Journal of Traffic and Transportation Management
Volume 3, No. 1 (2021) pp. 25-31

Modelling of a Stochastic Spatiotemporal Variable in Transport Domain

Peter Krammer [a] *, Marcel Kvassay [a] *, Ladislav Hluchý [a]

[a] *Institute of Informatics, Slovak Academy of Sciences, Bratislava, Slovakia, 845 07*

Abstract

In this article, building on our previous work, we engage in spatiotemporal modelling of transport demand in the Montreal metropolitan area over the period of six years. We employ classical machine learning and regression models, which predict bike-sharing demand in the form of daily cumulative sums of bike trips for each considered docking station. Hourly estimates of demand are then determined by considering the statistical distribution of demand across individual hours of an average day. In order to capture seasonal and other regular variation of demand, longer-term distribution characteristics of bike trips, such as their average number falling on each day of the week, month of the year, etc., were also used as input attributes. We initially conjectured that weather would be an important source of irregular variation in bike-sharing demand, and subsequently included several available meteorological variables in our models. We validated our models by Hold-Out and 10-Fold Cross-Validation, with encouraging results.

Keywords: *Machine learning, Data mining, Regression, Data distribution, Spatiotemporal data, Modelling, Regression tree*

1. Introduction

The popularity of bike and scooter sharing is on the increase, not least because of their eco-friendly nature. User-friendly technology, which enables people to unlock and use these vehicles through their smartphones, makes this type of transport economical and accessible to masses. It also simplifies the collection of data needed to estimate the transport needs in a given area and, if the data are representative, allows accurate modelling and prediction of local transport needs, thus helping to optimise the traffic system.

One disadvantage of a shared bike transport system is the need to regularly redistribute and rebalance the number of bikes among stations. This, in turn, requires a good working knowledge of bike-sharing demand, which depends on both location and time. Our main motivation was therefore to provide this information using statistical methods as well as machine learning in order to model and generalize bike-sharing demand evolution in time and space. It would enable the providers to improve their offer and make this mode of transport more readily available. We build on our previous work [1] and further enhance it by incorporating selected meteorological attributes available for the Montreal metropolitan area from the Weather and Climate website of the Canadian government [2].

2. Related Work

History, impacts, business models and future trends of bike sharing were explored in [3], which also mapped its expansion in Europe, North America and South-East Asia until about 2009. A newer article from 2015 [4] focused on the modelling of bike-sharing demand and estimated the impact and significance of various factors. As expected, the most significant factors included the hours of the day, the days of the week, and the year. Calendar months (including summer holidays) appeared to be only moderately significant. Somewhat surprisingly, rainfall levels did not seem to be significant and were overshadowed by other weather-related factors, such as temperature, atmospheric pressure or humidity. Typical models in [4] relied on classification and regression trees with boosting and operated on hourly data. The data used in [4] covered the area of Washington, DC and the years 2011-2015.

In [5], the authors opted for an unconventional approach. They built a weighted correlation network to capture the relationship among bike stations, and dynamically grouped neighbouring stations with similar bike usage patterns into clusters. In the next step, they used Monte Carlo simulation to predict the over-demand probability for each cluster. In their evaluation, they used real-world data from New York City and Washington, D.C.

* Corresponding author. Tel. / Fax: +421 2 5477 1004;
E-mail: peter.krammer@savba.sk, marcel.kvassay@savba.sk

DOI: 10.5383/JTTM.03.01.004

Lei Lin et al. in [6] employed Graph Convolutional Neural Network with Data-driven Graph Filter model capable of learning pairwise correlations between stations to predict station-level hourly demand in a large-scale bike-sharing network. Their model included the convolution and feedforward blocks as well as a recurrent block from the Long Short-term Memory neural network architecture to capture the temporal dependencies in their bike-sharing demand series.

A successful model of bike-sharing demand would enable the providers to solve the optimization problem related to the transfer of bikes from the stations with many deposited bikes to those where they are needed. This problem was tackled in [7], where the problem was defined rigorously through graph theory, and a mathematical model was proposed for the optimal strategy of distributing unused bikes among the docking stations.

In this article, we address the problem of spatiotemporal modelling of shared bike trips in Montreal between 2014 and 2019. Compared to [8], which tackled a similar problem by deep learning, we model longer periods and use different types of models and attributes. We also employ different accuracy metrics, more suited to the specific nature of our task. As a result, our models are simpler, more robust and require less time for training. Moreover, once they are trained, they have very modest hardware demands and can run even on mobile devices, which simplifies their eventual deployment.

3. Description of Data

In the first phase of our experiments, we considered two bike-sharing datasets from Canada, both freely available on Kaggle.com. One was from Toronto [9], the other from Montreal [10]. It turned out that the Montreal dataset was considerably more extensive, covering the years 2016-2019. Moreover, we were able to enrich it further with the data for the years 2014-2015 downloadable from the website of its original provider [11]. The final dataset that we used in this phase thus comprised more than 26 million records about bike trips in Montreal in the years 2014-2019 (size 1.3 GB), which helped us to separate long-term trends from seasonal variations.

Each record (table row) in this dataset contained the following attributes:

- start_date: Date and time of the start of the trip (YYYY-MM-DD hh:mm)
- start_station_code: Start station ID
- end_date: Date and time of the end of the trip (YYYY-MM-DD hh:mm)
- end_station_code: End station ID
- is_member: Type users. (1: Suscriber, 0: Non-suscriber)
- duration_sec: Total travel time in seconds

The records for each calendar year covered the months April to November (the period during which the service is standardly available), except for November 2019, for which no records were provided. Besides these records, the dataset also contained tables listing active docking stations for each calendar year (their number had grown over the years). Each docking station was characterised by the following attributes:

- GPS – its geographical position described by its longitude and latitude;
- station ID – its unique identifier (integer);
- station Name – its description (postal address).

Unfortunately, none of these tables listed the number of available bikes at these docking stations at any point. In the domain of public transport, we can define various tasks and apply different approaches. One of them is graph-based, with nodes representing docking stations and edges representing bike trips. Another possibility is a spatiotemporal model, which tries to capture the precise distribution of docking stations and bike trips in space and time. The latter approach corresponds more closely to our present intention to model and predict spatiotemporal aspects of bike-sharing demand.

A successful model would enable the providers of the service to optimise the operation of their bike-sharing networks. To put it simply, the providers have to transport bikes in a timely and regular manner from places where they are not needed to those where the current demand is high or is expected to increase soon. With the number of docking stations in the order of hundreds, such optimization might result in significant financial savings.

In this particular dataset there is one problem: lack of information about how many bikes were present at each docking station at any given time. For the sake of simplicity, we assume that the number of bike trips faithfully represents the demand for them, that is, we exclude from further consideration the possibility that somebody wanted to use a bike, but found no bike available in the closest docking station.

A similar task was already attempted in [8], where the authors had the data about the number of bikes in each docking station, but their approach differed and their bike-sharing dataset covered shorter period than ours did.

4. First Phase: Analysis and Modelling without Meteorological Variables

In our analysis, we concentrated on the docking stations that had been in operation throughout the period 2014-19. There were 456 such stations. The location of about ten of them had slightly shifted over the years (by about 300 meters), but their identifiers were kept the same. In such cases, we used the GPS coordinates valid for longer period as their location descriptor. We also removed a few records with invalid station ID from the dataset.

We analysed the data from the period 2014-18 and kept the year 2019 apart as an independent testing set for holdout validation. Our preliminary analysis revealed some interesting patterns:

- For example, the histogram of bike-trips by the day of the week in Figure 1 showed a relatively even temporal distribution with only minor drops of ridership over the weekends and Mondays.

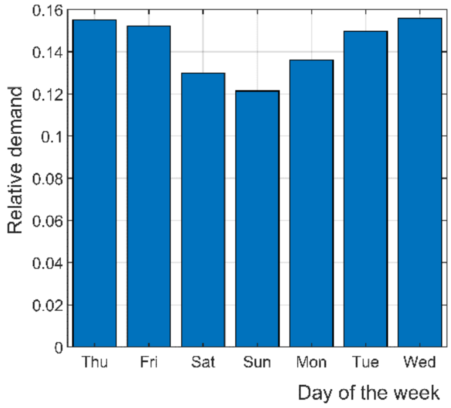

Fig. 1. Distribution of relative share of bike trips over the days of the week in Montreal

• The number of trips in a given calendar month was relatively stable across the years (apart from the general growing trend), but the data exhibited strong annual seasonality with peaks in the summer as shown in Figure 2. (Please note that this bike-sharing service operates only from April to November).

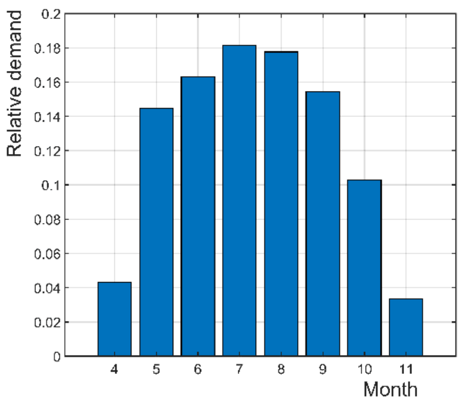

Fig. 2. Distribution of relative share of bike trips over calendar months in Montreal

These data distributions enabled us to determine the weights for each day of the week and for each month for our subsequent analyses.

• Next comes the annual trend shown in Figure 3, which is clearly increasing. In this case, in order to make it stand out more prominently, we also included our holdout set (2019 data) since our whole dataset only covered six calendar years. We approximated the missing November 2019 data based on the average ratio of November ridership to that of other months in the previous years.

We were able to approximate these annual data by the following linear function (1).

$$weight\ (year) = 5{,}575e5\ .\ year - 1{,}12e9 \qquad (1)$$

Of course, this is only a simplified approximation of the general trend in the recent past. Even in ideal conditions, the growth rate of ridership would eventually taper off, not to speak of dramatic reductions in mass mobility imposed by the current Covid pandemic. However, for the considered period 2014-19, a linear approximation is certainly appropriate.

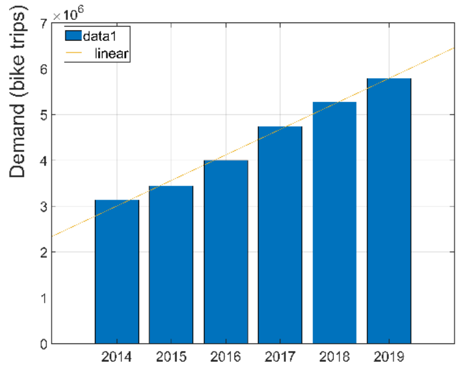

Fig. 3. Bike trips in Montreal by the year

Based on the above variables and relationships, we defined a group of attributes for our machine learning experiments as shown in Table 1. Besides the date and geographical position of docking stations from which the trips were initiated, we also used the weights of individual years, months and days of the week (derived from ridership statistics) as well as the distance of the docking stations from the city centre. In this form, one record no longer represents one individual bike trip but rather a cumulative number of such trips for one calendar day and one docking station. We thus condensed more than 26 million individual trips into about 600,000 cumulative daily records.

Table 1. List of attributes used in machine learning experiments

Attribute Name	Attribute Description
year, month, day	date – year, month and day
dayOfYear	number of the day of the year
gps_lon, gps_lat	GPS position of the docking station where the trip started
centre_dist	Distance of the docking station from the city centre
weightMonth	weight of a given month
weightYear	weight of a given year
weightDayOfWeek	weight of a given day of the week
DemandCount	(target attribute) number of bike trips starting at a given station in a given day

In our experiments, we have defined the city centre as the block with the highest density of initiated bike trips, which is at the same time situated roughly in the city centre at GPS position (45.51; -73.58). We have determined it empirically from the bivariate histogram in Figure 4.

Compared to the approach used in [8], which relied on deep learning with long short-term memory (LSTM), our techniques were more traditional and relied on classical machine learning. That is why we first undertook a preliminary analysis of our data including their statistical distribution over the days of the week, calendar months, years, etc. These distributions appeared to be sufficiently stable over the years and at the same time significant with respect to the target attribute. That enabled us to use them as weights for the days of the week and calendar months and to feed them as input attributes to our machine learning algorithms.

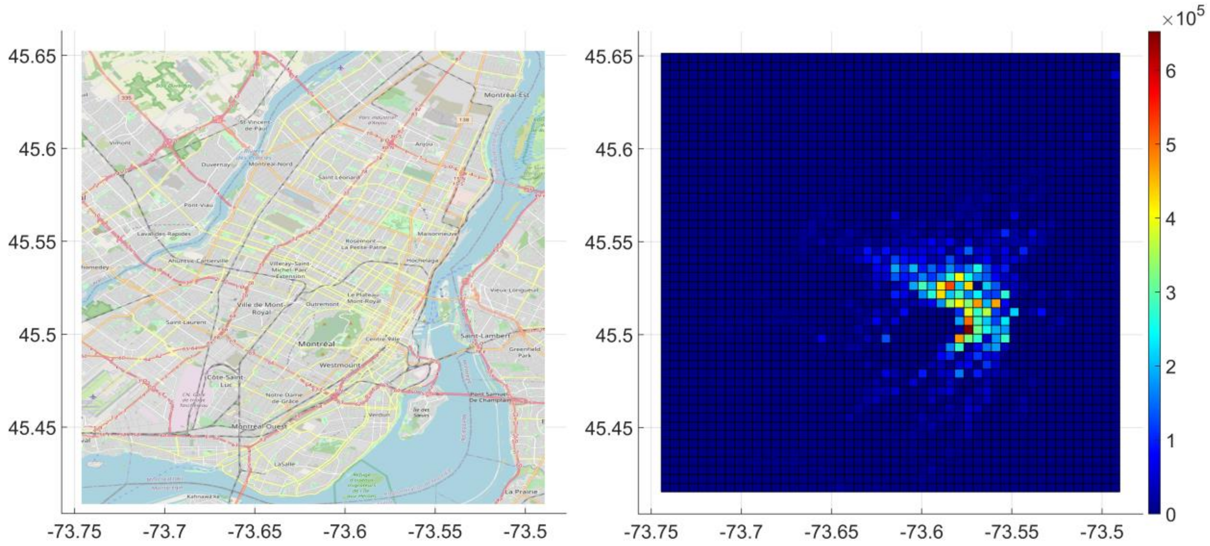

Fig. 4. Bivariate histogram of the number of bike trips grouped by the GPS coordinates of the starting station in Montreal (right) along with the corresponding map (left).

We modelled our data using various types of models including Gaussian Processes, Isotonic Regression, Regression Tree M5P, Random Forest, Radial Basis Function Regressor and MultiLayer Perceptron Regressor. We evaluated their accuracy by two validation methods. The first was holdout validation where the holdout set consisted of the data for the (most recent) year 2019. These data were not available to our models for training. Since the holdout set represented the latest data compared to the training set (2014-18), this made our modelling task more challenging. In effect, our training set consisted of 490,200 records, while the holdout set of 91,200 records. The second validation method that we used was 10-fold Cross-Validation, which is considerably more time-consuming but provides estimates that are more conservative.

We chose Correlation Coefficient and Relative Absolute Error as suitable validation criteria capturing our overall modelling accuracy. We considered them more suitable than Mean Average Percentage Error given that our target attribute included many zero values (attempts to calculate MAPE for such records would result in a division by zero). In our experiments, we achieved the best overall accuracy with M5P trees (with pruning and at least 13 samples per leaf) and with Random Forests consisting of 20 trees. Tables 2 and 3 list their observed accuracy.

Table 2. Modelling accuracy of our best models from the first phase as measured by 10-fold Cross-Validation

	Regression Tree M5P	Random Forest
Correlation Coefficient	0.91144	0.93561
Relative Absolute Error	0.35130	0.30644

Table 3. Modelling accuracy of our best models from the first phase as measured by Hold-Out Validation on 2019 data

	Regression Tree M5P	Random Forest
Correlation Coefficient	0.80260	0.80470
Relative Absolute Error	0.55194	0.55412

5. Second Phase: New Experiments with Meteorological Variables

In the second phase, we added daily weather information recorded on a nearby airport, which we obtained from the Weather and Climate website of the Canadian government [2]. Weather-related attributes in this dataset included maximum daily wind speed, minimum, maximum and average daily temperature, and cumulative daily precipitation. All these could potentially influence the bike-sharing demand: bad weather, i.e. low temperature, strong wind or intense precipitation should naturally dampen it. Of course, it would have been ideal to use the weather information directly from the locations of the concerned bike-sharing stations, but we had to make do with what was readily available.

There were quite a few missing values in this dataset. We filled them with interpolated data from the two closest days that had the corresponding fields filled in or, in the case of terminal values that only had one such neighbour, with a copy of the neighbour's value.

Apart from the above meteorological attributes, we also added another, called WeightStation, which represented a kind of "popularity" of each station. It was calculated as the proportion of bike-shares for a given station to all bike-shares in the period since its establishment until the end of 2018 (we excluded 2019 since it was used for Hold-Out validation).

Having incorporated the above attributes into our datasets, we repeated the training, testing and validation for the extended models using the same methodology as in the first phase. Subsequently, we observed an increase in accuracy for both 10-fold Cross-validation and Hold Out validation, as shown in Tables 4 and 5.

We can see an improvement in both Correlation Coefficient (CC) and Relative Absolute Error (REA), which is particularly noticeable for Random Forest and Hold Out validation, where CC increased from 0.8047 to 0.8411 and REA decreased from 0.55541 to 0.4868.

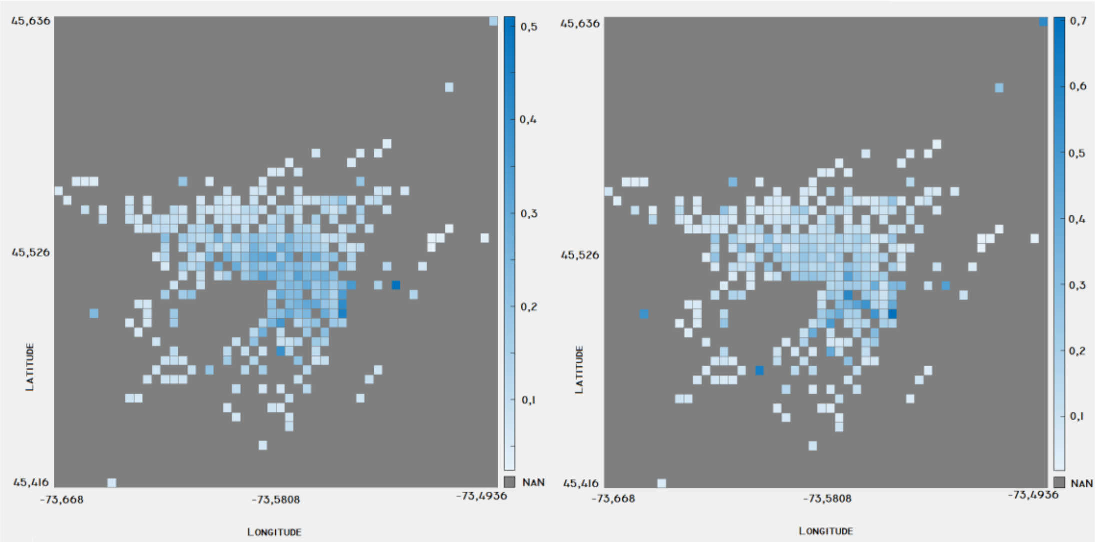

Fig. 5. Distribution of Relative Absolute Errors for our best Random Forest Model in the Montreal area achieved with Cross-Validation (left) and Hold-Out Validation (right).

As shown in Fig. 5, the distribution of Relative Absolute Errors for our best Random Forest Model depends on location. In general, Relative Absolute Errors tend to be greater for locations with higher bike sharing demand.

Table 4. Modelling accuracy of our best models (after adding new attributes) as measured by 10-fold Cross-Validation

	Regression Tree M5P	Random Forest
Correlation Coefficient	0.93640	0.95042
Relative Absolute Error	0.30614	0.27091

Table 5. Modelling accuracy of our best models (after adding new attributes) as measured by Hold-Out Validation on 2019 data

	Regression Tree M5P	Random Forest
Correlation Coefficient	0.82170	0.84110
Relative Absolute Error	0.51748	0.48678

We also evaluated input attribute importance by the criterion of average impurity decrease. The results are shown in Table 6, with the most important attributes listed first.

In general, the most important attributes were the various weights, in particular the weights of stations, years and calendar months. All of them scored impurity decreases of more than 6000. The weight of day of the week scored 2868, so it did not make it to the top, but still scored above average.

The next important variable was calendar year. Somewhat surprisingly, calendar month and the day of the week scored significantly lower than calendar year. The reason probably stems from the fact that the variation of daily bike-sharing demand over days of the week shown in figure 1 is smaller than its variation over calendar months and even years, as shown in figures 2 and 3. In other words, to reasonably estimate the number of bike trips on a given day, it is more important to know which month and year it is rather than which day of the week it is.

Table 6. Evaluation of input attribute importance by the criterion of average impurity decrease

Attribute Name	Attribute Importance
WeightStation	23403.54
WeightYear	7484.34
WeightMonth	6214.56
Year	5214.56
Centre_Dist	3966.40
Precipitation	3490.98
WeightDayOfWeek	2867.95
MaxDayTemp	2054.37
GPS_Lat	1794.81
DayOfYear	1246.18
AvgDayTemp	1166.74
DayOfWeek	1112.65
GPS_Lon	1110.08
Month	1086.72
MinDayTemp	787.89
DayOfMonth	712.07
MaxWindSpeed	567.67

Another interesting observation was that the weights of these date-related attributes scored significantly higher than the attributes themselves. Here the likely reason is that the weights contain additional statistical information about the bike-sharing demand, in particular its general trend and regular fluctuations.

Position-related attributes, such as the station's distance to the city centre and its GPS coordinates are widely dispersed in the list: while the distance to the centre turned out to be very important, its GPS coordinates were much less so. Nevertheless, they were not negligible, which means that bike-sharing demand is not symmetrically distributed with respect to the city centre – otherwise we should see concentric circles on the bivariate histogram in Fig.4. In reality, the demand appears to be much higher to the north of the city centre rather than to the south.

Some meteorological variables also came up as important, in particular daily total precipitation and maximum daily temperature. Surprisingly, other meteorological variables, such as average or minimal daily temperature and maximum wind speed were much less significant.

Thus, in certain respects, our results mirror those reported in [4]: both show high influence of calendar year and air temperature. Surprisingly (and somewhat counterintuitively), both also concur on the fact that daily total precipitation is not one of the most influential factors. Additionally, both show calendar month and wind speed as less significant, but not negligible.

Some variables used in [4], such as humidity and air pressure, were not available to us, so we could not verify their effect on bike-sharing demand. For the sake of completeness, let us also note that the research in [4] only modelled one specific location, so it had no need for GPS coordinates or other location-related attributes.

6. Discussion

Compared to [8], which modelled bike-sharing demand for one-minute intervals, we considered it both more appropriate and more practical to model it on an hourly or daily basis, which should reveal the macro-level usage patterns more clearly. With one-minute intervals, the stochastic component of the bike-sharing demand remains too prominent, which makes the prediction unnecessarily difficult, as demonstrated by the value of Mean Absolute Percentage Error (MAPE) reported in [8], which reached 290%.

At the same time, hourly predictions should also suffice for logistics purposes: it is unlikely that providers would actually shift bikes from places where they are deposited to those where they are needed more often than hourly. In our case, we decided to derive hourly predictions from the daily ones. In other words, our main models predict daily demand, from which hourly demand can be estimated, e.g. based on histograms of hourly ridership for each day of the week.

In the first phase of our experiments (without including meteorological variables in our models), we observed the values of Relative-Absolute Error (RAE) in the region of 0.3 to 0.55 (30% to 55%) and of the Correlation Coefficient in the range from 0.80 to 0.93. We used RAE rather than MAPE because a non-negligible number of our consolidated daily records showed zero daily trips for some docking stations: for them, MAPE could not be calculated at all, as it would involve a division by zero.

Our attempt at deriving hourly demand from daily totals also achieved promising results. Although the correlation coefficient of our hourly predictions decreased to 0.72 and their relative absolute error increased to 0.93, the value of the Mean Absolute Error, which is the average number of bikes by which our hourly predictions differed from the reality, was only 1.179. This level of accuracy might well suffice for bike-sharing service operators: assuming that they redistribute bikes among their stations on an hourly basis, they would just need to add one or two more bikes to the predicted hourly numbers to virtually guarantee that each station has some free bikes on offer throughout the day.

We would like to point out that we derived the hourly predictions from the daily ones based on the statistical distribution of bike-sharing demand over individual hours of an average day. It was therefore a very simple and straightforward calculation, based solely on the average hourly statistics. At the same time, our daily demand models, which served as input into these calculations, did take into account various spatial and temporal aspects like geographical location, day of the week, month, year, etc.

In the second phase of our experiments, with meteorological variables included in our models, we indeed observed an increase in accuracy, although it was less pronounced than we expected. One reason may be that typical weather for a season is already coded in the number and especially the weight of each calendar month. In other words, people already know what kind of weather to expect, say, in October, so any additional information as to whether it is going to rain on a given October day or not, will not modify their biking behaviour significantly. In addition, the fact that our weather variables were not measured directly at the docking stations but rather at a nearby airport may also have played a role.

Overall, the main advantage of our models is their simplicity and modest computing demands during prediction, which means that, after training, they could be employed even on mobile devices. Another advantage of our approach is its scalability: bike-sharing demand for different stations can be calculated independently on different, possibly mobile and widely distributed, computing nodes.

As for the limitations of our models, the main one is the linearly growing trend of bike-sharing demand over the years, which may be appropriate for interpolation over the covered period 2014-19 and for short-term prediction, but not for long-term extrapolation into the future. Even in ideal situations, a nonlinear (e.g. polynomial) extrapolation might be needed for medium-term future forecasts, not to speak of abrupt and dramatic reductions of mass mobility, such as the present one due to the Covid pandemic, which no model, however complex, can accurately predict in advance.

7. Conclusion

In this article, we have dealt with spatiotemporal modelling of shared bike trips in Montreal in the six-year period from 2014 until 2019. Compared to [8], which addressed a similar topic by deep learning, we adopted a different approach consisting of different accuracy metrics and different types of attributes and models. We also modelled longer periods. These choices influenced the speed of model training as well as model accuracy. In the first phase of our experiments, without incorporating meteorological variables, our trained models reached Relative Absolute Error of about 0.3 (for 10-Fold Cross-Validation) and 0.55 (for Hold-Out Validation). After recalculation to hourly values, they reached Mean Absolute Error 1.179, and Correlation Coefficient 0.72, which we considered sufficient for practical application.

When we incorporated available meteorological variables (temperature, precipitation, wind speed) and weighted individual stations by the relative share of bike trips starting in them, the accuracy of our models increased, although the increase was less marked than we expected. The highest increase was for Random Forest model and Hold out validation, where the correlation of the predicted with the actual bike-sharing demand grew from 0.80 to 0.8411 and the relative absolute error decreased from 0.5541 to 0.4868.

We also evaluated the importance of our attributes with respect to bike-sharing demand and confirmed several observations noted in [4], although our situation and approach slightly differed.

In the future we plan to incorporate more meteorological variables in our models, hopefully also from locations closer to the bike-sharing stations. We expect this will help us to increase the accuracy of our models even further. Potential sources for these data include NOAA [12] and ECMWF [13].

Acknowledgments

This research was supported by the projects VEGA 2/0125/20, and SVASAS OPVaI-MH/DP/2018/2.2.2-20.

References

[1] Krammer, Peter; Kvassay, Marcel; Hluchý, Ladislav. Spatiotemporal modelling of transport demand. In Procedia Computer Science, 2020, vol. 175, p. 349-356. https://doi.org/10.1016/j.procs.2020.07.050

[2] Historical Climate Data, available in March 2021 at: https://climate.weather.gc.ca/

[3] Paul DeMaio: Bike-sharing: History, Impacts, Models of Provision, and Future, Journal of Public Transportation, Vol. 12, No. 4, 2009, pp. 41 – 56, https://scholarcommons.usf.edu/jpt/vol12/iss4/3/ https://doi.org/10.5038/2375-0901.12.4.3

[4] Oriol Cosp Arqué: Demand forecast model for a bicycle sharing service, 2015, https://upcommons.upc.edu/bitstream/handle/2117/78121/Tesina.pdf?sequence=1&isAllowed=y

[5] Chen, L., Zhang, D., Wang, L., Yang, D., Ma, X., Li, S., Wu, Z., Pan, G., Nguyen, T.M.T., Jakubowicz, J., 2016. Dynamic cluster-based over-demand prediction in bike sharing systems, Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing, ACM, New York, NY, USA. pp. 841–852,https://dl.acm.org/doi/10.1145/2971648.2971652

https://doi.org/10.1145/2971648.2971652

[6] Lin, L., He, Z., Peeta, S., 2018. Predicting station-level hourly demand in a large-scale bike-sharing network: A graph convolutional neural network approach. Transportation Research Part C: Emerging Technologies 97, 258–276. https://doi.org/10.1016/j.trc.2018.10.011

[7] Chemla, Daniel; Meunier, Frédéric; Calvo, Roberto Wolfler: Bike sharing systems: Solving the static rebalancing problem. Discrete Optimization, 2013, 10.2: 120-146, https://www.sciencedirect.com/science/article/pii/S1572528612000771 https://doi.org/10.1016/j.disopt.2012.11.005

[8] Liu, Xu, et al. "Multi features and multi-time steps LSTM based methodology for bike sharing availability prediction." Procedia Computer Science 155 (2019): 394 - 401, on Future Networks and Communications in 2019. https://doi.org/10.1016/j.procs.2019.08.055

[9] Toronto Bikeshare Data, Bike Share Toronto Ridership, avail. March 2021, https://www.kaggle.com/jackywang529/toronto-bikeshare-data

[10] Bixi Montreal Bikeshare Data, Bikeshare information for Bixi Montreal, https://www.kaggle.com/jackywang529/bixi-montreal-bikeshare-data

[11] Montreal managed bike-sharing system, avail. March 2021, https://bixi.com/en/page-27

[12] National Oceanic and Atmospheric Administration, avail. March 2021, https://www.noaa.gov/

[13] European Centre for Medium-Range Weather Forecasts, avail. March 2021, https://www.ecmwf.int/en/about