UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

ESCOLA DE ENGENHARIA

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

Douglas Zechin

PROBABILISTIC TRAFFIC BREAKDOWN FORECASTING THROUGH BAYESIAN APPROXIMATION USING VARIATIONAL LSTMS

Porto Alegre

2023

Douglas Zechin

Probabilistic traffic breakdown forecasting through Bayesian approximation using Variational LSTMs

Tese submetida ao Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal do Rio Grande do Sul como requisito parcial à obtenção do título de Doutor em Engenharia, na área de concentração em Sistemas de Transportes.

Orientadora: Helena Beatriz Bettella Cybis, PhD

Porto Alegre

2023

Douglas Zechin

Probabilistic traffic breakdown forecasting through Bayesian approximation using Variational LSTMs

Esta tese foi julgada adequada para a obtenção do título de Doutor em Engenharia e aprovada em sua forma final pelo Orientador e pela Banca Examinadora designada pelo Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal do Rio Grande do Sul.

_____

Helena Beatriz Bettella Cybis, Dr.

Orientador PPGEP/UFRGS

_____

Prof. Alejandro Germán Frank

Coordenador PPGEP/UFRGS

Banca Examinadora:

Professor André Luiz Cunha, Dr. (USP)

Professor Felipe Caleffi, Dr. (UFSM)

Professor Michel José Anzanello, *Ph.D.* (PPGEP/UFRGS)

Dedicatória

Dedico esta tese à minha família, em especial a
meus pais João Roque e Magda e minha
namorada Bruna, que tanto me apoiaram.

AGRADECIMENTOS

Primeiramente agradeço o apoio conferido por toda minha família em minha jornada de estudos desde a infância, o qual culminou – e espero não esteja limitado – a esta tese. Merecem agradecimentos especiais meus pais João Roque e Magda, pela dedicação incondicional em me incentivar e apoiar nos estudos, e de minha namorada Bruna, pela companhia diária em todos os momentos deste trabalho, pela inspiração, pelas palavras de carinho proferidas nos momentos mais necessários, pelas ideias sensatas e pelo amor expresso em cada auxílio.

Agradeço também a todos os amigos que ao seu modo participaram desta jornada sem citar nomes para não cometer injustiças. Abro exceção para o amigo João, com quem tive a despretensiosa conversa que deu origem ao tema desta tese.

Agradeço a todos meus professores pelo conhecimento transmitido e que, consolidado, resultou neste trabalho. Agradeço especialmente à minha oritentadora Helena Cybis, que me acompanha desde a graduação e que tem desempenhado papel crucial em direcionar os rumos de meus trabalhos. Agradeço a meus colegas de laboratório Felipe e Maurício, que me acompanharam nos primeiros estudos envolvendo machine learning, e aos colegas Fagner, Yan, Laísa, Lucas, Francisco, Daniela, Rodrigo, Shanna, Tânia, Marcelle, Matheus e Daniel com quem pude trocar ricas experiências que contribuíram para minha formação.

ABSTRACT

Robust artificial intelligence models have been criticized for their lack of uncertainty control and inability to explain feature importance, which has limited their adoption. However, probabilistic machine learning and explainable artificial intelligence have shown great scientific and technical advances, and have slowly permeated other areas, such as Traffic Engineering. This thesis fulfils a literature gap related to probabilistic traffic breakdown forecasting. We propose a traffic breakdown probability calculation methodology based on probabilistic speed predictions. Since the probabilistic characteristic is absent in traditional formulations of neural networks, we suggest using Variational LSTMs to make the speed forecasts. This Recurrent Neural Network uses Dropout to produce a Bayesian approximation and generate probabilistic outputs. This thesis also investigates the effects of inclement weather on traffic breakdown probability and methods for identifying traffic breakdowns. The proposed methodology produces great control over the probability of congestion, which could not be achieved using deterministic models, resulting in important theoretical and practical contributions.

Key words: traffic breakdown, traffic forecasting, neural networks, inclement weather, Bayesian statistics.

SUMMARY

LIST OF FIGURES

5 COMPLEMENTARY MATERIALS

LIST OF TABLES

1 INTRODUCTION

Congestion is a traffic state with unreasonable performance due to excess demand in a given segment or network. It occurs at a certain level in most cities worldwide in both urban and suburban areas and might impose high social costs. A study in Brazil suggests that travels with more than 30 minutes, an ideal maximum value for commuting (Bertaud, 2018), take on average 114 minutes and generate a loss of productive potential greater than R$ 111 billion yearly. Porto Alegre, where the study site of this thesis is placed, is the fifth city in terms of productive potential loss between the capitals, reaching more than R$ 3.4 billion, or 2.9% of its IGP (FIRJAN, 2015).

The causes of traffic congestion vary from site to site. In this study, we strictly explore the phenomenon of congestion formation on highways. In this context, congestion can be generated due to bad geometric design, accidents, inclement weather, spillback from urban congestion nearby, constructions and, more recurrently, due to excessive demand. The excess of demand means that demand is higher than a certain threshold that the segment supports with reasonable traffic conditions, which is formally called capacity by the Traffic Engineering community (TRB, 2016).

Highway capacity is strongly related to traffic breakdown, which is the main subject of this thesis. Breakdown is the point from which the traffic flow begins transitioning from the free flow regime to a congested state. It usually happens when demand exceeds capacity and precedes a rapid speed and decrease in traffic flow. This phenomenon is depicted in **Figure 1**.

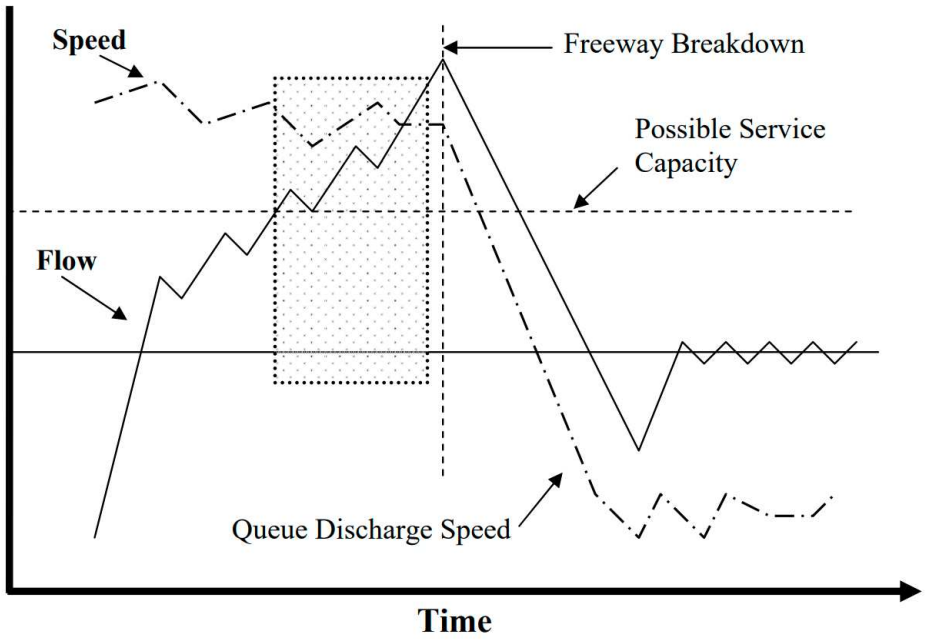

Figure 1: Traffic breakdown. Adapted from (Chaudhary *et al.*, 2004).

Traffic breakdown happens when a given segment is so saturated that an inaccurate lane change or breaking causes a cascade effect on the upstream traffic, leading to a great reduction in speed and even complete stops. From this point on, the traffic flow will hardly recover to a free-flow state until the demand also decreases. The exact value of traffic flow that causes the breakdown is not a fixed number related to the capacity. It might vary daily due to changes in the demand profile, weather conditions and even the attention and dexterity level of each driver. This imposes a stochastic characteristic to traffic breakdown and hence to highway capacity, which the Transportation Engineering Community has widely studied (Brilon, Geistefeldt e Regler, 2005; Chen e Ahn, 2018; Elefteriadou, Roess e McShane, 1995; Kondyli *et al.*, 2013; Persaud, Yagar e Brownlee, 1998; Qu, Zhang e Wang, 2017) and is called "classical understanding of stochastic highway capacity" (Kerner, 2019).

Congestion side effects can be mitigated through multiple strategies. The most expensive are changes in the infrastructure, such as the construction of extra lanes, which immediately increase capacity. Although straightforward and effective, this approach has some side effects, such as induced demand, increased walking distance for crossing and impacts on the surrounding area. An alternative and often complementary approach is the use of Active Traffic Management (ATM) strategies, such as ramp metering (Zechin, Cybis e Caleffi, 2016) and variable speed limits (Caleffi, Moisan e Cybis, 2016). ATM strategies aim to respond to changes in the traffic state and dynamically actuate to coordinate it, optimize traffic output, and, among others, detain the occurrence of traffic breakdown.

ATM strategies have increasingly incorporated forecasted data due to the abundant information available and the development of better forecasting models. Between the better-performing and most adopted forecasting models, LSTM (Long Short-Term Memory) neural networks and their variations have played a significant role (Akhtar e Moridpour, 2021). Therefore, anticipating traffic states, especially the occurrence of traffic breakdown, offers substantial contributions to ATM strategies and the mitigation of congestion effects.

This thesis proposes a methodology for traffic breakdown probability forecasting, which could be incorporated into multiple ATM strategies. For doing so, we suggested using Variational LSTMs, a variation of the classical LSTMs that can make probabilistic forecasts through Bayesian approximation. Our methodology uses the Variational LSTM to forecast sequences of speed distributions and, using a proposed formulation, calculate the breakdown probability for multiple future time steps.

1.1 **Theme and main goals**

This thesis is inserted in the area of Traffic Engineering dedicated to studies of the road environment and has an interface with the area of Intelligent Transportation Systems. The study's main objective is to propose a methodology for short-term probabilistic traffic breakdown forecasting on highways, and for that, we use a Variational LSTM neural network. The secondary objectives are:

a) Evaluate the effect of inclement weather on traffic breakdown probability;
b) Propose a forecasting model with probabilistic characteristics;
c) Increase the relative importance of predictions made during high-demand periods;
d) Evaluate the predictions on different traffic states to better understand its quality;
e) Propose a model optimization methodology.

1.2 **Justification**

This research is justified due to its methodological, theoretical, and practical contributions to the existing literature. As for the methodological contributions, we proposed using Variational LSTM neural networks to produce probabilistic traffic speed forecasts. Classical neural networks are deterministic and lack the probabilistic characteristics required for our study. Variational LSTMs use Dropout during inference to approximate Bayesian inference, giving it probabilistic characteristics (Fortunato, Blundell e Vinyals, 2017; Gal e Ghahramani, 2016a; b). Other relevant methodological contributions are increasing the relevance of high-demand periods during the model training, which biases the model to focus on critical traffic conditions, and evaluating the predictions for different traffic states, which enlarges the understanding of its performance and we could not find in past studies.

The theoretical contributions are mainly related to the proposed formulation for traffic breakdown probability calculation based on probabilistic speed forecasts. Although there is extensive literature on, e.g., speed, flow, congestion, and travel-time forecasting (Akhtar e Moridpour, 2021), we could not find studies that aim to forecast breakdown probability and explore the forecasting horizon as we proposed. There is strong support for using recurrent neural networks for general traffic forecasting purposes, but few alternatives to deal with it probabilistically. We understand that this happens since probabilistic approaches of neural networks are still under development, and this knowledge has not thoroughly permeated into the Traffic Engineering community, which produces a gap that we have hopefully partially fulfilled.

The practical contributions of our methodology are especially related to the road operator side. Since our methodology enables producing probabilistic breakdown forecasts, road operators can anticipate the occurrence of this phenomenon with more control than by using traditional deterministic methods. This methodology can also be incorporated into Active Traffic Management strategies such as ramp metering and variable speed limits to increase their effectiveness. Although the probable direct user of our methodology is the road operator, we algo expect road users to have a better driving experience.

1.3 **Structure**

This thesis was written in a three-articles format. The document comprises six main sections, which are an Introduction, a chapter for each article, Complementary Materials and a final chapter of Conclusions. The articles' sequence and themes were essential for the realization of this work. In the first article, we understood the characteristics of the data we deal with, the characteristics of the traffic and the influence that rain has on the occurrence of breakdown. In the second, we tested and developed the necessary skills to make good speed predictions. In the third, we combine the concepts covered in both articles by proposing a methodology for breakdown forecasts that respects their stochastic characteristics.

The first article, entitled "**Influence of Rain on Highway Breakdown Probability"**, was presented at the *Transportation Research Board* conference and later nominated for publication in the *Transportation Research Record* magazine. In this article, we studied the effect of rain on the probability of breakdown occurrence. We crossed rainfall data with traffic data and plotted breakdown probability curves as a function of traffic flow for different rainfall intensities. We plotted these curves using the Kaplan-Meier survival analysis model, which allowed us to show a significant effect of rain on breakdown occurrence.

The second article, entitled **"Forecast of traffic speeds with an encoder-decoder LSTM neural-network"**, was presented at the national congress of ANPET and nominated for publication in *Revista Transportes*. This article proposes a methodology for future traffic speed predictions of up to 25min using LSTM neural networks. Given the conclusions of the first article, we included rainfall data as a feature of the proposed model. Neural networks are flexible and can be structured in different ways. Because of this, we suggested using the neural network hyperparameter optimization called Hyperband. We validated the predictions by building breakdown probability curves with them and with data collected in the field and statistically testing their equivalence.

The third article, entitled **"Probabilistic traffic breakdown forecasting with Variational LSTM neural networks"**, was published in *Transportmetrica B*. In this article, we proposed the probabilistic breakdown forecasting methodology that gives name to this thesis, being a natural sequence of the other articles that compose it. The motivations for this article are: (i) breakdown probability estimation models are not suitable for forecasting, (ii) the most used and best-performing traffic forecasting models use neural networks, (iii) the classic network formulations neural networks are deterministic, not producing information about the credibility of individual forecasts, (iv) the few studies that approach traffic forecasts in a probabilistic way do not focus on the breakdown. Because of this, there is a large gap in the literature regarding ways to perform probabilistic breakdown forecasts. In this article, we propose a methodology that fills this gap with two main contributions: (i) a formulation for calculating the breakdown probability using probabilistic velocity forecasts and (ii) to perform probabilistic velocity forecasts, using a neural network of the type Variational LSTM. This made it possible to calculate the probability of a breakdown occurring for different time horizons. The Variational LSTM neural network uses Dropout during inference to give the outputs a probabilistic character through approximate Bayesian inference. The choice of this neural network was made based on a bibliographic review and also on account of its probabilistic characteristic. To test the quality of the forecasts, we compared them with forecasts made using other baseline models. The proposed methodology provided a high level of control over the occurrence of a breakdown, which would not be possible using deterministic forecasts.

We believe that this work resulted in interesting contributions to the theory and practice of Traffic Engineering. In addition to having achieved significant results, we believe that a good part of the contributions is related to the development of the articles. We seek to make clear the steps taken to achieve these results, avoiding the use of closed models and explaining how we optimized the chosen neural network model. We also made the used database public, which allows other researchers to test and propose the use of other models in future studies (Zechin, 2022). We understand that the use of neural networks is still incipient in the area of Traffic Engineering compared to its potential, partly because of their explainability restrictions. We hope this work contributes, even if discreetly, to the evolution of this field of studies.

1.4 Research stages

To achieve the results of these three articles, we adopted the sequence of research stages presented **Figure 2**. We detailed each of these stages in the following subsections.

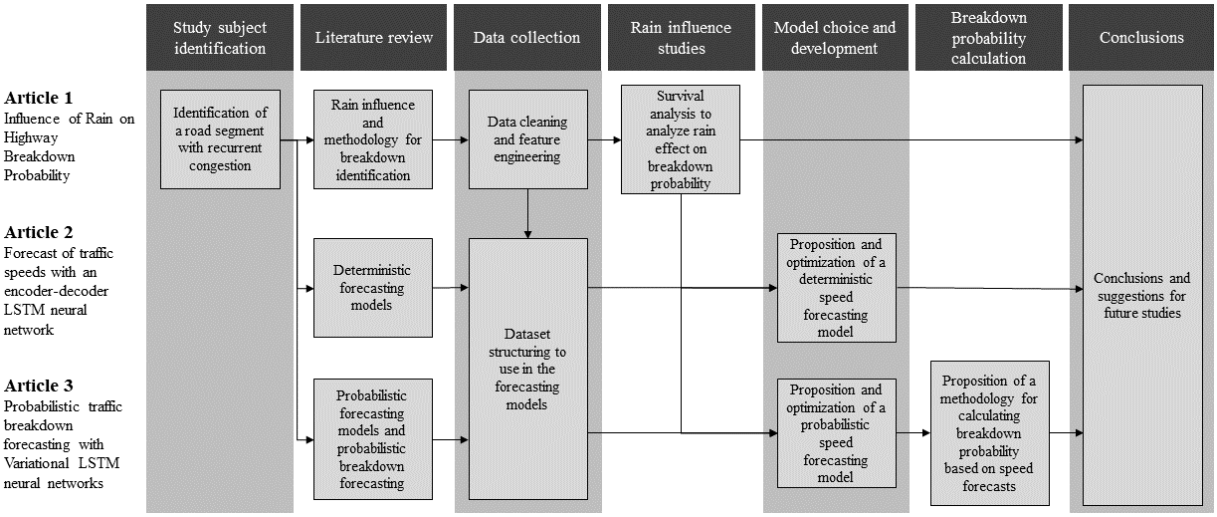

Figure 2: Research stages

1.4.1 *Study subject identification*

The study subject identification is a sequence of past studies in the region of the Guaíba Bridge, in the metropolitan region of Porto Alegre, Brazil. This region presents daily congestion during the morning due to the confluence of the traffic flow from the BR-290 highway with the flow from the bridge, as shown in **Figure 3**.

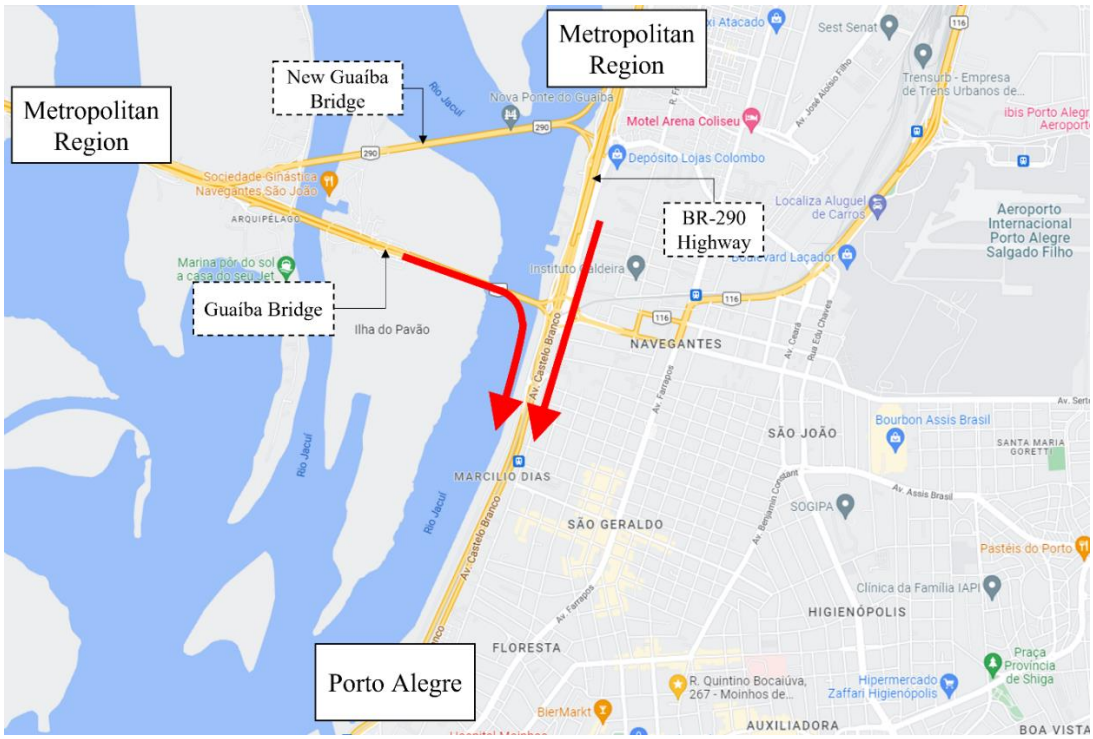

Figure 3: Study site. Source: Google Maps.

We used this region as a study site for several past studies of our research group. These studies include the effectiveness of active traffic management strategies, such as ramp metering and variable speed limits (Caleffi, 2018; Caleffi, Moisan e Cybis, 2016; Zechin, Cybis e Caleffi, 2016), influence of inclement weather on the occurrence of accidents (Caleffi *et al.*, 2016), evaluation of desired speed distributions (Galvan, Zechin e Cybis, 2019) and analysis of the impacts of the New Guaíba Bridge (Kappler, 2017), which was still under construction when the data of our study was collected.

The traffic breakdown subject surged as an approach to evaluating the congestion periods of this region. In the **first article** of this thesis, we discussed the influence of inclement weather on the traffic breakdown probability, which produced a better understanding of this phenomenon on the study site. This article raised the question of whether it was possible to forecast traffic speeds focusing on pre-breakdown periods, which we discussed in the **second paper**. Finally, the **third paper** united the knowledge created in the first two articles and proposed a methodology to forecast breakdown probability.

1.4.2 *Literature Review*

The scope of the literature review stage varies according to the article. In the **first paper**, we focus on the influence of inclement weather on traffic flow, traffic breakdown identification methodologies, traffic breakdown theory and calculation methodologies. The **second paper** focuses on forecasting methodologies and their implementations for traffic forecasting purposes. The **third paper** has a broader scope, and its literature review comprises traffic breakdown theory, traffic congestion forecasting, recurrent neural networks, and probabilistic approaches to recurrent neural networks.

1.4.3 *Data and study site*

The three articles used data collected on the BR-290 highway. The detectors used were located close to the access to the Guaíba Bridge and the data was collected between 2016 and 2017. The original traffic dataset was disaggregated so that each row represented an individual vehicle, timestamp, speed and vehicle type. Table 1 shows an excerpt of the original dataset.

Table 1: Original traffic dataset

Timestamp	Speed [km/h]	Lane	Type
2016-01-01 00:00:25.1	102	1	Car
2016-01-01 00:00:25.2	118	2	Car
2016-01-01 00:00:25.9	95	4	Car
2016-01-01 00:00:26.1	41	4	Car

2016-01-01 00:00:27.1	89	3	Car
2016-01-01 00:00:30.4	84	2	Car
2016-01-01 00:00:32.4	77	2	Truck
2016-01-01 00:01:04.4	108	3	Car
2016-01-01 00:01:16.6	99	3	Car
2016-01-01 00:01:20.5	115	2	Car
2016-01-01 00:01:39.2	56	4	Car
2016-01-01 00:02:21.9	64	2	Car
2016-01-01 00:02:29.7	101	2	Car

We aggregated the traffic dataset in 5min intervals. This approach produced a good balance between smoothening outliers and maintaining the level of detail required to characterize the breakdown. Due to our experience with this dataset, we understand that a smaller aggregation interval would introduce excessive variance to the final dataset. A bigger interval would hinder the correct description of the breakdown. In this region, a breakdown takes on average 4min to happen (Cybis *et al.*, 2013).

A data cleaning step was also necessary to produce a suitable dataset. The detectors presented malfunction during some periods, so those were mapped and removed from the final dataset. We also removed periods when the Guaíba Bridge was raised and when the traffic agency reported accidents in the surroundings.

Especially for Articles 2 and 3, during aggregation, we opted for producing different measures that could be used as input features for the models that captured as much information as possible. For example, instead of only calculating the average speed, we also calculated its variance, maximum speed, minimum speed, etc. This step is better described in each article. It is open for future studies discussing the effectiveness of this step for model performance and which features were more relevant, but we opted to limit the scope of our study at this point.

The rain dataset was acquired from a rain gauge 1.3km from the traffic detectors and accessed on the CEMADEN (Centro Nacional de Monitoramento e Alertas de Desastres Naturais) website under the code 431490215A. The precipitation readings were reported in up to 10min intervals during rainy moments and 1h intervals when no rain was measured.

The rain and traffic datasets have different aggregation intervals (10-60min and 5min intervals, respectively). Before crossing them, we opted for resampling the rain dataset so that it matched the 5min aggregation of the traffic dataset. For doing so we assumed a constant rain intensity for each reported rain interval by dividing the precipitation by the duration of the interval. Then we resampled the rain dataset in 5min intervals to match the traffic dataset and

joined them using time as key. The results obtained in the First Article suggest that this approach was enough to capture the influence of rain in traffic breakdown. However, we encourage future studies to explore approaches other than the assumption of a constant intensity, which is out the scope of this thesis. The final aggregated and cleaned dataset was made public so that other authors could replicate our studies (Zechin, 2022).

The observed congestion profile characterizes the occurrence of the phenomenon called breakdown, since (i) they occur due to the confluence of two important flows in an active bottleneck, (ii) there is the formation of queues upstream of the two approaches, (ii) no slowdown is observed downstream of the active bottleneck and (iv) a rapid and expressive drop in speed is observed through the collected data. **Figure 14** shows the speed profile of the study site, where we can observe the breakdown occurrence at around 7:30. When the breakdown happens, the speed drops from a free flow speed of circa 85km/h to approximately 25km/h.

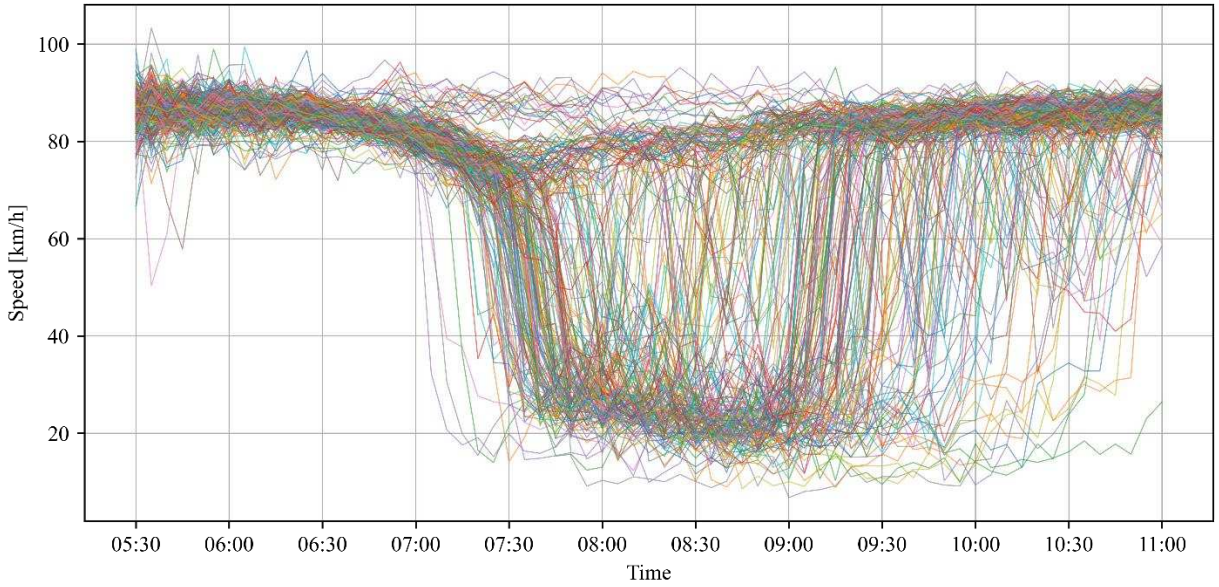

Figure 4: Speed profile at the study site. Each colour represents a different day.

1.4.4 *Rain influence studies*

This stage was crucial for this thesis since it produced evidence of the influence of inclement weather on the breakdown probability in this study site. Based on this study, we opted for using rain intensity as an input feature for the models proposed in the second and third articles.

1.4.5 *Model choice and development*

In both the second and third articles of this thesis, the model choice was primarily done based on the literature review. We opted for doing so since there is strong evidence on the

effectiveness of the chosen models and because we opted for focusing in the context on which the models were applied and not in the models per se. The LSTM model was the base model for both articles. We proposed using extra features such as hyperparameter tuning using the Hyperband technique, sample weighting to increase the relative importance of high-demand periods, encoder-decoder architecture, and bidirectional LSTM layers.

As a requirement of the methodology proposed in the third paper, we need a robust forecasting model also able to produce probabilistic outputs. We chose the Variational LSTM model due to literature convergence on the use of LSTM for forecasting purposes and its rare ability to make probabilistic forecasts, which is not present in the classical formulation of neural networks.

1.4.6 *Breakdown probability calculation*

The breakdown probability calculation stage is present only in the third paper and refers to using the probabilistic speed outputs of the Variational LSTM to forecast breakdown probability. This stage refers to the main contribution of our thesis.

1.4.7 *Conclusions*

The conclusions are presented in the three papers and as a section of this thesis. In each of them, we aimed to summarize the articles' findings and gave special attention to suggestions for future studies. We believe that the Traffic Engineering community still has a lot to benefit from the Machine Learning area and suggested some sequences of this study.

2 ARTICLE 1: INFLUENCE OF RAIN ON HIGHWAY BREAKDOWN PROBABILITY

Authors: Douglas Zechin, Felipe Caleffi, Helena Beatriz Bettella Cybis

Published at: Transportation Research Record I-9 (2020)

TRR
JOURNAL OF THE TRANSPORTATION RESEARCH BOARD

Research Article

Transportation Research Record
1–9

DOI: 10.1177/0361198120919754
journals.sagepub.com/home/trr
SAGE

Influence of Rain on Highway Breakdown Probability

Douglas Zechin[1], Felipe Caleffi[1], and Helena Beatriz Bettella Cybis[1]

Abstract

Capacity has been used to describe a deterministic value that represents the maximum volume of traffic supported by a road. Studies have pointed out the importance of not using a single value for capacity, but rather the concept of probability of occurrence of a traffic-flow breakdown. In this paper the probabilities of breakdown for a Brazilian highway under different weather conditions are compared. Data collected from inductive loop detectors and pluviometric data from automatic rain gauges are combined. Two methodologies of breakdown identification are then compared. The most consistent methodology for identifying breakdowns is used to generate breakdown probability distributions using the product limit and maximum-likelihood methods with the Weibull distribution. The results indicate significant differences in probability of breakdown for each studied climatic condition, including a maximum difference greater than 50% between dry and heavy rain conditions under the same traffic flow.

The term capacity has been used to describe a deterministic value that represents the maximum traffic volume supported by a highway. Although this concept has evolved over time, the convenience of use of a single value for this purpose overcomes the problems that this definition presents (*1*). The Highway Capacity Manual (HCM) (*2*) defines capacity as "the maximum sustainable hourly flow rate at which persons or vehicles reasonably can be expected to traverse a point or a uniform section of a lane or roadway during a given time period, under prevailing roadway, environmental, traffic, and control conditions." This definition is rather vague, disregarding the influence of external conditions on highway capacity. In addition, the HCM does not provide guidance on how highway capacity should be measured (*3*). In its most recent version, notions of capacity relating to occurrence of breakdown were added to this definition (*4*). Since it is a recent update, these probabilistic concepts are not yet in the mindset of most traffic engineers.

The first suggestions concerning the importance of not using a single value for capacity were made by Ponzlet (*5*), who suggested the use of different capacity values for different climatic conditions, periods of the day, and highway purposes. In other works, different capacity values were observed under constant conditions (*6–8*), which motivated the study by Brilon et al. (*1*) that introduced a probabilistic capacity analysis. In that paper, data from traffic detectors and pluviometry data were analyzed to determine the influence of different climatic conditions on breakdown probability.

In this study, two breakdown identification methodologies, from Brilon et al. (*1*) and Lu and Elefteriadou (*9*), are compared. The most consistent one was used for breakdown probability curve generation by means of the product limit method and maximum-likelihood methods with the Weibull distribution. The contribution of the present study is a better understanding of the application of the most recent probabilistic concepts of capacity to rainy conditions. This kind of evaluation can contribute to the fields of traffic simulation and traffic safety analysis, for example.

The remainder of this paper is organized as follows: in the following section the methodologies for breakdown identification and calculation of breakdown probability are presented; the third section presents the study site and data processing; the fourth section concludes the paper with a summary and recommendations for further work.

Methodology

The most recent definition of capacity is directly linked to the phenomenon of breakdown (*7*). This can be understood, on a highway, as the drop in speed and volume

[1]Department of Production and Transport Engineering, Federal University of Rio Grande do Sul, Porto Alegre, Brazil

Corresponding Author:
Douglas Zechin, douglaszechin@gmail.com

resulting from an excess of demand that leads to the transition from a free-flow to a congested regime. Since this phenomenon occurs in different traffic flows, capacity is treated as a probabilistic value. These values can be related to three periods of interest, depending on the purpose of the study: (i) before breakdown, (ii) immediately before breakdown and (iii) observed during congestion (*10*).

Although the influence of rainfall on traffic flow is a well-established concept, its impacts on breakdown probability call for further study. Since breakdown is a phenomenon that happens in a short period of time, and rain may not happen frequently, a large amount of detailed and closely located traffic and weather data are required. The studies by Kim et al. (*11*) and van Stralen et al. (*12*) address this topic and indicate that rainfall increases both the probability and duration of traffic breakdown. These authors also emphasize the need for more robust studies to compensate the volatility of inclement weather events and the consequences of different road configurations.

This section presents the methodologies for breakdown identification, as well as those used for the generation of breakdown probability distribution as a function of traffic flow. Coupled with meteorological data, it is possible to draw different probability distributions for dry and rainy weather conditions, and to analyze their differences.

Breakdown Identification Methods

Employing an adequate method to identify breakdown occurrence is decisive for the success of this study. The method chosen must be suitable for the analysis of breakdown under rainfall and for the available data format. Several methodologies have been proposed in the literature, so the one most appropriate for the specific characteristics of the traffic under analysis must be chosen from among them. This section presents the two methodologies used in the study, those presented by Brilon et al. (*1*) and Lu and Elefteriadou (*9*).

Methodology of Brilon et al. (1). Since breakdown events are related to a significant drop in speed, many studies identify breakdown events by establishing a speed threshold value (*3*). It is considered that a breakdown occurs in period i when the speed drops to a plateau below this threshold in period $i + 1$. This level can be established by observing time series of road speed and volume, and identifying abrupt speed decrease at a characteristic level of congested flow. A suggested speed threshold is one that is not verified during non-congested flow situations and that usually characterizes the transition to a congested regime.

Data from loop detectors are grouped in time intervals and arranged in chronological order. Each interval receives a classification according to the breakdown traffic condition {B}, free flow {F}, or congestion state {C}, according to the following criteria (*1*):

{B}: traffic flows well in time interval i, but the average speed drops below the threshold value in period $i + 1$, that is, breakdown occurs;
{F}: traffic flows well at time intervals i and $i + 1$, indicating that the capacity is greater than the volume observed at i;
{C}: (i) traffic is congested in time interval i, that is, the speed is below the threshold value or (ii) a breakdown is verified in the interval, that is, a downstream detector registers congestion in the interval i or $i–1$. In this case it is considered that the breakdown was because of the queue generated by this congestion, not the traffic conditions of the analyzed detector. This data should be discarded because it does not carry capacity information.

Traffic data may be grouped in small time intervals, usually 1–5 min, so that traffic fluctuations can be perceived (*1*). Brilon and Zuerlinden (*13*) concluded that traffic data should be aggregated at 5 min intervals for this methodology to produce good results.

Methodology of Lu and Elefteriadou (9). In this methodology, breakdown occurrence is defined as five or more consecutive intervals of 1 min with an average speed drop greater than 16 km/h (*9*). Therefore, the use of a speed limit, as imposed by the previous methodology, is not required. Although these criteria have been elaborated for the study of highway capacity before and during accidents, it is proposed to use them to calculate breakdown probability. Three criteria are adopted to identify this phenomenon correctly:

The speed difference between two consecutive minutes is negative.

$$\Delta S_i = S_i - S_{i-1} < 0 \tag{1}$$

The average speed over the previous 5 min is greater than the average speed over the next 5 min by at least 16 km/h.

$$\text{Mean}\{S_{i-5}, \ldots, S_{i-1}\} > \text{Mean}\{S_i, \ldots, S_{i+4}\} + 16\text{km/h} \tag{2}$$

The maximum speed during the next 10 min is lower than the speed before the speed drop.

$$\text{Max}\{S_i, \ldots, S_{i+9}\} < S_{i-1} \tag{3}$$

Unlike the method of Brilon et al., this method does not aim to classify the time intervals in {F} or {C}, it only presents tools for the classification of interval *i* when breakdown occurs, {B}. Intervals with flows lower than 1,000 vphpl (vehicles per hour per lane) were excluded from this part of the analysis because it was considered that they did not represent breakdown events, but fluctuations because of the speed variability in free-flow moments.

Methods for Calculation of Breakdown Probability

Although it may seem reasonable, the frequency with which breakdown events occur cannot be used to calculate breakdown probability. It is necessary to take into account the probability that a certain volume is observed, which demands more advanced mathematical treatment (*14*, *15*). Based on this, the purpose of these methods is to construct a cumulative capacity distribution function, $F_c(q)$, which allows probability of breakdown to be calculated based on the volume observed in the lane in the interval *i*. The most common methods used to construct this function are the limit-product estimator, or Kaplan-Meier estimator, and the adjustment to a Weibull cumulative distribution using the maximum-likelihood method. We adopt the internationally used nomenclature to designate these methods, which are product limit method and Weibull, respectively.

Product Limit Method. Product limit method (PLM) is a statistical method to estimate survival functions (*16*). Its main applications are associated with the durability of mechanical components and medicine, when the survival rates of individuals presenting specific clinical conditions or the undergoing new treatments are evaluated. Similarly, the transport engineering interest in this method refers to the formulation of a probability distribution of the non-occurrence of breakdowns (traffic survival) as a function of the observed traffic flow, $S(q)$, or of its complementary breakdown probability curve (traffic death), $F(q) = 1–S(q)$ (*1*, *14*)

This method is quite accurate for low volumes, and it is especially useful, for example, to calculate breakdown probability on a highway subject to different access flows (*14*). In these cases, breakdown probability is maintained at around 20% by means of ramp metering, a technique that aims to control the flow of on-ramps with traffic lights.

The survival function described by PLM is given by:

$$S(q) = \prod_{i:q_i \leq q} \frac{k_i - d_i}{k_i}, i \in \{B, F\} \quad (4)$$

where

q = flow (vph [vehicles per hour]);
q_i = flow in time interval i (vph);
k_i = number of time intervals with a volume q ⩾ q_i;
d_i = number of breakdowns with a flow q_i;
{B, F} = set of intervals when breakdown occurred or when traffic flow was non-congested.

The cumulative capacity distribution function is then given by:

$$F_c(q) = 1 - S(q) = 1 - \prod_{i:q_i \leq q} \frac{k_i - d_i}{k_i}, i \in \{B, F\} \quad (5)$$

This function accumulates in a value between 0 and 1. It will only reach value 1 when the largest sample volume, $q_{\max}$, corresponds to an event belonging to set {B}, otherwise accumulating in $F_c(q_{\max}) < 1$ (*1*).

The standard deviation of the PLM survival function $S(q)$ can be calculated, according to Greenwood (*17*), by:

$$\sigma_S(q) = S(q)\sqrt{\sum_{i:q_i \leq q} \frac{d_i}{n_i(n_i - d_i)}} \quad (6)$$

From the standard deviation one can calculate the function confidence interval with:

$$F_c(q) - z_{\alpha/2} \cdot \sigma_S(q) \leq F_c(q) \leq SF_c(q) + z_{\alpha/2} \cdot \sigma_S(q) \quad (7)$$

where $z_{\alpha/2}$ is the $\alpha/2$-th quantile of the normal distribution;and γ is the desired confidence level.

Maximum Likelihood—Weibull. Because breakdown probability does not always stack at 100% according to PLM, it is impossible to verify it for some volumes. To account for this problem, the maximum-likelihood method is used to fit an accumulated Weibull distribution to the data (*1*). For capacity analysis the maximum-likelihood function is:

$$L = \prod_{i=1}^{n} f_c(q_i)^{\delta_i} \cdot [1 - F_c(q_i)]^{1-\delta_i} \quad (8)$$

where

$f_c(q_i)$ = statistical capacity density function;
$F_c(q_i)$ = capacity cumulative distribution function;
n = number of time intervals;
δ_i = 1, if the range contains a non-censored value;
δ_i = 0, if the range contains a censored value.

To make the adjustment, however, the log-likelihood function L* is used since it has its maximum with the same parameters as the maximum likelihood, but it is computationally lighter. This function is given by:

$$L^* = \ln(L) = \sum_{i=1}^{n} \{\delta_i \cdot \ln[f_c(q_i)] + (1 - \delta_i) \cdot \ln[1 - F_c(q_i)]\} \quad (9)$$

As a capacity function distribution, we use the Weibull distribution function:

$$F_c(q) = 1 - e^{-\left(\frac{q}{\beta}\right)^{\alpha}} \quad (10)$$

where

$F_c(q)$ = capacity distribution function;
q = flow (vph);
α = shape parameter;
β = scale parameter (vph).

The form parameter α represents the distribution variance and is between 10 and 22 for the road capacity distribution function (*15*). The higher this value, the lower the variance. The parameter β is directly related to the problem shape, such as the number of lanes.

Study Site and Data Processing

The study site was a segment of Freeway BR-290/RS, which is the main access to the city of Porto Alegre in the state of Rio Grande do Sul, Brazil (see Figure 1). The segment is located in km 95 northbound, has five lanes in each direction, and presents daily breakdown congestion because of bottlenecks immediately downstream. The movable Guaíba Bridge, located immediately downstream of the loop detectors, causes frequent congestion because it is closed to traffic whenever it is necessary to allow ships to pass through. Single-vehicle data obtained from these loop detectors was correlated to a rain gauge 1.35 km distant from them, so that traffic was classified according to rainfall conditions. Similar studies used rain gauges with a distance of 5–20 km from their respective detectors (*5*).

Data from the loop detectors cover two years and three months of traffic data recording the speed (km/h) and the passage time of each vehicle with precision of 0.1 s. The traffic flow on this freeway is subject to several irregularities that hamper the study of the main flow, demanding data filtering. First, data from days with very low traffic volumes compared with average daily volumes observed—probably because of failure of a detector—and before 6:00 a.m. and after 10.00 p.m. were removed. Periods referring to accidents and activation of the movable bridge that obstructed the roadway, causing upstream traffic retention, were also removed. Finally, data from lanes whose detectors showed capture failures compromising the total freeway volume measurement were removed. This filtering process reduced the amount of data from 847 to 217 days of data.

Elefteriadou (*3*) suggests that the detectors should be located in the access for this type of analysis to be done successfully. Detectors located downstream of the bottlenecks tend not to represent the congested region of the road, whereas detectors located at long distances upstream do not represent the breakdown phenomenon, but rather the decrease in speed and volume resulting from the queue propagation generated.

Rain gauge data were obtained from the National Center for Natural Disasters Monitoring and Alert (Cemaden) database. They consist of precipitation records in millimeters per hour, with shorter measurement intervals during rain events, with a minimum of 10-min intervals.

Methodology for Identifying Breakdowns

The two methodologies presented were tested, and the most suitable to the study sequence was chosen. The method of Lu and Elefteriadou (*9*) was applied so that all time intervals i preceding breakdowns were identified. These points were inserted into the set of points {B}, while the 10 points before them were inserted into the set of points {F} (*3*). The method of Brilon et al. (*1*) was applied by first establishing the limit speed. Some authors use more sophisticated techniques to define this level, such as clustering (*18*); however, since just one loop detector was analyzed, it was decided to define it visually.

The traffic data were aggregated at intervals consistent with the breakdown identification methods described in the Methodology section (1 min and 5 min), generating two distinct databases. The road average speed was calculated by the arithmetic mean of the instantaneous speeds observed in all lanes; the flow was defined by the number of vehicles passing the detectors; and the density was calculated by dividing the flow observed in the time interval by the observed average speed.

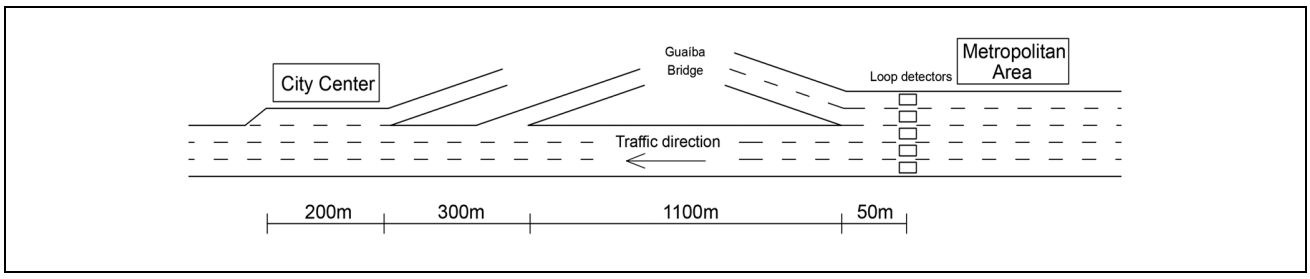

Figure 1. Study site.

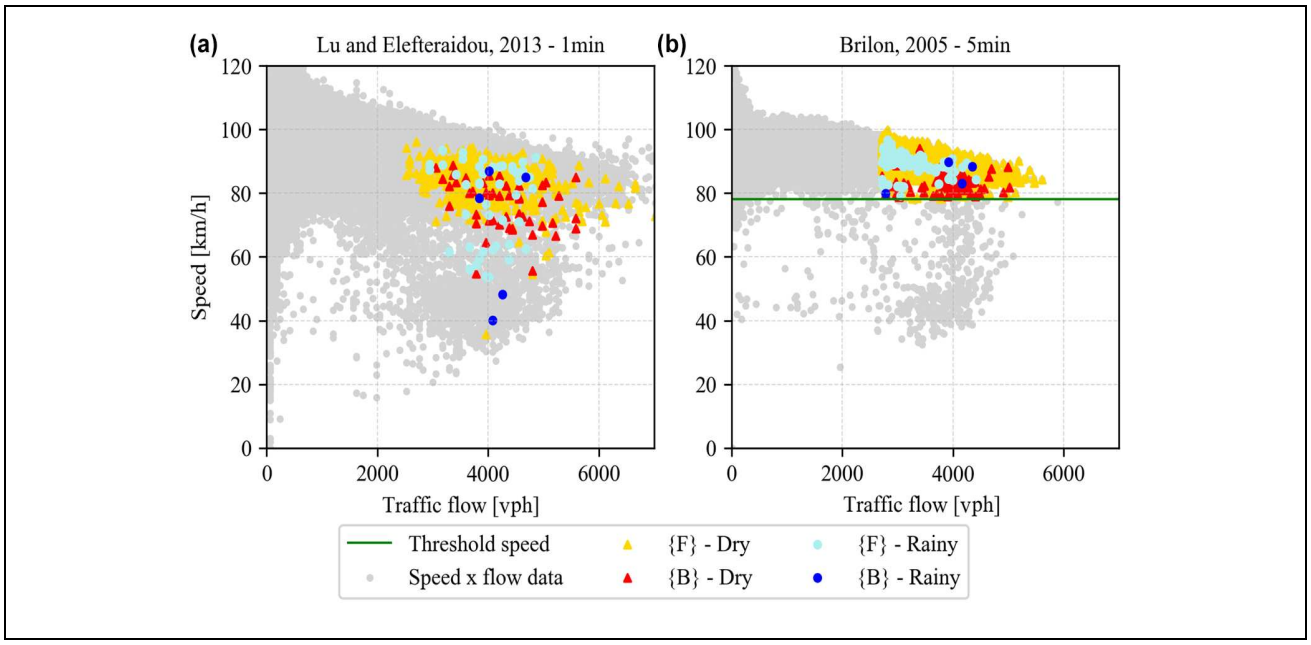

Figure 2. Comparison between breakdown identification methods: (*a*) Lu and Elefteriadou (*9*); (*b*) Brilon et al. (*1*).
Note: vph = vehicles per hour.

The rainfall data were given a binary classification, 1 or 0, depending on whether or not it was raining, respectively, in the given measurement range. They were then concatenated with the traffic databases, which received the binary observed in the same period in the rain database.

The points belonging to sets {B} and {F} were then classified as "dry" or "rain" according to the climatic condition in the corresponding time interval. Figure 2 depicts the results of application of both methodologies.

Although the amount of breakdown occurrences with each method was similar, it is observed that the conditions in which they occurred are different. In the Brilon et al. (*1*) methodology, the occurrences were limited to the established speed threshold and, because of this, they necessarily departed from a non-congested flow condition. In the Lu and Elefteriadou (*9*) methodology, breakdowns were identified in the transition region between free-flow and congested regimes, even in the congested region. Because of this, we chose the method of Brilon et al. (*1*) for use in the later stages of the study.

PLM Breakdown Probability

Breakdown probability can be calculated with the PLM and Weibull methods described in the Methodology section. The data set was first separated according to rain intensity, generating four different data sets. The grouping process was based on rain intensity to create groups with smaller discretization for lower rain intensities. According to Chung et al. (*19*), lower rain intensities are associated with higher marginal decline of road capacity. The resulting intervals were: 0 mm/h, 0–1.3 mm/h, 1.3–4 mm/h, and 4–17.5 mm/h, which were related to 349, 19, 6, and 7 breakdown events respectively.

The generated data sets were submitted to the Brilon et al. (*1*) method, chosen as the most efficient in the previous section. We kept the sets {B} and {F} resulting from this methodology and the data belonging to set {C} were excluded. The PLM was then applied to each of the data sets and the results are presented in Figure 3, with an 80% confidence interval.

The confidence intervals of the rainy and dry weather curves indicated that, in general, there was a statistically significant difference between them. The overlap of confidence intervals at low volumes indicated that, at these volumes, there was no significant difference in breakdown probability. Overlaps in larger volumes were expected and resulted from the size of the confidence intervals, which were related to the number of breakdown observations.

In Figure 2 it is possible to compare the breakdown probabilities for a specific volume or the volume related to a fixed breakdown probability for each rain intensity curve. Breakdown probability rose with the increase of rain intensity. However, the overlap between the curves with higher rain intensities indicated that differences became smaller as intensity increased.

The PLM method, therefore, does not allow accurate measurements to be made in relation to breakdown probability at high volumes, and this is one of the main criticisms of this method. The measurement of breakdown probabilities for higher volumes can be done by adjusting the data to a distribution such as Weibull's.

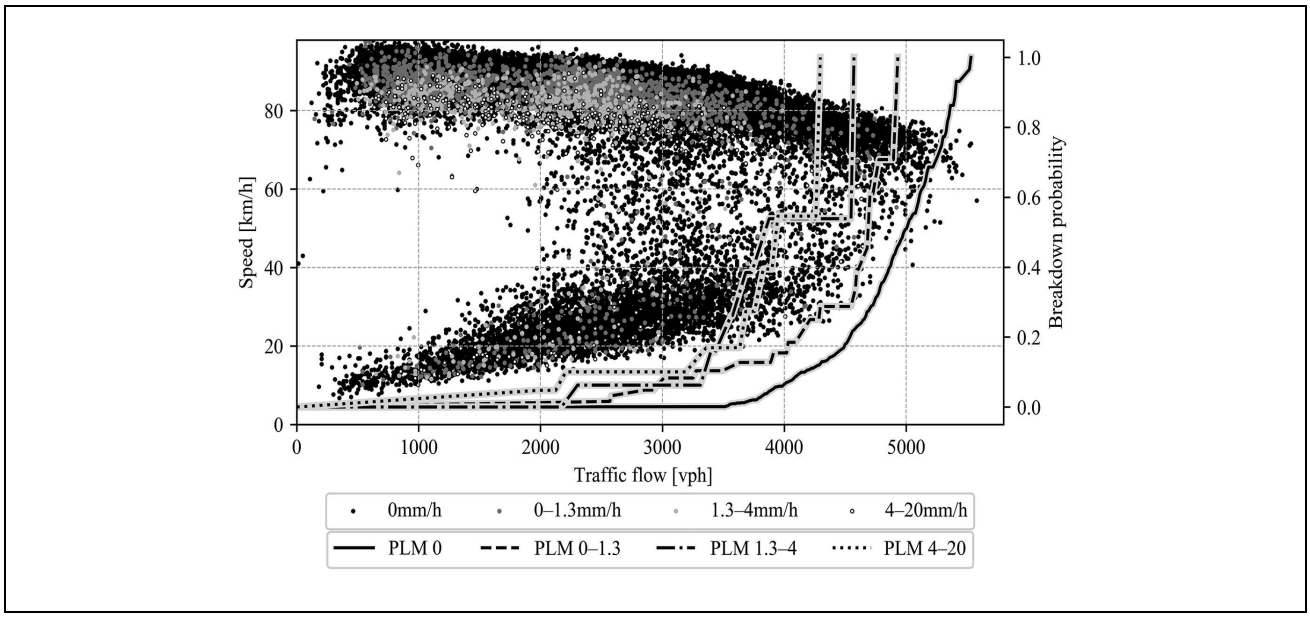

Figure 3. Speed-flow data and product limit method (PLM). The shadowed area represents the 80% confidence interval of the associated rain intensity group.
Note: vph = vehicles per hour.

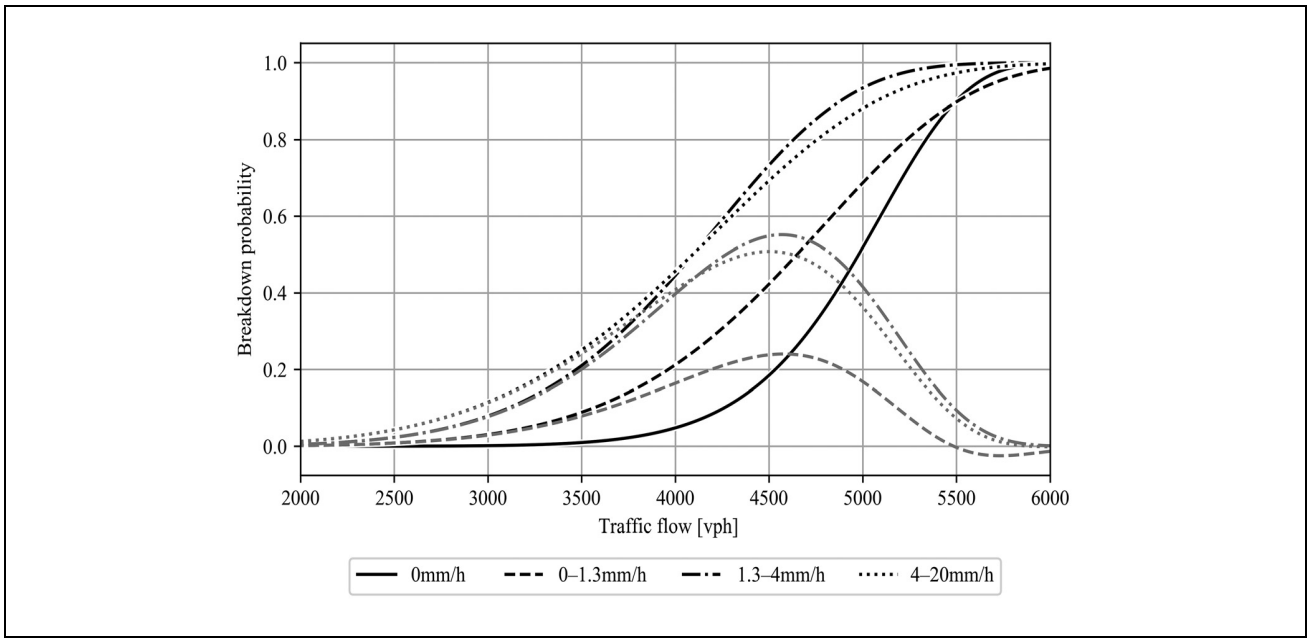

Figure 4. Comparison between product limit method (PLM) and Weibull distribution. Black lines refer to the Weibull distribution for each rain intensity. Gray lines refer to the difference between the corresponding Weibull distribution and the dry scenario.
Note: vph = vehicles per hour.

Breakdown Probability by Weibull Distribution

As expected, the distributions under analysis rejected the null hypothesis that they come from a normal distribution and are better fitted to the Weibull distribution (*1*, *20*). The adjustment to the Weibull distribution allowed the probabilities to be extrapolated and the volume corresponding to any probability to be calculated analytically.

Figure 4 depicts the Weibull distributions for each rain intensity. The differences between each rainy scenario and the dry scenario are also plotted and support the convergence hypothesis between the curves of the groups with higher rain intensities.

The representation used in Figure 4 is the traditional visualization of breakdown probability. However, when

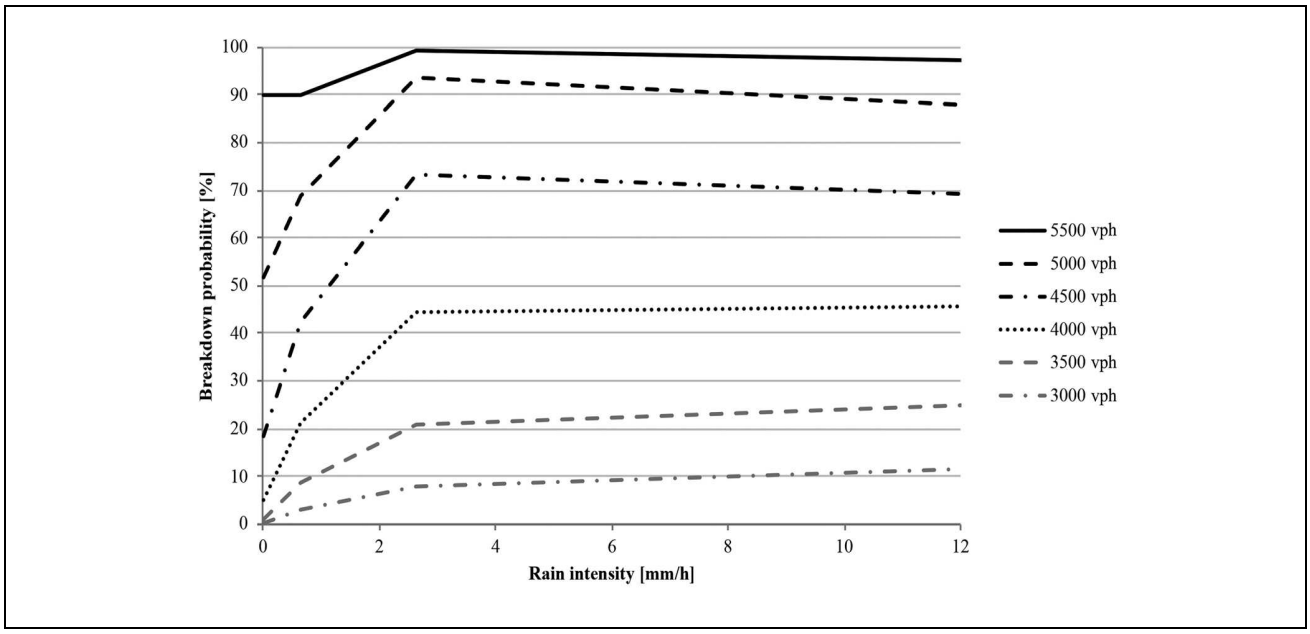

Figure 5. Relationship between breakdown probability and rain intensity for different traffic flows.
Note: vph = vehicles per hour.

more than one curve is presented, the existing differences between them are difficult to evaluate. An alternative for interpreting the results is to visualize the relations between breakdown probability and rain intensity for different traffic flows. This is depicted in Figure 5, where the mean value of rain intensity of each group is used in the horizontal coordinates.

Figure 5 indicates that the highest increases in breakdown probability were observed with lower rain intensities for the same traffic flow. An asymptotic convergence of breakdown probability was also observed for higher rain intensities. Lower traffic flows, like 3,000 vph, were less suitable to present breakdowns because the traffic state was closer to a free-flow condition. Therefore, rain intensity was less likely to cause breakdown. By contrast, high traffic flows, like 5,500 vph, were so likely to present breakdown that rain did not significantly affect the breakdown probability. However, intermediate traffic flows, ranging from 4,000 to 4,500 vph presented the greatest marginal increase, with breakdown probability increasing from 18% without rain to 73% with rain intensity ranging between 1.3 and 4.0 mm/h.

Conclusion

In this paper, the probability of traffic breakdown on a freeway under different climatic conditions is analyzed. For this study, two methodologies for breakdown identification were applied and the most consistent one was used for generation of breakdown probability curves by means of PLM and maximum-likelihood methods with the Weibull distribution.

The breakdown identification methodology presented by Brilon et al. (*1*) was more efficient and suitable than that presented by Lu and Elefteriadou (*9*), for the specific purposes of this paper. The Lu and Elefteriadou (*9*) methodology presents three generic, more complex and subjective criteria, which hinders its calibration in studies with many detectors. In contrast, the use of a speed threshold value is more adjustable and comprehensible, so it is possible to identify it with good specificity for each detector.

The PLM indicated a significant difference in breakdown probability with and without rainfall. This methodology is very efficient for the identification of breakdown probability for smaller volumes, being ideal for the joint application with ramp metering, to control access volumes and minimize breakdown probabilities. The 95% confidence intervals, however, indicated a greater uncertainty for larger volumes. This is because the frequency of observations in larger volumes is lower, especially in relation to rainfall, which in itself is less frequent than dry weather. To overcome this problem, the use of longer data periods is recommended, which was not feasible in this study.

The Weibull distribution adjustment to the data explained the differences between the two climatic conditions and allowed the calculation of breakdown probabilities for all volumes. The breakdown probability distributions of the detectors far from bottlenecks demonstrated that this methodology is not suitable for the capacity calculation in these cases. The differences in capacity and probability found at detectors near bottlenecks indicated that road operators must be aware of the

influence of rainfall on traffic, especially in locations where rainfall is highly frequent.

The results observed in this paper show that this methodology could be used to feed and improve existing active traffic management strategies that rely on breakdown probability, such as ramp metering and variable speed. Recommendations for future work focus on understanding the benefits of using this method in traffic management strategies and the effects on safety and operational conditions. Practical recommendations are the use of a larger data set, so that breakdown probability distributions can be made for shorter rain intensity ranges, and accounting for the period of the day. Other methodologies can also be used to define deterministic capacity values, such as those of Modi et al. (*20*), Kondyli et al. (*21*), and Van Arde and Rakha (*22*).

Acknowledgments

The authors would like to thank Triunfo-Concepa for the traffic data provided.

Author Contributions

The authors confirm contribution to the paper as follows: study conception and design: D. Zechin; data collection: D. Zechin; analysis and interpretation of results: D. Zechin, F. Caleffi, H.B.B. Cybis; draft manuscript preparation: D. Zechin, F. Caleffi. All authors reviewed the results and approved the final version of the manuscript.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Brazilian National Research Foundation CAPES.

References

1. Brilon, W., J. Geistefeldt, and M. Regler. Reliability of Freeway Traffic Flow: A Stochastic Concept of Capacity. *Proc., 16th International Symposium on Transportation and Traffic Theory*, College Park, MD, 2005, pp. 125–144.
2. *Highway Capacity Manual*. Transportation Research Board, National Research Council, Washington, DC, 2010.
3. Elefteriadou, L. *An Introduction to Traffic Flow Theory*. Springer, New York, 2014.
4. *Highway Capacity Manual*. Transportation Research Board, National Research Council, Washington, DC, 2016.
5. Brilon, W., and M. Ponzlet. Variability of Speed-Flow Relationships on German Autobahns. *Transportation Research Record: Journal of the Transportation Research Board*, 1996. 1555: 91–98.
6. Elefteriadou, L., R. P. Roess, and W. R. McShane. Probabilistic Nature of Breakdown at Freeway Merge Junctions. *Transportation Research Record: Journal of the Transportation Research Board*, 1995. 1484: 80–89.
7. Lorenz, M., and L. Elefteriadou. *Transportation Research Circular No. E-C018: A Probabilistic Approach to Defining Freeway Capacity and Breakdown*. Transportation Research Board of the National Academies, Washington, D.C., 2000, pp. 84–95.
8. Persaud, B., S. Yagar, and R. Brownlee. Exploration of the Breakdown Phenomenon in Freeway Traffic. *Transportation Research Record: Journal of the Transportation Research Board*, 1998. 1634: 64–69.
9. Lu, C., and L. Elefteriadou. An Investigation of Freeway Capacity before and during Incidents. *Transportation Letters*, Vol. 5, No. 3, 2013, pp. 144–153. https://doi.org/10.1179/1942786713Z.00000000016.
10. Elefteriadou, L., F. L. Hall, W. Brilon, R. P. Roess, and M. G. Romana. Revisiting the Definition and Measurement of Capacity. *Proc., 5th International Symposium on Highway Capacity and Quality of Service*, Yokohama, Japan. Vol. 2, 2006, pp. 391–399.
11. Kim, J., H. S. Mahmassani, and J. Dong. Likelihood and Duration of Flow Breakdown: Modeling the Effect of Weather. *Transportation Research Record: Journal of the Transportation Research Board*, 2010. 2188: 19–28.
12. Van Stralen, W. J. H., S. C. Calvert, and E. J. E. Molin. The Influence of Adverse Weather Conditions on Probability of Congestion on Dutch Motorways. *European Journal of Transport and Infrastructure Research*, Vol. 15, No. 4, 2015, pp. 482–500. https://doi.org/10.18757/ejtir.2015.15.4.3093.
13. Brilon, W., and H. Zuerlinden. *Ueberlastungswahrscheinlichkeiten und Verkehrsleistung als Bemessungskriterium fuer Verkehrsanlagen/Overload Probabilities and Traffic Activity as Design Criteria for Road Traffic Systems*. Forschung Straßenbau und Straßenverkehrstechnik, No. 870, 2003.
14. Elefteriadou, L., A. Kondyli, S. Washburn, W. Brilon, J. Lohoff, L. Jacobson, F. Hall, and B. Persaud. Proactive Ramp Management under the Threat of Freeway-Flow Breakdown. *Procedia - Social and Behavioral Sciences*, Vol. 16, 2011, pp. 4–14. https://doi.org/10.1016/j.sbspro.2011.04.424.
15. Geistefeldt, J., and W. Brilon. A Comparative Assessment of Stochastic Capacity Estimation Methods. In *Transportation and Traffic Theory: Golden Jubilee* (W. Lam, S. Wong, and H. Lo, eds.), Springer, Boston, MA, 2009, pp. 583–602.
16. Kaplan, E. L., and P. Meier. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, Vol. 53, No. 282, 1958, pp. 457–481. https://doi.org/http://www.jstor.org/stable/2281868.
17. Greenwood, M. A Report on the Natural Duration of Cancer. In *Reports on Public Health and Medical Subjects*. Ministry of Health. H.M.S.O., London, 1926, pp. 4–26.
18. Riente de Andrade, G., and J. R. Setti. *Transportation Research Circular No. E-C190: Speed-Flow Relationship*

and Capacity for Expressways in Brazil. Transportation Research Board of the National Academies, Washington, D.C., 2014, pp. 10–25.

19. Chung, E., O. Ohtani, H. Warita, M. Kuwahara, and H. Morita. Does Weather Affect Highway Capacity? *Proc., 5th International Symposium on Highway Capacity and Quality of Service*. *Country Reports and Special Session Papers*, Yokohama, Japan, 2006, pp. 139–146.
20. Modi, V., A. Kondyli, S. S. Washburn, and D. S. McLeod. Freeway Capacity Estimation Method for Planning Applications. *Journal of Transportation Engineering*, Vol. 140, No. 9, 2014, p. 05014004. https://doi.org/10.1061/(ASCE)TE.1943-5436.0000699.
21. Kondyli, A., L. Elefteriadou, W. Brilon, F. L. Hall, B. Persaud, and S. Washburn. Development and Evaluation of Methods for Constructing Breakdown Probability Models. *American Society of Civil Engineers*, Vol. 139, 2013, pp. 931–940. https://doi.org/10.1061/(ASCE)TE.1943-5436.0000574.
22. Van Aerde, M., and H. Rakha. Multivariate Calibration of Single Regime Speed-Flow-Density Relationships [Road Traffic Management]. *Proc., Pacific Rim TransTech Conference. 1995 Vehicle Navigation and Information Systems. 6th International VNIS. A Ride into the Future*, Seattle, WA, 1995, pp. 334–341. https://doi.org/10.1109/VNIS.1995.518858.

3 ARTICLE 2: FORECAST OF TRAFFIC SPEEDS WITH NEURAL NETWORK LSTM ENCODER-DECODER

Authors: Douglas Zechin, Matheus Basso do Amaral, Helena Beatriz Bettella Cybis

Published at: Transportes v. 30 n. 3 (2022)

TRANSPORTES
revistatransportes.org.br
anpet
associação nacional de pesquisa e ensino em transportes
anpet.org.br

Forecast of traffic speeds with neural network LSTM encoder-decoder

Previsão de velocidades de tráfego com rede neural LSTM encoder-decoder

Douglas Zechin[1], Matheus Basso do Amaral[2], Helena Beatriz Bettella Cybis[3]

[1]Federal University of Rio Grande do Sul, Rio Grande do Sul – Brazil, douglaszechin@gmail.com
[2]Federal University of Rio Grande do Sul, Rio Grande do Sul – Brazil, matheus.basso96@gmail.com
[3]Federal University of Rio Grande do Sul, Rio Grande do Sul – Brazil, helenabc@producao.ufrgs.br

Submitted:
17 de julho de 2021
Accepted for publication:
30 de agosto de 2022
Published:
14 de dezembro de 2022
Editor in charge:
Cira Souza Pitombo

Keywords:
Congestion.
Traffic forecasting.
Neural networks.

Palavras-chave:
Breakdown de tráfego.
Previsão de tráfego.
Redes neurais.

DOI:10.14295/transportes.v30i3.2660

ABSTRACT
This article proposes a speed prediction model for a highway segment in the city of Porto Alegre, which has daily traffic jams due to bottlenecks. We used traffic data and environmental variables, such as rainfall intensity, accidents and atypical events to make the forecasts. Then we proposed a neural network model with an encoder-decoder architecture and long short-term memory (LSTM) layers, which has the characteristic of establishing long-term relationships between the input variables, being relevant for applications in the Transportation area. As additional contributions, we evaluated the quality of forecasts for different prediction horizons and traffic regimes. We compared cumulative distribution functions (CDFs) generated using field and forecast data using a survival analysis method similar to the breakdown probability calculation. These CDFs represent the probability of a sudden speed drop due to the transition from the free-flow to the congested regime. The methodology presented a satisfactory performance based on both criteria, making good predictions even in critical traffic situations.

RESUMO
Este artigo tem como objetivo propor uma modelo de previsão de velocidades para um trecho de rodovia na cidade de Porto Alegre, que apresenta congestionamentos diariamente por conta de gargalos. Para realizar as previsões foram utilizados dados de tráfego e variáveis ambientais, como intensidade de chuva, acidentes e eventos atípicos. Propôs-se então um modelo de rede neural com arquitetura encoder-decoder e camadas long short-term memory (LSTM), que possuem a característica de estabelecer relações de longa dependência temporal entre as variáveis de entrada, sendo pertinentes para aplicações na área de Transportes. Como contribuições adicionais, avaliou-se a qualidade das previsões para diferentes horizontes de predição e regimes de tráfego, e comparou-se a capacidade e as curvas de probabilidade de breakdown calculadas com dados de campo e previstos. A metodologia apresentou desempenho satisfatório com base em ambos os critérios, sendo capaz de fazer boas previsões mesmo em situações críticas de tráfego.

1. INTRODUCTION

Traffic Engineering has received important contributions from recent technological advances in other areas, such as IoT (Internet of Things) and artificial intelligence. The intersection between these areas has led to the emergence of innovative fields of study, such as Smart Cities and autonomous vehicles, in addition to contributing to traditional areas, such as active traffic management (ATM), in which this study fits.

ATM has been around since the first half of the last century and traditionally proposes using simple algorithms and traffic and speed detectors to manage highway traffic operations. Although many traffic agencies still use these methods, ATM has received many contributions from data-driven approaches and seems to be increasingly merging with the concept of Smart Cities (Ma, Zhang and Ihler, 2020). An important feature made possible by more robust methods is improving traffic forecasts and anticipating undesired scenarios, such as congestion, accidents, and increased travel time.

In this paper, we propose using long short-term memory (LSTM) neural networks to perform speed predictions in the vicinity of a highway bottleneck located in the metropolitan region of Porto Alegre, Brazil. However, the proposed methodology aims to prioritize forecasts made close to the road capacity, which is the most critical moment for traffic management. The predictions consist of the expected average speed for the subsequent 5 time intervals of 5 minutes and are based on traffic data, precipitation, and other possibly relevant information such as the day of the week and detector malfunctions. We chose this approach because LSTMs can retain information by creating long-term dependencies, which generally results in better performance than parametric methods and standard neural networks for time series prediction.

Speed forecasting can lead to good results in terms of average error since traffic speed is mostly stable due to the existence of speed limits. However, a low average error can hide large forecast errors at critical times, such as during peak demand periods, where traffic characteristics change quickly. Other authors rarely address this problem, so we propose segregating the data into five sets with equivalent traffic characteristics and analyzing the model error for each one individually and for each forecast horizon.

We used Survival Analysis by the Kaplan-Meyer method to confirm the quality of traffic forecasts during peak periods close to road capacity. In this case, survival is related to the maintenance of a non-congested regime, and death is associated with the beginning of the transition to a congested regime. We statistically tested the similarity of cumulative distribution functions (CDFs) constructed with field and predicted data. Although the region presents breakdowns daily, the measured phenomenon was not treated as a breakdown because the detectors are located upstream of the active bottleneck. In this way, the CDFs represent the probability of starting the transition from the free-flow regime to the congested regime.

Until the conclusion of this article, the evaluation of the quality of traffic forecasting methodologies from the comparison of survival curves made with the forecasts and with field data had not been used in other researches. However, we understand that this produces a solid comparison, as these methods are already well established among the traffic engineering community and allow for the calculation of road capacity. Therefore, in addition to a detailed discussion about the model's error, we propose evaluating its effectiveness from this approach.

2. LITERATURE REVIEW

The development and improvement of traffic forecasting methods are alternatives for improving traffic management on urban highways and arterials (Vlahogianni, Karlaftis and Golias, 2014). Precise short-term and real-time predictions can be used as input into ATM algorithms, contributing to more efficient and responsive traffic management (Gu *et al.*, 2019). Traffic predictions are often a specific application of parametric time series prediction methods such as naïve and ARIMA. Although these methods have greater physical interpretability and

their solution is usually simpler (Fu, Zhang and Li, 2017), the computational capacity and the great availability of currently existing data allow the use of more robust models, such as neural networks.

Due to the dynamic nature of demand, non-linear non-parametric models tend to be better suited to capture traffic's spatial and temporal evolution to make good speed predictions. Recurrent Neural Networks (RNRs) adapt well to this type of problem, as they are a type of neural network capable of processing temporal sequences. However, there are different subtypes of RNRs with different purposes, and one of the best suited to this study is the LSTM. LSTMs can retain relationships with long temporal dependence, which is crucial for correctly interpreting traffic seasonality.

Hochreiter (1997) proposed the LSTM architecture with the main objective of modeling long dependencies, which is not possible with standard RNNs. Short-term traffic predictions can be defined as estimating the state of traffic for a close time in the future (Gu *et al.*, 2019). For this reason, accuracy and precision are essential aspects that must be considered. LSTM is a great candidate as it captures the non-linearity of traffic dynamics in an effective way across using memory blocks and thus has a superior capacity for predicting time series with long time dependencies (Ma *et al.*, 2015).

The ease of access to high-level neural network programming tools has enabled rapid assimilation of new techniques for specific applications (Chollet, 2018; Géron, 2019). Because of this, the use of LSTM neural networks has gained space for solving traffic problems, which are highly time-dependent and have multiple variables that are related in a complex way. Fu et al. (2017) showed that LSTM and GRU neural networks (Gated Recurrent Units) have similar performance for traffic flow prediction and perform better when compared to the ARIMA method. Laptev et al. (2017) proposed an application of an LSTM neural network with an encoder-decoder structure to forecast the travel demand of an urban private transport company and capable of making predictions with high quality. A comparison between FFN (Feed Forward Network), CNN (Convolutional Neural Network), and LSTM was made by Asplund (2019), who obtained better results using the LSTM neural network to predict traffic conditions using public transport traffic information as input data. As stated by Vlahogianni et al. (2014), the interest of researchers has shifted towards more responsive prediction methods and models for non-recurring traffic conditions through the development of prediction systems with high algorithmic complexity. Furthermore, do Amaral (2020) compared the quality of velocity predictions in the same locality using different predictive models and concluded that an LSTM neural network produced better predictions than traditional methods such as linear regression, ARIMA, and regular neural networks.

In this article, therefore, we propose making speed predictions in a segment of a suburban highway where breakdowns are observed daily due to the existence of a bottleneck. To make these predictions, we used environmental and traffic data collected with inductive loops upstream of the bottleneck. We chose as the model a LSTM neural network with encoder-decoder architecture to increase the predictive capabilities of LSTM neural networks pointed out in other studies. We assessed the quality of the predictions by comparing the error of the forecasts in traffic situations with similar characteristics and testing whether the CDF calculated with the predictions is equivalent to that calculated with field data.

3. METHODOLOGY

This article proposes using an LSTM neural network to make speed predictions using traffic data from a point on a Brazilian highway. Information on precipitation, road accidents, and atypical events were concatenated with traffic data and then grouped by lane at regular intervals to generate the input variables that feed the neural network. As input and output variables, we defined how much time in the past and the future the proposed network would consider to make predictions. After training the neural network, we evaluated the results for different regions of the fundamental diagram and compared them with the CDF obtained through the field data.

3.1. Study site

The study region comprises a section of the BR-209 highway in Porto Alegre, RS, selected due to the high traffic volumes in the morning peak period. The breakdown phenomenon occurs regularly on weekdays due to this great demand, bottlenecks in the approaches, and the lifting of the mobile span of the Guaíba Bridge downstream of the data detection location (Caleffi *et al.*, 2016; Caleffi, 2018; Zechin, Caleffi and Cybis, 2020), as shown in Figure 1.

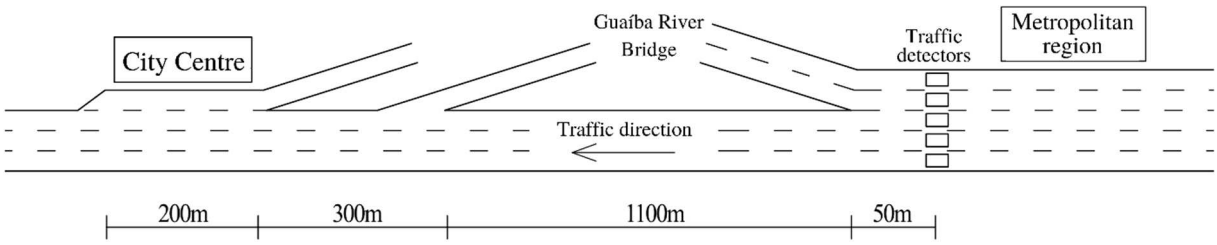

Figure 1. Study region

3.2. Traffic and environmental data

The data used in this article were made available by the company Triunfo Concepa, the concessionaire that operated the stretch of the highway. These data were collected using inductive loops located approximately 50 meters upstream of a fork that connects the road to the Guaíba Bridge. The data consists of two years (2016 and 2017) of disaggregated traffic counts with information on the instant of each vehicle's passage, speed, and lane. We only used data from the three lanes on the left since the others do not present congestion and connect the road to the bridge. We discarded data from days when the detectors malfunctioned, weekends, and days with accidents within a 5 km radius of the detectors, resulting in a useful sample of 263 days.

We also used environmental data to provide the network with as much useful information as possible. We obtained rainfall data from a rain gauge 500 m away from the inductive loops from the Cemaden (National Center for Monitoring and Alerts for Natural Disasters) online portal. We treated it as a continuous variable since rainfall intensity was provided at intervals of up to 10 minutes. We replicated the rainfall intensity calculated for a given instant to the previous data aggregation intervals used in the study until the time when another measurement was reported. This methodology is compatible with the data aggregation methodology used by Cemaden. In addition to rainfall data, we used the day of the week and bridge lifts as dummy variables.

In this region, the breakdown phenomenon occurs daily around 7:30am with no important exceptions. Because of this, we defined 4am to 11am as a suitable period for the analysis based on the speed profile of the highway. This covers the development of demand in the early morning, congestion, and the recovery of the free flow regime.

3.3. Generation of inputs and outputs

LSTM neural networks require data spaced in regular intervals to make adequate predictions, so we aggregated the data at 5 min intervals. Then, from the aggregated data, we created the variables volume, standard deviation of speed, average speed, minimum speed, median speed, and maximum speed per lane. We consolidated environmental variables and traffic variables, and continuous variables were normalized.

We defined the neural network inputs as 12 intervals in the past (60 min), each comprised of the previously created variables. For the outputs, we defined a forecast horizon of 25 min, corresponding to 5 intervals of 5 min, and the predicted variable was the average speed of the road. The first 80% of the data, in chronological order, was used for training and the remaining for testing. We did so to bring the study closer to an actual application, where past data would be used to predict unknown future events.

3.4. LSTM neural network with encoder-decoder architecture

Although neural networks with cells of the LSTM type have a remarkable ability to predict time series, relying on the ability to retain long-term information, there are network architectures that allow predictions to be even more accurate. In this work, we propose using the encoder-decoder architecture, as shown in Figure 2, which has shown promising results in applications in the transport area (Laptev *et al.*, 2017). This architecture interprets the information in two stages: the encoder processes the data, and the decoder computes the model outputs.

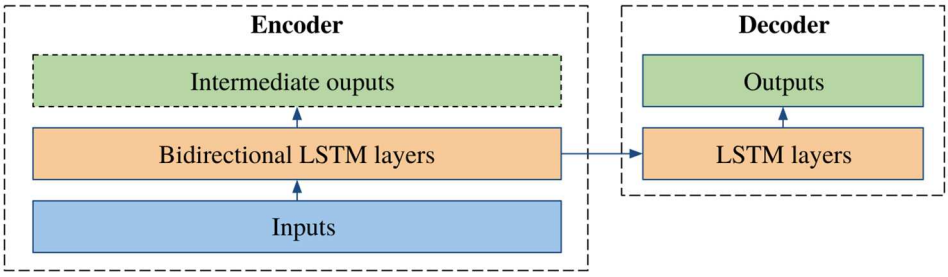

Figure 2. Encoder-decoder architecture with bidirectional LSTM layers

We inserted the input data into the neural network through the encoder. It passes through the bidirectional intermediate layers (Schuster and Paliwal, 1997), which make abstractions using LSTM cells. The computed information then follows two paths: (i) it is passed to a layer that generates intermediate outputs with the exclusive objective of increasing the assertiveness and stability of the model, and (ii) it is passed to the decoder, where it passes through intermediate layers before generating the outputs that are actually used as a forecast.

4. RESULTS

We created the proposed neural network model using the Keras (Chollet, 2015) and Tensorflow libraries in Python. As it is a relatively small neural network, it was possible to carry out the

training on the Google Colab cloud computing service, which has 12 GB of RAM memory and an NVIDIA Tesla P100 graphics card.

Neural network models have many parameters that can be adjusted to obtain better predictions. These parameters include the number of intermediate layers, number of neurons in each layer, activation functions, loss functions, optimization algorithms, and regularization algorithms. Although default values are used for general purposes, some parameters must necessarily be adjusted. These adjustments, in turn, can be made by trial and error or using some structured methodology. In this study, we used the hyperband technique (Li *et al.*, 2018), which has proven more time efficient and accurate than other techniques, such as grid search and random search. We also used the mean absolute error (MAE) of the decoder predictions as the objective function to be optimized. The optimization of the network hyperparameters took about 2 h. We present the optimized parameters and the respective optimal values in Table 1.

Table 1 – Optimized Parameters

Parameter	Tested values	Optimum value
Bidirectional LSTM layers of the encoder	0 – 3 bidirectional LSTM + 1 LSTM	2
LSTM layers of the decoder	1 - 5	1
Bidirectional LSTM layer neurons	32 - 512	512
LSTM layer neurons	32 - 512	256
Loss function	Mean square error; absolute mean error; percent average absolute error	Mean square error
Optimizer	Adam; RMSprop ; adagrad ; adadelta	RMSprop
Dropout	0.1 - 0.4	0.15

In addition to these parameters, we used a variable learning rate as a function of the number of training epochs of the neural network. It started with a learning rate of 10^{-3} and was divided by 10 every 20 training epochs.

Then we retrained the optimal model found with the hyperband technique for 60 epochs to achieve complete convergence. We used he model with the lowest MAE in the test portion for the following stages of the study since the use of many epochs can lead to overfitting (Chollet, 2015; Gal and Ghahramani, 2016).

4.1. Forecasts evaluation

The evaluation of the quality of traffic predictions on highways and arterials is not trivial since it does not have uniform characteristics in time and space. Traffic on these roads is usually classified as free flow or congested, and traffic behavior in each of these situations is entirely different and requires different and specific strategies. The transitions between these regimes also present peculiarities and are of particular interest for traffic management since they are linked to the operational capacity of the roads.

With this in mind, we proposed segregating the data into analysis regions with similar traffic characteristics from the flow-speed diagram. In this way, the error can be compared by analysis region and forecast horizon, as shown in Figure 3. We created the proposed regions empirically according to the following criteria: (R1) free flow; (R2) drop in speed due to proximity to capacity; (R3) transition to the congested state; (R4) congestion; and (R5) free flow recovery.

The MAE of the forecasts was 5.40 km/h globally. However, we observed that the error differs in order of magnitude when comparing different traffic regions and forecast horizons:

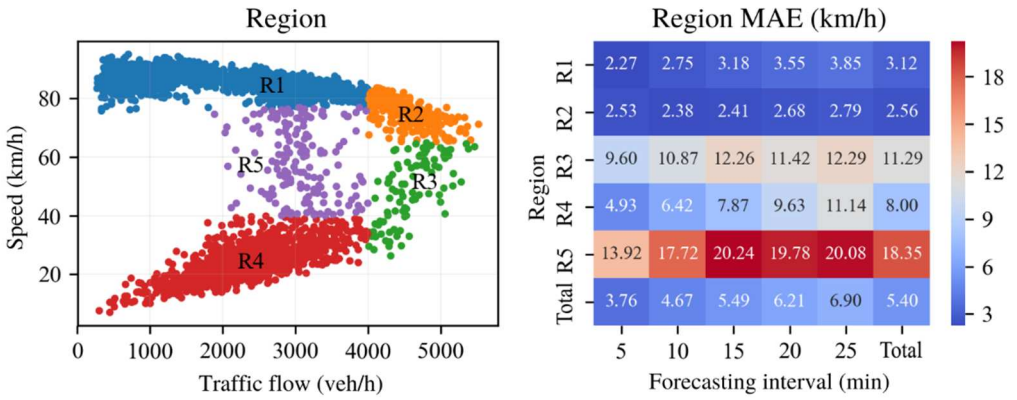

Figure 3. Regions of analysis and MAE by region and forecast horizon

- R1: In this region, vehicles travel at speeds limited by the legal limits of the road. Because of this, the MAE is expected to be low, basically resulting from different individual desired speed choices (Galvan, Zechin and Cybis, 2019). A low MAE was achieved by the proposed model, with little error increase even for the maximum forecast horizon;
- R2: this was the region where the model made the most accurate predictions, which is interesting since it precedes the beginning of the transition to the congested state. In this region, there is greater speed homogeneity resulting from the increase in traffic flow. However, the speed profile does not follow a stable pattern like the R1 region. Good predictions, especially for longer horizons, indicate that the model is capable of predicting the onset of congestion;
- R3: This region refers to the transition from the free flow to the congested state. In this region, a sudden drop in speed is observed, and the calculated average velocity depends significantly on the instant within the aggregation interval (5min in this study) in which this phenomenon occurred. Because of this, there is great speed variability in this region, and it is natural that larger errors are observed proportionally to the size of the chosen data aggregation interval. Thus, in this region, it is expected that the model is able to capture the rapid downward trend even with larger errors than in the other regions. Based on this, we understand that the errors found are compatible with expectations;
- R4: vehicles travel in a stop-and-go motion in this region, and the speed variability is more significant. This happens mainly because data was collected with inductive loops, which measure the instantaneous speed of vehicles. The model errors are smaller for shorter prediction horizons and are close to the errors measured in the R1 region, but increase for larger horizons. There is less interest in obtaining highly accurate predictions in this region since the possibilities of acting on traffic are lower during congestion due to the high density and low speed of vehicles;
- R5: Although predictions during congestion are not very interesting, the possibility of predicting free flow recovery is interesting, and this is done in the R5 region. However, this is the region where the model incurred the most significant errors. The probability that the congestion will end depends on the volume upstream of the bottleneck approaches decreasing, which cannot be measured with just one detector, especially during congestion. Because of this, we expected forecasts in this region to be reactive, respond to measured velocity variations, and have a low anticipation capacity. We stated that this occurred, since the error is high and increases as the forecast horizons increase.

To support the interpretations above, we propose evaluating how the error behaves as a function of time. Figure 4 shows the speed profile used in the test portion of the neural network along with the predictions made, the error of each prediction, and the volume used as weight during training. To evaluate the quality of predictions in the future, we present the predictions made for the first (5 min) and fifth (25 min) predicted interval. As the test portion is large, we show a sample of 200 predicted sequences, where some important phenomena can be observed.

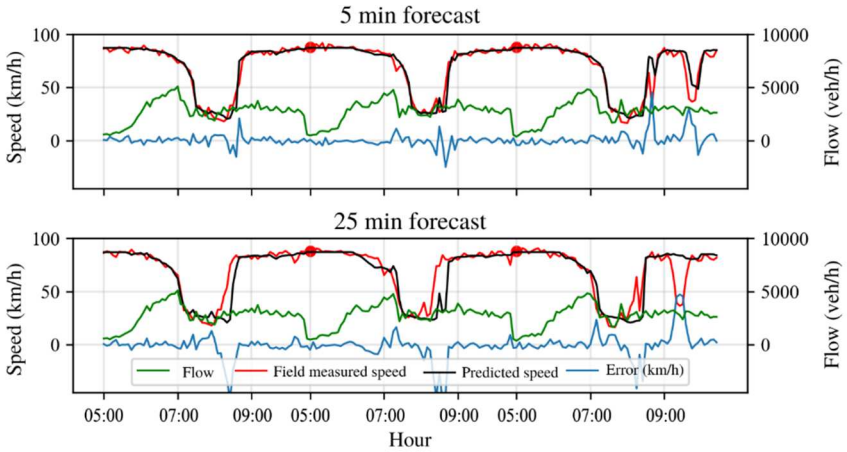

Figure 4. Speed predictions over time. The volumes are scaled to correspond to the vertical axis. Red signals indicate the start of a new morning.

Although distant in time, we observe that the predictions made for 5 and 25 min in the future are similar in terms of error and have good adherence to the speeds measured in the field. The error is noticeably smaller in the regions close to the transition from the free flow to the congested regime since the volume was used as weight during the training process, increasing the relative importance of these intervals. This is a highly desired effect since good forecasts close to capacity are necessary to anticipate the beginning of the transition to the congested regime. In free-flow moments, speed variability is greater since the volume is low, and most vehicles travel unimpeded. It is interesting to note that the model converged to linear predictions in these situations since the main trend is stable and the weights are smaller because they are proportional to the volume. In the congested regime, both forecast horizons have larger and similar errors due to the speed fluctuations that occur during the stop-and-go motion. The biggest difference between the predicted intervals happens in the transition from the congested to the free flow regime; in this case, the speed forecasts seem to react to changes on the road without anticipation of speed recovery. This is clear by looking at the delay between forecasts and field measurements, which is even more significant in the 25 min forecast. As expected, the predictions regarding the recovery of free flow are more erratic than the others since they are highly dependent on the flow of vehicles upstream of the bottleneck under analysis. As this information does not exist in this study, it is natural that the observed error is greater.

4.2. Validation using the predictions to calculate CDFs

The analyses indicate that the proposed model performs satisfactorily for the speed prediction task, especially in regions of particular interest for active traffic management.

Model validation was performed by calculating and statistically comparing CDFs constructed with field and predicted data. These curves were estimated using the breakdown probability calculation methodology suggested by Brilon et al. (2005) to provide robustness to the model validation (Han and Ahn, 2018). Then we statistically tested the hypothesis that the CDFs generated with measured and predicted speeds in the field are different. In this study, we did not use the term breakdown to refer to the measured phenomenon due to the unfavorable position of the detectors. However, the survival analysis does not make this distinction. It is sufficient that we compute the observed phenomenon the same way using predicted and field data for the statistical analysis to be valid.

The methodology for breakdown probability calculation used by Brilon et al. (2005) is widely recognized for its effectiveness and simplicity, having also been used in several studies that followed it (Andrade and Setti, 2014; Elefteriadou et al., 2011, 2014). The original methodology defines a speed threshold, so that the interval preceding a drop in speed that exceeds this limit is considered a breakdown. This interval is censored (received a 1 marker) and the intervals preceding it receive a 0 marker. We discarded intervals following the breakdown. Then we sorted the markers and their respective volumes from the entire database by volume and applied them to the non-parametric Kaplan-Meier model (Kaplan and Meier, 1958) to generate breakdown probability curves as a function of volume. In this study, we considered that the beginning of the transition to the congested regime is analogous to the breakdown phenomenon treated in these studies. We adapted the methodology by Brilon et al. (2005), adding as a criterion for identifying a censored interval the need for 2 consecutive intervals to be below the established speed threshold. We did so to reduce the likelihood of identifying false positives.

Although the breakdown probability curve provides a stochastic view of the road's capacity, traffic managers tend to prefer to use a deterministic value for it. Shojaat et al. (2016) proposed the sustainable flow index (SFI) to meet this demand without giving up the information offered by the probability distribution. This metric originates from the concept of risk, defined as the multiplication of the probability of an adverse event occurring and the damage caused by it. In the context of traffic engineering, and more precisely of the occurrence of a breakdown , the SFI represents the volume that transits through a road and is calculated by the product between the volume and the complementary probability of the occurrence of a breakdown.

The capacity, therefore, is obtained by maximizing the SFI. As an example, the SFI curves, the CDFs made with the predictions, and the speeds measured in the field are shown in Figure 5. We used the last predicted interval (25 min in the future) and a speed threshold of 65 km/h.

Then we investigated the quality of the predictions applied to this methodology by varying the threshold velocity to identify the highest threshold velocity that (i) generates statistically identical probability curves and (ii) produces similar capabilities. The hypothesis that the generated curves are identical was tested by fitting the Cox survival model (Cox., 1972) to the volume data, from the binary marker of early transition to the congested regime calculated previously and an accessory variable that indicates whether the data refers to a prediction or a field measurement. We tested the significance of the accessory variable in the model through the likelihood ratio test, so that p-values greater than an assigned acceptance limit $\alpha = 0.05$ do not allow rejecting the null hypothesis that the curves are identical with 95% confidence, which is desirable in this study. Figure 6 shows the p-values obtained in comparing the curves generated for different speed thresholds and each forecast horizon. Note that we only created 4 curves, since we considered making two predictions lower than the established speed

threshold an identification criterion for the beginning of the transition to the congested regime. We also present the calculated capacities.

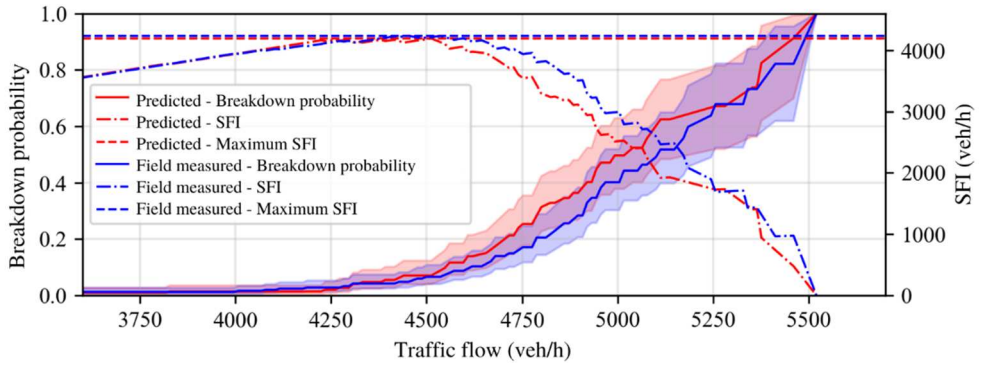

Figure 5. CDFs of the beginning of the transition to the congested regime and SFI for speed threshold = 65 km/h with 25 min prediction data

Speeds greater than 70 km/h generally produce p-values below the limit α, where the null hypothesis that the distributions are identical is rejected. However, we note that the threshold speed from which the p-values become greater than this threshold decreases as the forecast horizon increases. We understand that this occurs because the forecasts are more imprecise the longer the forecast horizon, and the assertiveness of the forecasts increases when there are clearer signs that the speed drop has started and lower speeds are measured.

Visual inspection in Figure 6 shows a convergence between the calculated capacity values for values close to 65 km/h, where there is a maximum absolute difference below 200 veh/h. As we observe convergence between capacities for this speed threshold and all p-values are greater than 0.05, we understand that the neural network well represents both the beginning of the transition to the congested regime and capacity.

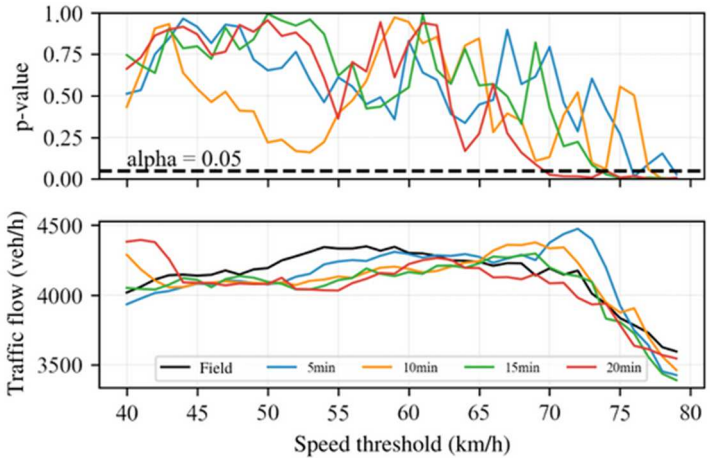

Figure 6. p-value of the accessory variable in the Cox survival model and capacity for different limit speeds

In this application, the speed threshold of 65 km/h could be suggested to characterize the beginning of the transition to the congested regime from speeds predicted by the real-time model in practical applications. However, it is noteworthy that this value is suggested based on the data of this specific case study, so that the ideal speed threshold may differ in other locations due to geometric and behavioral specificities and peculiarities in the demand profile.

5. CONCLUSIONS

In this article, we proposed using an LSTM neural network with encoder-decoder architecture to perform speed predictions of a road segment where breakdowns are observed daily due to a bottleneck. We used rainfall and traffic data collected with inductive loops, including road accidents and lifting information from the mobile span of a bridge, to aggregate as much information relevant to the neural network as possible. We evaluated the forecast results for different traffic states to detail the model's quality. We also validated the results by applying predictions in the calculation of CDFs that represent the probability of the beginning of the transition to the congested regime.

With an MAE of 5.40 km/h, the forecast errors obtained in the regions of greatest interest showed satisfactory results for all predicted intervals, but it is noted that the error increases with the forecast horizon. The use of volumes as a sample weight allowed the reduction of prediction errors in situations where traffic is close to capacity. Because of this, we observed convergence between the probability curves calculated with field and predicted data, indicating that the model can also make good predictions at critical moments for traffic.

Practical applications of the proposed methodology must consider the peculiarities of the used data. The hyperparameters found during the neural network optimization process may differ depending on factors such as the amount of data, the number of variables created, data aggregation, and traffic characteristics in the studied region. The suitability of the methodology for the chosen region can also be verified through the generation and statistical comparison of CDFs .

We suggest for future work using data from detectors located closer to the bottlenecks, so that the breakdown characterization can be performed with greater precision, and to assess whether the location of the detectors significantly influences the results. The use of data from multiple sections of the segment, especially upstream, would also be interesting, as it would allow the model to consider the local traffic state and the volume of vehicles that will pass through the section in the future. Traffic has a stochastic nature, so the probabilistic prediction of speeds may be a more appropriate tool (Fortunato et al., 2017; Kendall and Gal, 2017). Making predictions using adaptations of LSTM neural networks compatible with disaggregated traffic data can also contribute to maximizing the use of information (Neil, Pfeiffer and Liu, 2016). Neural network models are often considered black-box models. However, recent advances indicate ways to create visualizations for humans (Arras *et al.*, 2019). Crossing traffic data with other databases can add even more information to the network, such as the use of traffic images, Bluetooth data, telephony data, and integrations with mobility applications, as well as a previous study of the significance of the variables, to reduce the number of variables used and provide only relevant information. Other models of neural networks, such as transformers networks, also seem promising for solving traffic problems (Wu et al., 2020) but still demand more studies.

REFERENCES

do Amaral, M.B. (2020) *Previsão de velocidades de tráfego com rede neural LSTM*. Bachelor thesis. Departmento de Engenharia Civil: Universidade Federal do Rio Grande do Sul. Porto Alegre, Brazil.

Andrade, G.R. and Setti, J.R. (2014) Speed–Flow Relationship and Capacity for Expressways in Brazil, in *Innovative Applications of the Highway Capacity Manual 2010*. DOI: http://onlinepubs.trb.org/onlinepubs/circulars/ec190.pdf.

Arras, L. *et al.* (2019) Explaining and Interpreting LSTMs, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11700 LNCS(2019), pp. 211–238. DOI: 10.1007/978-3-030-28954-6_11.

Brilon, W., Geistefeldt, J. and Regler, M. (2005) Reliability of Freeway Traffic Flow: A stochastic Concept of Capacity, *Proceedings of the 16th International Symposium on Transportation and Traffic Theory*, (July), pp. 125–144.

Caleffi, F. *et al.* (2016) Influência das condições climáticas e de acidentes na caracterização do comportamento do tráfego em rodovias, *Transportes*, 24(4), p. 57. DOI: 10.14295/transportes.v24i4.1104.

Caleffi, F. (2018) *Proposição de um método de harmonização da velocidade baseado em modelo de previsão de conflitos veiculares*. PhD Thesis. Departamento de Engenharia de Produção e Transportes: Universidade Federal do Rio Grande do Sul. Porto Alegre, Brazil.

Chollet, F. (2018) *Deep Learning with Python*. 1st edn. Shelten Island, NY: Manning Publications.

Elefteriadou, L. *et al.* (2011) Proactive ramp management under the threat of freeway-flow breakdown, *Procedia - Social and Behavioral Sciences*, 16, pp. 4–14. DOI: 10.1016/j.sbspro.2011.04.424.

Elefteriadou, L. *et al.* (2014) Enhancing ramp metering algorithms with the use of probability of breakdown models, *Journal of Transportation Engineering*, 140(4), pp. 1–9. DOI: 10.1061/(ASCE)TE.1943-5436.0000653.

Fortunato, M., Blundell, C. and Vinyals, O. (2017) Bayesian Recurrent Neural Networks, pp. 1–14. DOI: 10.48550/arXiv.1704.02798.

Fu, R., Zhang, Z. and Li, L. (2017) Using LSTM and GRU neural network methods for traffic flow prediction, *Proceedings - 2016 31st Youth Academic Annual Conference of Chinese Association of Automation, YAC 2016*, pp. 324–328. DOI: 10.1109/YAC.2016.7804912.

Gal, Y. and Ghahramani, Z. (2016) A theoretically grounded application of dropout in recurrent neural networks, *Advances in Neural Information Processing Systems*, pp. 1027–1035. DOI: 10.48550/arXiv.1512.05287.

Galvan, Y.T., Zechin, D. and Cybis, H.B.B. (2019) Utilização do método Kaplan-Meier na estimativa das distribuições de velocidade desejada, *Anais do 33º Congresso de Pesquisa e Ensino em Transportes - 2019*. Rio de Janeiro: Associação Nacional de Pesquisa e Ensino em Transportes.

Géron, A. (2019) *Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow. 2nd edn*. Sebastopol: O'Reilly Media

Gu, Y. *et al.* (2019) Short-term prediction of lane-level traffic speeds: A fusion deep learning model, *Transportation Research Part C: Emerging Technologies*, 106(July), pp. 1–16. DOI: 10.1016/j.trc.2019.07.003.

Han, Y. and Ahn, S. (2018) Stochastic modeling of breakdown at freeway merge bottleneck and traffic control method using connected automated vehicle, *Transportation Research Part B: Methodological*, 107, pp. 146–166. DOI: 10.1016/j.trb.2017.11.007.

Hochreiter, S. and Schmidhuber, J. (1997) Long Short-Term Memory, *Neural Computation*, 9(8), pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735.

Iung, B. (2013) Cœur et grossesse, *EMC - Traité de médecine AKOS*, 8(2), pp. 1–4. DOI: 10.1016/s1634-6939(13)59289-1.

Kendall, A. and Gal, Y. (2017) What uncertainties do we need in Bayesian deep learning for computer vision?, *Advances in Neural Information Processing Systems*, 2017-December(Nips), pp. 5575–5585. DOI: 10.48550/arXiv.1703.04977.

Laptev, N. *et al.* (2017) Time-series Extreme Event Forecasting with Neural Networks at Uber, *International Conference on Machine Learning - Time Series Workshop*, pp. 1–5. Available at: <http://www.cs.columbia.edu/~lierranli/publications/TSW2017_paper.pdf>. Accessed at: 12/08/2022.

Li, L. *et al.* (2018) Hyperband: A novel bandit-based approach to hyperparameter optimization, *Journal of Machine Learning Research*, 18, pp. 1–52. DOI: 10.48550/arXiv.1603.06560.

Li, Y., Wang, N. and Carroll, R.J. (2010) Generalized functional linear models with semiparametric single-index interactions, *Journal of the American Statistical Association*, 105(490), pp. 621–633. DOI: 10.1198/jasa.2010.tm09313.

Ma, X. *et al.* (2015) Long short-term memory neural network for traffic speed prediction using remote microwave sensor data, *Transportation Research Part C: Emerging Technologies*, 54, pp. 187–197. DOI: 10.1016/j.trc.2015.03.014.

Ma, Y., Zhang, Z. and Ihler, A. (2020) Multi-Lane Short-Term Traffic Forecasting with Convolutional LSTM Network, *IEEE Access*, 8, pp. 34629–34643. DOI: 10.1109/ACCESS.2020.2974575.

Neil, D., Pfeiffer, M. and Liu, S.C. (2016) Phased LSTM: Accelerating recurrent network training for long or event-based sequences, *Advances in Neural Information Processing Systems*, (Nips), pp. 3889–3897. DOI: 10.48550/arXiv.1610.09513.

Pujol, B.S., Asplund, M. and Serra, J.C. (2019) *Machine learning for early detection of traffic congestion using public transport traffic data*. Bachelor Thesis. Department of Computer and Information Science Machine: Linköping University. Linköping, Sweden.

Vlahogianni, E.I., Karlaftis, M.G. and Golias, J.C. (2014) Short-term traffic forecasting: Where we are and where we're going, *Transportation Research Part C: Emerging Technologies*, 43, pp. 3–19. DOI: 10.1016/j.trc.2014.01.005.

Zechin, D., Caleffi, F. and Cybis, H.B.B. (2020) Influence of Rain on Highway Breakdown Probability, *Transportation Research Record*, 2674(8), pp. 687–695. DOI: 10.1177/0361198120919754.

4 ARTICLE 3: PROBABILISTIC TRAFFIC BREAKDOWN FORECASTING THROUGH BAYESIAN APPROXIMATION USING VARIATIONAL LSTMS

Authors: Douglas Zechin, Helena Beatriz Bettella Cybis

Published at: Transportmetrica B: Transport Dynamics

TRANSPORTMETRICA B: TRANSPORT DYNAMICS
https://doi.org/10.1080/21680566.2022.2164529

Taylor & Francis
Taylor & Francis Group

Probabilistic traffic breakdown forecasting through Bayesian approximation using variational LSTMs

Douglas Zechin and Helena Beatriz Bettella Cybis

Department of Industrial and Transportation Engineering, Federal University of Rio Grande do Sul, Porto Alegre, Brazil

ABSTRACT

This paper proposes a framework for short-term traffic breakdown probability calculation using a Variational LSTM neural network model. Considering that traffic breakdown is a stochastic event, this forecast framework was devised to produce distributions as outputs, which cannot be achieved using standard deterministic recurrent neural networks. Therefore, the model counts on the robustness of neural networks but also includes the stochastic characteristics of highway capacity. The framework consists of three main steps: (i) build and train a probabilistic speed forecasting neural network, (ii) forecast speed distributions with the trained model using Monte Carlo (MC) dropout, and therefore perform Bayesian approximation, and (iii) establish a speed threshold that characterizes breakdown occurrence and calculate breakdown probabilities based on the speed distributions. The proposed framework produced an efficient control over traffic breakdown occurrence, can deal with many independent variables or features, and can be combined with traffic management strategies.

ARTICLE HISTORY
Received 17 July 2022
Accepted 29 December 2022

KEYWORDS
Traffic breakdown; traffic forecasting; neural networks; Bayesian statistics; machine learning

Introduction

Traffic engineering has received valuable contributions from technological advances like the internet of things and artificial intelligence. The intersection between these areas has led to emerging innovative fields and contributed to traditional areas, such as active traffic management (ATM), in which this study fits. Traffic forecasting improvements and anticipation of unwanted scenarios, such as congestion, accidents, and increased travel-time, have been enabled by using more robust methods (Li, Abdel-Aty, and Yuan 2020).

Traffic forecasting is usually treated as a time series prediction problem and solved with parametric methods such as ARIMA (Vlahogianni, Karlaftis, and Golias 2014). These methods have great physical interpretability, require small amounts of data, and their solution is usually simple and demands low computational power. At the same time, recent advances in machine learning suggest that Recurrent Neural Networks (RNNs) adapt well to this type of problem. They can process temporal sequences and are more suitable for capturing traffic's spatial and temporal characteristics, considering the dynamic nature of demand. As of this paper's writing, LSTM (Long Short-Term Memory), proposed by Hochreiter (1997), has proven to outperform other RNN subtypes for most applications (Akhtar and Moridpour 2021). For that reason, we chose LSTMs to conduct this study and apply the proposed methodology.

Most models that use deep learning methods to forecast traffic speed or flow do not consider its stochastic nature and only perform deterministic predictions, mainly because standard NNs cannot

CONTACT Douglas Zechin douglaszechin@gmail.com Department of Industrial and Transportation Engineering, Federal University of Rio Grande do Sul, Porto Alegre, Brazil

capture model uncertainty (Gal and Ghahramani 2016b). However, the relevance of the probabilistic characteristic of highway capacity has been widely studied by the Transportation Engineering community (Brilon, Geistefeldt, and Regler 2005; Chen and Ahn 2018; Elefteriadou, Roess, and McShane 1995, 2011; Kondyli et al. 2013; Persaud, Yagar, and Brownlee 1998; Qu, Zhang, and Wang 2017), resulting in what is called the classical understanding of stochastic highway capacity (Kerner 2019).

This paper proposes a framework for forecasting short-term traffic breakdown probability that attempts to address multiple gaps. Firstly, we propose a formulation for calculating traffic breakdown probability based on samples of speed distributions generated by a probabilistic forecasting model. This formulation enables the calculation of the breakdown probability of individual events, differing from past studies that usually use survival analysis to produce time-independent traffic breakdown probability distributions. Survival analysis-based models define breakdown probability as a function of traffic flow and are unsuitable for short-term forecasting purposes. Secondly, we adopted a novel probabilistic RNN approach for traffic forecasting, from which we sampled the aforementioned speed distribution forecasts. RNNs stand out from other models for traffic forecasting and are widely adopted for multiple purposes in past studies (Akhtar and Moridpour 2021). However, the traditional formulation of these networks cannot produce probabilistic forecasts and is unsuitable for the needs of our framework. To address this gap and represent model uncertainty we propose adopting Variational LSTMs (Gal and Ghahramani 2016b) for speed forecasting, by using dropout as a Bayesian approximation of a Gaussian process.

This study also proposes additional contributions by suggesting approaches to improve the forecasting quality in a traffic breakdown context. Firstly, since this paper is related to traffic breakdowns on highways, we are interested in making good predictions during high-demand periods. In most traffic prediction models that use RNNs, the loss function weights each sample of the training set equally, and there is no concern regarding the criticality of the evaluated traffic state. Although critical, high-demand, pre-breakdown periods are less frequent than uncongested situations and generate a relatively small number of observations in datasets. To overcome this particular issue, the proposed methodology adopts traffic flow as the sample weight during the training process of the neural network to increase their relative importance and force the model to make better predictions in these situations. Secondly, the error of the predictions is evaluated individually for 5 different regions of the fundamental diagram to account for differences in the quality of the predictions in different traffic states. This is fundamental to prevent a large number of good predictions, usually made during stable and more frequent traffic conditions, from masking bad predictions made during critical situations. Thirdly, we propose the use of an encoder–decoder neural network model with an extra output to increase the model training performance. The extra output is positioned after the encoder, in parallel with the decoder, and is used only during the training phase of the NN model.

The complete methodology for probabilistic short-term traffic breakdown predictions presented in this paper consists of three main parts:

(1) Variational LSTM neural network model:
    (a) Data preparation;
    (b) Neural network architecture;
    (c) Sample weighting;
    (d) Hyperparameter tuning and training.
(2) Speed forecasting and evaluation:
    (a) Speed distribution forecasting using Monte Carlo Dropout;
    (b) Model evaluation for different traffic states to guarantee that traffic breakdown periods are well represented.

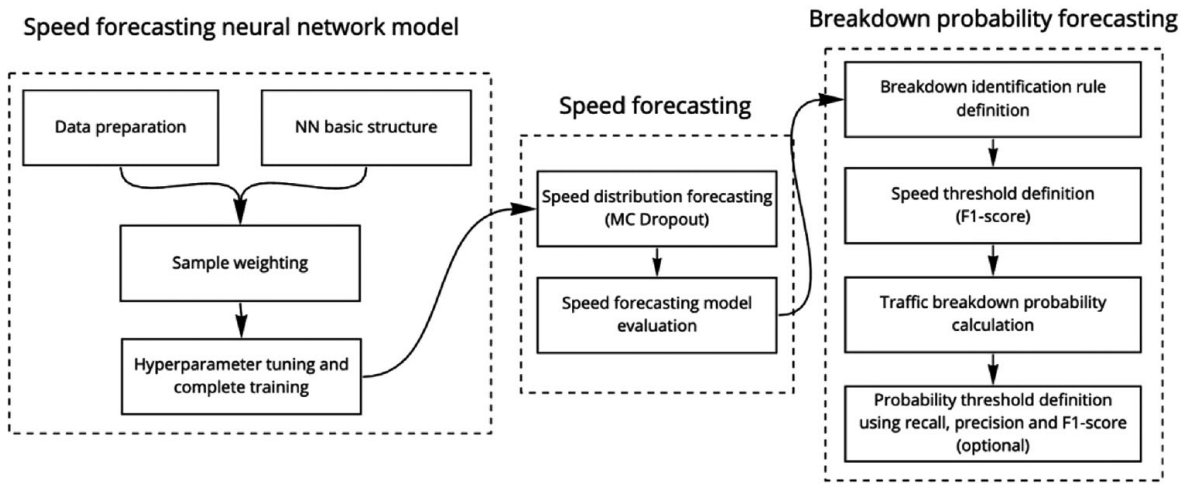

Figure 1. Methodology scheme.

(3) Traffic breakdown probability forecasting:
   (a) Speed and breakdown probability thresholds;
   (b) Calculate traffic breakdown probability for each forecasted time step.

This methodology is schematically presented in Figure 1.

Theoretical background

The methodology presented in this paper consists of three main parts. The first and second parts refer to forecasting speed distributions with credible intervals using a Variational LSTM neural network model. The third part presents the methodology for identifying traffic flow breakdowns from the forecasted speed distributions. In the following sections, we present the theory that supports these topics and a literature review concerning applications of these methods to similar problems.

Traffic breakdown

In the context of highways, traffic breakdown refers to a sudden speed reduction that makes traffic transition from proper operation to non-acceptable flow conditions. This phenomenon occurs as a consequence of the traffic flow achieving the highway capacity in a particular segment, most commonly on an active bottleneck. Therefore, breakdown and capacity are deeply related (Brilon, Geistefeldt, and Regler 2005).

The conditions that trigger the occurrence of a traffic breakdown differ from site to site and include geometry, weather conditions, demand profile, and fleet and driver characteristics. The maximum observed traffic flow that precedes a breakdown also varies under similar conditions, which indicates that highway capacity is a probabilistic measure (Kerner 2019).

The most common approach to traffic breakdown probability calculation is using survival-analysis or hazard models. These models are frequently used in medicine to describe the survival probability of a group using a specific medication or with a particular disease, in mechanics to describe engine durability, and in business to calculate customer churn probability. In traffic engineering applications, the breakdown probability – analogous to traffic death – has not been related to time but traffic flow. The general understanding is that breakdown probability is higher with higher traffic flows as well as the probability of an engine failing increases as time passes (Brilon, Geistefeldt, and Regler 2005; Elefteriadou, Roess, and McShane 1995; Kidando, Moses, and Sando 2019; Lu and Elefteriadou 2013; Persaud, Yagar, and Brownlee 1998).

The non-parametric Kaplan-Meier estimator is a commonly adopted hazard model. This model requires traffic flow and speed data aggregated in time intervals that range from 1 to 5 min. The first step is producing a binary feature that indicates if a breakdown occurred in each time-step, which is usually done by observing a sudden speed drop. This binary variable and the traffic flow are then applied to the following formulation to produce a breakdown probability curve as a function of traffic flow.

$$F_c(q) = 1 - S(q) = 1 - \prod_{i:q_i \leq q} \frac{k_i - d_i}{k_i}, \quad i \in \{B, F\}$$

Where: q: flow [veic / h]; q_i: flow in time interval i [veic / h]; k_i: number of time intervals with a volume $q \geq qi$; d_i: number of breakdowns with a flow qi; {B, F}: set of intervals when breakdown occurred or when traffic flow was non-congested.

A successful implementation of this formulation depends on correctly identifying breakdown occurrence. Among the traffic breakdown identification methodologies presented in the literature, one of the most studied and well-accepted is to consider that a breakdown happens when the average speed measured on a highway is higher than a speed threshold, s_{th}, in a time step t and lower in the subsequent time step $t + 1$ (Brilon, Geistefeldt, and Regler 2005; Elefteriadou, Roess, and McShane 1995; Kidando, Moses, and Sando 2019; Lu and Elefteriadou 2013; Persaud, Yagar, and Brownlee 1998). Although mathematically more complex models such as wavelet decomposition (Ke et al. 2018) have also been reported to be reliable in identifying traffic breakdowns, we used the speed threshold approach in this paper.

Although the Kaplan-Meier estimator is a simple and widely used method, it only describes breakdown probability as a function of traffic flow. Authors have opted to use the Cox proportional hazard model to increase robustness and explainability (Guo, Wang, and Bubb 2013; Asgharzadeh and Kondyli 2020; Li et al. 2022). This model uses similar concepts to the Kaplan-Meier estimator but enables incorporating information from external covariates, resulting in a richer model.

The methodologies used to describe traffic breakdown probability have the primary goal of stochastically defining the capacity of a highway segment based on historical data, which might be relevant to evaluate geometry adjustments, for example. However, although these models have a probabilistic approach to breakdown, they are not designed for forecasting purposes and are unsuitable for our needs. In the following sections, we present traffic forecasting approaches and propose a methodology that enables forecasting respecting the stochastic nature of traffic breakdown.

Traffic breakdown forecasting

Traffic forecasting is a broad expression that can refer to multiple purposes. Forecasts are applicable in various contexts, such as urban roads, rural roads, arterial roads, and highways. The forecasted values also differ, and some of the most common are speed, flow, occupancy and incident probability.

Traffic forecasting is essentially a time-series problem. This problem category can be solved sufficiently well using baseline models such as Multiple Linear Regression (MLR) and autoregressive integrated moving average (ARIMA) models. However, these models have lost space due to the rapid development of Machine Learning and the increased availability of data and computational resources. Akhtar and Moridpour (2021) made a broad systematic review of traffic congestion forecasting, the specific topic where our study lies. They concluded that machine learning models already prevail over more traditional ones and that there is still a wide range of algorithms to be tested for this purpose. LSTM neural networks and their variations are among the most supported models for traffic prediction on highways.

Applications of LSTMs for traffic forecasting consist of variations of LSTMs and their conjoint use with different types of NNs (Luo and Zhou 2021). Among several applications, it is worth mentioning works like Gu et al. (2019), which proposed a deep learning model for short-term prediction of lane-level traffic speeds. Their model included stacked LSTM and GRU (Gated Recurrent Units) layers and

proved superior to other popular time series forecasting models. Ma, Zhang, and Ihler (2020) proposed mixed CNN and LSTM models to predict flow and speed, and Li et al. (2020) to predict congestion. Aiming for increased accuracy and scalability using big data, Xia et al. (2021) proposed a NAW-DBLSTM model on Spark that uses a bidirectional LSTM with an attention mechanism to perform traffic flow forecasting.

Although these models produce accurate predictions, they do not offer a manner to measure the confidence of individual predictions and are unsuitable for our study. Our methodology utilises probabilistic speed predictions to forecast traffic breakdown probability, and as for the writing of this paper, we could not find studies that propose a similar solution for this problem. Therefore, we propose developing our methodology over a specific neural network called Variational LSTM. This kind of neural network has the required probabilistic characteristics, has strong literature support and has proven more effective than baseline non-probabilistic models for speed forecasting, as presented later in the text.

Variational LSTMs

LSTMs were devised to mitigate the vanishing gradient problem that causes simpler RNNs of a certain size to become untrainable (Hochreiter 1997). LSTMs use memory cells with internal recurrence and grant the possibility of controlling the information that should be retained or forgotten. This characteristic allows LSTMs to outperform standard NNs and RNNs on time series problems, improving their ability to retain long-term dependencies, preventing older signals from vanishing during the training process, and improving training convergence (Goodfellow, Bengio, and Courville 2016). We have omitted the mathematical background on LSTMs since it is widely available in machine learning books.

The predictive power of variations of LSTMs for traffic forecasting is already well established. However, the uncertainty of the predictions might be of great concern for critical applications of machine learning-based strategies, and the calculation of confidence intervals is not possible with the traditional formulation of neural networks. To address this specific limitation, variational inference was suggested to approximate posterior distributions using neural networks and generate Bayesian NNs (Graves 2011). Although this enables the production of outputs with credibility intervals, this technique has proved computationally complex (Gal and Ghahramani 2016b).

A more straightforward and reliable method to approximate Bayesian NNs was proposed by Gal and Ghahramani (2016b) using *dropout*, a technique commonly used for NN regularization. Dropout consists of zeroing the weights of a percentage of the neurons to avoid overfitting (Srivastava et al. 2014) and is usually applied only during the training process of NNs. However, when used during inference, dropout imposes a Bernoulli distribution over the NN's weights, and each prediction turns out to be slightly different. Sampling from the NN using dropout during inference results in a distribution of outputs and is called *Monte Carlo (MC) dropout*. It produces a Bayesian approximation, as mathematically demonstrated for vanilla NNs by Gal and Ghahramani (2016b), and therefore a distribution that can be used for probability calculations.

Gal and Ghahramani (2016a) later suggested that, with additional considerations, MC dropout can also be applied to LSTMs, resulting in NNs that are able to perform approximate variational inference and called Variational LSTMs. For Variational LSTMs to have these Bayesian properties, the same network units (same mask) are dropped at each time step for inputs, outputs, and recurrent layers. We depict the configuration of a simple two-layer Variational LSTM in Figure 2 and represent dropouts between cells as arrows.

In contrast, in naïve applications of dropout on LSTMs, different dropout masks are sampled at each time step for the inputs and outputs, and no dropout is used on the recurrent connections. Fortunato, Blundell, and Vinyals (2017) proved mathematically that Variational LSTMs produce correct posterior distributions. This study uses speed distributions forecasted using Variational LSTMs to calculate breakdown probability, as shown in the following section.

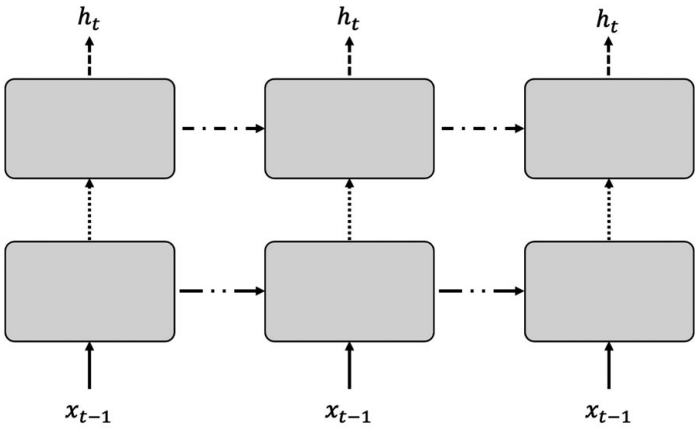

Figure 2. LSTM example with 2 layers (horizontal rows) and 3 cells each layer. Vertical arrows represent conventional dropout, and horizontal arrows represent recurrent dropout. Each arrow dash style is a different mask.

Proposed traffic breakdown forecasting methodology

Before using the neural network outputs to calculate traffic breakdowns, it is first necessary to define how to identify them based on aggregated data and to define a calculation methodology. In this study, we adopted the speed threshold strategy presented in the Traffic Breakdown section to identify breakdowns due to its simplicity, interpretability, and efficiency (Lu and Elefteriadou 2013). We added a complementary condition that the speed remains under the speed threshold for a minimum number of time steps to avoid false positives, which has been widely used for this purpose.

In this paper, the evaluation is done within an evaluation window, which comprises a set of $N+1$ time steps. The first time step, t, aims to identify a pre-breakdown situation by checking if the average speed $\boldsymbol{s_t}$ is greater than a speed threshold $\boldsymbol{s_{th}}$. The subsequent N time steps, $t+n$, represent the number of time steps required to present the average speed under the speed threshold in order to correctly characterize a breakdown event. If, and only if, all of these criteria are matched, we consider that a breakdown happens in the time step t of the evaluation window.

Breakdown identification is trivial when analysing field data and we can assume that the probability that an event B refers to a breakdown b during the time step t is binary since we observed it. In this case, the breakdown probability on a time interval t can be calculated by:

$$\mathrm{P}(B = b|\boldsymbol{s_{th}}, \boldsymbol{s}, \boldsymbol{N}, \boldsymbol{t}) = \begin{cases} 1, & (\boldsymbol{s_t} \geq \boldsymbol{s_{th}}) \; and \; (\boldsymbol{s_{t+n}} < \boldsymbol{s_{th}}) \\ 0, & otherwise \end{cases} \quad \boldsymbol{for\ n} = 1, \ldots, \boldsymbol{N} \tag{1}$$

However, the outputs of the proposed NN are not discrete, and the speed predictions of each forecasted time step t are associated with a probability distribution. Therefore, there is an associated uncertainty on whether the forecasted speed dropped below the speed threshold. The Bayesian approximation consists of replicating the predictions several times to produce a distribution of outputs, so we applied the same breakdown identification rule for the predicted speed sequences of each replication. We calculated the breakdown probability of a time step t as the ratio between the number of sequences that satisfy the breakdown identification rule and the total number of replications. Considering R, the number of replications, and r, the subindex denoting a single replication, the breakdown probability at a time step t is given by

$$P(B = b|\boldsymbol{s_{th}}, \boldsymbol{s}, \boldsymbol{N}, \boldsymbol{t}, \; \boldsymbol{R}) = \frac{\sum_{\boldsymbol{r}=1}^{\boldsymbol{R}} \begin{cases} 1, & (\boldsymbol{s_{t_r}} \geq \boldsymbol{s_{th}}) \; and \; (\boldsymbol{s_{t+n_r}} < \boldsymbol{s_{th}}) \\ 0, & otherwise \end{cases}}{\boldsymbol{R}} \quad \boldsymbol{for\ n} = 1, \ldots, \boldsymbol{N} \tag{2}$$

In the context of our framework, the sequences of speed distributions are generated by the Variational LSTM neural network for a desired number of forecasted time steps. We applied the aforementioned breakdown probability formulation to these speed distributions and calculated the

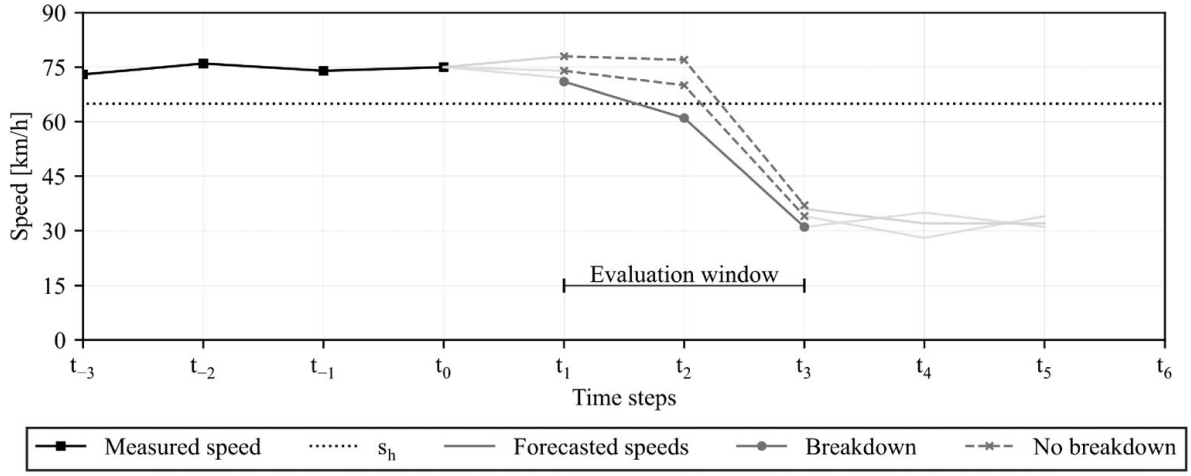

Figure 3. Visualization of the methodology for traffic breakdown probability calculation.

final breakdown probability by defining the speed threshold, $\boldsymbol{s_{th}}$. This produces a traffic breakdown probability for each of the forecasted time steps.

Figure 3 elucidates the proposed methodology with a simplified example. After reading average speeds up to t_0 we predicted speeds for future time steps t_1, t_2, t_3, t_4 and t_5, or a forecast horizon of five time steps. For simplicity and clarity, we adopted a total number of replications $R = 3$ in this example, and a speed threshold $s_h = 65$ km/h. We also set the value of N to 2, which means that a breakdown is predicted when a speed higher than the speed threshold is followed by two time steps with speeds under the speed threshold. In this example, the evaluation window is 3 time steps. Supposing that we are interested in evaluating the window between t_1 and t_3, only the sequence represented by the grey line with circle markers satisfies the breakdown identification rule. Using Equation 2, we have that the breakdown probability for this evaluation window is $B = 1/3$.

Most authors define $\boldsymbol{s_{th}}$ visually, either by observation of the fundamental diagram or by the traffic speed time series (Kidando, Moses, and Sando 2019). In the Third Part of the methodology, we propose using recall, precision, and the F1-score to define an optimum value for $\boldsymbol{s_{th}}$.

The traffic breakdown probability might not be a satisfactory answer depending on the requirements of the practical application of this methodology. It might be necessary to define whether a breakdown is expected to happen or not. A traffic breakdown probability threshold, $\boldsymbol{b_{th}}$, can be used to define the minimum probability for a positive event to be considered. In the Third Part of this framework, we propose a methodology for determining both $\boldsymbol{s_{th}}$ and $\boldsymbol{b_{th}}$.

First part: variational LSTM neural network model

The first part of the methodology consists of building and training a Variational LSTM neural network model, which we used to produce speed forecasting. This session describes the several steps developed to build this model. The first step involves the data gathering and preparation effort, aiming to present as much information as possible to the neural network to optimize its prediction capabilities. We defined the neural network structure in the second step and set up sample weighting in the third step using traffic flow values to increase the relative importance of breakdown periods during the training process. In the fourth step, we submitted the model to hyperparameter tuning to optimize its structure and improve its prediction capability, and then we trained it upon convergence.

Data preparation

This paper focuses on forecasting highway speeds close to a bottleneck, particularly in performing good predictions during traffic breakdowns. Traffic data was collected upstream of a highway segment

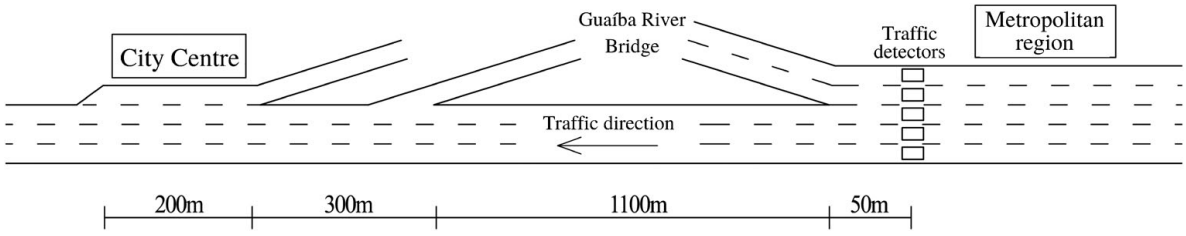

Figure 4. Study site and traffic detector's location.

Table 1. Input and output variables of the NN.

Purpose	Type	Feature	Number of features	Format	Description
Input	Traffic data	speed_avg_*n*	3	Continuous	Average speed on lane *n*.
		speed_med_*n*	3	Continuous	Median speed on lane *n*.
		speed_min_*n*	3	Continuous	Minimum speed on lane *n*.
		speed_max_*n*	3	Continuous	Maximum speed on lane *n*.
		speed_std_*n*	3	Continuous	Standard deviation of the speed on lane *n*.
		volume_l*n*	3	Continuous	Volume on lane *n*.
		detected_*n*	3	Binary	Indicates if any vehicle was detected on lane *n* during a given time step. Mitigates the effects of null average speeds.
	Environment	day_week_*d*	5	Binary	Indicates the day of the week.
		precipitation	1	Continuous	Average rain intensity in mm/h.
Output	Traffic data	speed_*f*	5	Continuous	Average track speed of the *f*th time step in the future.

with multiple on-ramps and daily morning congestion in the metropolitan area of Porto Alegre, Brazil, as shown in Figure 4. The bottleneck happens due to traffic originated in the metropolitan region of the city that is accessing the main road through an on-ramp (depicted as Guaíba Bridge in Figure 4) and a secondary access with minor traffic originated in the local neighbourhood. Due to fleet heterogeneity and intense and concentrated demand, traffic management is very challenging in this region, and some active traffic management strategies have been studied to mitigate congestion and the occurrence of crashes (Caleffi, Moisan, and Cybis 2016; Zechin, Caleffi, and Cybis 2020).

We gathered disaggregated traffic data from the loop detectors with information on passing vehicles' timestamps, speed, and lane for a two-year period. The Guaíba Bridge is a movable bridge, and the upstream traffic must be interrupted frequently due to the bridge lift. We created a binary variable indicating bridge interruption events for each evaluated time step, since removing data from the entire day with bridge lifting would compromise and reduce the dataset. The dataset also comprised a binary variable to indicate periods when light crashes occurred. The data set included rain intensity data collected from a pluviometer 500 m distant from the traffic detectors to allow the model to account for the influence of the weather. We also removed periods with severe crashes, malfunction of detectors, and weekends, resulting in 246 days of useful data. Since this region presents traffic breakdowns only in the morning, the analysis period was limited to between 4 and 12 am.

LSTMs require inputs organized in regular time intervals, demanding data aggregation. We performed feature engineering to create the inputs of the NN while preserving essential traffic characteristics and losing as little information as possible, and aggregated the data in 5 min intervals per lane. For the outputs, we calculated the average speed of the whole segment in 5 min intervals up to 25 min in the future. Feature engineering resulted in the features presented in Table 1, and we detail the NN structure in the next section.

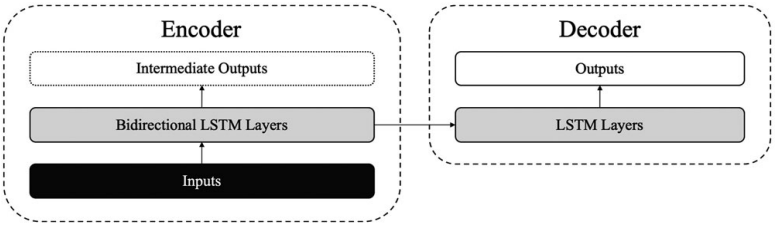

Figure 5. Encoder–decoder LSTM architecture.

Neural network architecture

Concerning the NN model, we propose an encoder–decoder architecture and LSTM layers for traffic speed predictions. The encoder–decoder architecture comprises two main parts: (i) the encoder is responsible for interpreting the time series given as input to the NN, and (ii) the decoder receives this information and generates the outputs. This architecture has achieved better results than standard NN with LSTM layers (Laptev et al. 2017). We used bidirectional LSTM layers in the encoder since they led to better results in preliminary tests and were also suggested by other authors (Cui, Ke, and Wang 2018). We stacked multiple bidirectional and traditional LSTM layers in both the encoder and the decoder to expand the learning capacities of the model (Chollet 2015; Géron 2019). Figure 5 presents a visual representation of the proposed model.

The proposed NN inputs are time series, so we used a *lookback* of 12 time steps of 5 min intervals (1 h) for each sample, and each time step contains all the *n_features* presented in Table 1 as inputs. When structured, the inputs have a final shape (*n_features*, *n_samples*, *lookback*), where *n_samples* is the number of samples in the dataset. This NN architecture is very flexible, and there is no consensus in the literature regarding optimal construction. We propose a neural network model with two outputs:

- Decoder output: this is the main output, which is used during the training and testing phases and to make final speed predictions. Its shape is (*n_samples*, *forecast*), where *forecast* is the number of forecasted time steps.
- Encoder output: we set this output on top of the encoder to stabilize convergence and increase the decoder output accuracy. Its shape is (*forecast*, *n_samples*, *lookback*), so for each of the 12 time steps in the past, the output is the 5 subsequent time steps. This output is also taken into account in the loss function and used for backpropagation, so the encoder receives closer updates during training and the model converges faster and becomes more stable. We tested several approaches beforehand and used this in our study since it performed better. The encoder output is used only during the training phase and is removed from the model during inference. Therefore, although it increases the complexity of the model during training, the complexity during inference is the same as if it were not used.

We built the Variational LSTM neural network model with the Keras/Tensorflow Python library since it has implementations that enable its representation, especially regarding the peculiarities of the dropout. Building the neural network from scratch was also crucial to model the encoder–decoder architecture with two outputs, sample weighting using traffic flow, and perform hyperparameters tuning.

Sample weighting

It is reasonable to assume that the relative importance of these outputs is not the same. To address this problem, we used the weighted sum of the Mean Squared Error (MSE) of each output as the final loss function. Using weights equal to 3 and 1 for the decoder and encoder outputs, respectively, produced the best results in preliminary tests and was kept during the training processes.

Table 2. Hyperparameter tuning.

Hyperparameter	Tested values	Optimum value
Bidirectional LSTM layers on the encoder (+ 1 LSTM layer)	[0;1;2;3]	3
LSTM layers on the decoder	[1;2;3;4;5]	1
Number of neurons on encoder layers	[32;64;128;256;512]	512
Number of neurons on decoder layers	[32;64;128;256;512]	256
Loss function	Mean Squared Error (MSE); Mean Absolute Error (MAE); Mean Absolute Percentage Error (MAPE)	MSE
Optimizer	Adam; RMSProp; Adagrad; Adadelta	RMSProp
Dropout and recurrent dropout rate	[0,1 – 0,4]	0,15

The importance of making good predictions varies according to traffic conditions since traffic characteristics are entirely different at moments with low traffic flow and during peak hours (Caleffi, Moisan, and Cybis 2016). These situations should somehow have different relative importance during the NN training process. However, traditional loss functions such as the Mean Average Error, Root Mean Squared Error, or Mean Squared Error consider that all samples have the same importance except for the magnitude of the respective error. Therefore, less frequent events like breakdowns, which are of the utmost importance for practical reasons, tend to be neglected in favour of reducing the error in the abundant and less important periods with low traffic flow when speed variance and errors tend to be higher (Goodfellow, Bengio, and Courville 2016).

We proposed using sample weights during the training process to address this problem as we understand that the importance of the predictions is highly correlated to traffic flow as traffic breakdown and traffic flow are also highly correlated. We weighted errors in both outputs by the traffic flow 10 min before the first predicted time step to deal with the imbalance between breakdown and non-breakdown periods (Yang et al. 2019). Considering that q_n is the traffic flow 10 min before the first predicted time step of a sample n out of N samples, $L_n$is the non-weighted loss calculated for sample n, and p is the proportion of non-zero traffic flow values, the final loss function L^* will be:

$$L^* = \frac{1}{N}\sum_{n=1}^{N}\frac{q_n L_n}{p} \tag{9}$$

The final loss function L^* will depend on the chosen non-weighted loss function L, chosen via hyperparameter tuning and presented in the next section.

Hyperparameter tuning and training

For the training process, we split the dataset into a training and a test set to guarantee that the model can perform well on unseen data. Therefore, we used the training set during the training process and evaluated its performance using both sets. This study deals with time series data so we did not shuffle the data to build these sets. The training set was built using the first 80% of the dataset after cleaning and the remaining data was left for the testing set.

NN models have many parameters that can be adjusted to obtain better predictions. Although these adjustments could have been made by trial and error in this study, we used the hyperband technique (Li et al. 2018), which has proved to be more efficient and accurate than traditional methods such as grid search and random search. We used the mean absolute error (MAE) of the decoder predictions as the optimization objective function. The optimum values of each tuned hyperparameter are shown in Table 2.

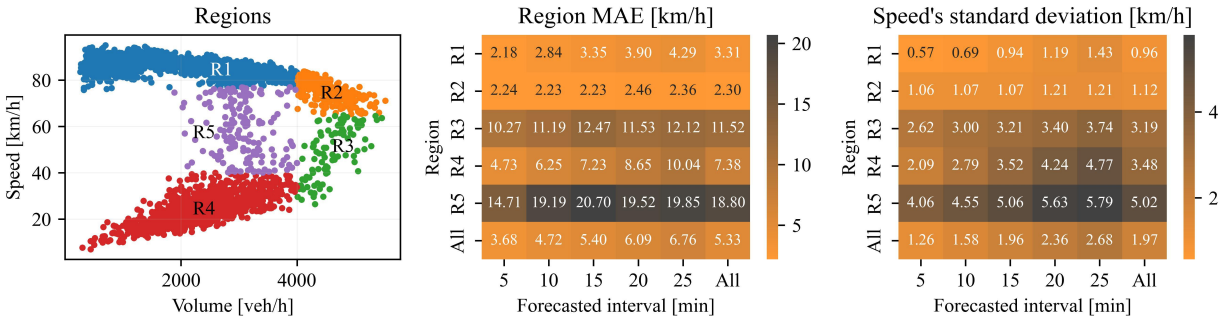

Figure 6. Traffic regions, MAE, and speed's standard deviation per region and forecasted interval.

Second part: speed forecasting and evaluation

The second part of the methodology involves two steps: prediction generation and results evaluation. Although we used the neural network model at this stage, we preferred to present it separately because once the model has been successfully trained, only its structure and weights are relevant to the next steps.

The proposed methodology for speed distribution forecasting using MC Dropout consists of running the neural network over the test dataset inputs 1000 times (Gal and Ghahramani 2016a). This process produces slightly different outputs in each iteration due to the use of dropout with random masks during inference, resulting in a distribution that approximates a Bayesian inference. Each input sample of the test dataset produced 5 speed distributions, corresponding to a sequence of 5 average speed time steps aggregated over 5 min intervals.

For evaluation purposes, we segregated the data into analysis regions with similar characteristics based on the flow-speed diagram. This was because highway traffic presents completely different characteristics during free and congested flow, and the transition between these regimes also presents peculiarities that are strongly related to capacity. Therefore, the MAE and forecasted speed's standard deviation (SSD) can be compared by analysing each chart region and forecasting time step, as shown in Figure 6. MAE measures how close the average predicted speed is to the field measurements, and SSD indicates how spread the speed distributions made by the NN are. We created the proposed analysis regions visually according to the following criteria: (R1) free flow; (R2) drop in speed due to the proximity of capacity; (R3) transition to congested state; (R4) congestion, and (R5) recovery to free flow. A more robust segregation method could have been used for this purpose, but the visual approach seemed more reasonable due to its simplicity and since this analysis is not part of the proposed methodology.

The global MAE of the forecasts was 7.92 km/h, and the global standard deviation of the speed was 1.79 km/h. However, we observed that these metrics differ when comparing different traffic regions and forecasted time steps:

- R1: in this region, vehicles travel at speeds limited by the legal limits of the road. MAE is expected to be low, varying mainly due to fluctuations in desired speed. This was achieved by the model predictions, with a minor increase in error and SSD even for the maximum forecast horizon;
- R2: in this region, there is a greater speed homogeneity resulting from increased traffic synchrony. Good forecasts, especially for longer forecasting horizons, indicate that the model can predict the onset of congestion. Although the observed MAE is considerably slower in R2 than R1, the SSD is more than double, indicating that the model is more accurate but has more uncertainty in this region. This shows the effectiveness of using volume as the sample weight to prioritize high-demand time steps, which were assumed to be more relevant. The higher SSD also evidences the stochastic characteristics of traffic breakdown;
- R3: this region refers to the transition from the free flow to the congested regime through a breakdown. Breakdowns trigger an abrupt drop in speed so that the average speed calculated during the

Table 3. Baseline models used for speed forecasting comparison.

Model	Characteristics
Multiple Linear Regression (MLR)	-
Neural Network	1x 512 neurons layer + 1x 256 neurons layer; Adam optimizer; MSE loss function
LSTM (simple)	1x 512 neurons layer; Adam optimizer; Dropout = 0.1 MSE loss function
Variational LSTM (simple)	1x 512 neurons layer; Adam optimizer; Dropout = 0.1 MSE loss function
ARIMA	p = 3; d = 0; q = 1
XGBoost	Estimators = 100; Learning rate = 0.1

transition depends on the time within the aggregation interval (5 min in this study) when this phenomenon occurred. This region has great speed variability due to the nature of traffic behaviour, and the larger the aggregation interval, the greater the observed errors. Thus, the model is expected to capture the rapid downward trend in speed so that a future breakdown can still be characterized along with R2's forecasts. Interestingly, the SSD was slower for longer forecasting horizons because the model's inability to make these predictions is embedded in the high errors observed in this situation, resulting in low variance and high bias. Based on this, we understand that the errors and SSDs found are compatible with expectations;

- R4: vehicles travel in stop-and-go movements and the average speed is slow in this region, resulting in great speed variability. The model errors are larger than in regions R1 and R2 and close to those from R3, but they increase for larger horizons. There is less interest in obtaining highly accurate predictions in this region since the possibilities of traffic actuation are smaller during congestion due to the high density and low speeds. A lower calculated SSD compared to R3 accompanied by a large MAE indicates that the model has high bias and low variance in this region;
- R5: Although accurate speed forecasts may not be particularly interesting during highly congested periods, the possibility of predicting speed recovery may be useful, and this is an outcome of the analysis of the R5 region. However, this is the region where the model made the biggest mistakes. While the probability of breakdowns can be treated as a function of the traffic flow (Brilon, Geistefeldt, and Regler 2005), the probability of ending congestion depends on whether the volume upstream of the bottleneck decreases, which cannot be measured with only one measurement location. On account of the configuration of the study area, we expected the forecasts in this region to be reactive to measured speed variations and have low anticipation capacity. These expectations were met since the error is high and increases as the forecast horizons increase alongside a reduction in SSD, as occurred in R3.

The primary goal of our paper is to propose a breakdown probability forecasting methodology. For doing so, we used a well-established neural network that has the required probabilistic characteristics and trained it focusing on good predictions during high-demand periods. To demonstrate where our model stands in relation to baseline models, we propose a comparative analysis of its forecasting quality with the baseline models presented in Table 3.

With the MLR we aim to set a baseline; the simpler LSTMs and the NN put our Encoder-Decoder architecture into perspective; the ARIMA model is widely used for time-series problems; and the XGBoost is a state-of-the-art Machine Learning model based on decision trees that uses extreme gradient boosting for performance enhancement and has also been proven effective to solve time-series

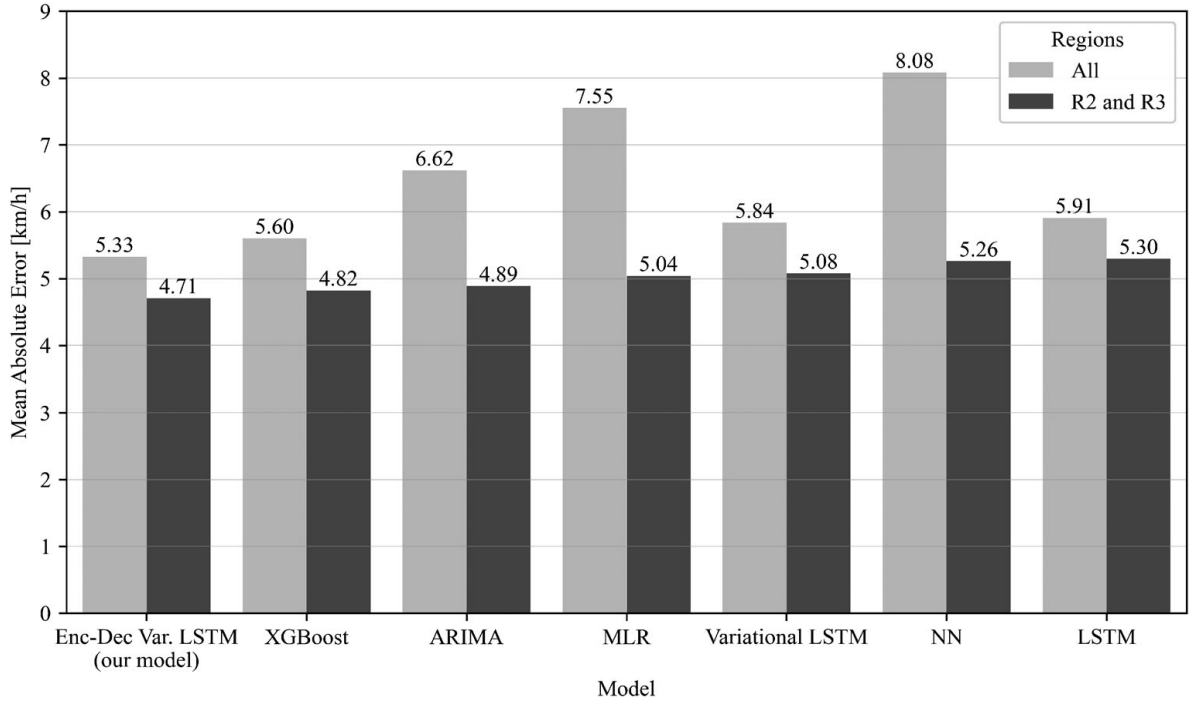

Figure 7. Forecasting error comparison with baseline models.

problems, outperforming even Deep Learning models (Fang et al. 2022). Except for the simpler Variational LSTM, these models do not produce probabilistic forecasts, so we used the average values of our predictions to compare them.

We compared the MAE of the predictions made in the R2 and R3 regions (pre-breakdown and transition to congested state) and the overall performance. The error distribution of all models passed a test for normality using the Kolmogorov–Smirnov test with a 95% confidence interval (p-value < 0.05). We present the comparisons in Figure 7.

Our model has the slowest MAE compared to all the other models for the R2 and R3 regions, which are the main focus of our study and for which our model was trained to perform better. The Neural Network and the simpler LSTMs performed poorly, which evidences the importance of fine-tuning and correctly scaling deep learning models as presented in our study. Our model outperformed the XGBoost model, which was the best performing model among the ones chosen for comparison.

Besides outperforming the other models, our model can also forecast speed sequences with credible intervals, an essential feature for our methodology. These results state the quality of our model and its suitability to being used in our methodology.

We did not compare our model with other Bayesian models since they require a series of assumptions, such as the definition of prior distributions for each input feature. Since we are proposing and applying a methodology, we understand that these discussions would be beyond our scope. Still, we encourage future studies to explore these comparisons and we suggest the use of other models with the same probabilistic characteristics.

Third part: traffic breakdown probability forecasting

The third part of the breakdown forecasting method is the estimation of the breakdown probability from the speed predictions of the neural network model. In this study we used an evaluation window of 3 time steps to define a breakdown event, which comprises a set of 1 time step with average speed above and 2 time steps with average speed under the speed threshold, $\boldsymbol{s_{th}}$. This procedure produces a breakdown probability for each forecasted time step. Such probabilities can then be compared to a

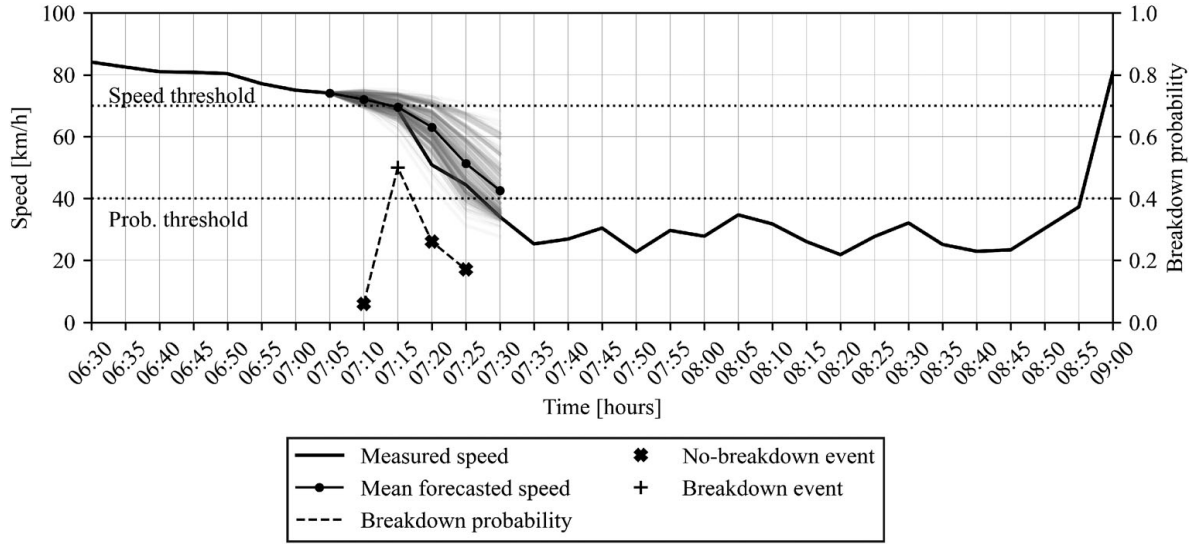

Figure 8. Speed and breakdown probability thresholds example. The cross signs along with the probability line represent non-breakdown intervals based on the probability threshold, and the plus sign represents the opposite.

breakdown probability threshold, $\boldsymbol{b_{th}}$, to decide whether the time step will be considered a breakdown or not.

Figure 8 shows an example of the interaction of $\boldsymbol{s_{th}}$ and $\boldsymbol{b_{th}}$ along with the speed predictions. In this example, the current time is 07:05 and the breakdown probability forecasting horizon is 25 min in the future, or 5 time steps of 5 min intervals. The grey lines represent forecasted speed distributions associated with each of these time steps, and the chosen speed threshold, $\boldsymbol{s_{th}}$, is 75 km/h. Applying the abovementioned methodology to calculate the breakdown probability resulted in the values presented by the dashed line. The breakdown probabilities can then be used as inputs for another application or, upon definition of a probability threshold $\boldsymbol{b_{th}}$, state whether the breakdown will occur in each of the forecasted time steps or not. In this example, a $\boldsymbol{b_{th}}$ of 0.4 indicates that a breakdown should happen in the second upcoming time step, which indeed occurred, as we can observe based on the measured speed line.

Defining which thresholds should be used is, in turn, not trivial. Although empirically determined values, such as visually defining the speed threshold, could lead to satisfactory results, we propose a structured method to define both thresholds and discuss their implications. Using the breakdown identification methodology, we calculated the breakdown probability for each sample and forecasting time steps of the test dataset. Then we used standard machine learning metrics to test whether the prediction quality is suitable for this purpose and to define speed and breakdown probability thresholds. These metrics are adequate for this purpose since they mainly focus on the true events, representing traffic breakdowns in this study. The metrics used were precision, recall, and F1-score, which are calculated as follows:

$$Precision = (True\ Positive)/(Predicted\ Positive) \tag{10}$$

$$Recall = (True\ Positive)/(Actual\ Positive) \tag{11}$$

$$F1 = 2\,(Precision \times Recall)/(Precision + Recall) \tag{12}$$

where: True Positive = correctly predicted *true* (breakdown) events. Predicted Positive = total of predicted *true* events. Actual Positive = total of actual *true* events.

Tuning a model based solely on recall or precision penalizes its performance in terms of the other metric. When precision is increased, for example, the total number of *true* predictions decreases, so only the most certain ones are considered. On the other hand, when recall rises, the total number of *true* predictions also increases, and, in a limiting case, all labels could be considered *true* for a perfect

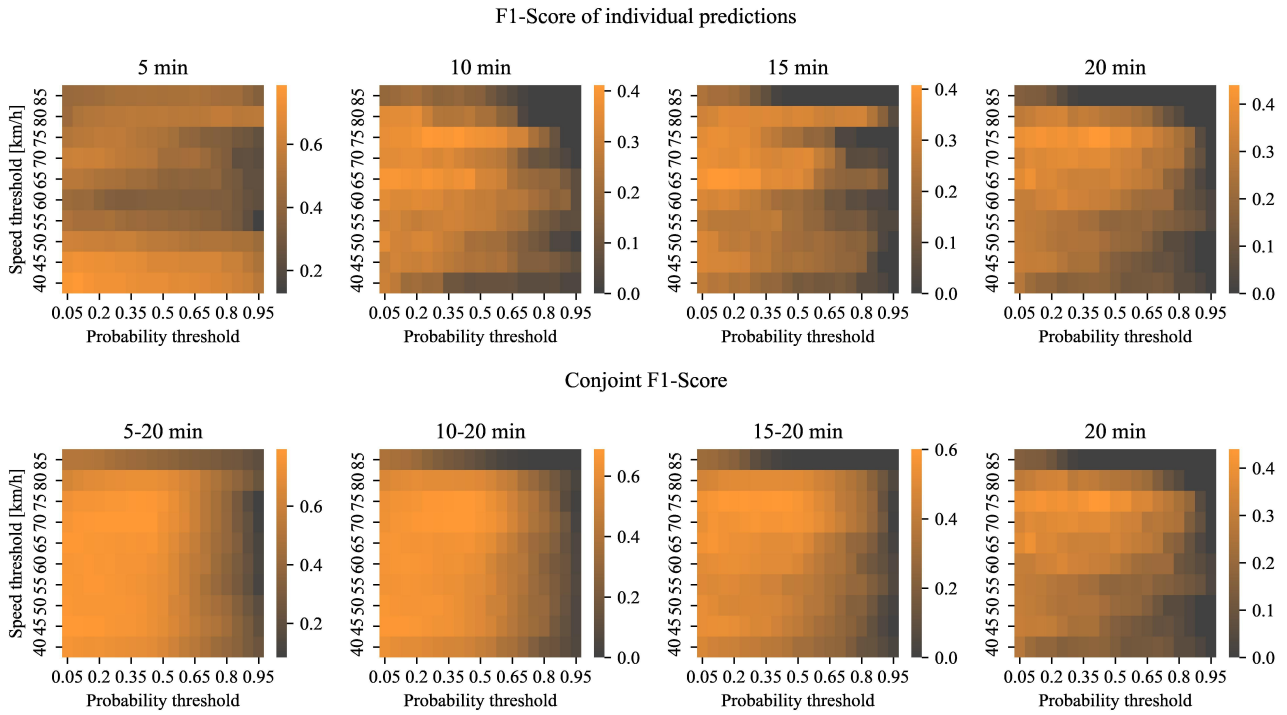

Figure 9. F1-score of breakdown predictions with varying speed and probability thresholds.

recall to be achieved. However, when the number of True Positives is low, both recall and precision will also be low. The F1-score was proposed to balance these metrics, which is calculated by their harmonic mean. We varied the speed and probability thresholds to produce greater insight into the optimal values for both these parameters and calculated the F1-score for each pair.

It is also interesting to analyse whether the predictions were effective along each sequence but not necessarily precise in predicting the exact moment the breakdown occurs within the forecasting horizon and the time limits up to which the neural network can correctly predict this phenomenon. Therefore, while the normal F1-scores are calculated for each evaluation window of the predicted sequence, we propose a conjoint F1-score that uses the maximum breakdown probability between these evaluation windows. We compared the conjoint score with the occurrence or not of a breakdown in the same period, as presented in Figure 9. We gradually removed the shorter-term evaluation windows, aiming to evaluate the conjoint quality of the longer-term forecasts.

This visualization indicates that a speed threshold of 75 km/h combined with a probability threshold of approximately 0.4 produces the most accurate predictions based on the F1-score. 5 and 10 min forecasting horizons are adequate for predicting breakdowns, but 15 and 20 min present poor F1-scores for this scenario. However, the conjoint score of the 15–20 min produces results that are even better than the 10 min horizon alone. Since uncertainties are more significant as the forecasting horizon increases, it is reasonable to use this conjoint score for the 15 and 20 min intervals instead of the individual predictions. Therefore, in this study case, the decision-maker could use 5 and 10 min predictions and the conjoint analysis made for 15–20 min to make the final predictions for practical purposes.

Discussion

In the presented methodology, we proposed using a probability threshold that optimizes the F1-score to assess the occurrence of traffic breakdowns in future time steps. However, depending on the purpose of the application, a different weighting between recall or precision – or even just one of them – could be used. For example, assuming that the application does not tolerate false positives, only very likely breakdown events should be considered. For that, a high probability threshold should be used,

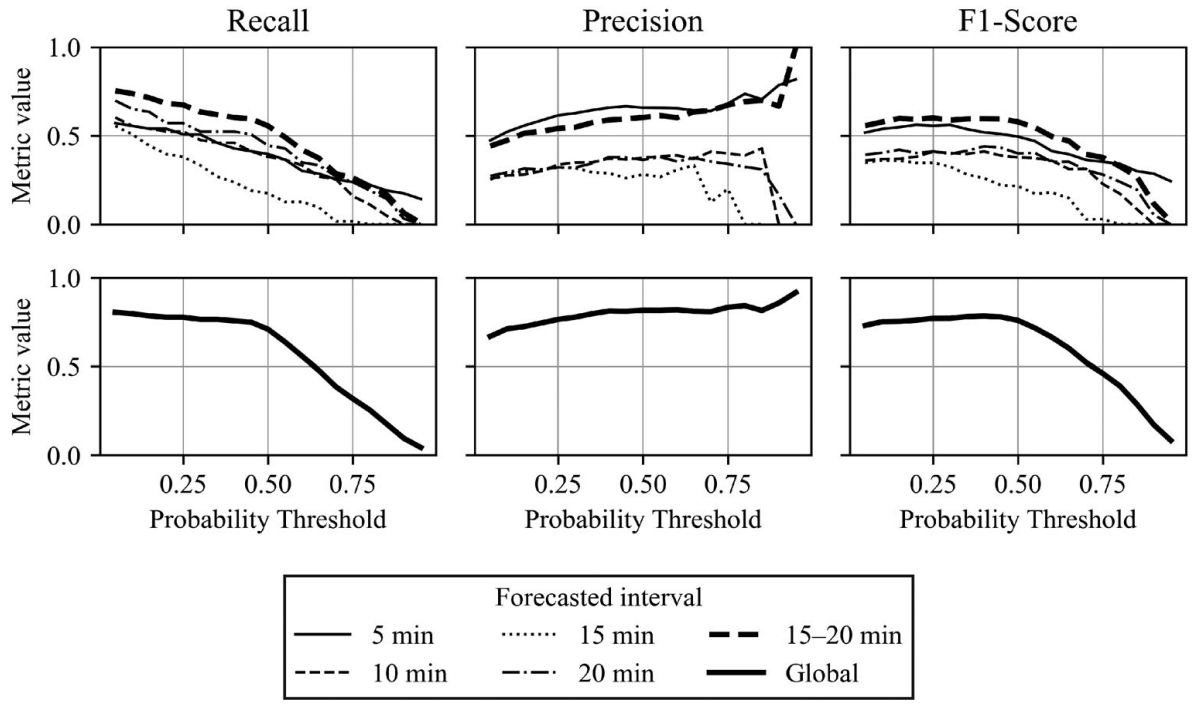

Figure 10. Breakdown forecasting evaluation using recall, precision, and F1-score.

resulting in a higher precision. To illustrate this statement, we fixed the speed threshold at 75 km/h, varied the probability threshold from 0.05 to 0.95, and calculated the recall, precision, and F1-score, as presented in Figure 10.

Based on the global F1-score, we would reach the same conclusions concerning an optimal probability threshold as obtained based on Figure 9 and choose a value close to 0.4. This result could satisfy a general purpose. However, practical applications of this methodology could prioritize recall or precision, or even focus on better predictions for a specific forecasting horizon. The richness of these results illustrates the benefits of Variational RNNs in this context, enabling the decision-maker to create much more controlled policies when compared to predictions made with neural networks without this probabilistic characteristic.

In practical implementations of this framework, breakdown probabilities can be calculated in real time based on field measurements, and the breakdown probability of any fixed moment in time can be continuously monitored. The results above indicate a higher precision for shorter forecasting horizons, which is reasonable for a highly stochastic phenomenon such as traffic breakdown. Therefore, a stronger signal indicating the occurrence or non-occurrence of a breakdown will be produced, the closer the monitored moment is to the present.

The proposed probability threshold approach allows for a concrete decision about the prediction of a breakdown event, which can benefit real-time ATM decisions. Furthermore, both speed distributions and breakdown probabilities could be incorporated into strategies such as dynamic speed limits, adaptive ramp metering, lane use control, hard shoulder running, and improved traveller information.

The position and quality of the collected field data is crucial for a successful implementation of this framework. In this study we used data from loop detectors on a single location of the studied segment. This imposes some challenges compared to studies that use data from multiple segments, mainly because upstream traffic flow is important to characterize the upcoming demand on the bottlenecks. Even so, due to the ability to learn long – and short-term rules of the LSTM, the final model could make reasonable traffic breakdown predictions. Also, a common problem when dealing with traffic data is the quality of the available information. It is not rare for loop detectors to malfunction and a lot of potentially useful information is lost, which happened in this study. Future studies could

consider approaches to aggregate missing data in the methodology (Tian et al. 2018) and test it in a broader dataset.

Concerning the maintenance of this framework, practical implementations can expect the occurrence of model drift, with degradation of the model performance over time. As for most forecasting models, this happens due to changes in traffic behaviour, in traffic demand or in the surrounding infrastructure, which is known as data drift. The quality of the predictions should be monitored over time and the model should be retrained when its performance matches a certain criterion. The cause, magnitude and frequency of model drift will vary according to each location, and it is possible that even the speed and breakdown thresholds should be recalculated. The model should also be retrained when new features are added, such as data from upstream traffic detectors or a weather station.

Conclusions

This study proposed a framework for traffic breakdown probability forecasting on a freeway segment with daily traffic breakdowns. The methodology presents an approach for calculating the breakdown probability based on sequences of forecasted speed distributions and proposes the use of a Variational LSTM neural network model to produce the forecasts. This type of neural network produces credible intervals, a fundamental characteristic for our framework that is absent in the classical implementations of neural networks that are traditionally used for traffic forecasts.

The quality of speed predictions was adequate for pre-breakdown conditions, and traffic breakdowns could be reasonably predicted for forecasting horizons up to 15 min. The paired tuning of the breakdown probability threshold using precision, recall and the F1-score produced great control over the framework results, so that the predictions can be interpreted according to each application requirement.

Although NNs have proven predictive capabilities, they are usually treated as black-box models and the abstractions captured by the models can be hard to understand. In this sense, Explainable AI (XAI) has gained space among the deep learning community by proposing methods that allow humans to visualize these abstractions in a human-friendly way. XAI is a relatively new field, and some advances have already been made for LSTMs (Arras et al. 2019). Applying these methods to transportation problems is an open field and should be considered in future studies.

Besides enhancing interpretability, different models have also been suggested for time series forecasting purposes. LSTM has the limitation of dealing solely with data spaced at regular time intervals and much information is lost during the aggregation process. To account for that, some adjustments to the traditional formulation of LSTMs have been made in other knowledge fields, for example, to create a model that makes disease predictions for patients with irregularly spaced appointments (Baytas et al. 2017) and deal with signal processing (Neil, Pfeiffer, and Liu 2016). Beyond LSTMs, Transformers (Vaswani et al. 2017), which use attention mechanisms, have been widely studied and should be considered in future studies.

The methodology presented in this study uses well-studied concepts of traditional traffic breakdown probability models to produce a model that is able to forecast traffic breakdown probabilities using deep learning. It relies on traffic and environmental data to support its training process and predictions and can be set to operate in real time if there is immediate data availability. The methodology is also very flexible, and the model produced can be tuned to meet specific site needs by adapting both speed and traffic breakdown probability thresholds. The final product is a data-driven approach that creates a flexible methodology that is suitable for real-time applications, captures modern advances made in the field of machine learning, and is also grounded on traffic engineering developments made over the last decades.

Acknowledgements

This research is supported by grants from CAPES (Coordination of Superior Level Staff Improvement), Brazil.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on contributors

Douglas Zechin Doctoral candidate and Master in Transport Systems at the Federal University of Rio Grande do Sul (UFRGS) and graduated in Civil Engineering at the same university with an emphasis on Transport Systems and Civil Structures. He did an exchange program at the Technical University of Munich (Germany) focused on Transport Systems and participated in the HyperloopTT pre-feasibility study at Serra Gaúcha, Brazil. He is experienced in traffic simulation, active traffic management, and machine learning applications in both fields.

Helena Beatriz Bettella Cybis Graduated in Civil Engineering at the Federal University of Rio Grande do Sul (1980), Master in Transport - University of Leeds (1989), Ph.D. in Transport - University of Leeds (1993), post-doctorate at the University of Berkeley (2010). She is a full professor at the Department of Production and Transport Engineering at the Federal University of Rio Grande do Sul. She was president of the National Association for Research and Education in Transport from 2017 to 2020, having been Scientific Director and President of the Scientific Committee of the Association's annual congresses from 2007 to 2008 and 2011 to 2014 and Vice-president of the areas of Traffic Engineering and Safety of the Pan American Congress of Traffic Engineering, Transport and Logistics from 2007 to 2014. She has experience in transport engineering and transport planning, acting on the following topics: traffic engineering, traffic allocation models, traffic models and traffic simulation, and the study of pedestrian behaviour.

Funding

This work was supported by CAPES: [Grant Number].

ORCID

Douglas Zechin http://orcid.org/0000-0002-0645-5415
Helena Beatriz Bettella Cybis http://orcid.org/0000-0001-8363-8154

References

Akhtar, M., and S. Moridpour. 2021. "A Review of Traffic Congestion Prediction Using Artificial Intelligence." *Journal of Advanced Transportation* 2021. doi:10.1155/2021/8878011.

Arras, L., et al. 2019. "Explaining and Interpreting LSTMs." In: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* 211–238. doi:10.1007/978-3-030-28954-6_11.

Asgharzadeh, M., and A. Kondyli. 2020. "Effect of Geometry and Control on the Probability of Breakdown and Capacity at Freeway Merges." *Journal of Transportation Engineering, Part A: Systems* 146 (7): 1–11. doi:10.1061/JTEPBS.0000381.

Baytas, I. M., et al. 2017. Patient Subtyping via Time-Aware LSTM Networks, *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Part F1296, pp. 65–74. doi:10.1145/3097983.3097997.

Brilon, W., J. Geistefeldt, and M. Regler. 2005. "Reliability of Freeway Traffic Flow: A Stochastic Concept of Capacity." *Proceedings of the 16th International Symposium on Transportation and Traffic Theory* July: 125–144.

Caleffi, F., Y. Moisan, and H. B. B. Cybis. 2016. "Analysis of an Active Traffic Management System Proposed for A Brazilian Highway." *International Journal of Emerging Technology and Advanced Engineering* 6 (4): 10–17.

Chen, D., and S. Ahn. 2018. "Capacity-Drop at Extended Bottlenecks: Merge, Diverge, and Weave." *Transportation Research Part B: Methodological* 108: 1–20. doi:10.1016/j.trb.2017.12.006.

Chollet, F. 2015. *Deep Learning with Python*. 2nd ed. Shelter Island: Manning Publications Co.

Cui, Z., R. Ke, and Y. Wang. 2018. Deep Bidirectional and Unidirectional LSTM Recurrent Neural Network for Network-wide Traffic Speed Prediction, pp. 1–11. http://arxiv.org/abs/1801.02143.

Elefteriadou, L., et al. 2011. "Proactive Ramp Management Under the Threat of Freeway-Flow Breakdown." *Procedia - Social and Behavioral Sciences* 16: 4–14. doi:10.1016/j.sbspro.2011.04.424.

Elefteriadou, L., R. P. Roess, and W. R. McShane. 1995. "Probabilistic Nature of Breakdown at Freeway Merge Junctions." *Transportation Research Record* 1484 (1484): 80–89.

Fang, Z. G., et al. 2022. "Application of a Data-Driven XGBoost Model for the Prediction of COVID-19 in the USA: A Time-Series Study." *BMJ Open* 12 (7): e056685–8. doi:10.1136/bmjopen-2021-056685.

Fortunato, M., C. Blundell, and O. Vinyals. 2017. Bayesian Recurrent Neural Networks, pp. 1–14. doi:10.48550/arXiv.1704.02798.

Gal, Y., and Z. Ghahramani. 2016a. "A Theoretically Grounded Application of Dropout in Recurrent Neural Networks." *Advances in Neural Information Processing Systems*, 1027–1035. doi:10.48550/arXiv.1512.05287.

Gal, Y., and Z. Ghahramani. 2016b. "Dropout as A Bayesian Approximation: Representing Model Uncertainty in Deep Learning." *33rd International Conference on Machine Learning, ICML 2016* 3: 1651–1660.
Géron, A. 2019. *Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow*. 2nd ed. Sebastopol: O'Reilly Media.
Goodfellow, I., Y. Bengio, and A. Courville. 2016. *Deep Learning, CrossRef Listing of Deleted DOIs*. Cambridge, MA: The MIT Press. doi:10.2172/1462436.
Graves, A. 2011. Practical Variational Inference for Neural Networks, in *NIPS 11: Proceedings of the 24th International Conference on Neural Information Processing*, pp. 2348–2356.
Gu, Y., et al. 2019. "Short-Term Prediction of Lane-Level Traffic Speeds: A Fusion Deep Learning Model." *Transportation Research Part C: Emerging Technologies* 106 (July): 1–16. doi:10.1016/j.trc.2019.07.003.
Guo, H., W. Wang, and H. Bubb. 2013. Modeling of Traffic Behavior in Traffic Safety Using A Reliability Approach, pp. 243–265. doi:10.2991/978-94-91216-80-0_13.
Hochreiter, S. 1997. "Long Short-Term Memory." *Neural Computation*, 1735–1780. doi:10.1162/neco.1997.9.8.1735.
Ke, R., et al. 2018. "New Framework for Automatic Identification and Quantification of Freeway Bottlenecks Based on Wavelet Analysis." *Journal of Transportation Engineering, Part A: Systems* 144 (9): 4018044. doi:10.1061/JTEPBS.0000168
Kerner, B. S. 2019. *Breakdown in Traffic Networks*. Berlin: Springer Nature. doi:10.1007/978-3-662-54473-0.
Kidando, E., R. Moses, and T. Sando. 2019. "Bayesian Regression Approach to Estimate Speed Threshold Under Uncertainty for Traffic Breakdown Event Identification." *Journal of Transportation Engineering, Part A: Systems* 145 (5), doi:10.1061/JTEPBS.0000217.
Kondyli, A., et al. 2013. "Development and Evaluation of Methods for Constructing Breakdown Probability Models." *Journal of Transportation Engineering* 139 (September): 931–940. doi:10.1061/(ASCE)TE.1943-5436.0000574.
Laptev, N., et al. 2017. "Time-series Extreme Event Forecasting with Neural Networks at Uber." *International Conference on Machine Learning - Time Series Workshop*, 1–5. http://www.cs.columbia.edu/ lierranli/publications/TSW2017_paper.pdf.
Li, L., et al. 2018. "Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization." *Journal of Machine Learning Research* 18: 1–52. doi:10.48550/arXiv.1603.06560.
Li, T., et al. 2020. "Short-Term Traffic Congestion Prediction with Conv–BiLSTM Considering Spatio-Temporal Features." *IET Intelligent Transport Systems* 14 (14): 1978–1986. doi:10.1049/iet-its.2020.0406.
Li, C., et al. 2022. "Survival Analysis of the Likelihood and Duration of Traffic Flow Breakdown at Freeway Merge Bottlenecks." International Conference on Transportation and Development 2022 70–81. doi:10.1061/9780784484333.007.
Li, P., M. Abdel-Aty, and J. Yuan. 2020. "Real-Time Crash Risk Prediction on Arterials Based on LSTM-CNN." *Accident Analysis & Prevention* 135 (July 2019): 105371. doi:10.1016/j.aap.2019.105371.
Lu, C., and L. Elefteriadou. 2013. "An Investigation of Freeway Capacity Before and During Incidents." *Transportation Letters* 5 (3): 144–153. doi:10.1179/1942786713Z.00000000016.
Luo, Q., and Y. Zhou. 2021. Spatial-Temporal Structures of Deep Learning Models for Traffic Flow Forecasting: A Survey, pp. 187–193s. doi:10.1109/icoias53694.2021.00041.
Ma, Y., Z. Zhang, and A. Ihler. 2020. "Multi-Lane Short-Term Traffic Forecasting with Convolutional LSTM Network." *IEEE Access* 8: 34629–34643. doi:10.1109/ACCESS.2020.2974575.
Neil, D., M. Pfeiffer, and S. C. Liu. 2016. "Phased LSTM: Accelerating Recurrent Network Training for Long or Event-Based Sequences." *Advances in Neural Information Processing Systems, (Nips)*, 3889–3897. doi:10.48550/arXiv.1610.09513.
Persaud, B., S. Yagar, and R. Brownlee. 1998. "Exploration of the Breakdown Phenomenon in Freeway Traffic." *Transportation Research Record: Journal of the Transportation Research Board* 1634: 64–69. doi:10.3141/1634-08.
Qu, X., J. Zhang, and S. Wang. 2017. "On the Stochastic Fundamental Diagram for Freeway Traffic: Model Development, Analytical Properties, Validation, and Extensive Applications." *Transportation Research Part B: Methodological* 104: 256–271. doi:10.1016/j.trb.2017.07.003.
Srivastava, N., et al. 2014. "Dropout: A Simple Way to Prevent Neural Networks from Overfittin." *The Journal of Ma- Chine Learning Research* 15 (1), doi:10.5555/2627435.2670313.
Tian, Y., et al. 2018. "LSTM-based Traffic Flow Prediction with Missing Data." *Neurocomputing* 318: 297–305. doi:10.1016/j.neucom.2018.08.067.
Vaswani, A., et al. 2017. "Attention is all you need." *Advances in Neural Information Processing Systems* 2017-Decem (Nips): 5999–6009.
Vlahogianni, E. I., M. G. Karlaftis, and J. C. Golias. 2014. "Short-Term Traffic Forecasting: Where We are and Where We're Going." *Transportation Research Part C: Emerging Technologies* 43: 3–19. doi:10.1016/j.trc.2014.01.005.
Xia, D., et al. 2021. "A Parallel NAW-DBLSTM Algorithm on Spark for Traffic Flow Forecasting." *Neural Computing and Applications* 34), doi:10.1007/s00521-021-06409-5.
Yang, B., et al. 2019. "Traffic Flow Prediction Using LSTM with Feature Enhancement." *Neurocomputing* 332: 320–327. doi:10.1016/j.neucom.2018.12.016.
Zechin, D., F. Caleffi, and H. B. B. Cybis. 2020. "Influence of Rain on Highway Breakdown Probability." *Transportation Research Record: Journal of the Transportation Research Board* 2674: 687. doi:10.1177/0361198120919754.

5 COMPLEMENTARY MATERIALS

This chapter aims to present complementary content that did not fit the articles but also brings important contributions. In the following sections we present charts that better depict the predictions, the code used to generate the proposed Variational LSTM model, and the limitations and recommendations for future studies.

5.1 Speed forecasting (Third article)

This section presents some speed forecasting analyses made during the comparison between our model and the benchmarks that did not fit the third article but produced a better understanding. **Figure 1**, for example, presents the error distribution of all the compared models. We can observe that our model (Enc-Dev Var. LSTM) has a higher kurtosis than the benchmarks and a distribution well centered in the origin.

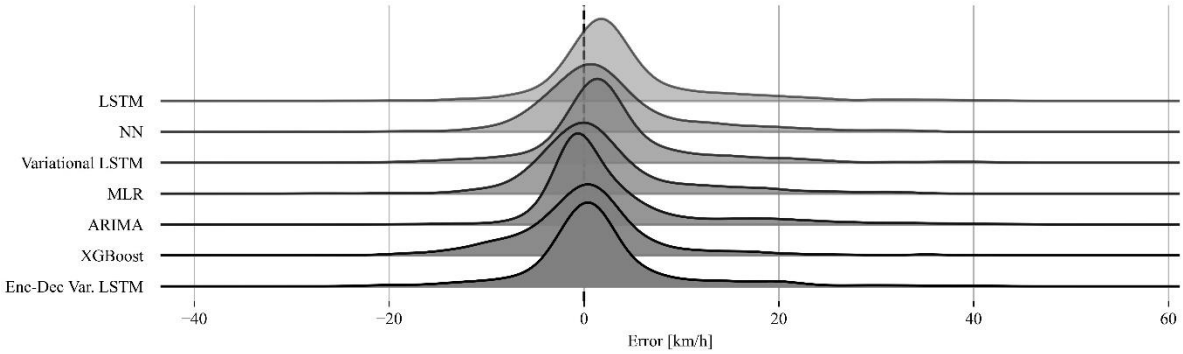

Figure 1: Ridgeplot of the speed forecasting error distribution of the proposed model and the benchmarks

Another interesting visualization for comparing the models is plotting the speed profile over time for each. In **Figure 2** we plotted the ground truth speed values in black and each predicted sequence with a colourful line starting with a circular marker. The background colours indicated the current region of the fundamental diagram, as suggested in the articles. This figure shows how much better the proposed model performs compared to the benchmarks, especially during the breakdown. The performance is similar during high-speed (free flow) time intervals, as traffic speeds have small variability. All models fail to anticipate the end of the congestions (region 4, red background) due to the lack of information regarding upstream traffic. Our model had a better performance during this period, however this discussion was out of the scope of this study. Future studies should specifically consider exploring this topic.

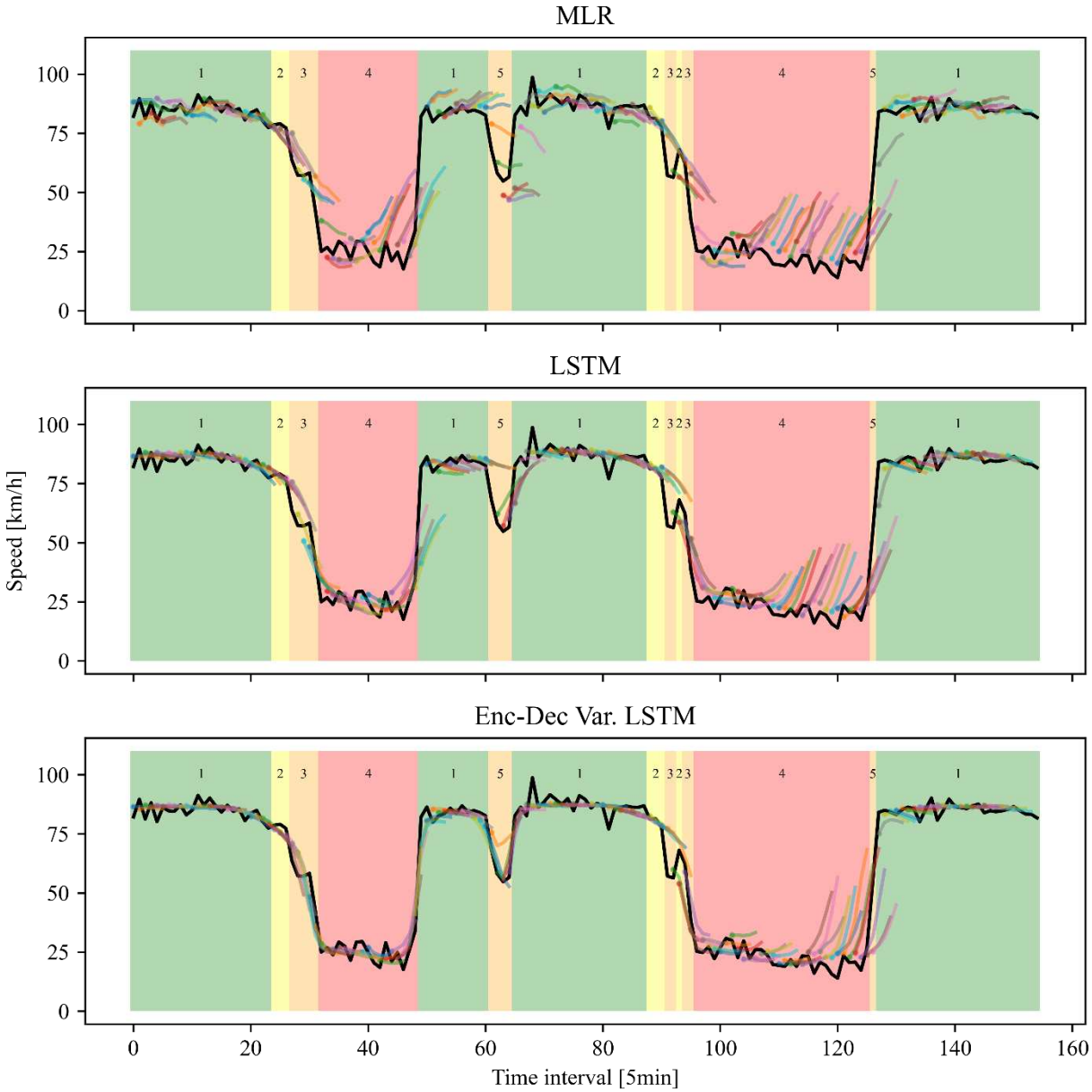

Figure 2: Comparison of speed forecasts between our model (last) and other benchmarks. The black line represents the ground truth, and each colourful line represents a sequence of speed forecasts. The background colours and numbers represent the regions of the fundamental diagram used in the articles.

5.2 Code for the Variational LSTM model (Third article)

We developed the code for the Variational LSTM model in Python using the Tensorflow/Keras library. We trained the model using the Google Colab platform, which, as the writing of this thesis, offers good computational power with GPUs for free. The dataset was also made public (Zechin, 2022). The code for the neural network is as follows:

```python
# LIBRARIES IMPORTING ---------------------------------------------------------
import numpy as np
import tensorflow as tf
from os.path import join
from tensorflow.keras.layers import Dense, LSTM, Input, RepeatVector, Bidirectional
from tqdm import tqdm


# DATA LOADING ----------------------------------------------------------------
folder = "path/to/folder/containg/data/"

X_train = np.load(join(folder, "X_train.npy"))
Y_train = np.load(join(folder, "Y_train.npy"))
vol_train = np.load(join(folder, "vol_train.npy"))
X_test = np.load(join(folder, "X_test.npy"))
Y_test = np.load(join(folder, "Y_test.npy"))
vol_test = np.load(join(folder, "vol_test.npy"))
u = np.load(join(folder, "u.npy"))
s = np.load(join(folder, "s.npy"))
dias_train = np.load(join(folder, "days_train.npy"))
dias_test = np.load(join(folder, "days_test.npy"))

# Denormalize data
X_test_desnorm = X_test * s + u
Spd_test_desnorm = (
    X_test_desnorm[:, -1, 0] * X_test_desnorm[:, -1, 15]
    + X_test_desnorm[:, -1, 1] * X_test_desnorm[:, -1, 16]
    + X_test_desnorm[:, -1, 2] * X_test_desnorm[:, -1, 17]
) / (X_test_desnorm[:, -1, 15] + X_test_desnorm[:, -1, 16] + X_test_desnorm[:, -1, 17])
Vol_test_desnorm = (
    X_test_desnorm[:, -1, 15] + X_test_desnorm[:, -1, 16] + X_test_desnorm[:, -1, 17]
)

# Create the ouput for the decoder
Y_train_decoder = Y_train[:, -1, :].reshape(Y_train.shape[0], Y_train.shape[2], 1)
Y_test_decoder = Y_test[:, -1, :].reshape(Y_test.shape[0], Y_test.shape[2], 1)
vol_train_decoder = vol_train[:, -1, :].reshape(
    vol_train.shape[0], vol_train.shape[2], 1
)
vol_test_decoder = vol_test[:, -1, :].reshape(Y_test.shape[0], vol_test.shape[2], 1)

# NEURAL NETWORK DEFINITION ---------------------------------------------------

# Constants
inp_shape = (X_train.shape[1], X_train.shape[2])
dropout_rate = 0.15
weight_decay = 1e-5
latent_dim = 256
```

```
window_len = X_train.shape[1]
n_total_features = X_train.shape[2]
forecast = Y_train.shape[2]
n_deterministic_features = X_train.shape[2]


# Encoder
encoder = Input(shape=(window_len, n_total_features), name="Input")
encoder1 = Bidirectional(
    LSTM(512, dropout=dropout_rate, return_sequences=True, name="Encoder_LSTM_1")
)(encoder, training=True)
encoder2 = Bidirectional(
    LSTM(512, dropout=dropout_rate, return_sequences=True, name="Encoder_LSTM_2")
)(encoder1, training=True)
encoder3 = Bidirectional(
    LSTM(512, dropout=dropout_rate, return_sequences=True, name="Encoder_LSTM_3")
)(encoder1, training=True)
encoder_outputs, state_h, state_c = LSTM(
    256,
    dropout=dropout_rate,
    return_state=True,
    return_sequences=True,
    name="Encoder_LSTM_4",
)(encoder3, training=True)
encoder_out = Dense(forecast, activation="linear", name="Output_encoder")(
    encoder_outputs
)

# Decoder
decoder = RepeatVector(forecast, name="Repeat_Vector")(encoder_outputs[:, -1, :])
decoder1 = LSTM(
    256, dropout=dropout_rate, return_sequences=True, name="Decoder_LSTM_1"
)(decoder, initial_state=[state_h, state_c], training=True)
decoder2 = tf.keras.layers.Dense(1, activation="linear", name="Output_decoder")(
    decoder1
)

# Final model definition
model = tf.keras.models.Model(inputs=encoder, outputs=[encoder_out, decoder2])

# NEURAL NETWORK COMPILING ----------------------------------------------------------------

# Function for calculating the last time step MAE
def last_time_step_mae(Y_true, Y_pred):
    return tf.keras.metrics.mean_absolute_error(Y_true[:, -1, :], Y_pred[:, -1, :])


# Compiling the model
optimizer = tf.keras.optimizers.RMSprop(10e-2)
```

```
loss = tf.keras.losses.MeanSquaredError()
model.compile(
    loss=[loss, loss],
    optimizer=optimizer,
    metrics={"Output_encoder": last_time_step_mae, "Output_decoder": "mae"},
    loss_weights=[1, 3],
)

# NEURAL NETWORK TRAINING ----------------------------------------------------------------

# Function for decaying the learning rate and callback to use it
def decay(epoch):
    if epoch < 40:
        return 0.004
    elif epoch < 60:
        return 0.0004
    else:
        return 0.00004


call_decay = tf.keras.callbacks.LearningRateScheduler(decay)

# Callback to save the best model
call_check = tf.keras.callbacks.ModelCheckpoint(
    filepath="Variational_LSTM",
    monitor="val_Output_decoder_mae",
    verbose=0,
    save_best_only=True,
    save_weights_only=True,
    mode="auto",
    save_freq="epoch",
)

# Sample weights using the volume
sample_weight = np.roll(vol_train[:, -1, 0], 2)

# Training the model
r = model.fit(
    x=X_train,
    y=[Y_train, Y_train_decoder],
    epochs=100,
    verbose=1,
    validation_data=(X_test, [Y_test, Y_test_decoder]),
    shuffle=False,
    batch_size=512,
    sample_weight=sample_weight,
    callbacks=[call_check, call_decay],
)  #
```

```
# INFERENCE ------------------------------------------------------------

n_inference = 1000  # Number of inference samples
pred = []

for _ in tqdm(range(n_inference)):
    pred.append(model.predict(X_test)[1])

pred = np.asarray(pred)

# Save the predictions
with open("Y_pred_new.npy", "wb") as f:
    np.save(f, pred)
```

5.3 Complementary discussions

Due to the format of this thesis some discussions did not fit the articles but are worth mentioning and therefore presented in this section.

5.3.1 *Speed profile smoothing*

As observed in Figure 2, the speed forecasting model fits the speed profile but tends to smooth its oscillations. We understand this has two main reasons. First, we only use data from a single traffic detector, so the predictions are entirely made based on the historical trends learned during training (weights and biases) and the traffic data of the previous time steps (inputs). Traffic data of an additional upstream detector could provide more information on the future traffic state and help better represent speed oscillations.

Secondly, during training, the model learns by minimizing prediction errors. Since the dataset has a high variance due to the traffic's stochastic nature and information suppression due to the aggregation, the model converges to a more conservative approach that prioritizes representing the main trends and, on average, minimizes the error.

5.3.2 *Forecasting horizon*

In this study, we used a forecasting horizon of up to 25min. We understand this horizon is adequate compared to the present literature, where we mainly observe 1-10min forecasting horizons for studies that deal with short-term traffic predictions (Akhtar e Moridpour, 2021; Vlahogianni, Karlaftis e Golias, 2014).

From a practical perspective, the forecasting horizon will depend on the applications of the presented methodology. For example, active traffic management strategies used in a highway breakdown context usually aim to postpone the occurrence of the breakdown as much as possible. In this context, the 25min horizon seems adequate, and the predictions could support strategies such as variable speed limits and ramp metering.

Finally, the 25min horizon used in this study was chosen due to its fit with the literature and a reasonable idea of possible applications of the methodology but is by no means fixed. The quality of the predictions might vary according to the quality of the data and traffic and infrastructure characteristics. Our methodology accounts for the flexibility of neural networks, so this value should be changed as needed.

5.3.3 *Free-flow speed recovery*

As depicted in Figure 2, the speed forecasting model can not produce good predictions regarding the end of congestions and the recovery of the free-flow speed. This topic is not related to our primary goals but is worth mentioning.

During the speed forecasting model training, we used sample weighting to increase the relative importance of traffic breakdown periods. As a side effect, this approach penalizes other periods by decreasing their importance during training. Although this approach can contribute to reducing the performance during free-flow recovery, our model steel outperforms the benchmarks.

Therefore, we mainly relate the insufficient prediction quality during free-flow recovery to the lack of upstream data. As depicted in Figure 4 of the Introduction, the moment when the recovery happens has a much greater variance than the traffic breakdown. The lacking information regarding upcoming flow is determinant for this case since it will indicate if there is enough clearance between vehicles for the average speed to be higher.

This topic is an interesting follow-up of this thesis and should be considered for future studies, as mentioned in the following section, along with other suggestions.

5.4 **Limitations and recommendations for future studies**

This study uses traffic data collected in a highway segment with daily congestion due to a bottleneck. This is an optimum scenario for this study since traffic breakdowns happen approximately in the same period every day, resulting in a good and consistent traffic sample. The dataset was disaggregated so that each line of the dataset is a passing vehicle with its

characteristics. It enabled us to aggregate the data most adequately and produce features such as the speed variance, which could not be achieved with already aggregated data. This might be a limitation for future studies that aim at replicating our methodology since disaggregated data is seldom stored by traffic agencies.

The traffic dataset comprises a single segment of the studied highway. Due to that, the forecasting model has no information regarding upcoming traffic, which could improve its performance. We highly recommend that future studies explore the usage of upstream and downstream traffic data and measure the benefits of using and not using them. Predictions of the end of congestion and not only their beginning should also be explored. In our study, the detectors were positioned slightly upstream of the active bottleneck. Future studies could study the effects of the position of the sensors on the model performance.

This study also accounts for rain data obtained from a rain gauge close to the traffic detectors. In its raw format, the rain data was aggregated in a minimum interval of 10 min and maximum of 60 min, according to the measured rain intensity. We consider this a good-quality weather dataset since most are aggregated hourly or daily. The proximity of the rain gauge and the data frequency were beneficial for our study and might impose limitations on future studies.

We chose the Variational LSTM model for this study since it accounts for the probabilistic approach, rare among other neural network models. Also, LSTMs have strong literature support as one of the best-performing models in traffic forecasting. Besides that, we also proposed improvements over the traditional model, such as using the encoder-decoder architecture and sample weighting to improve the relative importance of high-demand periods. To test the quality of the proposed model, we compared its speed predictions with predictions made with other benchmark models. As a recommendation for future studies, we suggest searching for other forecasting models with probabilistic characteristics and comparing their performance with the performance of the proposed model. This was not done in our study since our main contribution is the proposed method; however, it is an interesting topic to study further.

An interesting topic that could also be addressed in future works is feature importance. The machine learning area called Explainable Artifical Intelligence has made good progress by proposing ways of producing reasoning and explainability for neural networks. The lack of these features has imposed resistance in their adoption, which is increasingly starting to loosen. This also comes with the usage of data from different sources, such as users reported, Bluetooth

and mobile devices data and metadata, which could be used to enrichen forecasting models and produce more insights on explainability.

6 CONCLUSIONS

This thesis proposes a framework for probabilistic traffic breakdown forecasting in a highway segment. The study comprises three articles and is structured in six chapters: an Introduction, a chapter for each article, a chapter with Complementary Materials and this Conclusion.

The **first article**, entitled *Influence of Rain on Highway Breakdown Probability,* analyses the probability of traffic breakdown on a freeway under different climatic conditions. The speed threshold was found to be the most suitable methodology for breakdown identification. We used the Weibull distribution to generate breakdown probability curves with traffic data gathered under different rain intensities. We found that the breakdown probability is significantly higher during rainfall and can increase up to 50% when the traffic flow is close to capacity. The results suggest that this methodology could be used to improve existing traffic management strategies. Recommendations for future studies include using a larger data set and accounting for the time of day.

The **second article**, entitled Forecast of Traffic Speeds with Neural Network LSTM Encoder-Decoder, used an LSTM neural network to perform speed predictions on a road segment with daily congestions using rainfall and traffic data. The model had satisfactory results with an MAE of 5.4 km/h for all predicted intervals. Using volumes as sample weight helped reduce prediction errors when traffic was close to capacity. Comparing the breakdown probability curves obtained from the predictions and field data showed that the model captured the transition from free flow to congested traffic.

The **third article**, entitled *Probabilistic Traffic Breakdown Forecasting through Bayesian Approximation Using Variational LSTMs*, is a sequence of the previous ones and benefits from their methodology, results, and conclusions. In this study, we proposed a methodology for traffic breakdown probability calculation based on probabilistic speed forecasts. For that, we developed a machine learning model that uses Dropout to approximate a Bayesian inference and produce probabilistic outputs. The proposed model had better speed forecasting performance during high-demand periods when compared to baseline models. The produced breakdown probability forecasts enable a level of control over highway operations that could be achieved using deterministic forecasts. Besides its theoretical contributions, this methodology could also be used in practical applications to improve the effectiveness of traffic management strategies.

This thesis fulfils a literature gap related to probabilistic traffic breakdown forecasting by developing a traffic breakdown probability calculation methodology using speed forecasts. The probabilistic forecasts were possible due to recent contributions in the Probabilistic Machine Learning area. We hope this study will positively influence future works regarding the absorption of concepts from both the Probabilistic Machine Learning and Explainable Artificial Intelligence areas so that the Traffic Engineering community sustainably and confidently incorporates robust and more general models.

REFERENCES

AKHTAR, M.; MORIDPOUR, S. A Review of Traffic Congestion Prediction Using Artificial Intelligence. **Journal of Advanced Transportation**, v. 2021, 2021.

BERTAUD, A. **Order Without Design: How Markets Shape Cities**. 1st. ed. Cambridge, MA: MIT Press, 2018.

BRILON, W.; GEISTEFELDT, J.; REGLER, M. Reliability of Freeway Traffic Flow: A stochastic Concept of Capacity. **Proceedings of the 16th International Symposium on Transportation and Traffic Theory**, n. July, p. 125–144, 2005.

CALEFFI, F. *et al.* Influência das condições climáticas e de acidentes na caracterização do comportamento do tráfego em rodovias. **Transportes**, v. 24, n. 4, p. 57, 2016.

___. **Proposição de um método de harmonização da velocidade baseado em modelo de previsão de conflitos veiculares**. [s.l.] Universidade Federal do Rio Grande do Sul, 2018.

CALEFFI, F.; MOISAN, Y.; CYBIS, H. B. B. Analysis of an Active Traffic Management System Proposed for A Brazilian Highway. **International Journal of Emerging Technology and Advanced Engineering**, v. 6, n. 4, p. 10–17, 2016.

CHAUDHARY, N. *et al.* Ramp metering algorithms and approaches for Texas. **FHWA/TX-05/0-4629-1, Technical Report 0-4629-1**, v. 7, n. 2, 2004.

CHEN, D.; AHN, S. Capacity-drop at extended bottlenecks: Merge, diverge, and weave. **Transportation Research Part B: Methodological**, v. 108, p. 1–20, 2018.

CYBIS, H. B. B. *et al.* **Concepção de Sistema de Gerenciamento Ativo de Tráfego**. [s.l: s.n.]. Disponível em: <https://www.gov.br/antt/pt-br/assuntos/rodovias/concessionarias/lista-de-concessoes/concepa/relatorios/relatorios-de-pesquisa-rdt/concepcao-de-sistema-de-gerenciamento-ativo-de-trafego.pdf/view>.

ELEFTERIADOU, L.; ROESS, R. P.; MCSHANE, W. R. Probabilistic nature of breakdown at freeway merge junctions. **Transportation Research Record**, v. 1484, n. 1484, p. 80–89, 1995.

FIRJAN. O Custo Dos Deslocamentos Nas Principais Áreas Urbanas Do Brasil. p. 1–6, 2015.

FORTUNATO, M.; BLUNDELL, C.; VINYALS, O. Bayesian Recurrent Neural

Networks. p. 1–14, 2017.

GAL, Y.; GHAHRAMANI, Z. A theoretically grounded application of dropout in recurrent neural networks. **Advances in Neural Information Processing Systems**, p. 1027–1035, 2016a.

___. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. **33rd International Conference on Machine Learning, ICML 2016**, v. 3, p. 1651–1660, 2016b.

GALVAN, Y. T.; ZECHIN, D.; CYBIS, H. B. B. **Utilização do método kaplan-meier na estimativa das distribuições de velocidade desejada**Anais do congresso ANPET. **Anais**...Camboriú: 2019

KAPPLER, L. B. Análise dos benefícios da implantação da segunda ponte do guaíba. 2017.

KERNER, B. S. **Breakdown in Traffic Networks**. [s.l: s.n.].

KONDYLI, A. *et al.* Development and Evaluation of Methods for Constructing Breakdown Probability Models. **American Society of Civil Engineers**, v. 139, n. September, p. 931–940, 2013.

PERSAUD, B.; YAGAR, S.; BROWNLEE, R. Exploration of the Breakdown Phenomenon in Freeway Traffic. **Transportation Research Record: Journal of the Transportation Research Board**, v. 1634, p. 64–69, 1998.

QU, X.; ZHANG, J.; WANG, S. On the stochastic fundamental diagram for freeway traffic: Model development, analytical properties, validation, and extensive applications. **Transportation Research Part B: Methodological**, v. 104, p. 256–271, 2017.

TRB. Highway Capacity Manual. **Transportation Research Board, National Research Council, Washington, D.C.**, 2016.

VLAHOGIANNI, E. I.; KARLAFTIS, M. G.; GOLIAS, J. C. Short-term traffic forecasting: Where we are and where we're going. **Transportation Research Part C: Emerging Technologies**, v. 43, p. 3–19, 2014.

ZECHIN, D. **Traffic speed forcasting dataset**, 2022. Disponível em: <https://doi.org/10.6084/m9.figshare.16574390.v1>

ZECHIN, D.; CYBIS, H. B. B.; CALEFFI, F. Avaliação de Ramp Metering na BR-

290/RS utilizando o algoritmo ALINEA. **XXX ANPET**, 2016.