Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 180 (2021) 366–370

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Industry 4.0 and Smart Manufacturing (ISM 2020)

Route Duration Prediction in a Stochastic and Dynamic Vehicle Routing Problem with Short Delivery Deadlines☆

Nikolaus Frohner[a], Matthias Horn[a], Günther R. Raidl[a]

[a]*Institute of Logic and Computation, TU Wien, Favoritenstrae 9–11/192–01, 1040 Vienna, Austria*
Email addresses: {nfrohner,horn,raidl}@ac.tuwien.ac.at

Abstract

We are facing a real-world vehicle routing problem where orders arrive dynamically over the day at an online store and have to be delivered within short time. Stochastic information in form of the expected number and weight of orders and the traffic congestion level is available upfront. The goal is to predict the average time needed to deliver an order for a given time and day. This information is desirable for both routing decisions in the short horizon and planning vehicle drivers' shifts with just the right capacity prior to the actual day.

We compare a white box linear regression model and a neural network based black box model on historic route data collected over three months. We employ a hourly data aggregation approach with sampling statistics to estimate the ground truth and features. The weighted mean square error is used as loss function to favor samples with less uncertainty. A mean validation R^2 score over 10×5-fold cross-validations of 0.53 indicates a substantial amount of unexplained variance. Both predictors are slightly optimistic and produce median standardized absolute residuals of about one.

© 2021 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the scientific committee of the International Conference on Industry 4.0 and Smart Manufacturing

Keywords: Performance Prediction; Dynamic and Stochastic VRP with Time Windows; Weighted Least Squares

1. Introduction

We consider a dynamic and stochastic vehicle routing problem arising in a real-world application where orders arrive over the day at an online store and are to be delivered within short time from a single depot. Decisions have to be performed in real-time regarding the clustering of orders, the routes for the vehicles, when to start these routes, and when to send drivers home from their shifts with some flexibility concerning shift ends.

Stochastic information per hour regarding the number of arriving orders over the day, their expected weight classes, and the traffic congestion levels is available upfront. The goal is to facilitate this information to avoid myopic decision for the aforementioned aspects and account for the dynamism of the problem. Additionally, the prior shift planning itself relies on this as we need to have the right amount, possibly including a safety buffer, of driving capacity available over the day.

Our approach is to condense all this stochastic information into one single relation, the mean order delivery time per hour $\phi_t(a)$, which maps for a given day a and hour t to the time needed to serve an order. Conversely, the reciprocal value is the number of orders that can be served per hour.

☆ This project is partially funded by the Doctoral Program "Vienna Graduate School on Computational Optimization", Austrian Science Foundation (FWF) Project No. W1260-N35.

1877-0509 © 2021 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the scientific committee of the International Conference on Industry 4.0 and Smart Manufacturing
10.1016/j.procs.2021.01.175

In this work, we train and evaluate corresponding predictors $\tilde{\phi}_t(a)$ in a supervised learning setting. Different white box and black box models are studied. Error analysis is performed to allow selection between more conservative and more aggressive use of the prediction that naturally comes with uncertainties.

Data of the daily activity has been collected over three months and is used both for training and evaluation in 5-fold cross-validations. Data acquisition was subject to substantial measurement noise due to manual logging of route legs by the drivers. Both, weighted least squares regression and a neural network (NN) with a weighted sample loss function resulted in the best results on the collected data.

In Section 2, we discuss related work on which ours is based. Section 3 introduces our linear white box model, whereas Section 4 describes the black box model with the neural network. Results are presented in Section 5 after which we conclude in Section 6.

2. Related Work

The dynamic and stochastic vehicle routing problem we consider is presented in detail together with an adaptive large neighborhood search routing algorithm and large horizon planning approach for handling dynamic shift ending flexibilities in [1]. There, also first steps towards predicting order delivery time on simulated data without traffic and stop time variability are performed.

We make use of a classical result by Beardwood et al. [2] from the late 50s which proves asymptotic dependence $\sim k_l \sqrt{n\mathcal{A}}$ of the length of an optimal traveling salesperson problem tour on the number of cities n, distributed on some geometric area $\mathcal{A}$; k_l is an empirical constant depending on the distribution of the cities and the metric.

An extension to the capacitated vehicle routing problem (CVRP) is presented by Daganzo in 1984 [3], accounting for the capacity C of vehicles and the average distance to customers $\bar{r}$ in a delivery area. The suggested model for the length is $2\bar{r}n/C + k_l \sqrt{n\mathcal{A}}$. An intuitive explanation is that n/C routes are needed and that each vehicle has to drive to a dense region of customers yielding round trip distance $2\bar{r}n/C$ and intra-customer distance $k_l \sqrt{n\mathcal{A}}$, following [2].

Figliozzi [4] builds upon these results and proposes further refined white box models to estimate route lengths of classical CVRP benchmarks with and without time windows and a real-world freight forwarding problem. Results with high estimation accuracy are obtained for corner-depot instances by making use of the known number of routes m and adding an additional term $k_b \sqrt{\mathcal{A}/n}$. The problem is similar to ours with the main difference that we do not know the number of routes in advance. The same author describes a traffic-aware tour model [5] with optimization constraints that link the number of vehicles needed and the variability induced by congestion to ensure service guarantees with a certain confidence. From the latter we adopt the approach to model the increase of free-flowing travel times due to congestion by a corresponding factor.

3. White Box Model

Let $R_t(a)$ be the set of all routes that start within the hour t on day a. We consider one route $r \in R_t(a)$ with $l_r \geq 1$ served customers $v_1, \ldots, v_{l_r}$. Its total associated duration Δ_r is assumed to consist of the loading time of the goods, the total travel time, and the total stop times at the customers, formally expressed by

$$\Delta_r = \Delta_r^{\text{load}} + \Delta_r^{\text{travel}} + \Delta_r^{\text{stop}}. \tag{1}$$

We assume that the mean time per customer Δ_r/l_r is a random variable drawn from an unknown distribution $\mathcal{D}$ with mean $\phi_t(a)$—our ground truth.

As stated before, we seek to create a predictor $\tilde{\phi}_t(a)$. To this end, we propose three features observed in the hour the route is started t and the subsequent hour $t+1$ that have likely the strongest influence on Δ_r: The number of orders due n, the traffic congestion level ξ as average relative increase of base travel times within the delivery area, and the orders' weight w, each in the given hour.

We assume that the spatial distribution of the orders is stationary over the day. Then the features n and ξ have supposedly the strongest impact on Δ_r^{travel} but are negligible for the rest. In contrast, w determines the time needed for loading Δ_r^{load} and delivering the goods to the customers at the stops Δ_r^{stop}. Predicting these features are difficult problems in their own right. For this work, we assume corresponding predictors to be given and do not explicitly analyze the uncertainties they introduce.

Adapting [4, 5] and normalizing the delivery area to $\mathcal{A} = 1$, we propose the following model with parameters to be estimated by (weighted) linear regression:

$$\tilde{\phi}_t^{wb}(a) = k_c + \sum_{t' \in \{t, t+1\}} k_w^{t'} w^{t'} + k_m^{t'} \xi^{t'} + k_l^{t'} \frac{\xi^{t'}}{\sqrt{n^{t'}}} \tag{2}$$

The superscript t' indicates the feature with corresponding parameter for the given hour. We omit this for convenience everywhere else, the use of the features in

the current hour t and subsequent hour $t+1$ is henceforth implicit.

Parameter k_c represents the average loading and stop time, independent of any feature. Additionally, k_w accounts for an additional increase in those by the weight or bulkiness of the delivered goods.

Parameter k_m belongs to a traffic proportional term arising from travel times to the first and back from the last customer, loosely corresponding to the term $2\bar{r}/C$ from [3]. We make the simplifying assumption that on average, the maximum number of customers in a route l_r implied by the delivery deadlines is constant for our problem. The actual relation is likely to be more complicated since we expect it to increase with larger n, as more efficient routes can be created and the number of required routes decreases. However, l_r are limited by the finite stop times needed for each customer and the guaranteed delivery times.

Parameter k_l accounts for the travel times between the customers that are expected to increase with traffic but decrease with the number of orders $\sim 1/\sqrt{n}$, leading to the feature $\xi/\sqrt{n}$.

Significance and contribution to the prediction quality of all the coefficients will be studied in Section 5 with linear regression using scikit-learn [6].

4. Black Box Model

In contrast to the white box model, we consider a general function with learnable parameters θ

$$\tilde{\phi}_t^{\mathrm{bb}}(a;\theta) = g(w^t, \xi^t, n^t, w^{t+1}, \xi^{t+1}, w^{t+1}; \theta) \tag{3}$$

which is realized by a fully connected feed-forward neural network with two hidden layers of 64 neurons each and ReLU activation functions This network should allow for a piecewise linear approximation of $\phi_t(a)$ with reasonably high resolution which was confirmed by preliminary tests to define its architecture. The traffic congestion factor lies approximately in the range from 1.0 to 1.6, corresponding from free-flowing traffic to heavy congestion with 60% increase of travel time.

As loss function, the weighted mean squared error is to be minimized, where S^{T} denotes a training set of day-hour tuples

$$L(\theta) \propto \sum_{(a,t)\in S^{\mathrm{T}}} \left(\bar{\phi}_t(a) - \tilde{\phi}_t^{\mathrm{bb}}(a;\theta)\right)^2 / s^2_{\bar{\phi}_t(a)} \tag{4}$$

and $\bar{\phi}_t(a)$ denotes the sample mean by averaging over the Δ_r/l_r occurring in (a,t). According to the central limit theorem, the sample standard error $s_{\bar{\phi}_t(a)}$ decreases with the inverse square root of the number of routes. Since $\phi_t(a)$ itself is unknown, hours, where less variance and more routes are observed are given more weight in the loss function than those with higher uncertainty.

We implemented the neural network approach using TensorFlow [7] with the Adam [8] optimizer. Early stopping monitoring the performance on 30% validation data is used for regularization. As initial learning rate we choose 0.01 with dynamic updates by factors of one-tenth down to 10^{-4} when plateaus are encountered.

5. Results

We have prepared data by aggregating routes for day-hour tuples collected over three months. Routing is performed in a semi-automatized way by experienced dispatchers with the help from a routing software. Travel times to the customers and stop times at the customers are logged by the drivers, the loading time and the travel time back to the depot are estimated.

The features vector is given as $(w, \xi, n)_{(a,t)}$, where w is estimated by calculating the hourly average for the same weekdays (or holidays as own type) over the month, ξ is determined by comparing the base driving times as provided by a public routing API with the logged driving times of the routes and taking the average for each (a,t), and n is taken to be the number of orders due in (a,t). Since routes started at the end of an hour are likely to be more influenced by the subsequent hour, we always present $(w, \xi, n)_{(a,t+1)}$ to our models as well.

Routes with implausible logs were discarded, and a quantile cut of $[0.02, 0.98]$ on the congestion level ξ is applied to remove routes with unrealistic driving time logs and outliers. This reduces the number of routes used for aggregation down to 87%.

The labels are calculated as

$$\bar{\phi}_t(a) = \frac{1}{|R_t(a)|} \sum_{r\in R_t(a)} \Delta_r/l_r. \tag{5}$$

The sample standard error $s_{\bar{\phi}_t(a)} = s_{\phi_t(a)}/\sqrt{|R_t(a)|}$ is used for weighting the labels as done in Eq. (4).

As first sanity checks, Pearson correlation coefficients are calculated yielding moderate $\rho_{1/\sqrt{n},\bar{\phi}_t(a)} \approx 0.6$ and $\rho_{\xi,\bar{\phi}_t(a)} \approx 0.5$. Correlation between the weight w and $\bar{\phi}_t(a)$ is much smaller, therefore we further check for significance in a t-test of a linear regression $\left(\overline{\Delta^{\mathrm{stop}}/l}\right)_t(a) \sim k_c + k_w w + k_m \xi$. We include also the traffic to check how it contributes to the explanation of the stop times per customer when competing with the weight. For k_w this reveals significance with a t-score of almost 10 and no significance for k_m with a t-score

Table 1. Standardized residuals' R^2 scores, medians (Med), and median absolute deviation from the medians (MAD) averaged over ten 5-fold cross-validations with different models and ablation.

model	$\overline{R^2_{\text{train}}}$	$\overline{R^2_{\text{val}}}$	$\overline{\text{Med}^{\text{res}}_{\text{train}}}$	$\overline{\text{Med}^{\text{res}}_{\text{val}}}$	$\overline{\text{MAD}_{\text{train}}}$	$\overline{\text{MAD}_{\text{val}}}$
WLS-$\xi/\sqrt{n}$,ξ,w	0.52	0.50	-0.34	-0.34	**1.37**	1.39
WLS-$\xi/\sqrt{n}$,ξ	0.52	0.50	-0.34	-0.34	**1.37**	1.39
WLS-$\xi/\sqrt{n}$,w	0.47	0.45	-0.43	-0.43	1.44	1.45
WLS-$\xi/\sqrt{n}$	0.47	0.45	-0.43	-0.43	1.44	1.44
WLS-$1/\sqrt{n}$	0.36	0.34	-0.52	-0.52	1.51	1.52
WLS-ξ	0.28	0.27	-0.63	-0.63	1.46	1.46
NN-n,ξ,w	**0.53**	**0.53**	**-0.32**	**-0.32**	1.39	**1.38**
NN-n,ξ	0.52	0.51	-0.33	-0.33	1.41	1.39
NN-n	0.40	0.39	-0.47	-0.46	1.51	1.50

only slightly above zero. We are in the range where the t-score approximately equals the z-score.

Table 1 compares average training and validation performance values on 10×5-fold cross-validations (hence $N = 50$). Weighted least squares (WLS), the white box model, is compared with the neural network based black box model. Only pairs (a, t) with at least four routes are considered to have somewhat valid sample statistics, leading to the total batch size of 1081. We measure weighted R^2 scores, median residuals to measure prediction bias, and median absolute deviation of the median (MAD). We make use of robust statistics since our problem is subject to outliers. As an ablation study, we remove each feature individually and observe that for WLS the order weights w have no impact on our figures. This is backed up by the fact that t-tests on the significance of coefficients in the linear model yield scores close to zero for k_w when combined with and seemingly dominated by the other features. In contrast for the NN, we see a slight increase in the performance values when going from n, ξ to n, ξ, w.

As expected, using either n or ξ without the other feature, hurts the performance substantially. The neural network achieved the best performance averaged over the 10×5-fold cross-validation runs with an R^2 score of 0.53, median of -0.32 and median deviation from the median of 1.38 on the standardized residuals.

By comparing training and validation figures, we observe very little, if any, overfitting. One reason for this is conjectured to be the substantial amount of noise in the data which makes overfitting with the given small amount of features and small (WLS) to moderate (NN) model capacity quite difficult. Possible noise sources are manifold and include the driver bias (each driver has a different pace), non-stationarity of the spatial distribution of the orders, traffic variability over the delivery area, manual leg duration measurement error, and the hourly estimation of the features.

The models consistently underestimate the time needed to serve an order, as indicated by the negative

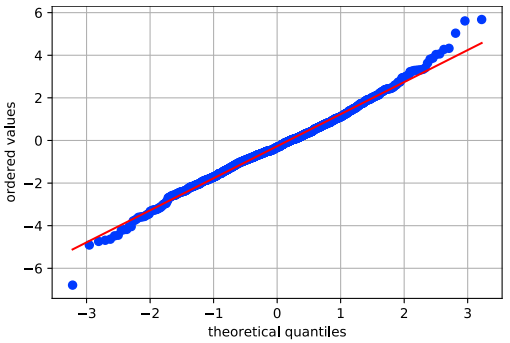

Fig. 1. QQ-plot for standardized residuals of neural network prediction on whole data batch.

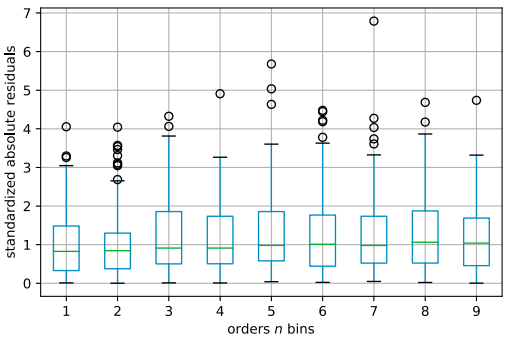

Fig. 2. Standardized absolute residuals of neural network prediction on whole data batch over bins of number of orders n.

bias. Figure 1 depicts a QQ-plot of the standardized residuals (using the sample standard errors) vs. a normal distribution. We observe good resemblance with a normal distribution (linear fit with $R^2 = 0.996$) with the mentioned slight negative bias. The standardized absolute residuals over the number of orders n bins is depicted in Fig. 2. We observe quite stable performance over n with median standardized absolute residuals around one.

In conclusion, both model types with all features have comparable performance, whereas the NN model shows marginally but consistently better validation performance. This is illustrated by a comparison of both predictors in action on an example day in February 2020 in Fig. 3. Some outliers regarding the mean order delivery time are observed on this day, smoothed out by the predictors. The behavior is not exactly the same, yet they are quite similar. This leads us to the conclusion that data and feature bias are dominating and both models act approximately equivalent in our high-noise setting on our historical data. Still, the neural network does not show reasonable (asymptotic) behavior for unseen regions of the features, in particular n, which is the advantage of the white box approach.

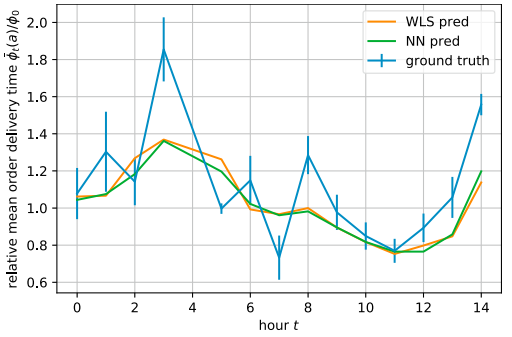

Fig. 3. Example day in February 2020, comparing the ground truth in form of the relative mean order delivery time $\bar{\phi}_t(a)/\phi_0$, with ϕ_0 a total average, with the NN and the WLS predictions.

6. Conclusions

We analyzed the performance of a linear white box model and a neural network based black box model on a route duration prediction problem in a stochastic and dynamic vehicle routing problem. It is derived from a real-world setting where orders arrive in an online store dynamically over the day and have to be delivered to customers within short due times. The goal is to predict the mean order delivery time for a given day and hour. This information is required for improved routing decisions and planning the number of required drivers over the day. Stochastic information per hour is given regarding the number of arriving orders, their average weight, and the traffic congestion level.

Data has been collected over three months by manual logging of travel and stop times by the drivers, subject to substantial measurement noise. The ground truth and the traffic feature were accessed by sampling statistics and subject to noise themselves, depending on the number of routes in a given hour. Hence we chose the mean squared error weighted with the inverse sample standard errors. Based on results from the literature, we proposed a linear model with parameters to be tuned by weighted linear regression and a small two-layer neural network.

In an ablation study, we showed the importance of combining both the number of orders and traffic congestion level as features. All features were presented to the models for the current and the subsequent hours. Order weights have significant impact on the stop times at the customers, but did contribute little to the overall predictive performance.

In the end, the neural network achieved the best performance with the cost of unreasonable predictions for regions where no or not much training data was available. Smoother approximation with reasonable asymptotic behavior is achieved with the weighted linear regression approach, which also gives comparable performance. Standardized absolute residuals are for both stable with a median about one over increasing numbers of orders.

The real-world performance of the predictors is an open question. High quality data collection including traffic information from a third-party source would be recommendable to resolve travel time and stop time effects. The number of orders that are visible when a route is started should be logged to learn how the number of orders feature could be improved. One possible criticism for the suggested weighted loss functions is that hours with more routes are likely hours with better performance, which is conjectured to result in the negative (optimistic) bias of the predictors. Naturally, the less busy hours are harder to predict, since less data is available and the variance is intrinsically higher. This can be mitigated by employing a safety buffer based on the error analysis of the predictors.

Further work is concerned with refining the models on simulated data where noise sources can be excluded and further validation on newly collected data.

References

[1] N. Frohner, G. R. Raidl, A double-horizon approach to a purely dynamic and stochastic vehicle routing problem with delivery deadlines and shift flexibility, in: P. D. Causmaecker, E. Özcan, G. V. Berghe (Eds.), Proceedings of the 13th International Conference on the Practice and Theory of Automated Timetabling - PATAT 2021: Volume I, Bruges, Belgium, 2020.

[2] J. Beardwood, J. H. Halton, J. M. Hammersley, The shortest path through many points, in: Mathematical Proceedings of the Cambridge Philosophical Society, Vol. 55, Cambridge University Press, 1959, pp. 299–327.

[3] C. F. Daganzo, The distance traveled to visit n points with a maximum of c stops per vehicle: An analytic model and an application, Transportation science 18 (4) (1984) 331–350.

[4] M. A. Figliozzi, Planning approximations to the average length of vehicle routing problems with varying customer demands and routing constraints, Transportation Research Record 2089 (1) (2008) 1–8.

[5] M. A. Figliozzi, The impacts of congestion on commercial vehicle tour characteristics and costs, Transportation research part E: logistics and transportation review 46 (4) (2010) 496–506.

[6] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[7] TensorFlow: Large-scale machine learning on heterogeneous systems, software available from tensorflow.org (2015). URL https://www.tensorflow.org/

[8] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980.