} log_2 \frac{n}{p+n} \tag{3}$$

being p and n, respectively, positive and negative instances of a dichotomous classification variable [Long et al. 1993].

Random Forest is an assembly-based algorithm that generates a forest of sub trees. These trees are generated by a random factor and the combination of learning models increases the overall result [J.Ham and Kamber 2011].

4.3. Training and Testing

To test and compare the algorithms, a training/test method based on cross validation was applied [Kassambara 2018]. This approach divides the dataset in k parts and performes training and test iterations with each one. Each iteration separates the k-th part to test the trained model. Then the training part (k_i) is returned to the dataset while the next part (k_{i+1}) is separated for training. Thus, at the end of the process, all k parts are trained and tested. For this work, each algorithm was tested using cross-validation with 10 parts.

In order to assess the best model for purpose of this work, we considered the following metrics to calculate the effectiveness of the algorithms:

RSquare ($\mathbf{R^2}$) corresponds to an adjustment measure of a generalized linear statistical model in relation to the observed values, varying from 0 to 1. This metric represents how much the model can explain the observed values. The higher this metric, the more explanatory the model will be, and the model better describes the samples [Kassambara 2018].

Mean Absolute Error (MAE) measures the average magnitude of errors in a set of predictions. As shown in Equation 4, this metric refers to the average over the test sample of the absolute differences between the prediction (y_j) and the real observation ($\hat{y}_j$), in which all individual differences have equal weight [Kassambara 2018].

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j| \tag{4}$$

Root Mean Square Error (RMSE) is a quadratic scoring rule that also measures the mean magnitude of the error. It represents the square root of the mean square differences between the prediction and real observation (Equation 5).

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2} \tag{5}$$

As shown in Equation 5, it is important to note that the square root of the mean squared errors has some interesting implications, since the errors are high before the average, assigning a relatively high weight to large errors. In this sense, this justifies the utility of this metric in our approach, once substantial errors are undesirable [Kassambara 2018].

Both MAE and $RMSE$ express the average predictive model error. They can range from 0 to ∞ and are indifferent to the error directions. In addition, they have negatively oriented scores. So, since they represent error rates, lower values correspond to better models. Table 3 shows the performance of each algorithm in relation to the metrics described.

Table 3. Comparison among predictive models.

Algorithm	R^2	MAE	RMSE
Linear Regression	0.3825272	242.9403	320.544
Decision Tree	0.5607824	191.7552	270.3072
Random Forest	0.9560645	53.45990	87.02509

According Table 3, Linear Regression results were the worst in all metrics. We expected that because rental flow does not obey an expressive linearity, as shown in figure 4 and 5. It was possible to perceive the dynamicity of the flow in relation to several attributes, mainly $hour$ and $weekday$, that present a certain non-linear pattern. On the other hand, the algorithms based on the gain of information showed better results, mainly the Random Forest, since it was able to explain 95% of the explanatory variables. The information gain strategy adopted by tree-based algorithms has shown promise for the problem faced. Therefore, despite having a higher computational cost, the model of prediction using Random Forest was chosen to infer the hourly rental flow prediction.

In addition, from the random trees generated by the algorithm, we could calculate the relevance of the attributes by the percentage of Increased of Mean Square Error ($IncMSE$), that describes the predictive capacity of the Mean Square Error with variables randomly exchanged. If this permutation drastically changes the predicted value, then the variable is considered critical.

The Increased Impurity Node was also calculated ($IncNodePurity$) that measures the loss function when the best tree nodes are selected, revealing variables that are most significant for the prediction [Echeverry-Galvis et al. 2014]. Figure 9 shows the relevance of the variable in the predictive model constructed with Random Forest algorithm.

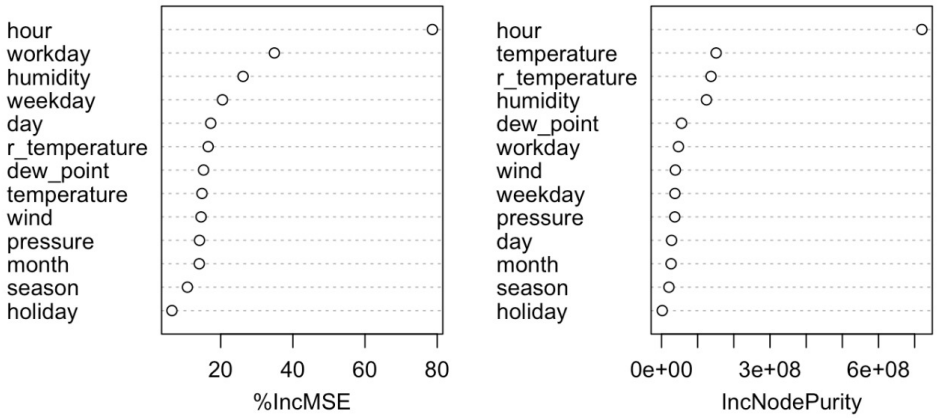


Figure 9. Relevance of the variables in predictive model providing by Random Forest algorithm.

As can be seen, the attributes identified as relevant in the pattern indetification process (Section 4.1) are the most influential in the decision tree. It can also be seen that the values of $IncMSE$ and $IncNodePurity$ differ greatly between themselves, with the exception of $hour$, which is the tree root. That explains why the Random Forest had much better results than Decision Tree.

In this work, we also generated station rental flow prediction and daily rental flow prediction, but the results achieved using the algorithms presented in this work were not as significant as hourly rental flow prediction. From the obtained predictive model, a Web Scraping script was built for obtaining real-time weather information and so predict the amount of rents at a given time.

5. Conclusions

In this current work, Machine Learning algorithms and Data Science techniques supported the understanding of CBS dataset and the evaluation of bicycle rental flows. Certainly, this work approach can support decision making in BSSs, and it can be also applied in other systems that deal with urban mobility, such as rentals car and motorcycles. In addition, applying context information benefits in identifying patterns in these systems, besides enabling to generate rental flow predictive models. In this sense, patterns were observed in the relation between the temporal factors and the rental flow, mainly, if considered the hour that the rent occurred. For an example of practical application, these patterns can assist the managers of these systems in identifying better times for bicycle maintenance, considering less flowing hours.

In relation to the Machine Learning techniques tested in this work, Random Forest algorithm proved to be possible to apply temporal and climatic variables to create efficient predictive models. In our analyzes, this algorithm demonstrated a better result for the purposes of this work, explaining 95% of the explanatory variables.

As future works, we intend to perform a cluster analysis to describe which additional factors influence the relationship between $m\_distance$ and $m\_duration$. Finally, we intend to apply Artificial Neural Networks to infer the rental flow, expecting to enhance our prediction by day, season and year.

Acknowledgment

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

References

Borgnat, P., Robardet, C., Rouquier, J.-B., Abry, P., Fleury, E., and Flandrin, P. (2011). Shared Bicycles in a City: A Signal Processing and Data Analysis Perspective. *Advances in Complex Systems*, 14(3):415–438.

Chen, L. and Jakubowicz, J. (2015). Inferring bike trip patterns from bike sharing system open data. In *2015 IEEE International Conference on Big Data (Big Data)*, pages 2898–2900.

Correa, R., da Cunha, K. B., and Boareto, R. (2010). *A bicicleta e as cidades: como inserir a bicicleta na política de mobilidade urbana*. Instituto de energia e meio ambiente, 2 edition. [In Portuguese].

Düsing, R. (2000). Knowledge discovery in databases. *Wirtschaftsinformatik*, 42(1):74–75.

Echeverry-Galvis, M., Peterson, J., and Caceres, R. (2014). The social nestwork: Tree structure determines nest placement in kenyan weaverbird colonies. *PloS one*, 9:e88761.

Fanaee-T, H. and Gama, J. (2013). Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, pages 1–15.

Fishman, E., Washington, S., and Haworth, N. (2013). Bike share: A synthesis of the literature. *Transport Reviews*, 33(2):148–165.

Georgescu, M., Pavaloaia, V., Popescul, D., and Tugui, A. (2015). The race for making up the list of emergent smart cities. an eastern european country's approach. *Transformations in Business and Economics*, 14:529–549.

Hamilton, T. L. and Wichman, C. J. (2018). Bicycle infrastructure and traffic congestion: Evidence from dc's capital bikeshare. *Journal of Environmental Economics and Management*, 87:72 – 93.

J.Ham and Kamber, M. (2011). *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers is an imprint of Elsevier, 3 edition.

Jäppinen, S., Toivonen, T., and Salonen, M. (2013). Modelling the potential effect of shared bicycles on public transport travel times in greater helsinki: An open data approach. *Applied Geography*, 43:13 – 24.

Kaltenbrunner, A., Meza, R., Grivolla, J., Codina, J., and Banchs, R. (2010). Urban cycles and mobility patterns: Exploring and predicting trends in a bicycle-based public transport system. *Pervasive and Mobile Computing*, 6(4):455 – 466. Human Behavior in Ubiquitous Environments: Modeling of Human Mobility Patterns.

Kassambara, A. (2018). *Machine Learning Essentials: Practical Guide in R*. STHDA.

Long, W. J., Griffith, J. L., Selker, H. P., and D'agostino, R. B. (1993). A comparison of logistic regression to decision-tree induction in a medical domain. *Computers and Biomedical Research*, 26(1):74–97.

Moncayo-Martínez, L. A. and Ramirez-Nafarrate, A. (2016). Visualization of the mobility patterns in the bike-sharing transport systems in mexico city. In *2016 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, pages 1851–1855.

Mátrai, T. and Tóth, J. (2016). Comparative assessment of public bike sharing systems. *Transportation Research Procedia*, 14:2344 – 2351. Transport Research Arena TRA2016.

Randhawa, A. and Kumar, A. (2017). Exploring sustainability of smart development initiatives in india. *International Journal of Sustainable Built Environment*, 6(2):701 – 710.

Razzaque, M. A. and Clarke, S. (2015). Smart management of next generation bike sharing systems using internet of things. In *2015 IEEE First International Smart Cities Conference (ISC2)*, pages 1–8.

Souza, L. C. and Gomes, E. T. A. (2014). O uso da bicicleta como meio de transporte: Mobilidade urbana na cidade do recife. In *Anais do I Congresso Brasileiro de Geografia Política, Geopolítica e Gestão do Território*, pages 384–395. Letra1. [In Portuguese].

System, C. B. S. (2018). Capital bike sharing trip history data.

Vogel, P., Greiser, T., and Mattfeld, D. C. (2011). Understanding bike-sharing systems using data mining: Exploring activity patterns. *Procedia - Social and Behavioral Sciences*, 20:514 – 523.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wickham, H. and Grolemund, G. (2017). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly Media, 1 edition.

Zheng, Y., Capra, L., Wolfson, O., and Yang, H. (2014). Urban computing: Concepts, methodologies, and applications. *ACM Trans. Intell. Syst. Technol.*, 5(3):38:1–38:55.