UNIVERSIDADE DE SÃO PAULO
Escola de Engenharia de São Carlos
Departamento de Engenharia de Transportes
Programa de Pós-Graduação em Engenharia de Transportes

Helena Stein Stefani

Método de previsão de fluxo de tráfego em rodovias urbanas a partir de dados de velocidade de fontes *online*

São Carlos
2021

Helena Stein Stefani

Método de previsão de fluxo de tráfego em rodovias urbanas a partir de dados de velocidade de fontes *online*

Programa de Pós-graduação em
Engenharia de Transportes da EESC-USP

Exemplar definitivo (corrigido). O exemplar original está disponível na CPG da EESC-USP

São Carlos, 04/10/2021

RESOLUÇÃO CoPGr Nº 6018, DE 13 DE OUTUBRO DE 2011, artigo 5º

Dissertação de Mestrado submetida à Escola de Engenharia de São Carlos, da Universidade de São Paulo, como parte dos requisitos para a obtenção do título de *Mestre em Ciências*, Programa de Pós-Graduação em Engenharia de Transportes. Área de Concentração: Planejamento e Operação de Sistemas de Transportes

Orientador: Prof. Dr. André Luiz Barbosa Nunes da Cunha

São Carlos

2021

AUTORIZO A REPRODUÇÃO TOTAL OU PARCIAL DESTE TRABALHO, POR QUALQUER MEIO CONVENCIONAL OU ELETRÔNICO, PARA FINS DE ESTUDO E PESQUISA, DESDE QUE CITADA A FONTE.

Ficha catalográfica elaborada pela Biblioteca Prof. Dr. Sérgio Rodrigues Fontes da EESC/USP com os dados inseridos pelo(a) autor(a).

Stefani, Helena Stein
S816m Método de previsão de fluxo de tráfego em rodovias urbanas a partir de dados de velocidade de fontes online / Helena Stein Stefani; orientador André Luiz Barbosa Nunes da Cunha. São Carlos, 2021.

Dissertação (Mestrado) - Programa de Pós-Graduação em Engenharia de Transportes e Área de Concentração em Planejamento e Operação de Sistemas de Transporte -- Escola de Engenharia de São Carlos da Universidade de São Paulo, 2021.

1. Previsão de fluxo veicular. 2. Inferência bayesiana. 3. Fontes de dados online. I. Título.

Eduardo Graziosi Silva - CRB - 8/8907

FOLHA DE JULGAMENTO

Candidata: Bacharela **HELENA STEIN STEFANI.**

Título da dissertação: "Método de previsão de fluxo de tráfego em rodovias urbanas a partir de dados de velocidade de fontes online".

Data da defesa: 05/08/2021.

Comissão Julgadora	**Resultado**
Prof. Dr. **André Luiz Barbosa Nunes da Cunha (Orientador)** (Escola de Engenharia de São Carlos – EESC/USP)	APROVADA
Prof. Dr. **José Elievam Bessa Júnior** (Universidade Federal de Minas Gerais/UFMG)	APROVADA
Prof. Dr. **Manoel Mendonça de Castro Neto** (Universidade Federal do Ceará/UFC)	APROVADA

Coordenador do Programa de Pós-Graduação em Engenharia de Transportes:
Prof. Associado **Adalberto Leandro Faxina**

Presidente da Comissão de Pós-Graduação:
Prof. Titular **Murilo Araujo Romero**

Agradecimentos

Agradeço aos meus pais, Joice e Fábio, que sempre me apoiaram e me deram condições pra que eu pudesse concretizar meus desejos e deveres. Agradeço também ao meu irmão, Adriano, por ser meu maior exemplo de paciência e ponderação.

Agradeço ao Departamento de Engenharia de Transportes e à Universidade de São Paulo, por terem confiado na minha capacidade e fornecido toda a infraestrutura necessária à realização deste trabalho, e aos professores, por terem disponibilizado seu tempo, conhecimento, e experiências. Ao CNPq e à CAPES, pela bolsa de estudos concedida.

Agradeço ao meu orientador, André Luiz Barbosa Nunes da Cunha, que me acolheu desde o início, por ter me delegado o desafio de trabalhar com um tema que não apenas ampliou meu conhecimento técnico, mas também me ajudou a acreditar muito mais em mim mesma.

Agradeço aos meus colegas do grupo de pesquisa, Alceu, André, Elaine, Gabriel, Leandro e Paola, pela parceria e trocas de ideias. Em especial, agradeço ao André por ter me ajudado com os primeiros códigos em *Python* e por ter voluntariamente proposto alternativas quando eu não encontrava soluções para a pesquisa. Obrigada pelas noites de terça-feira colonizando territórios imaginários durante as partidas de *Catan*.

Agradeço aos colegas com quem dividi os primeiros perrengues e *coffee-breaks* do mestrado: Breno, Luíza Rijo, Luiza Gagno, Marina, Renan, Gustavo e Felipe. Agradeço também a cada um dos meus colegas do STT, com quem dividi vários e vários e vários cafés, por terem contribuído para tornar meus dias mais divertidos.

Agradeço em especial à Luíza Rijo, com quem compartilhei muitas histórias, *stories*, e o por do sol alaranjado da janela da cozinha, e ao Douglas, meu grande amor, conselheiro e

companheiro de vida, que não contente em ser minha fortaleza, assumiu nas horas vagas as atribuições de técnico de informática e coorientador-não-oficial.

Por fim, e principalmente, agradeço a Deus, que sempre esteve por perto, mesmo quando eu caminhava de olhos fechados.

"Terei que correr o sagrado risco do acaso.
E substituirei o destino pela probabilidade."
(Clarice Lispector)

Resumo

Stefani, H. S. **Método de previsão de fluxo de tráfego em rodovias urbanas a partir de dados de velocidade de fontes *online***. 100 p. Dissertação de Mestrado – Escola de Engenharia de São Carlos, Universidade de São Paulo, 2021.

Previsão de fluxo veicular. Inferência bayesiana. Fontes de dados online

A previsão de fluxo veicular é uma ferramenta útil para avaliar as condições de tráfego em uma rodovia e pode ser utilizada para promover maior eficiência no planejamento de sistemas de transportes. A geração contínua de dados de tráfego por sistemas tradicionais de monitoramento no decorrer de longos períodos resultou em grandes volumes de dados históricos, o que fomentou o desenvolvimento de modelos de previsão de fluxo de tráfego. Atualmente, esses modelos buscam atingir altos níveis de precisão, e apesar de a acurácia ser um indicador difundido para aferir a qualidade das previsões, ela não considera a variabilidade desses valores. Nesse contexto, uma ferramenta que permite obter não apenas valores pontuais, mas também intervalos de credibilidade, é a estatística *bayesiana*. Em paralelo a isso, empresas especializadas em serviços de Internet passaram a fornecer informações de tráfego *online* em tempo real com base no deslocamento de seus usuários. Um desses serviços é a *Google Distance Matrix API*, que fornece tempos de viagem em tempo real. Essas fontes de dados têm o potencial de complementar as análises de tráfego, principalmente quando os sistemas tradicionais de coleta falham ou são inexistentes. Nesse sentido, o objetivo deste trabalho foi propor um método de previsão de fluxo de tráfego que fornecesse distribuições de probabilidade a partir de dados de velocidade obtidos de fontes *online*. Desse modo, o trabalho foi

dividido em duas etapas. Primeiramente, estruturou-se um método obtenção de velocidades no espaço a partir de tempos de viagem *online*. A etapa seguinte consistiu em utilizar essas velocidades para condicionar a agregação de fluxos veiculares históricos representativos das condições segundo as quais foram gerados. Esses subconjuntos de fluxos foram inseridos em um modelo hierárquico *bayesiano*, a partir do qual foram geradas mais de 27000 previsões de fluxo veicular na forma de distribuições de probabilidade, considerando diferentes condições de ocorrência e horizontes de previsão para dez seções de rodovias. O modelo proposto se mostrou robusto o suficiente para gerar previsões que incorporassem a variabilidade e incerteza sem a necessidade de grandes quantidades de dados, e seu grande trunfo foi possibilitar, por meio da incorporação da informação a *priori* do método *bayesiano*, que as previsões pudessem ser realizadas apesar da restrição da quantidade de informações disponíveis. Adicionalmente, uma das consequências diretas desse trabalho foi a validação das fontes de dados *online*, que podem ser utilizadas de forma complementar em casos de falhas nos sistemas tradicionais de coletar, ou como fonte de dados principal em regiões onde não há infraestrutura de monitoramento disponível. Por fim, o modelo pode ser adotado tanto individualmente quanto como recurso adicional a outros métodos de previsão, a fim de garantir que valores previstos por outros métodos se encontrem num intervalo de valores provável.

Palavras-chave: Previsão de fluxo veicular. Inferência *bayesiana*. Fontes de dados *online*.

Abstract

Stefani, H. S. **Urban roadway traffic flow prediction from crowdsourced speed data**. 100 p. Master Thesis – São Carlos School of Engineering, University of São Paulo, 2021.

Traffic flow prediction is a useful tool for assessing traffic conditions on a roadway and can be used to promote greater efficiency in the planning of transportation systems. The continuous generation of traffic data by traditional monitoring systems over long periods has resulted in large volumes of historical data, which has fostered the development of traffic flow prediction models. Currently, these models seek to achieve high levels of accuracy, and although it is a widespread indicator to gauge the quality of forecasts, it does not consider the variability of these values. In this context, a tool that allows obtaining not only point predictions but also credibility intervals is *Bayesian* statistics. In parallel to this, companies specialized in Internet services started to provide real-time traffic information based on the displacement of their users. One such service is the *Google Distance Matrix API*, which provides travel times in real-time. These data sources have the potential to complement traffic analyses, especially when traditional collection systems fail or are non-existent. In this sense, the goal of this work was to propose a traffic flow prediction method that would provide probability distributions from speed data obtained from *online* sources. Thus, the work was divided into two stages. First, we structured a method to obtain space speeds from *online* travel times. The next step consisted of using these speeds to condition the aggregation of historical vehicle flows representative of the conditions under which they were generated. These subsets of flows were fed into a hierarchical *Bayesian* model, from which more than 27000 vehicle flow predictions were generated in the form of probability distributions, considering different occurrence conditions

and prediction horizons for ten highway sections. The proposed model proved to be robust enough to generate forecasts that incorporated variability and uncertainty without the need for large amounts of data, and its greatest asset was to enable, through the incorporation of the *priori* information of the *Bayesian* method, that forecasts could be made despite the restricted amount of information available. Additionally, one of the direct consequences of this work was validation of the *online* data sources, which can be used in a complementary way in cases of failures in traditional collecting systems, or as the main data source in regions where no monitoring infrastructure is available. Finally, the model can be adopted both individually and as an additional resource to other forecasting methods, to ensure that values predicted by other methods are in a likely range of values.

Keywords: Traffic flow prediction. Bayesian Inference. Crowdsourced Data.

Sumário

1 Introdução **15**

1.1 Objetivos 18

1.2 Justificativa 18

1.3 Estrutura do trabalho 19

2 Revisão da literatura **21**

2.1 Parâmetros de tráfego 21

2.2 Fontes de dados de tráfego 23

2.3 Previsão do fluxo de tráfego 26

2.3.1 Modelos de tráfego 28

2.3.2 Modelos baseados em dados 29

2.4 Inferência *bayesiana* 33

2.4.1 Estruturação de modelos *bayesianos* 35

2.4.2 Diretrizes para definição das prioris e verossimilhança 36

2.4.3 Avaliação dos resultados da distribuição posterior 38

2.5 Considerações finais 39

3 Banco de dados **41**

3.1 Seleção de locais para coleta de dados de tráfego 41
3.2 Coleta de dados 44
3.2.1 Dados históricos de sensores 44
3.2.2 Dados da API 45
3.3 Validação entre dados de velocidade da API e de sensores 46
3.3.1 Modelo *bayesiano* para validação de velocidades da API 48
3.4 Considerações finais 51

4 Previsão de fluxo de tráfego 53
4.1 Agrupamento dos dados de fluxo 53
4.2 Estruturação do modelo de previsão 56
4.2.1 Definição da distribuição de verossimilhança 56
4.2.2 Definição das distribuições a *priori* e modelo *bayesiano* 59
4.3 Obtenção das distribuições posteriores 61
4.4 Análise das distribuições posteriores 62
4.5 Considerações finais 63

5 Resultados 65
5.1 Análise de convergência das cadeias 65
5.2 Implicações práticas da distribuição posterior 70
5.3 Distribuição dos parâmetros μ, σ e ν 72
5.4 Previsões por sensor 73
5.5 Intervalos de previsão 76
5.6 Acurácias 77
5.7 Considerações finais 78

6 Conclusões e sugestões 81
6.1 Conclusões 81
6.2 Recomendações para trabalhos futuros 83

Referências 85

Apêndices **93**

APÊNDICE A Previsões **95**

Capítulo 1

Introdução

Fluxo de tráfego veicular é um parâmetro tipicamente utilizado para avaliar as condições e a qualidade de tráfego em um trecho viário. Analogamente, outras medidas, como velocidades e tempos de viagem, podem ser utilizados para detectar os estados de tráfego em maiores ou menores horizontes de tempo. O acesso a tais informações permite aos usuários das vias planejar melhores rotas e organizar suas atividades diárias e, de forma indireta, poupa o estresse de lidar com congestionamentos, promove a economia de tempo e dinheiro e melhora qualidade de vida (POLSON; SOKOLOV, 2017; ABDOLLAHI; KHALEGHI; YANG, 2020). Caso esses parâmetros indiquem que uma determinada via não consegue atender às demandas de seus usuários, seja no curto, médio ou longo prazo, cabe aos profissionais da área de planejamento de tráfego utilizá-los para delinear soluções estruturais ou operacionais a fim de sanar os problemas existentes.

Com o intuito de analisar as condições de tráfego em tempo real, foram desenvolvidas, ao longo das últimas décadas, diversas abordagens para prever parâmetros de tráfego. Esses métodos dependem fortemente de dados históricos e em tempo real, coletados de diferentes fontes de dados, tais como laços indutivos, radares, monitoramento por câmeras, sistemas de posicionamento global (GPS), contribuição coletiva, mídias sociais, dentre outros, que passaram a ser aplicados à coleta de velocidades, fluxos e outras medidas de tráfego. A geração de dados por meio da infraestrutura desses sistemas de monitoramento no decorrer de longos períodos resultou em grandes volumes de dados históricos, o que deu origem ao *Big Data* aplicado a transportes (LV et al., 2014). A combinação entre a disponibilidade de dados sem precedentes e a capacidade de processamento rápido ocasionou o desenvolvimento e a aceitação das tecnologias de Sistemas Inteligentes de Transportes (do inglês, *Intelligent Transportation Systems*,

ou ITS) (VLAHOGIANNI; KARLAFTIS; GOLIAS, 2014).

Nesse contexto, a previsão de tráfego pode ser conduzida com base em modelos de tráfego ou por análise de dados históricos (SEO et al., 2017). Modelos de tráfego se baseiam na teoria de fluxo de tráfego para encontrar relações matemáticas entre os principais parâmetros que descrevem a corrente veicular. Entretanto, esses modelos têm a desvantagem de que a condição de tráfego que eles buscam reproduzir deve ser muito similar à situação real para resultar em previsões precisas, já que esses modelos propõem simplificações da vida real e são limitados a um número restrito de variáveis (MORI et al., 2015). Ao mesmo tempo, métodos baseados em análise de dados históricos vêm crescendo sistematicamente, acompanhando a difusão expressiva da inteligência artificial (VLAHOGIANNI; KARLAFTIS; GOLIAS, 2014). Esses métodos contam com abordagens estatísticas que buscam extrair padrões em séries temporais de dados, como médias móveis auto regressivas e filtros de Kalman, e técnicas de aprendizado de máquina, associadas essencialmente ao uso de redes neurais.

Apesar das diferenças existentes entre diferentes métodos de previsão de fluxo de tráfego, eles convergem em uma direção: para validá-los, na maioria das vezes são utilizadas medidas de acurácia, que consistem em avaliar em quanto um valor previsto e um valor observado divergem entre si. Atualmente, os modelos competem entre si a fim de gerar previsões que sejam o mais precisas possíveis, e embora a acurácia seja um indicador difundido e pareça satisfatória para aferir a eficácia de um modelo preditivo, ela por si só não é suficiente para aferir sua qualidade (MORI et al., 2015). Uma das desvantagens de avaliar modelos apenas com base na acurácia é que essa medida é baseada em estimativas pontuais, e não em intervalos de valores. Em outras palavras, a incorporação dos intervalos de previsões é frequentemente negligenciada, embora também deva ser incluída no processo de análise. Desse modo, é possível estender a análise e verificar se as estimativas pontuais se encontram dentro de intervalos de probabilidade previstos, de modo a complementar a interpretação sobre os valores pontuais.

Outra desvantagem dos modelos de previsão existentes é que a sua natureza baseada em dados costuma exigir grandes volumes de dados de entrada (MORI et al., 2015). Essas abordagens tendem a funcionar bem para localidades em que tais informações estão amplamente disponíveis. Em função disso, esses modelos não funcionam bem se os dados são limitados. Nesse sentido, torna-se relevante contar com modelos robustos o suficiente para lidar com a restrição de dados. Nesse contexto, uma ferramenta que permite obter intervalos de credibilidade e lidar com pequenas quantidades de dados ao mesmo tempo é a estatística *bayesiana*. Sua principal ideia consiste em incorporar conhecimento prévio aos dados na forma de distribuições de probabilidade. O resultado final se resume a uma distribuição de probabilidade cujas propriedades descrevam um ou mais parâmetros previstos. Em função de sua versatilidade, muitos profissionais e pesquisadores da área de transportes começaram a utilizar métodos estatísticos *bayesianos* nos últimos anos (GHOSH; BASU; O'MAHONY, 2007; QUEEN; ALBERS, 2009; GHOSH; BASU; O'MAHONY, 2010; WANG; DENG; GUO, 2014; KIDANDO et al., 2017; ZHU et al., 2019)

em detrimento de métodos estatísticos frequentistas ou de aprendizado de máquina.

Paralelamente a isso, um dos adventos das tecnologias recentes é que os deslocamentos das pessoas podem ser traduzidos em *bits* de informação, desde que se porte um *smartphone* que as conecte a uma rede global. Nesse cenário, empresas especializadas em serviços de Internet, como Google, HereMaps e TomTom, dentre outras, oferecem seus serviços em troca dos dados de seus usuários. Elas se encarregam de coletar, processar e retornar esses dados na forma de novos serviços ou dados comercializáveis. Neste último caso, para tornar o processo mais direto, essas companhias desenvolveram suas próprias Interfaces de Programação de Aplicativos, ou API's, que consistem em um conjunto de rotinas que possibilitam aos usuários lançar mão desses serviços ou dados compilados.

Um serviço que segue essas diretrizes é a *Google Distance Matrix API*, fornecida pela Google. Por meio dessa aplicação, seus usuários podem solicitar o tempo de viagem em tempo real entre duas localidades que possuem cobertura de serviços de Internet. Para isso, é necessário que alguns usuários da plataforma estejam naquela rota, ou pelo menos em parte dela, de modo que a API calcula e retorna o tempo de viagem ao usuário solicitante. Caso não existam usuários naquela rota naquele dado momento, a empresa utiliza sua base de dados histórica para fornecer esses valores. Esses serviços tornaram ubíquos dados como tempos de viagem, e sua disponibilidade pode ser aproveitada para realizar análises de tráfego e servir como fonte alternativa de dados quando os métodos tradicionais de coleta falham.

Dado o cenário anterior, nota-se o potencial de utilizar tanto tempos de viagem obtidos *online* quanto dados históricos de sensores para realizar previsão de fluxo de tráfego. Por um lado, dispor de um método alternativo de previsão, que considere tempos de viagem em curso em uma dada seção viária onde sensores podem falhar é de interesse para profissionais da área e autoridades competentes. Por outro lado, seria razoável utilizar um modelo complementar a outras abordagens de previsão de tráfego que quantifique as incertezas na previsão.

Estes questionamentos levaram à proposta de um método que se utiliza de informações de sensores de rodovias urbanas e de tempos de viagem fornecidos *online* e dados históricos de sensores para estimar fluxos de tráfego. Tendo em mente a natureza estocástica dos estados de tráfego veicular e a necessidade de prover intervalos de credibilidade que incorporem tanto a variabilidade quanto erros de medida, este trabalho propõe uma abordagem probabilística para estimar fluxos de tráfego. Para isso, são utilizadas ferramentas de inferência *bayesiana* para estruturar um modelo que incorpore ao mesmo tempo dados observados e conhecimentos prévios acerca desses parâmetros. Dessa forma, o método deste trabalho foi elaborado para responder às seguintes perguntas de pesquisa:

1. Como dados de tempo de viagem *online* podem ser incorporados às análises de tráfego de modo complementar aos dados de sensores?

2. Como prever intervalos de credibilidade de fluxos de tráfego a partir de dados de velo-

cidade coletados *online* e dados históricos de sensores?

1.1 Objetivos

As questões de pesquisa levaram ao objetivo geral deste trabalho, que consiste em propor um método de previsão de fluxo de tráfego que forneça intervalos de credibilidade às previsões, a partir de velocidades, obtidas indiretamente a partir de tempos de viagem de fontes *online*, utilizadas como uma das entradas do modelo. Os objetivos específicos são:

1. Propor um método para validar a aderência entre dados de velocidades *online* e dados de sensores, a fim de gerar compatibilidade entre ambas as fontes e facilitar a fusão de dados;
2. Propor método de agregação de dados históricos de sensores de fluxo e velocidade veicular a partir de velocidades *online* para gerar distribuições de fluxo;
3. Verificar a aplicabilidade de inferência *bayesiana* à previsão de fluxos viários, a fim de obter intervalos de previsão, além de estimativas pontuais de fluxos.

1.2 Justificativa

Obter informações de tráfego é uma necessidade inerente tanto ao usuário do sistema de transportes para a definição de suas rotas quanto para autoridades responsáveis pela fiscalização e manutenção do espaço viário, bem como para profissionais envolvidos com planejamento de transportes. Usualmente, tais informações têm fornecimento restrito e estão limitadas apenas aos locais de coleta. Utilizar fontes de dados *online* é uma alternativa para descentralizar essa informação que na maioria das vezes fica restrita às entidades que se apropriam desses dados com exclusividade e dificilmente disponibilizam seu acesso, seja por meios comercializáveis ou gratuitos.

Adicionalmente, dado o aumento da oferta de dados de tráfego e com a expansão da capacidade computacional nos últimos anos, torna-se relevante verificar a aplicabilidade de dados de fontes *online* para a previsão de fluxo de tráfego. A análise de fluxos tende a ser mais abrangente do que a análise de velocidades, já que, para uma mesma velocidade, existe uma gama enorme de fluxos possíveis, e por este motivo, dá-se preferência ao estudo de fluxos neste trabalho.

Instrumentos de medição estão sujeitos a falhas, descalibração e erros de medida, e por isso podem muitas vezes não representar com fidelidade o fenômeno ocorrido. Assim, um método alternativo de previsão que utilize como variável de entrada um dado coletado de fonte *online*, aliado a dados históricos, pode ser de utilidade para tornar as previsões independentes de um sensor sujeito a mau funcionamento. Paralelamente, conferir intervalos de credibilidade, além das medidas tradicionais de acurácia, agrega informação às previsões, de forma que se possa medir a variabilidade sobre as expectativas dos valores previstos.

Como contribuições, espera-se tornar o acesso à informação de tráfego mais difundido e promover uma análise alternativa do fenômeno de tráfego, além de facilitar o planejamento operacional e estratégico para profissionais da área de transportes, de forma a melhorar a qualidade de vida dos usuários das vias de transporte indiretamente. Futuramente, espera-se criar alternativas confiáveis e menos custosas para planejamento de tráfego.

1.3 Estrutura do trabalho

Este trabalho está dividido em seis partes, cujo presente capítulo introduz a relevância da estimação de parâmetros de tráfego e apresenta suas pretensões e objetivos, além de sinalizar as ferramentas que serão utilizadas. O Capítulo 2 trata da revisão bibliográfica, que aborda os conceitos necessários à compreensão do presente estudo, compreendendo parâmetros de tráfego, ferramentas de coleta, os métodos de estimação de fluxo veicular e os principais conceitos inferência *bayesiana* e suas aplicações. O Capítulo 3 trata da obtenção e organização do banco de dados, além de apresentar uma abordagem para validação entre velocidades de sensores e da API. O Capítulo 4 apresenta o método principal deste trabalho, que consiste na obtenção de conjuntos de fluxo de tráfego em função de velocidades coletadas pela API e outras variáveis de entrada, a serem utilizados para previsão. Em seguida, no Capítulo 5 são apresentados os resultados principais e discussões. Por fim, no Capítulo 6 são apresentadas as conclusões e sugestões para trabalhos futuros.

Capítulo 2

Revisão da literatura

Este capítulo está divido em quatro partes. A primeira parte introduz os principais parâmetros que descrevem a corrente de tráfego e para quais níveis de análise são utilizados. Em seguida, são exploradas as principais fontes de dados existentes para coleta de parâmetros de tráfego. Na terceira parte, foram relacionados os principais modelos de tráfego já desenvolvidos e abordagens de previsão de tráfego recorrentes na literatura. Por fim, são apresentados os conceitos e ferramentas aplicados em inferência *bayesiana*.

2.1 Parâmetros de tráfego

Parâmetros de tráfego são grandezas físicas utilizadas para avaliar as condições de tráfego em uma seção viária em um dado intervalo de tempo. A corrente de tráfego pode ser classificada em função do controle operacional como interrompida ou não interrompida. A corrente não interrompida ocorre quando se percorre o trajeto sem interrupções geradas pela sinalização, existência de cruzamentos ou movimentos conflitantes, ou então, quando a distância entre interrupções for superior a três quilômetros (ROESS; PRASSAS; MCSHANE, 2010). No âmbito deste trabalho, são abordados fluxos não interrompidos de trechos urbanos de rodovias. As definições dos parâmetros de tráfego abordados em seguida estão referenciadas em Roess, Prassas e McShane (2010).

Os parâmetros de tráfego podem ser utilizados para análise em nível macroscópico, que considera a corrente veicular como um todo, ou em nível microscópico, que leva em consideração as interações em nível veicular. Fluxo, velocidade e densidade são os principais parâmetros macroscópicos de tráfego. O volume de tráfego (Q) pode ser definido como o número de veículos (n) que passam por uma determinada seção, pista ou sentido de deslocamento, em um

dado intervalo de tempo (Δt), e é medido em unidades veiculares, embora seja comumente expresso em veículos por hora. O fluxo (q), mostrado na Equação 2.1, está relacionado ao volume e também é expresso em veículos por hora, porém é observado em intervalos inferiores à duração de uma hora.

$$q = \frac{n}{\Delta t} \tag{2.1}$$

A velocidade é definida como o deslocamento de um veículo ou da corrente de tráfego em um dado intervalo de tempo e é expressa em unidades de espaço pelo tempo. Pode ser medida em relação ao tempo ou espaço. A velocidade média no tempo (u_t) na Equação 2.2 corresponde à média aritmética das velocidades instantâneas dos veículos passando por uma seção em um dado instante.

$$u_t = \frac{1}{n} \sum_{i}^{n} u_i \tag{2.2}$$

Já a velocidade média no espaço (u_s) considera a velocidade de cada veículo ao longo de um trecho de via e consiste no somatório de todos os espaços sobre todos os tempos de viagem. De forma alternativa, pode ser calculada como a média harmônica das velocidades de todos os veículos que transitaram pelo trecho considerado no respectivo intervalo de tempo. É a medida de velocidade comumente usada no estudo dos fenômenos de tráfego. A fim de considerar a velocidade ao longo de uma seção, a velocidade média no espaço deve ser utilizada (MORI et al., 2015).

$$u_s = \frac{n}{\sum_{i}^{n} \frac{1}{u_i}} \tag{2.3}$$

Por serem de naturezas distintas e úteis para diferentes tipos de análise, algumas relações que podem explicar as as variações entre as velocidades no tempo e no espaço foram propostas. O primeiro modelo desenvolvido para obter velocidades no tempo a partir de velocidades no espaço foi proposto por Wardrop (1952). Esta relação é expressa pela Equação 2.4, que acrescenta ao valor da velocidade média no espaço o produto entre seu inverso e sua variância, de modo a obter a velocidade média no tempo, o que implica em velocidades no tempo numericamente maiores do que suas velocidades equivalentes no espaço.

$$\overline{u}_t = \overline{u}_s + \frac{\sigma_s^2}{\overline{u}_s} \tag{2.4}$$

Essa relação, entretanto, apresenta limitações práticas, já que grande parte dos métodos tradicionais de coleta obtém velocidades no tempo, de modo que tanto velocidades no espaço e suas variâncias não são facilmente disponibilizadas. Com essa limitação em mente, Rakha e Zhang (2005) provaram a relação complementar, conforme a Equação 2.5, que utiliza apenas velocidades no tempo e suas variâncias. Ainda assim, embora velocidades no tempo sejam mais comuns, as variâncias dessas velocidades raramente são registradas pelos sistemas de coleta de dados, o que limita a aplicação dessa relação para análises de tráfego.

A fim de criar subsídios para utilização da Equação 2.5, Soriguera e Robusté (2011) propuseram um método para extrair variâncias diretamente dos intervalos de agregação de dados históricos de velocidade no tempo. Sob a suposição de que os conjuntos de velocidades de veículos se comportam segundo uma distribuição normal, os autores consideraram apenas a velocidade média no tempo e contagens de veículos para diferentes limites de velocidades, para estimar a variância da velocidade sobre a média aritmética das velocidades individuais. Uma abordagem semelhante foi proposta por Martínez-Díaz e Pérez (2015), que assumiram uma distribuição log-normal de velocidades veiculares. Ambos os procedimentos, entretanto, baseiam-se na disponibilidade do número de veículos que passa na seção de análise dentro do período considerado, de modo que ambas as abordagens se tornam mais indicadas para condições estacionárias de tráfego e com maior disponibilidade de dados.

$$\overline{u}_s = \overline{u}_t - \frac{\sigma_t^2}{\overline{u}_t} \tag{2.5}$$

O último parâmetro macroscópico é a densidade (k), expressa pela Equação 2.6 e definida como o número de veículos que ocupam um trecho de via (ΔL). É um dos mais importantes parâmetros de tráfego, pois reflete diretamente na qualidade do nível de serviço ao indicar o grau de proximidade entre veículos, o que influencia na liberdade de manobras e também em aspectos psicológicos inerentes à condução dos motoristas (ROESS; PRASSAS; MCSHANE, 2010). Sua mensuração de forma direta é de difícil obtenção, e por isso, busca-se obtê-la a partir de outros parâmetros de tráfego.

$$k = \frac{n}{\Delta L} \tag{2.6}$$

2.2 Fontes de dados de tráfego

Dados de tráfego são o ponto de partida para o desenvolvimento de diversos modelos e teorias que permitem o estudo da dinâmica veicular. Pesquisadores e profissionais da área de transportes buscam obtê-los de forma a garantir quantidade e qualidade de informação com os menores custos possíveis. Para atender à demanda de melhorias nos sistemas de gerenciamento de tráfego, os métodos de coleta de dados têm evoluído continuamente e o acesso a informações de tráfego em tempo real se tornou rotineiro (LEDUC, 2008).

A coleta tradicional de dados de tráfego pode ter abordagens baseadas na via ou no veículo (MORGUL et al., 2014). Dispositivos de tecnologias baseadas nas vias são mais utilizados e podem estar embutidos na estrutura do pavimento, como laços indutivos. Eles são instalados em uma configuração de bobina em espiral que gera um campo magnético quando um corpo metálico se desloca sobre eles, de modo que a informação de passagem é transmitida para um dispositivo de contagem instalado ao lado da rodovia. Os laços indutivos podem ser simples ou duplos. No primeiro caso, os principais parâmetros coletados são o número de veículos e ocupância, enquanto os últimos registram também a velocidade instantânea e, em alguns ca-

sos, a composição veicular da corrente de tráfego. Essas medidas vêm agregadas em intervalos de tempo que podem variar de alguns segundos a vários minutos. Após agregação, as contagens se convertem em fluxos veiculares e velocidades instantâneas em velocidades médias no tempo, no caso dos laços indutivos duplos.

A quantidade de dados gerados por esses sensores estacionários nem sempre é suficiente para o controle de tráfego em função de seu emprego limitado a distâncias que podem variar de alguns metros a dezenas de quilômetros. Além disso, muitas rodovias não contam com esse tipo de infraestrutura, especialmente em trechos rurais e em países em desenvolvimento (SEO et al., 2017). Embora seja um método de coleta amplamente utilizado, ele tem altos custos de instalação e de manutenção. Além disso, a obtenção de dados confiáveis desses sensores pode ser um desafio, porque esses equipamentos podem sofrer danos pela passagem de veículos pesados e ficar sujeitos erros de funcionamento, o que pode comprometer as coletas ou resultar em medidas de má qualidade (MORI et al., 2015).

Dos dispositivos baseados em veículos, destaca-se o uso de veículos de prova, utilizados especificamente para a obtenção de dados de tráfego em tempo real. Sua operação tem como princípio coletar dados em tempo real por meio da localização do veículo via telefone móvel ou aparelhos GPS ao longo da rede. As características mais importantes dos veículos de prova são a taxa de penetração e a frequência de amostragem. A primeira descreve a proporção de veículos de prova na corrente de tráfego, enquanto a frequência de amostragem consiste no intervalo de tempo entre o envio de informações dos sensores. Apesar de apresentar vantagens como seu custo reduzido e ocorrência ao longo de toda a rede viária, como os dados móveis obtidos por trajetórias de GPS contém apenas velocidades médias ou amostras de dados de trajetórias, é preciso fazer suposições adicionais sobre os dados gerados por esse meio, já que não é possível extrapolar de forma direta outras variáveis de tráfego, como fluxo e densidade (SEO et al., 2017).

A popularização dos métodos de coleta de informações de tráfego promoveu o surgimento de grandes volumes de dados disponibilizados em tempo real. Estes sustentam o desenvolvimento dos Sistemas Inteligentes de Transportes, ou *ITS*, amplamente utilizados para controle de tráfego e sistemas de informação ao usuário (MORGUL et al., 2014). Com altas taxas de atualização de informações, sistemas de processamento de dados tradicionais se tornam ineficientes por não comportar o crescimento da quantidade e complexidade desses dados. Nesse sentido, análises em *Big Data*, termo usado para caracterizar o processamento e armazenamento de grandes conjuntos de dados, fornecem as ferramentas ideais para aplicações em *ITS* (WU et al., 2018).

Nesse cenário, as contínuas gerações de dados possibilitaram o desenvolvimento de tecnologias de mapeamento *online* que se tornaram populares a partir da implementação das Interfaces de Programação de Aplicativos ou API's (do inglês *Apllication Programming Interface*). Uma API permite que componentes de um *software* se comuniquem entre si a partir de um

conjunto de rotinas para obter dados de determinado serviço em diversos formatos disponíveis. Em razão de sua funcionalidade, os avanços em tecnologias de mapeamento resultaram em novos nichos de mercado com a venda, compra e compartilhamento de dados geográficos. Atualmente, diversas empresas como a *Google*, *Here*, *TomTom* e *Bing*, dentre outras, utilizam suas API's para disponibilizar serviços de mapeamento *online* e informações de tráfego em tempo real.

Desde 2005, a *Google* distribui dados de mapeamento *online* por meio de suas API's (WANG; XU, 2011). A plataforma concentra atualmente dados oriundos da contribuição de 1 bilhão de usuários ativos por mês, que concordam em compartilhar seus dados de geolocalização em troca dos serviços da empresa, com 25 milhões de atualizações diárias em tempo real e cobertura de praticamente toda a superfície terrestre (GOOGLE, 2021a). Um de seus produtos é a *Distance Matrix API*, que, dentre outras funcionalidades, fornece em tempo real os tempos de viagem entre pares origem-destino, definidos por suas coordenadas geográficas.

Isso é possível pois usuários da versão gratuita da plataforma, o *Google Maps*, fornecem à empresa as informações emitidas pelos aparelhos GPS embutidos em seus *smartphones* quando em deslocamento. O usuário recebe então informações de sua localização, e pode solicitar outros serviços da plataforma, como o tempo de viagem entre sua origem e destino, obter o nível de carregamento das vias, caracterizado por uma escala de cores que varia em intensidade e indica as faixas de velocidade em operação, dentre outras funcionalidades.

Embora a *Google* não divulgue o método utilizado para prever tempos de viagem, os algoritmos desenvolvidos pela empresa identificam os padrões de deslocamentos dos usuários da plataforma ao longo das redes viárias e, a partir de dados históricos e em tempo real coletados pela empresa (GOOGLE, 2021b), retornam as solicitações ao usuário pagante da *Distance Matrix API* na forma de tempos de viagem.

O resultado dessa dinâmica de coleta pode ser comparada ao de um veículo de prova, que instrumentado com um GPS, efetua a coleta de dados de localização de forma passiva e, como hoje em dia a maioria dos veículos é equipada com um ou mais *smartphones*, sua utilização passa a servir como instrumento de coleta de informações sobre a corrente de tráfego (LEDUC, 2008). Abdollahi, Khaleghi e Yang (2020) ressaltam que dados de GPS são mais indicados para previsão do tempo de viagem por sua capacidade de armazenar informações úteis como coordenadas geográficas, data e hora, velocidade e outros metadados. Assim, o processamento realizado pela empresa caracteriza a utilização de *Big Data* aplicado a *ITS*, cujos dados podem ser utilizados por pesquisadores e profissionais da área de transportes para desenvolvimento de seus projetos.

Para utilizar esse serviço, é necessário ter uma conta na plataforma do *Google Cloud*, onde a *Distance Matrix API* e demais produtos da empresa são disponibilizados. Como forma de promover a ambientação com a plataforma, o *Google* fornece, até a data de realização desta pesquisa, uma cota de 200 dólares americanos para teste de seus produtos, durante o período

de um ano. O serviço é cobrado por requisição, ou seja, no caso da solicitação de tempo de viagem em tempo real à *Distance Matrix API*, são descontados 10 dólares da conta do usuário a cada 1000 pares origem-destino requisitados. Conforme destacado por Lana et al. (2018), aqueles que, no futuro, possuírem e puderem analisar *Big Data*, possivelmente terão vantagens competitivas significativas.

Enquanto os sensores são limitados espacialmente e demandam altos custos de instalação e manutenção, a coleta de dados de tráfego por GPS de celulares é tecnicamente viável e surge como alternativa de custo reduzido, e apesar de suas vantagens claras, os dados de GPS não vão substituir os sensores fixos, mas sim funcionar como fonte de dados complementar (LEDUC, 2008). Um grande desafio, entretanto, consiste em compatibilizar as diferentes fontes de dados e extrair informações de modo orientado às aplicações em *ITS* (VLAHOGIANNI; KARLAFTIS; GOLIAS, 2014). Algumas abordagens propuseram a compatibilização entre dados de sensores e dados de veículos de prova para utilização em análises de tráfego (CACERES et al., 2012; BACHMANN et al., 2013; LI et al., 2013; HE et al., 2016). Sendo assim, as fontes de dados são um elemento importante a se considerar no estudo de parâmetros de tráfego, porque é necessário considerar esses aspectos ao se comparar teorias fundamentadas a partir de tipos de dados distintos (LANA et al., 2018).

2.3 Previsão do fluxo de tráfego

O desequilíbrio entre oferta e demanda da infraestrutura viária resulta frequentemente no esgotamento de sua capacidade. Isso gera congestionamentos que implicam em perdas nos âmbitos social, econômico e ambiental e, em função disso, organizações públicas e privadas têm se dedicado nas últimas décadas à compreensão dos fenômenos de tráfego a fim de propor soluções para esses problemas (LANA et al., 2018). As principais abordagens que podem ser empregadas, isoladamente ou em conjunto, a fim de mitigar tais ocorrências, são a expansão da capacidade viária, a implantação de políticas públicas de transporte, e o gerenciamento dos fluxos de tráfego (AREM et al., 1997).

Nesse último caso, a previsão de parâmetros de tráfego é utilizada por autoridades e engenheiros de tráfego para promover o planejamento e uso eficiente e seguro de redes viárias e também por usuários do sistema de transportes que buscam informações sobre o estado de tráfego a fim de auxiliar no planejamento de seus deslocamentos. Nesse contexto, a previsão de fluxo de tráfego tem o potencial de ajudar os usuários da rede de transportes decidir melhores rotas de viagem, aliviar congestionamentos e reduzir emissões de carbono, além de contribuir para melhorar a eficiência das operações de trânsito (LV et al., 2014).

As premissas para definição dos modelos de previsão de fluxo de tráfego devem levar em conta os horizontes de previsão, que podem ser de longo ou curto prazo. Horizontes de curto prazo variam de poucos segundos a algumas horas no futuro com base em informações atuais

e históricas (VLAHOGIANNI; KARLAFTIS; GOLIAS, 2014). Além dos horizontes de previsão, devem ser consideradas também a variedade e a qualidade de fontes de dados e sua área de abrangência, que podem, neste último caso, estar vinculadas a pontos ou segmentos rodoviários, ou a redes urbanas como um todo (MORI et al., 2015). A maior parte dos modelos de previsão de tráfego veicular se aplica a dados de rodovias (VLAHOGIANNI; KARLAFTIS; GOLIAS, 2014), considerando que o tráfego urbano apresenta maior nível de complexidade. Para contornar essa dificuldade, observou-se nos últimos anos o surgimento de métodos capazes de extrair as dependências temporais e espaciais entre arestas da malha viária para melhor compreender a dinâmica de tráfego veicular em contexto urbano (ERMAGUN; LEVINSON, 2018; RYU et al., 2018).

Além da extensão espacial à qual se aplicam as previsões, alguns autores levam em consideração a ocorrência apenas de dias úteis (GUO et al., 2018; RYU et al., 2018) ou o conjunto de dados sem feriados (DAI et al., 2019; DO et al., 2019). Em uma análise sobre os padrões de fluxo veicular para um conjunto de sensores localizados no entorno da cidade de São Paulo, por exemplo, Cunha (2007) e Ribeiro (2017) concluíram que dias de terças-feiras, quartas-feiras e quintas-feiras apresentaram perfis de volumes semelhantes e, por isso, sugeriram classificá-los como dias típicos e remover feriados do conjunto de dados. Tratamento análogo foi utilizado por Li et al. (2020) para previsão de velocidades. Outros autores, por outro lado, incluem em suas análises finais de semana (WU et al., 2018; SEKUŁA et al., 2018) e feriados (SEKUŁA et al., 2018; LIN et al., 2018).

Outros fatores que podem interferir na dinâmica da corrente veicular e ser incorporados aos modelos para refinar os resultados das previsões são a inclusão de variáveis que considerem as condições climáticas (POLSON; SOKOLOV, 2017; SEKUŁA et al., 2018), a ocorrência de eventos atípicos (POLSON; SOKOLOV, 2017; LIN et al., 2018) ou de incidentes (GUO et al., 2018). Nesse contexto, Lana et al. (2018) afirmam que os principais desafios levantados pelas revisões da literatura mais recentes sobre o tema devem incluir técnicas de gerenciamento da fusão de diferentes fontes de dados de tráfego, de modo a considerar seu bom nível de agregação e de resolução. Adicionalmente, Mori et al. (2015) concluíram que é necessário considerar, para o desenvolvimento de modelos de previsão futuros, a fusão de diferentes fontes de dados, a inclusão de variáveis exógenas como eventos climáticos, incidentes e congestionamentos e por fim, a inclusão de intervalos de confiança.

Outra grande dificuldade dos pesquisadores e estudiosos da área reside na complexidade da comparação entre diferentes abordagens de previsão propostas ao longo das últimas décadas. A maioria dos modelos utiliza medidas de acurácia para alegar a superioridade de seus métodos. Entretanto, Vlahogianni, Karlaftis e Golias (2014) apontam que é importante considerar, também, as características não lineares e não estacionárias dos modelos de previsão de curto prazo. Nessa mesma linha, os autores destacam que é possível que um modelo demonstre bom ajuste aos dados sem que necessariamente apresente alta acurácia, e que por outro lado, é comum a ocorrência de modelos que tenham boa acurácia, mas falhem em um ou

mais diagnósticos estatísticos de análise de erros, o que mostra que os erros produzidos não são aleatórios, mas apresentam um padrão recorrente não captado pelos modelos, o que pode tornar as previsões não confiáveis.

De modo geral, a previsão de parâmetros de tráfego costuma ser conduzida com base ou no desenvolvimento de modelos de tráfego, ou por análise de dados históricos (SEO et al., 2017). Neste último caso, os métodos propostos se dividem em duas frentes, sendo a primeira delas baseada em técnicas de inferência estatística, e a segunda em técnicas de aprendizado de máquina, com predominância do uso de redes neurais. As subseções seguintes descrevem os principais aspectos referentes a ambas as abordagens.

2.3.1 Modelos de tráfego

Os primeiros esforços empregados para explicar os fenômenos de tráfego buscaram soluções analíticas com base em relações empíricas entre seus parâmetros. Os parâmetros macroscópicos de fluxo, velocidade e densidade podem ser medidos de forma independente uns dos outros e se relacionam entre si, o que dá origem aos modelos matemáticos que sustentam as teorias de fluxo de tráfego (WOLSHON; PANDE et al., 2016). A equação geral que relaciona esses parâmetros, conhecida como relação fundamental de tráfego, é descrita pela Equação 2.7, em que cada variável depende de características da rodovia, do veículo, do motorista e de fatores ambientais (GARBER; HOEL, 2008):

$$q = k \cdot u \tag{2.7}$$

A partir da relação fundamental podem ser derivadas relações entre outros parâmetros, como densidade e fluxo. Observa-se que a Equação 2.7 permite obter a densidade veicular de forma indireta, e pela maior facilidade de medir parâmetros de fluxo e velocidade, é utilizada frequentemente para tal finalidade. A Figura 2.1 mostra configurações usuais das relações de tráfego em função desses diferentes parâmetros.

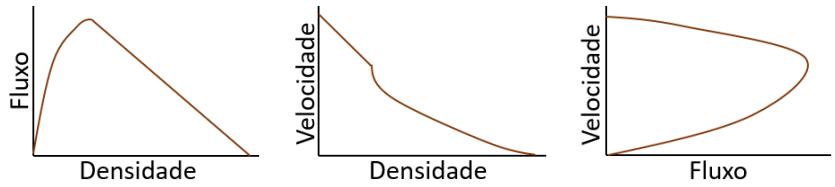


Figura 2.1: Relações fundamentais para diferentes parâmetros. Adaptado de: Wageningen-Kessels et al. (2015)

Assim, os modelos de fluxo de tráfego se sustentam na suposição de que existem relações entre a distância entre os veículos e sua velocidade. Além disso, as relações fundamentais são essenciais para o desenvolvimento desses modelos, de modo que o tráfego se encontre em

algum estado da relação fundamental (WAGENINGEN-KESSELS et al., 2015). Revisões detalhadas sobre os modelos existentes podem ser encontradas em Carey e Bowers (2012), Wageningen-Kessels et al. (2015) e Seo et al. (2017).

O diagrama fundamental pode ser descrito pela Figura 2.1a e é um dos principais elementos da teoria de fluxo de tráfego. Ele descreve a relação entre a densidade e fluxo num determinado intervalo de espaço e de tempo em uma rodovia. Modelos recentes voltados à previsão de tráfego com foco em diagramas fundamentais macroscópicos podem ser vistos em Ambühl e Menendez (2016), Qian et al. (2017), Gu et al. (2018) e Saffari, Yildirimoglu e Hickman (2020).

2.3.2 Modelos baseados em dados

Ao contrário dos conceitos apresentados na Seção 2.3.1, métodos de previsão baseados em dados não requerem modelos de tráfego explícitos, de modo que não é necessário ter conhecimento prévio sobre a teoria de fluxo de tráfego para propor modelos preditivos. Segundo Lv et al. (2014), a previsão de tráfego por análise de dados pode ser dividida entre técnicas paramétricas, com modelos de séries temporais e filtros de Kalman, métodos não paramétricos, como os métodos do vizinho mais próximo (k-NN, do inglês *k-nearest neighbor*) e redes neurais artificiais (ANN, do inglês *artificial neural network*), e simulações.

Dentre as técnicas paramétricas, destacam-se os modelos auto-regressivos integrados de médias móveis (ARIMA), aplicados pela primeira vez para previsão de tráfego por Ahmed e Cook (1979) a fim de capturar os padrões em séries temporais. Desde então, foram propostas variantes com técnicas multivariadas (WILLIAMS, 2001; CHANDRA; AL-DEEK, 2009; GHOSH; BASU; O'MAHONY, 2009), o modelo ARIMA sazonal (SARIMA) (WILLIAMS; DURVASULA; BROWN, 1998; WILLIAMS; HOEL, 2003) e métodos de agregação de vários modelos ARIMA (SHAHRIARI et al., 2020). A fim de considerar também a componente de variância das séries temporais de fluxo de tráfego, foram desenvolvidas abordagens híbridas, consistindo de junções entre os modelos ARIMA e o método autorregressivo de heteroscedasticidade condicional (GARCH) (CHEN et al., 2011; WANG; YE, 2016) e de programação genética (XU; LI; WANG, 2016). Uma outra família de modelos paramétricos que se popularizou foram os filtros de Kalman, aplicados pela primeira vez à previsão do estado de tráfego por Okutani e Stephanedes (1984) e utilizados em outras abordagens por Wang e Papageorgiou (2005), Hinsbergen et al. (2011), Xu et al. (2017), Kumar (2017) e Wang et al. (2018).

Já as abordagens não paramétricas se baseiam em ferramentas de aprendizagem de máquina, com o predomínio de redes neurais (VLAHOGIANNI; KARLAFTIS; GOLIAS, 2014). Estes algoritmos usam arquiteturas em uma ou mais camadas para extrair características do mais baixo ao mais alto nível e descobrir estruturas escondidas no conjunto de dados (LV et al., 2014). Modelos aplicados a previsão de tráfego inicialmente apresentavam arquiteturas mais simples e tendiam a ser estruturados em apenas uma camada (SMITH; DEMETSKY, 1994; LEDOUX, 1997). Posteriormente, foram introduzidas técnicas mais complexas, como redes otimizadas em múltiplas

camadas (VLAHOGIANNI; KARLAFTIS; GOLIAS, 2005) e um modelo neural *fuzzy* (YIN et al., 2002). A fim de melhorar a qualidade das previsões, foram propostos também modelos híbridos, que incorporavam tanto técnicas estatísticas quanto de aprendizado de máquina (ZHENG; LEE; SHI, 2006; TAN et al., 2009; LIPPI; BERTINI; FRASCONI, 2013). Para descobrir estruturas existentes nos dados do mais baixo ao mais alto nível, alguns autores utilizaram arquiteturas de aprendizagem profunda (*deep learning*) (POLSON; SOKOLOV, 2017; WU et al., 2018), autocodificadores variacionais (*variational autoencoder*) (BOQUET et al., 2020) e autocodificadores profundos (*deep autoencoders*) (LV et al., 2014). As abordagens mais recentes têm utilizado modelos espaço-temporais, capazes de incorporar às previsões informações sobre as características espaciais (DO et al., 2019) e espaço-temporais, cujo método mais recorrente é o dos k-vizinhos mais próximos (HABTEMICHAEL; CETIN, 2016; GUO et al., 2018; RYU et al., 2018).

De modo geral, modelos estatísticos tendem a superar os modelos de aprendizado de máquina quando existem informações a *piori* que explicam as relações funcionais das variáveis do problema e a interpretabilidade dos resultados e causalidades são importantes, além de possibilitarem verificar as propriedades estatísticas do mecanismo que gerou o cenário de previsão (KARLAFTIS; VLAHOGIANNI, 2011). Por outro lado, as redes neurais são mais úteis se a obtenção de resultados precisos é relevante, sem que haja necessidade de conhecer o processo gerador dos dados, ou quando essa atividade for de difícil identificação e as hipóteses de modelos estatísticos não forem válidas. Redes neurais têm a vantagem de não exigir domínio sobre as teorias de fluxo de tráfego e gerar bons resultados, porém é necessário ter em mãos dados de qualidade e em grandes quantidades, além de seu processo de calibração e treinamento poder ser custoso e os modelos serem de difícil interpretação (MORI et al., 2015). Como é necessário obter as condições históricas e de tráfego em tempo real, é necessário retreiná-los periodicamente para incorporar as mudanças inerentes à dinâmica de tráfego ao longo do tempo. Em resumo, enquanto os modelos estatísticos têm fundamentação matemática sólida e conseguem fornecer respostas sobre os mecanismos que originam os dados, modelos de aprendizado de máquina falham em explicar os processos de transformação dos dados.

Apesar das diferenças entre cada abordagem, ambas convergem quanto ao modo de análise de suas previsões. Para validar os resultados dos modelos, geralmente são empregadas medidas de erro. O erro absoluto médio percentual (*MAPE*, do inglês *mean average percent error* e a raiz do erro quadrático médio (*RMSE*, do inglês *root mean squared error* são duas medidas amplamente utilizadas para avaliar a qualidade de previsões e, por isso, são observadas frequentemente na literatura sobre previsão de parâmetros de tráfego. A Tabela 2.1 mostra as acurácias obtidas para o MAPE de alguns dos métodos de previsão abordados. Ambas as medidas são calculadas respectivamente pelas Equações 2.8 e 2.9 e relacionam os valores reais (O) e valores previstos (P), em que n é o número de previsões realizadas.

$$MAPE = \frac{1}{n} \sum_{i}^{n} \left| \frac{P_i - O_i}{P_i} \right| \tag{2.8}$$

Autor	Abordagem	Método	MAPE (%)
Hinsbergen et al. (2011)	St	Kalman	12.8
Habtemichael e Cetin (2016)	AM	NN	5.3
Xu et al. (2017)	St	Kalman	13.18 a 19.85
Kumar (2017)	St	Kalman	10.0
Sekuła et al. (2018)	AM	NN	24.59
Guo et al. (2018)	AM	NN	6.28 e 36.10 (incidente)
Wu et al. (2018)	AM	NN	7.0 a 9.0
Ryu et al. (2018)	AM	NN	7.53
Do et al. (2019)	AM	NN	17.0
Boquet et al. (2020)	AM	NN	<10.0
Shahriari et al. (2020)	St	ARIMA	10.5
Yao, Zhang e Zhang (2020)	St	ARIMA	~15.0

Tabela 2.1: Erros médios absolutos percentuais de métodos de previsão de fluxo de tráfego encontrados na literatura. As abordagens se diferem entre estatística (St) e de aprendizado de máquina (AM), e os métodos aqui listados podem ser filtros de Kalman, redes neurais (NN) e modelos ARIMA.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i}^{n}(P_i - O_i)^2} \tag{2.9}$$

Embora a maioria dos estudos ainda validem seus modelos pela aferição de erros médios quadráticos e de erros médios absolutos percentuais, Mori et al. (2015) reforçaram a necessidade de quantificar a incerteza, já que poucos estudos até agora fornecem os intervalos de credibilidade de suas previsões. Com isso em vista, Guo, Huang e Williams (2014) propuseram um filtro de Kalman adaptado capaz de atualizar as variâncias no processo para uma estrutura SARIMA-GARCH (do inglês *seasonal autoregressive integrated moving average plus generalized autoregressive conditional heteroscedasticity*). Para avaliar os intervalos, foram utilizadas a porcentagem de *kickoff*, indicador que mede a relação entre o número de observações ocorridas fora do intervalo de previsão determinado e o número total de previsões, e a amplitude sobre a taxa de fluxo média, que mede a razão entre o tamanho do intervalo de previsão e o fluxo observado. Nessa mesma linha, Zhang, Zhang e Haghani (2014) utilizaram um modelo híbrido formado por três componentes, sendo elas a tendência diária obtida por análise espectral, a componente determinística capturada por um modelo ARIMA, e a componente de volatilidade, obtida pelo modelo GARCH. Para análise dos intervalos de credibilidade, ao contrário da porcentagem de *kickoff*, os autores utilizaram a probabilidade de cobertura do intervalo de previsão (PCIP), descrita pela Equação 2.10, em que c_i é considerado no somatório apenas se o fluxo observado pertence ao intervalo previsto, e a amplitude média do intervalo de previsão (AMIP), representada pela Equação 2.11, que considera a média das n diferenças entre os

valores superior ($S(x_i)$) e inferior ($I(x_i)$) do dado intervalo de credibilidade.

$$PCIP = \frac{1}{n}\sum_{i}^{n} c_i \tag{2.10}$$

$$AMIP = \frac{1}{n}\sum_{i}^{n}(S(x_i) - I(x_i)) \tag{2.11}$$

Ao contrário dessas abordagens probabilísticas, Lin et al. (2018) buscaram prever intervalos de volumes horários de curto prazo a partir de um modelo híbrido, estruturado por uma rede neural de única camada escondida, com otimização de pesos sinápticos pelo método de enxame de partículas (*Particle Swarm Optimization* ou PSO), para apoiar o processo de escalação de funcionários na alfândega entre Canadá e Estados Unidos. Este modelo não forneceu previsões pontuais, apenas intervalos de credibilidade. Seus intervalos foram comparados com outros modelos estatísticos (ARMA, um modelo híbrido de análise espectral e filtros de Kalman tradicional e adaptado) supondo normalidade dos parâmetros previstos e utilizando a variância gerada por esses modelos para definir os intervalos de credibilidade a 90%, 95% e 99%, sendo esta a primeira aplicação de um modelo de aprendizagem de máquina a fornecer intervalos de previsão para volume de tráfego.

Uma forma direta de fornecer intervalos de credibilidade e, portanto, a incerteza de previsões, é por meio de distribuições de probabilidade. Entretanto, distribuições de fluxos de tráfego são pouco comuns na literatura, sendo mais frequentes a utilização de distribuições de *headways* (LI; CHEN, 2017) e tempos de viagem (GUESSOUS et al., 2014; DUAN et al., 2019; GHADER; DARZI; ZHANG, 2019). A primeira, e até então única aplicação de distribuições de fluxos de tráfego na pesquisa em transportes foi proposta por Djenouri, Zimek e Chiarandini (2018) para detecção de *outliers* em séries temporais. Essas distribuições foram geradas considerando a sequência de fluxos de tráfego observados durante um intervalo de tempo, como por exemplo, horas de pico às segundas-feiras.

Uma ferramenta que lida diretamente com distribuições de probabilidade é a estatística *bayesiana*. Sua aplicação à previsão de parâmetros de tráfego pode ser vista em Kidando et al. (2017) para previsão de tempos de viagem por meio de modelos hierárquicos *bayesianos* e em Fei, Lu e Liu (2011) por meio de um modelo linear dinâmico *bayesiano*. Sua utilização já é bastante difundida em transportes, cobrindo outras subáreas de pesquisa, como alocação de tráfego (ZHU et al., 2019), calibração dos parâmetros da curva fluxo-velocidade (OLIVEIRA; CUNHA, 2019) comprimentos de filas em interseções (MEI et al., 2019) e análise de padrões de deslocamento por caronas (ZHU et al., 2021).

Abordagens híbridas com uso desta ferramenta aplicadas à previsão de fluxo de tráfego também são recorrentes, com exemplo de estimação de parâmetros de um modelo SARIMA (GHOSH; BASU; O'MAHONY, 2007) e otimização de pesos para combinação entre modelos distin-

tos de previsão (WANG; DENG; GUO, 2014). Ghosh, Basu e O'Mahony (2010) combinaram o uso de *wavelets* com método de Bayes hierárquico, que considera que a variância das observações de fluxo variam conforme o período do dia. Já Queen e Albers (2009) propuseram um modelo dinâmico *Bayesiano* multivariado para prever fluxo de tráfego mesmo em casos de atividades atípicas na rodovia, como incidentes de tráfego. Por fim, Tebaldi, West e Karr (2002) propuseram um modelo hierárquico de regressão linear *bayesiana* considerando a interdependência temporal e espacial entre seis sensores consecutivos em um segmento de rodovia.

2.4 Inferência *bayesiana*

A inferência estatística busca informações sobre quantidades não observadas a partir de dados numéricos observados, e ela pode ser tratada tanto pela inferência clássica quanto pela inferência *bayesiana*. A inferência clássica consiste em considerar a variabilidade que seria observada caso um experimento fosse realizado sob as mesmas condições um grande número de vezes. Assim, as conclusões acerca do parâmetro de interesse são feitas com base na amostra observada. A inferência *bayesiana*, por outro lado, utiliza as probabilidades para quantificar a incerteza sobre eventos que ainda não ocorreram.

Os três objetivos principais da inferência *bayesiana* consistem em estimar valores de parâmetros, prever valores ou comparar modelos (KRUSCHKE, 2014). Uma de suas características é que, uma vez obtida a distribuição a *posteriori* de um conjunto de dados, é possível utilizá-la para gerar dados, ou seja, fazer previsões (MARTIN, 2016). Como a análise *bayesiana* não requer amostras grandes, ao contrário dos modelos de aprendizado de máquina, conjuntos de dados menores podem ser analisados sem perder o poder explicativo enquanto mantém precisão (SCHOOT et al., 2015). A utilização de inferência *bayesiana* para aplicação em análises complexas de dados só se tornou possível devido à produção de computadores rápidos e baratos acessíveis a um público amplo (KRUSCHKE, 2014; MARTIN, 2016). Dessa forma, seu objetivo é combinar os eventos futuros com os dados observados, a partir do Teorema de Bayes.

O Teorema de Bayes enuncia que a probabilidade de um evento está condicionada ao conhecimento que se tem a *priori* sobre ele e pode ser expresso pela Equação 2.12:

$$p(A|B) = \frac{p(A)p(B|A)}{p(B)} \tag{2.12}$$

Em que:

- $p(B|A)$: é a probabilidade de ocorrência do evento B, dado que o evento A é verdadeiro, chamada de verossimilhança dos dados observados;
- $p(A)$: é a probabilidade a *priori*, ou seja, a probabilidade de ocorrência do evento A antes de se saber sobre a ocorrência do evento B;

- $p(A|B)$: é a distribuição de probabilidade a *posteriori*, ou seja, a probabilidade posterior de ocorrência do evento A, dado que o evento B ocorreu;

- $p(B)$: é a probabilidade de ocorrência do evento B.

Assume-se que o processo de geração dos dados seja representado por uma distribuição de probabilidades, a qual pode ser descrita por uma função, a verossimilhança, que represente a variabilidade da observação. Essa função implica na escolha de um modelo estatístico ao qual esteja associada uma família de distribuições, por meio do qual sejam efetuadas as hipóteses e obtidos os parâmetros sobre os dados já conhecidos. A informação da distribuição a *priori* é atualizada pela função de verossimilhança e leva à distribuição a *posteriori* via Teorema de Bayes. Assim, a obtenção da distribuição $p(A|B)$ possibilita realizar as inferências sobre a variável desejada e descrever seu comportamento. Por se tratar de uma distribuição, costuma ser representada pela média, mediana ou moda, podendo também ser descrita pelas variâncias ou quantis, que quantificam a incerteza (BETANCOURT, 2017).

Muitas vezes, a distribuição a *posteriori* não possui solução analítica e, nesses casos, pode ser obtida por métodos numéricos computacionais. Estes podem ser markovianos ou não-markovianos. Dos métodos não-markovianos, destacam-se a computação em grelha, aproximação quadrática e os métodos variacionais. Do outro grupo, destacam-se o método de Metropolis-Hastings e o método de Monte Carlo via cadeias de Markov (MCMC). O método atualmente mais utilizado, o *No U-Turn Sampler* (NUTS), é uma variação de MCMC, utilizando como base a física hamiltoniana, que consiste na manutenção da energia total de um sistema quando ocorre variação entre estados, sendo fortemente recomendado por Salvatier, Wiecki e Fonnesbeck (2016) para modelos de variáveis contínuas.

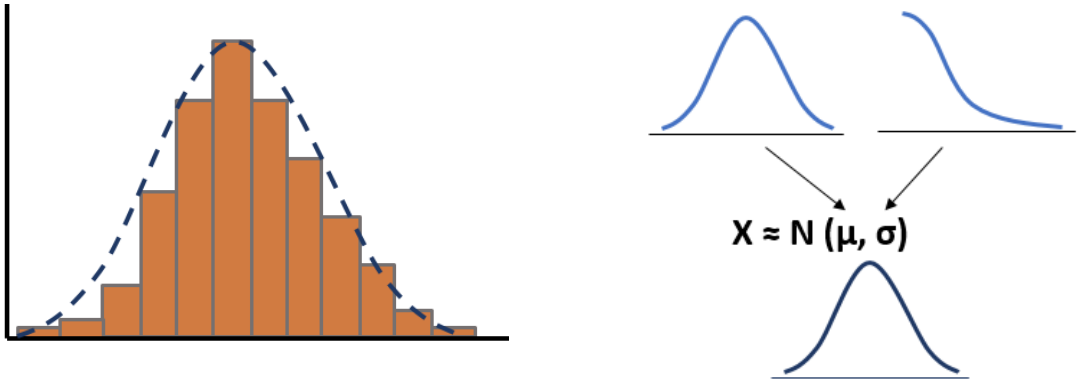
Uma cadeia de Markov implica na transição entre estados, ou seja, a mudança para um novo estado depende do estado atual. Sua utilização consiste na ideia de que, havendo uma distribuição com transições proporcionais à distribuição alvo, basta retirar amostras dessa distribuição. No método NUTS, essa transição deve respeitar a uma taxa de aceitação definida pelo analista, que deve ser comparada com a razão entre a probabilidade dos estados anterior e sucessor. Uma taxa de aceitação de 0%, por exemplo, significa aceitar todos os pontos propostos, o que resultaria em amostras não representativas da distribuição posterior, enquanto uma taxa de 100% restringiria a amostragem apenas à distribuição real geradora dos dados.

A implementação dessas técnicas é facilitada pela existência de pacotes de programação probabilística que possuem algoritmos nativos específicos para inferência *bayesiana*. Um desses pacotes é o *PyMC3* (SALVATIER; WIECKI; FONNESBECK, 2016), voltado para programação em linguagem *Python*, que permite aos seus usuários focar no seu problema científico em vez de destinar seu tempo ao estudo e implementação de métodos computacionais para resolvê-lo, além de incluir funcionalidades como o diagnóstico dos modelos e resumo dos resultados.

2.4.1 Estruturação de modelos *bayesianos*

Modelos buscam descrever sistemas ou processos e, para construí-los, são necessários dados e hipóteses sobre os dados que os descrevem (MARTIN, 2016). A análise visual ou revisão sistemática sobre esses dados permite criar suposições sobre sua configuração e propor funções que o descrevam, o que implica na definição da função de verossimilhança dos dados. Para fins de ilustração, na Figura 2.2a é mostrada uma distribuição de frequências hipotética de um conjunto de dados, cujo processo gerador se aproxima de uma distribuição normal. Essa é a verossimilhança, que após escolhida, pode ser definida em função de seus parâmetros. Esses parâmetros, por sua vez, podem ser descritos por outras distribuições de probabilidade, as distribuições a *priori*, que servem para incorporar informações sobre o problema em análise.

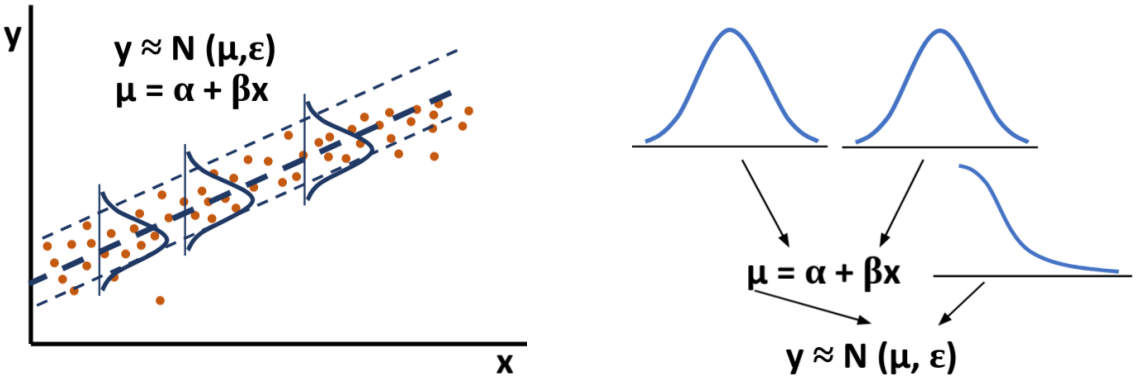
Esse tipo de estrutura é conhecido como modelo hierárquico, por conter parâmetros que dependem de parâmetros vindos de outras distribuições. A Figura 2.2b mostra um esquema da estrutura de um modelo de uma camada, em que os parâmetros da verossimilhança foram definidos por outras distribuições de probabilidade. Estes valores devem concentrar o máximo de informação possível que se detém sobre aquele parâmetro, sobre o qual há um grau maior ou menor de certeza. Neste caso, seria possível atribuir tantos níveis hierárquicos quanto desejado, de forma que os parâmetros das *prioris* também tenham suas próprias *prioris*, e assim por diante (MARTIN, 2016).



(a) Distribuição de frequências hipotética (b) Estrutura do modelo

Figura 2.2: Composição de um modelo hierárquico *bayesiano*: os dados no histograma fornecem indícios acerca da verossimilhança, cujos parâmetros deverão ser representados por outras distribuições de probabilidade

A inferência *bayesiana* também pode ser utilizada para regressões lineares, conforme esquema mostrado na Figura 2.3a. Uma regressão linear *bayesiana* é conduzida de modo que a variável prevista y seja descrita em relação ao preditor x segundo uma distribuição de probabilidade que contenha variabilidade residual em y. O modelo linear generalizado é descrito na Figura 2.3 de forma que y assuma uma distribuição normal com média μ e desvio ϵ, com μ seguindo uma relação linear de modo que $\mu = \alpha + \beta x$. Essa configuração implica que os dados estejam normalmente distribuídos ao redor da média (KRUSCHKE, 2014).



(a) (b)

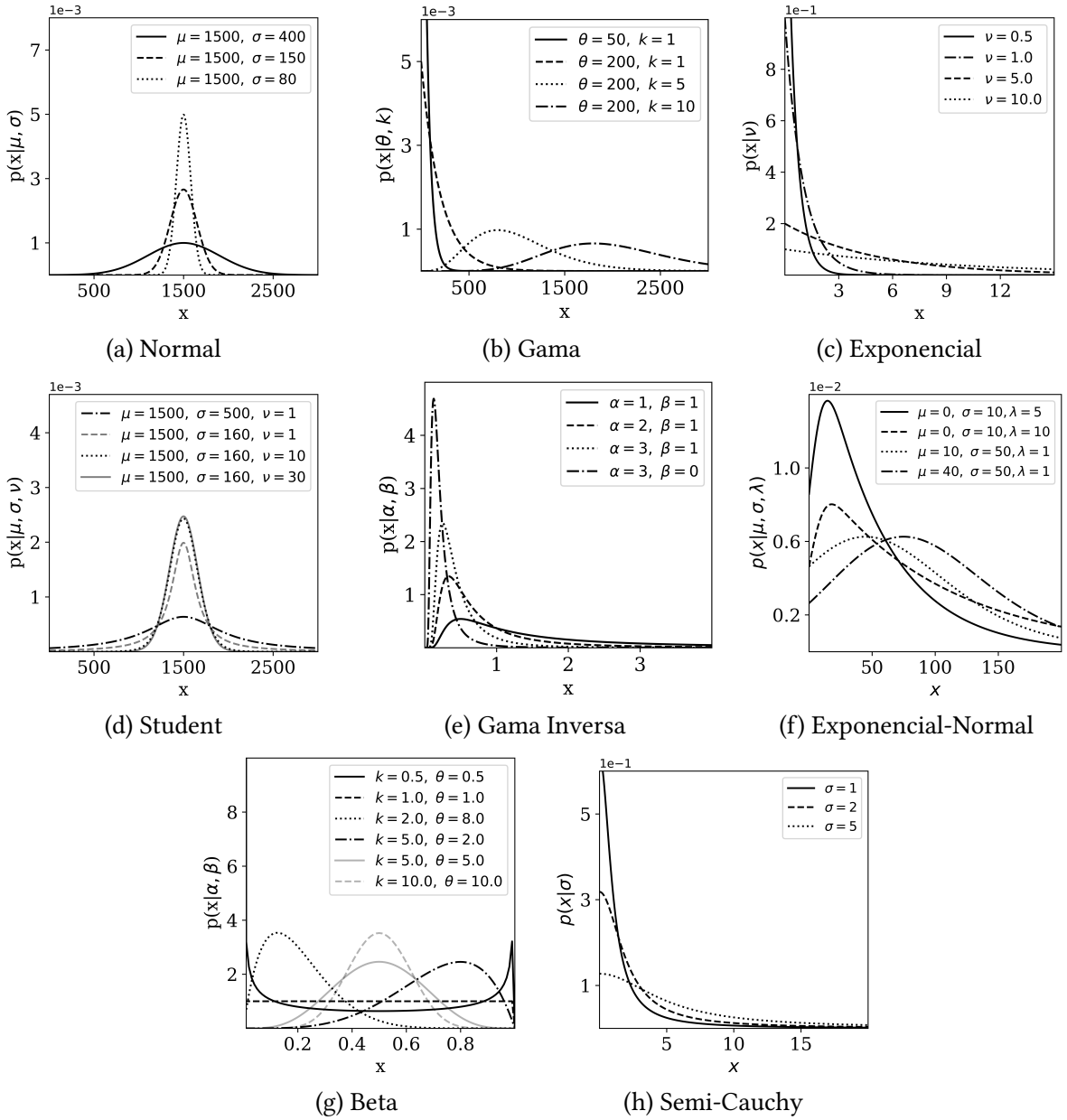
Figura 2.3: Esquema de implementação de uma regressão linear *bayesiana*

Algumas das distribuições de probabilidade mais usuais são mostradas na Figura 2.4. Na Figura 2.4a foram traçadas três distribuições normais, centralizadas em um valor comum μ e com desvios-padrão variáveis σ. De forma análoga, esses parâmetros também descrevem a distribuição de Student na Figura 2.4d, que conta também com o parâmetro ν de normalidade. Quanto menor o valor desse parâmetro, mais distantes se tornam as caudas da distribuição em relação à abscissa.

A distribuição gama apresentada na Figura 2.4b pode ser representada pelos parâmetros de forma k e de escala θ. Também é comum que esses parâmetros sejam chamados de α e β, em que $\beta = 1 / \theta$. É geralmente utilizada em modelagem *bayesiana* para descrever variâncias. Seu inverso resulta na distribuição gama inversa, conforme Figura 2.4e. A distribuição gama é de extrema relevância pois dela derivam outras distribuições, como a de Erlang, qui-quadrado e exponencial. Esta última, cuja forma está representada na Figura 2.4c, depende apenas do parâmetro ν. A distribuição normal exponencialmente modificada, na Figura 2.4f, apresenta a cauda mais longa para direita e é um convolução das distribuições normal e exponencial, dependendo dos parâmetros μ, σ e λ. A distribuição beta dependente dos parâmetros α e β, e é limitada ao intervalo [0,1], podendo assumir diversas configurações possíveis, como se pode observar na Figura 2.4g. Por fim, a distribuição de semi-Cauchy, dependente do parâmetro de escala σ, é usualmente utilizada para representar variâncias e consta na Figura 2.4h.

2.4.2 Diretrizes para definição das prioris e verossimilhança

O conhecimento para definir as distribuições a *priori* geralmente vem de revisões sistemáticas da literatura, da meta-análise dos dados ou de estudos sobre dados similares (O'HAGAN et al., 2006). Conforme destacado por Schoot e Depaoli (2014), o conhecimento de referência para as prioris pode ser incorporado na análise de forma que o método não dependa de grandes conjuntos de dados. Por este motivo, a inferência *bayesiana* pode lidar com situações em que apenas pequenas amostras estejam disponíveis. Outra funcionalidade das distribuições a *priori* se resume a estabilizar as inferências em espaços paramétricos complexos ou ou a acrescentar ruído ao modelo para agregar incerteza por meio do Teorema de Bayes (GELMAN; SIMPSON;



(a) Normal (b) Gama (c) Exponencial

(d) Student (e) Gama Inversa (f) Exponencial-Normal

(g) Beta (h) Semi-Cauchy

Figura 2.4: Distribuições de probabilidade comumente utilizadas em modelos hierárquicos *bayesianos*

BETANCOURT, 2017).

Quando o intervalo de variação dos parâmeros descritos pelas distribuições a *priori* é ilimitado, estas são chamadas de distribuições impróprias. Muitas vezes, prefere-se utilizar distribuições próprias por facilitar o uso de ferramentas inerentes à análise *bayesiana*, como o fator de Bayes (GELMAN et al., 2006), que consiste na razão entre duas distribuições posteriores. Os mesmos autores destacam que a distribuição a *priori* é considerada fracamente informativa se ela for configurada de forma que não acrescente qualquer informação prévia que se conheça para ser incorporada ao modelo.

Existem algumas diretrizes e recomendações para escolha das distribuições em modelos *bayesianos*. A distribuição de Student é útil como função de verossimilhança para a modelagem de *outliers* nos dados observados, sendo útil também em níveis hierárquicos mais altos de um modelo como distribuição a *priori* (KRUSCHKE, 2014). Já para a estimação de parâmetros de variância, é comum considerar distribuições impróprias, como a uniforme restrita a todos os valores positivos, e a gama-inversa com parâmetros alfa e beta tendendo a zero (GELMAN et al., 2006). Para evitar seu uso, Gelman et al. (2006) recomendam limitar a distribuição uniforme ao intervalo [0, 100], ou utilizar a distribuição normal-dobrada centrada em zero com $\sigma^2 = 100^2$. Caso seja possível propor uma distribuição a *priori* não informativa para a variância, Gelman et al. (2006) sugerem utilizar uma distribuição uniforme para conjuntos com mais do que 5 observações. Se o modelo contiver muitos parâmetros, é possível utilizar uma distribuição de Cauchy dobrada (também chamada de semi-Cauchy) com hiperparâmetros estimados do banco de dados.

É importante lembrar que duas distribuições a *priori* diferentes sempre convergirão para a mesma distribuição a *posteriori*, quanto maior for a quantidade de dados (MARTIN, 2016). Por fim, o analista deve usar a priori que melhor se ajuste ao seu caso, se ele obtiver prioris de referência (SIMPSON et al., 2017). Se a distribuição a *posteriori* não aparentar coerência ou aderência aos dados, supõe-se que haja conhecimentos prévios que não foram incluídos no modelo (GELMAN et al., 2006), de modo que seja necessário aprofundá-lo ou alterá-lo.

2.4.3 Avaliação dos resultados da distribuição posterior

Um modelo é considerado adequado se os dados recaem majoritariamente dentro da zona prevista e estão distribuídos aproximadamente da forma como o modelo prevê, e as discrepâncias observadas nas previsões são aleatórias e não sistemáticas. Se o conjunto de dados for muito grande e dominar a priori, ou se a priori for muito fraca ou informada de resultados anteriores e, por isso, for consistente com os dados recentes, então o ponto máximo da verossimilhança será próximo ao pico da distribuição posterior (KRUSCHKE, 2014).

Para verificar a qualidade dos modelos, Gabry et al. (2019) defendem que as ferramentas de visualização devem ser utilizadas em várias etapas da análise. Isso pode ser feito, por exemplo, por meio de análise exploratória dos dados para auxiliar na estruturação do modelo inicial,

ou por meio de verificações preditivas posteriores para observar a aderência do modelo aos dados. Além disso, é importante avaliar a qualidade das cadeias de Markov geradas. Para isso, em vez de simular apenas uma cadeia de previsão, são geradas múltiplas cadeias, inicializadas em diferentes pontos do espaço paramétrico, a partir das quais é calculada a estatística $\hat{R}$, que quantifica a variância entre elas. Segundo Vehtari et al. (2020), é recomendado utilizar pelo menos quatro cadeias e $\hat{R}$ não deve ultrapassar 1.01.

Outra estatística de referência é tamanho efetivo da amostra. Como o processo de Monte Carlo fornece elementos correlacionados, o tamanho efetivo da amostra fornece indícios acerca do grau de correlação entre as amostras. Esse valor é sempre menor do que o número absoluto amostrado e, por essa razão, é utilizado também para avaliar a convergência do modelo. Se o tamanho efetivo da amostra for muito pequeno, as expectativas calculadas provavelmente não são boas aproximações da distribuição posterior (VEHTARI et al., 2020). Os autores recomendam que o tamanho efetivo da amostra seja superior a 400. Tanto o $\hat{R}$ quanto o tamanho efetivo da amostra, bem como outros diagnósticos sobre os modelos, podem ser calculados em *Python* pela biblioteca ArviZ (KUMAR et al., 2019), desenvolvida especificamente para analisar modelos *bayesianos* estruturados com o *PyMC3*.

Caso não atendam aos limites especificados, essas estatísticas podem indicar problemas de amostragem nas cadeias, em que o algoritmo encontra dificuldades de se adaptar a regiões muito pequenas de alta densidade e baixo volume, ou a regiões muito amplas de grande volume e baixa densidade. Um outro problema comum em alguns modelos é a ocorrência de regiões achatadas da densidade posterior. Isso ocorre devido ao uso de distribuições a *priori* fracamente informativas, o que pode gerar uma cadeia de Markov que oscile muito entre valores extremos (MCELREATH, 2020). Isso é passível de ocorrer quando há uma pequena quantidade de dados. Nesses casos, uma alternativa para lidar com a autocorrelação consiste em aumentar a taxa de aceitação para possibilitar melhor exploração do espaço paramétrico.

Por fim, a análise *bayesiana* se resume aos processos de escolha de um modelo probabilístico, chamado de verossimilhança, com definições de distribuições para as quantidades observadas, os dados, e não observadas, as *prioris*, à obtenção da distribuição a *posteriori* dos valores não observados e à verificação do ajuste do modelo aos dados. Realizadas as etapas anteriores, a distribuição preditiva a *posteriori* permite realizar inferências sobre os valores não observados e hipóteses elaboradas, bem como estimativas e previsões sobre seus intervalos de credibilidade.

2.5 Considerações finais

Este Capítulo abordou os principais conceitos necessários à compreensão deste trabalho. Para tanto, foram abordados os principais parâmetros que descrevem a corrente de tráfego e para quais níveis de análise são utilizados. Em seguida, foi feita uma análise sobre as principais fontes de dados existentes para coleta de parâmetros de tráfego. A terceira parte tratou dos

modelos de tráfego já desenvolvidos e abordagens de previsão recorrentes na literatura e, por fim, foram apresentados os conceitos e ferramentas aplicados em inferência *bayesiana*.

Capítulo 3

Banco de dados

Este capítulo descreve o processo de coleta e organização dos dados e propõe uma forma de compatibilizar velocidades de sensores e da API, que possuem naturezas distintas, integradas num banco de dados a partir da hora e data de coleta comuns. A primeira etapa consistiu na seleção das seções de rodovias, de forma a garantir dados em quantidade suficiente para previsão e que contemplassem diferentes regimes de tráfego. Em seguida, foram tratados os dados históricos referentes aos sensores escolhidos. Posteriormente, foram estabelecidos critérios para a coleta de dados de tempo de viagem pela API, a fim de calcular a velocidade de forma indireta. Por fim, propôs-se um método de compatibilização entre as velocidades coletadas pela API e por sensores, a partir de uma regressão linear *bayesiana*.

3.1 Seleção de locais para coleta de dados de tráfego

A composição do banco de dados foi pensada de forma a possibilitar a integração entre dados históricos e dados *online*. Sendo assim, a escolha dos locais para coleta de dados de tráfego se limitou a trechos de rodovias com existência de infraestrutura de coleta por sensores, a fim de garantir a obtenção de dados históricos. Além disso, foi necessário garantir que os sensores estivessem em operação no momento das coletas em tempo real, a fim de possibilitar a posterior validação entre velocidades de sensores e velocidades obtidas indiretamente de fontes *online*. Em função da disponibilização dessas informações pela Agência de Transporte do Estado de São Paulo (ARTESP), a seleção dos locais de coleta se restringiu a rodovias compreendidas nesta região.

Optou-se por utilizar a *Distance Matrix API* do Google para coleta de dados de tempo de viagem *online*, pelo fato de ser uma plataforma mais difundida e, consequentemente, alcan-

çar grande número de usuários. Para escolha dos pontos de coleta, verificou-se, inicialmente, a existência de mais de 300 sensores alocados ao longo de toda malha rodoviária do estado de São Paulo. Entretanto, coletas de dados pela API têm custo por elemento requisitado e, embora a licença fornecida pela plataforma permitisse um número limitado de requisições gratuitas, seria inviável coletar dados referentes a todos os sensores disponíveis. Além disso, para contemplar o objetivo deste trabalho, seria interessante utilizar trechos que compreendessem diferentes regimes de tráfego, em que se observassem sobretudo a ocorrência de períodos congestionados, com o intuito de desempenhar análises mais abrangentes.

A fim de atender às restrições da cota gratuita fornecida pela API para criação do banco de dados, foi necessário realizar uma análise exploratória inicial do comportamento do tráfego na região de abrangência de alguns sensores. Dos mais de 300 pontos existentes, foram obtidos os dados de fluxo e velocidade de 74 sensores de rodovias paulistas em 133 sentidos de tráfego, cuja região de ocorrência está apontada na Figura 3.1, agregados em intervalos de 15 minutos, no mês de março de 2019. A fim de acrescentar ao processo de seleção dos locais de coleta um recurso para melhor visualização dos dados fornecidos pela fonte de informação escolhida, foi criada em linguagem *Python* uma rotina de coleta de imagens da aplicação do *Google Maps*, centralizadas na coordenada geográfica desses sensores, também a cada 15 minutos, que registra a cor do *pixel* mais próximo ao sensor. Assim, foi atribuída ao par fluxo-velocidade do sensor a cor da camada de tráfego disponibilizada pela aplicação *Google Maps*.

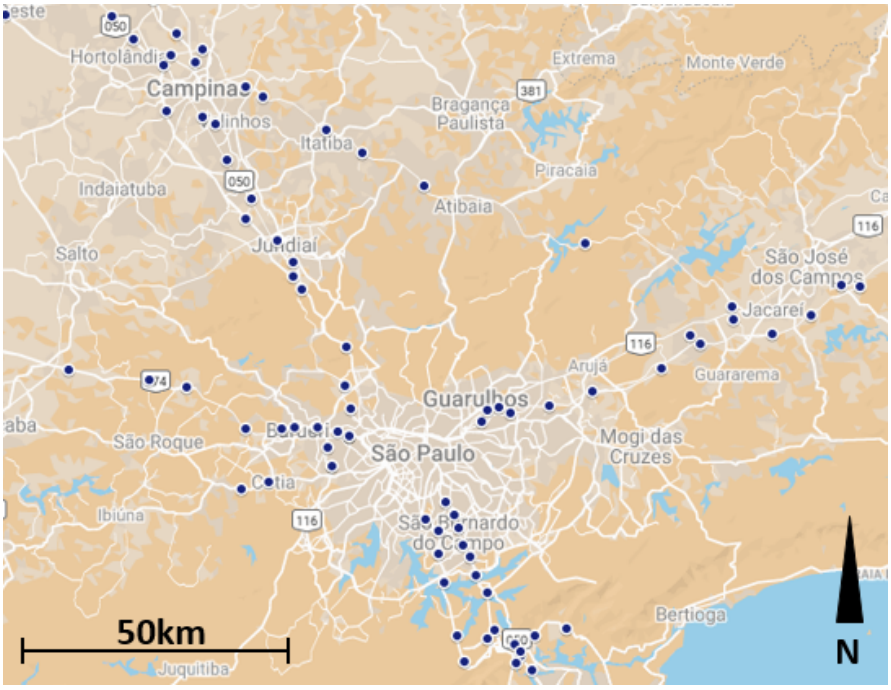
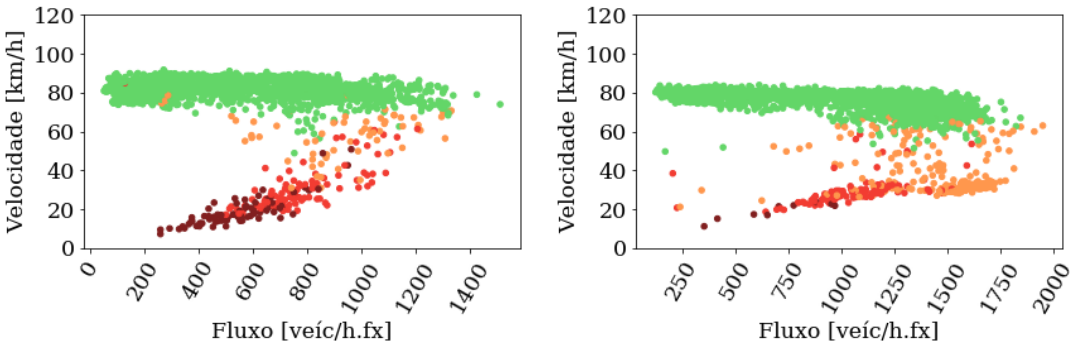


Figura 3.1: Localização dos sensores pré-selecionados no Estado de São Paulo. Adaptado de: Google Maps

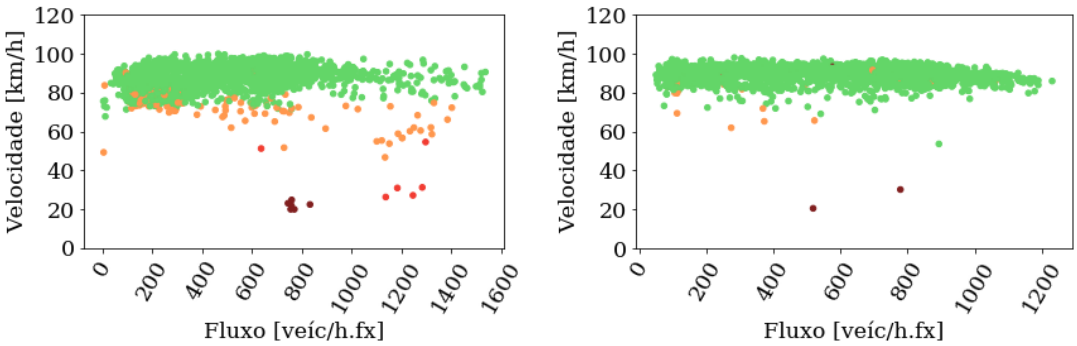
As Figuras 3.2 e 3.3 mostram quatro diagramas obtidos para o banco de dados de março de 2019, em que as cores dos pares fluxo-velocidade descrevem o estado do tráfego mostrado pela aplicação do Google Maps no mesmo instante de registro da coleta pelo sensor. Observa-se

que das quatro cores fornecidas pela aplicação do *Google*, o marrom e o vermelho tendem a se concentrar em regiões de fluxo congestionado, o verde em regiões de fluxo livre e o laranja em fluxos de transição, geralmente atingindo valores que se aproximam da capacidade.



(a) SP280, km 29.5, sentido leste (b) SP021, km 16.08, sentido leste

Figura 3.2: Diagramas fluxo-velocidade de sensores escolhidos para coleta pela API.

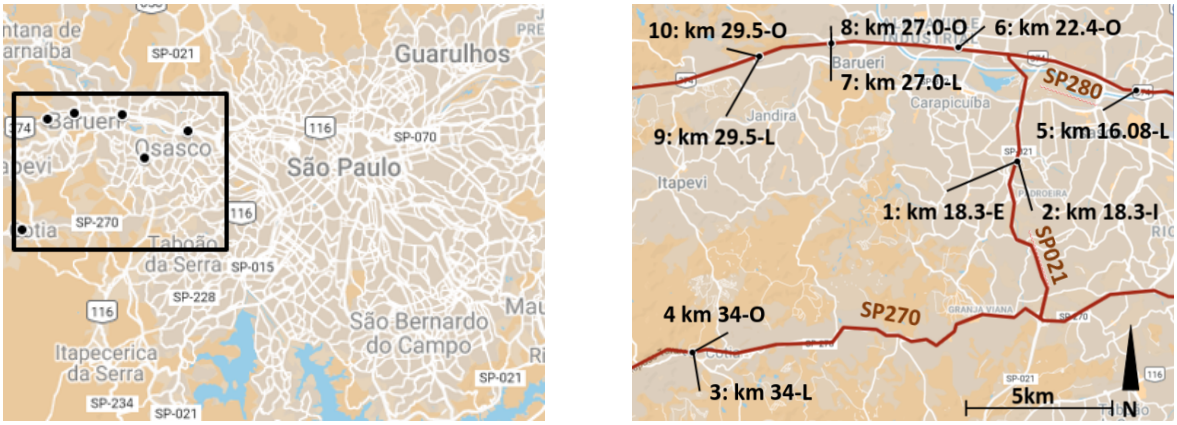


(a) SP280, km 36.9, sentido leste (b) SP021, km 16.08, sentido oeste

Figura 3.3: Diagramas fluxo-velocidade de sensores não escolhidos para coleta pela API.

A Figura 3.2 contém diagramas fluxo-velocidade para sensores selecionados para coleta pela API, onde se observou a ocorrência de fluxos congestionados, o que não ocorreu em grandes volumes de observações dos sensores da Figura 3.3. Nota-se também que as Figuras 3.2b e 3.3b se referem a sentidos opostos do mesmo sensor. Entretanto, o sentido leste apresentou fluxo direcional mais intenso no período avaliado e, por este motivo, foi selecionado para coleta pela API, enquanto o sentido oeste não foi contemplado.

Após a análise visual das curvas fluxo-velocidade, foram selecionados 10 trechos de tráfego referentes a 6 sensores, cujas localizações em rede estão indicadas na Figura 3.4. Todos os sensores se concentraram na região metropolitana da cidade de São Paulo. A fim de pressupor equivalência entre dados para períodos distintos, verificou-se que as rodovias não passaram por alterações significativas ao longo dos anos, como alteração das velocidades de fluxo livre ou relacionadas ao aumento de capacidade, de forma a manter a homogeneidade dos parâmetros de tráfego coletados. A Tabela 3.1 contém os identificadores para cada sentido utilizado, bem como as velocidades máximas permitidas para veículos leves e pesados em cada trecho viário.



(a) (b)

Figura 3.4: Localização dos sensores selecionados próximo à cidade de São Paulo. Adaptado de: Google Maps

Tabela 3.1: Velocidades máximas de operação (V_{op}) para rodovias dos sensores escolhidos

ID	Rodovia	km	Sentido	V_{op} [km/h] Veíc. Leves	V_{op} [km/h] Veíc. Pesados
1	SP021	018+300	Externo	100	80
2	SP021	018+300	Interno	100	80
3	SP270	034+000	Leste	80	80
4	SP270	034+000	Oeste	80	80
5	SP280	016+080	Leste	100	90
6	SP280	022+400	Oeste	100	90
7	SP280	027+000	Leste	100	90
8	SP280	027+000	Oeste	100	90
9	SP280	029+500	Leste	120	90
10	SP280	029+500	Oeste	120	90

3.2 Coleta de dados

O processo de coleta de dados foi divido em duas etapas. Primeiramente, reuniram-se dados históricos de sensores, que agregam tanto fluxos quanto velocidades, e em seguida, procedeu-se à coleta sincronizada de dados da API e dados de sensores para promover a integração entre dados da API e dados históricos de sensores por meio das velocidades.

3.2.1 Dados históricos de sensores

Prever demandas futuras requer a obtenção e análise de dados históricos e, portanto, coletar quantidade razoável de observações dos sensores, do tipo laço indutivo duplo, selecionados na Seção 3.1, foi uma das etapas deste trabalho. Pela indisponibilidade de dados mais recentes, o banco de dados históricos compreendeu o período entre 2011 e 2014, com valores médios de fluxos e velocidades agregados para todas as faixas de cada trecho em intervalos de cinco minutos.

Por ser comum a ocorrência de erros de medição dos equipamentos, excluíram-se os dias em que ocorreram erros evidentes de coleta ou longos períodos de mau funcionamento. Esses erros foram caracterizados pela ocorrência de registros zerados ou parcialmente existentes, ou fora dos limites aceitáveis para os parâmetros estudados. Isso foi feito com base em análise visual das séries temporais no período de um dia de curvas fluxo-velocidade, para todos os sensores ao longo de todo o período. Dias de operação atípica, com redução percebida da velocidade de fluxo livre em relação à velocidade máxima permitida, também foram removidos do banco de dados, já que no escopo deste trabalho, buscou-se dar ênfase à análise de operações típicas.

3.2.2 Dados da API

A coleta de dados pela *Distance Matrix API* do Google ocorreu no período compreendido entre os dias 1/7/2019 e 14/7/2019, das 6 às 20 horas para os trechos selecionados. Uma rotina em linguagem *Python* foi elaborada para efetuar as requisições à API a cada quinze minutos, segundo intervalo de agregação dos dados de sensores disponíveis neste mesmo período. Os parâmetros obrigatórios solicitados na requisição foram a chave de acesso e coordenadas de origem e destino. A Figura 3.5 mostra um esquema de como foi estabelecida a distância para definição das coordenadas para solicitação à API. A coordenada de destino é considerada alinhada ao sensor e a coordenada de origem deve ser escolhida em função da distância de percurso. Para definir essa distância, foi necessário levar em conta o fato de que os tempos de viagem são retornados pela API em segundos inteiros, com omissão da primeira casa decimal. Sendo assim, a informação omitida da casa decimal pode gerar conjuntos de velocidades numericamente equivalentes se calculadas a partir de distâncias muito curtas.

Para ilustrar esse efeito, escolheram-se duas distâncias hipotéticas, com comprimentos de 50 e 700 metros. Um veículo que, na prática, atravessa ambos os trechos com velocidade média de 100 km/h, demoraria 1.8 segundos no menor trecho e 25.2 segundos no maior. Entretanto, ao fazer a solicitação à *Distance Matrix API* entre o par origem-destino, o usuário recebe apenas a parte inteira do tempo de viagem, ou seja, 1 segundo e 25 segundos, respectivamente. Devido ao truncamento, pode-se pensar que o valor de 1 s tenha sido truncado a partir de 1.00 s ou 1.99 s, para a menor distância, ou então, 25.00 s ou 25.99 s, para a maior distância. Ao extrapolar a velocidade indiretamente pela razão entre distância e tempo de viagem, o intervalo de velocidades possíveis poderia variar de 90.4 a 180.0 km/h no primeiro caso, o que representa uma diferença de quase 100%. Já para a distância de 700 metros, esse intervalo varia de 97.0 a 100.8 km/h.

Para reduzir o efeito dos erros de truncamento, buscou-se uma distância que apresentasse erros relativos de no máximo 10%. Para isso, testou-se, segundo a mesma linha de raciocínio do exemplo anterior, o erro relativo para faixas de velocidades no intervalo de 20 a 120 km/h, em função de diferentes distâncias. A diferença entre ambas as velocidades em relação

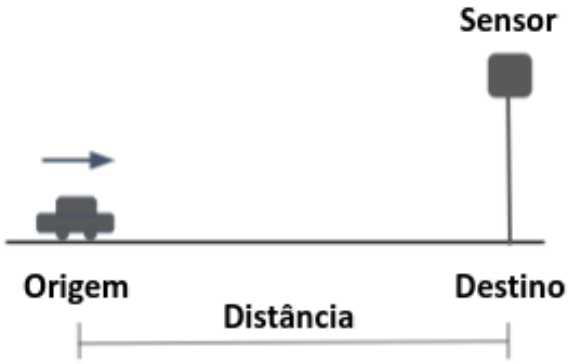


Figura 3.5: Posição das coordenadas de origem e destino para requisição à API

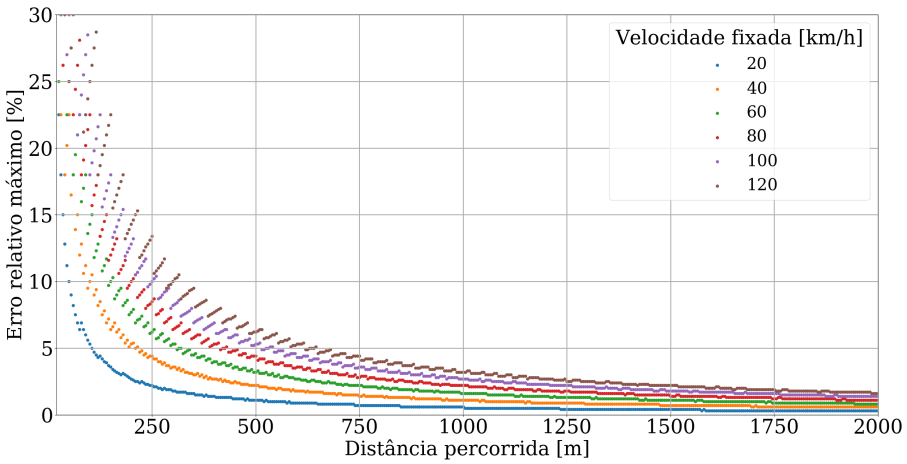


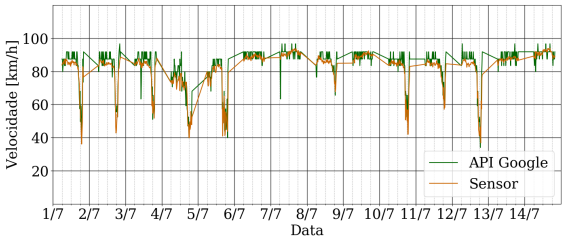
Figura 3.6: Erros relativos entre as velocidades máximas e mínimas possíveis para velocidades pré-fixadas, em função da variação das distâncias de percurso.

à velocidade de referência representam os valores dos erros relativos para cada distância. A Figura 3.6 mostra no eixo y os maiores erros relativos possíveis que podem ser obtidos para as velocidades calculadas indiretamente a partir dos tempos de viagem truncados fornecidos pela API. Observa-se que, para qualquer velocidade calculada com base em uma distância de 250 metros, os erros máximos variam de 2.5% para 20 km/h a 12.5% para 120 km/h, com tendência à redução dos erros à medida em que se aumenta a distância entre coordenadas de origem e destino. Enquanto distâncias muito curtas abrem margem grande para erros, distâncias muito longas, por outro lado, podem introduzir variações provenientes de outros fenômenos não relacionados à região de coleta do sensor. Dessa forma, decidiu-se adotar distâncias de aproximadamente 500 m para definir as coordenadas de origem e destino, cujos erros devem variar a não mais que 8%.

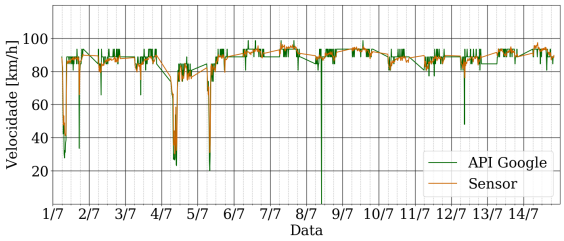
3.3 Validação entre dados de velocidade da API e de sensores

A fim de possibilitar a compatibilização entre os dados de sensores e da API, foram calculadas as velocidades no espaço pela razão entre as distâncias entre os pares O-D e tempos de viagem

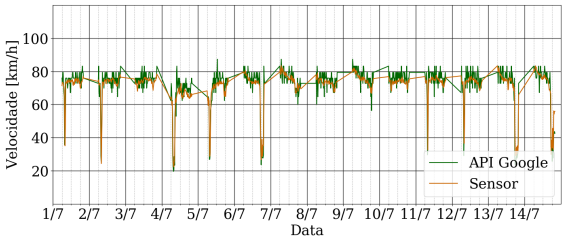
retornados pela API. Na Figura 3.7 são apresentadas as séries temporais das velocidades da API e de sensores para todas as 10 seções no período estudado. Observações subsequentes foram conectadas por linhas contínuas para facilitar a visualização da variação entre os registros de velocidades. Os períodos correspondem ao intervalo das 6 às 20 horas de cada dia e os pares de velocidades foram compatibilizados a partir do valor da velocidade média ao fim de cada período de agregação do sensor com o horário da velocidade coletada pela API.



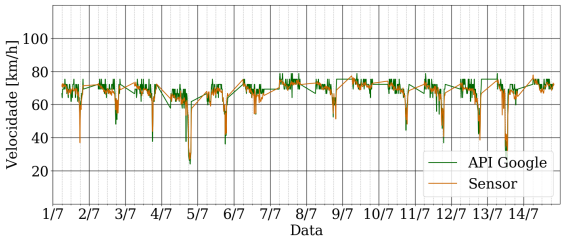
(a) Sensor 1 na SP-021, km 18.3, externo



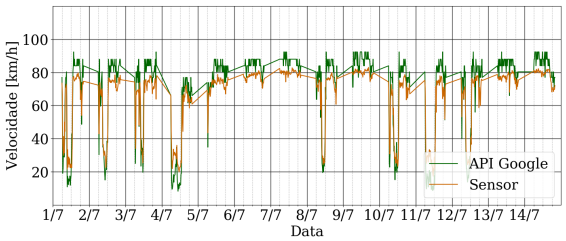
(b) Sensor 2 na SP-021, km 18.3, interno



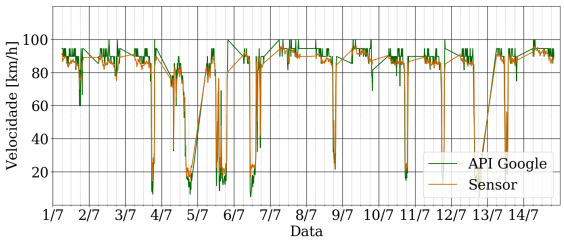
(c) Sensor 3 na SP-270, km 34.0, leste



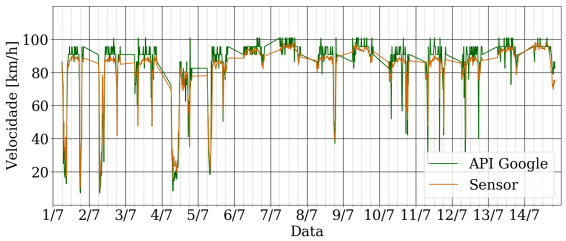
(d) Sensor 4 na SP-270, km 34.0, oeste



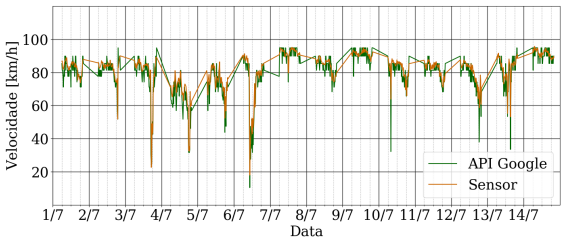
(e) Sensor 5 na SP-280, km 16.08, leste



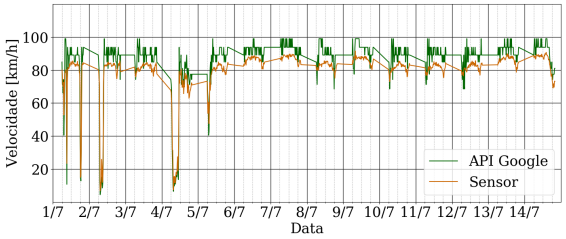
(f) Sensor 6 na SP-280, km 22.4, oeste



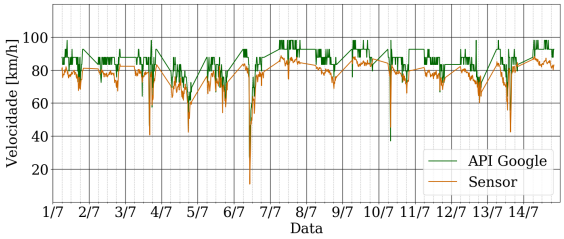
(g) Sensor 7 na SP-280, km 27.0, leste



(h) Sensor 8 na SP-280, km 27.0, oeste



(i) Sensor 9 na SP-280, km 29.5, leste



(j) Sensor 10 na SP-280, km 29.5, oeste

Figura 3.7: Séries temporais das velocidades calculadas a partir dos tempos de viagem da *Distance Matrix API* do Google e de sensores

Em função dos métodos distintos de coleta, velocidades coletadas pela API e por sensores

podem ser classificadas como grandezas no espaço e no tempo, respectivamente. Segundo as Equações 2.4 e 2.5, espera-se que velocidades no tempo resultem em valores maiores que velocidades no espaço, considerando a mesma região e período de ocorrência. Assim, uma velocidade no espaço, que é o caso da API, deve ser numericamente menor do que o valor de uma velocidade no tempo, que é o caso do sensor do tipo laço indutivo duplo. Esse princípio é observado em todos os casos para faixas de velocidades em regime congestionado. Entretanto, para velocidades próximas à de fluxo livre, observa-se o fenômeno inverso, e de modo mais pronunciado para os sensores 1, 5, 7, 9 e 10. Conforme mostrado na Figura 3.6, velocidades maiores tendem a ser mais sensíveis ao erro resultante do truncamento do tempo de viagem para a velocidade calculada pela API e, por este motivo, supõe-se a ocorrência do comportamento contrário ao esperado para faixas de velocidades próximas à de fluxo livre. Outro fator que também pode contribuir para gerar diferenças numéricas é o fato de que qualquer instrumento de medida utilizado para aferir grandezas físicas possui limitações e, portanto, erros de medida ou de precisão com relação à coleta de velocidades, tanto de sensores quanto pela API, também podem ocorrer.

3.3.1 Modelo *bayesiano* para validação de velocidades da API

Para validar as velocidades calculadas pela API, propôs-se uma regressão linear *bayesiana* a fim de correlacionar velocidades de sensores e da API. Dessa forma, buscou-se extrair as variações inerentes às naturezas no tempo e no espaço de ambas as fontes de velocidades, bem como incorporar o caráter probabilístico dos erros de medição. A ideia central do modelo proposto consistiu em obter variâncias a partir das velocidades observadas e validá-las segundo as Equações 2.4 e 2.5. Para isso, estruturou-se o modelo hierárquico *bayesiano* representado na Figura 3.8. Nesta estrutura, tanto as velocidades da API como as dos sensores foram consideradas alternadamente preditores e variáveis previstas.

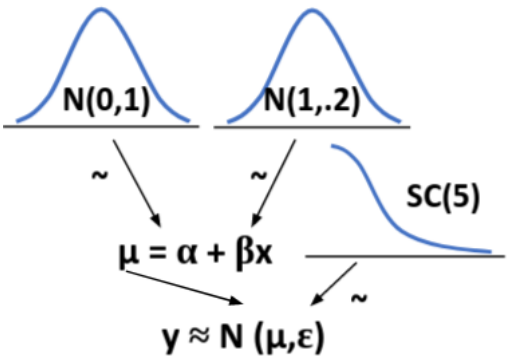


Figura 3.8: Modelo *bayesiano* para regressão linear entre velocidades de sensores e da API. As distribuições utilizadas são a normal (N(μ, σ) e semi-Cauchy (SC(σ)

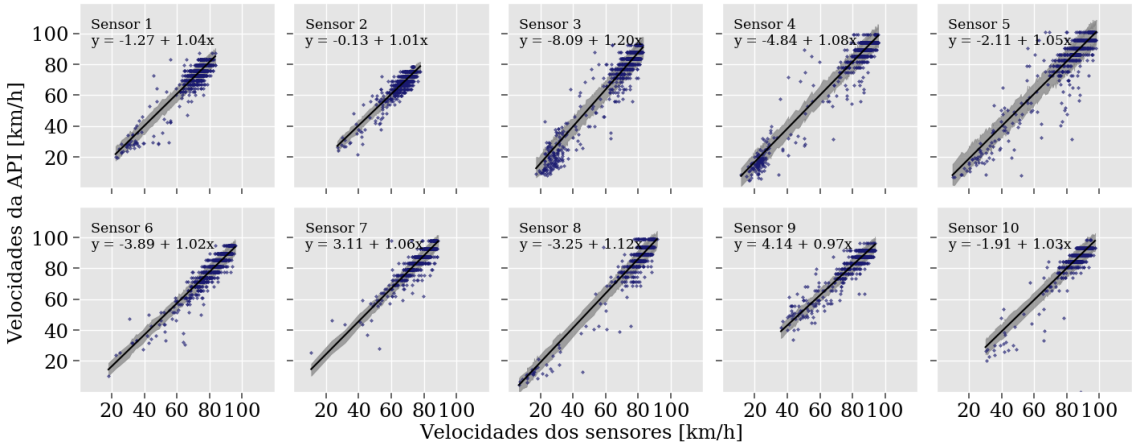
Para o valor de *alfa*, adotou-se uma distribuição a *priori* normal com média zero e desvio padrão igual a um. Esta estrutura pressupôs ignorância acerca das naturezas distintas entre ambas as velocidades e sugeriu que elas fossem fortemente correlacionadas. A fim de reforçar

tal hipótese, o valor unitário atribuído ao desvio padrão restringiu a variação do coeficiente linear em apenas uma unidade. Ao parâmetro *beta* também foi atribuída uma distribuição normal, com média igual a um, considerando que a variação em uma unidade do preditor correspondesse à variação em uma unidade na variável prevista. O desvio padrão de 0.2 unidades conferiu baixa variabilidade ao coeficiente angular e reforçou a suposição de que ambas as velocidades fossem fortemente correlacionadas. A partir dos parâmetros *alfa* e *beta*, foi obtido um valor determinístico *mu* para a velocidade média, ao qual foi incorporada a incerteza a partir de uma distribuição semi-Cauchy a *priori*, com parâmetro de escala igual a cinco.

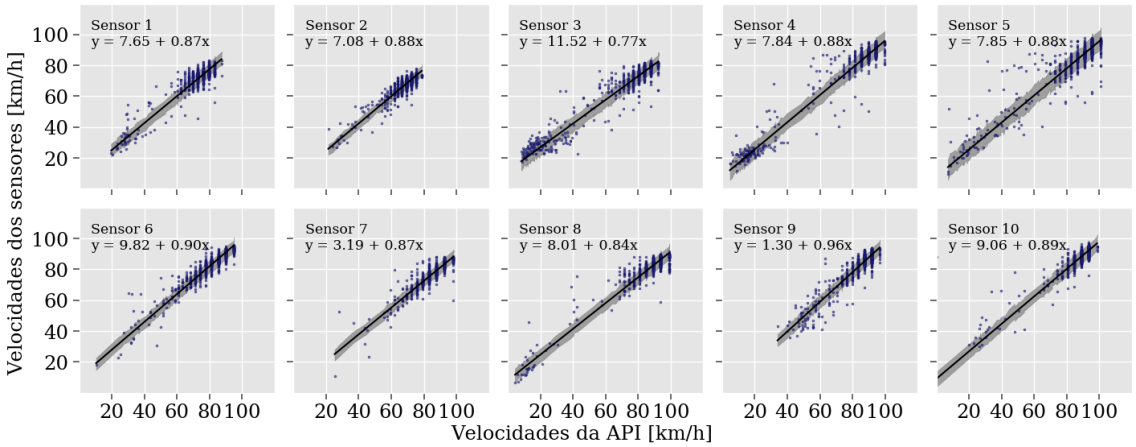
Após definição do modelo de regressão linear *bayesiana*, a integração do espaço posterior foi feita usando a biblioteca PyMC3 (SALVATIER; WIECKI; FONNESBECK, 2016), na versão 3.7, desenvolvida especificamente para estruturar e estimar modelos *Bayesianos* em Python, na versão 3.6.5. Outras bibliotecas utilizadas foram a Arviz 0.5.1 (KUMAR et al., 2019) para gerar estatísticas de modelos, e Theano 1.0.4 (Theano Development Team, 2016), utilizada pelo PyMC3 para resolver expressões matemáticas em espaços multidimensionais. Para a integração do espaço paramétrico e inicialização das cadeias, foi utilizada a configuração automática fornecida pelo PyMC3. Duas cadeias de 3000 amostras da distribuição posterior foram simuladas com mais 1000 amostras de *tuning*, totalizando 8000 iterações para cada sensor. Para fins de verificação, os traços de α, β e ϵ foram visualizados para conferir a convergência das cadeias de Markov.

A Figura 3.9 apresenta as regressões obtidas para todos os sensores. A Figura 3.9a refere-se às regressões em que as velocidades dos sensores foram utilizadas como preditores e as velocidades da API como variáveis previstas, enquanto a Figura 3.9b apresenta as regressões em que as velocidades da API foram utilizadas como preditores e velocidades dos sensores as variáveis previstas. Para obter as variâncias das velocidades a fim de utilizá-las nas Equações 2.4 e 2.5, foram calculados os desvios-padrão referentes aos intervalos de probabilidade de 68%, destacados em escala de cinza. Esse valor de 68% foi escolhido por representar uma unidade do desvio-padrão da distribuição normal. Os coeficientes para α e β são mostrados em cada regressão e correspondem aos valores médios das distribuições posteriores.

Como exemplo, observam-se na Figura 3.10 os traços posteriores gerados para o Sensor 1, considerando as velocidades da API como preditores, cujos valores médios de 1.30 para α e 0.96 para β foram os mesmos exibidos na linha de regressão para o mesmo sensor na Figura 3.9. Assim como neste exemplo, os traços posteriores para os três parâmetros dos demais sensores também assumiram configuração normal. A fim de analisar a convergência do modelo, notou-se que as duas cadeias concordaram entre si e os traços apresentaram boa mistura para todos os parâmetros, o que indica uma exploração adequada do espaço paramétrico. Os gráficos de autocorrelação, como o da Figura 3.11 referente ao sensor utilizado como exemplo, também reforçam a convergência das cadeias, uma vez que seus valores tenderam a zero à medida que o número de iterações aumentou. Estes padrões de convergência foram observados para todas as regressões realizadas.



(a) Regressões utilizando velocidades dos sensores como preditores e velocidades da API como variáveis previstas



(b) Regressões utilizando velocidades da API como preditores e velocidades de sensores como variáveis previstas

Figura 3.9: Linhas de regressão com intervalos a 68% de probabilidade em escala de cinza. Os pontos em azul se referem aos dados observados

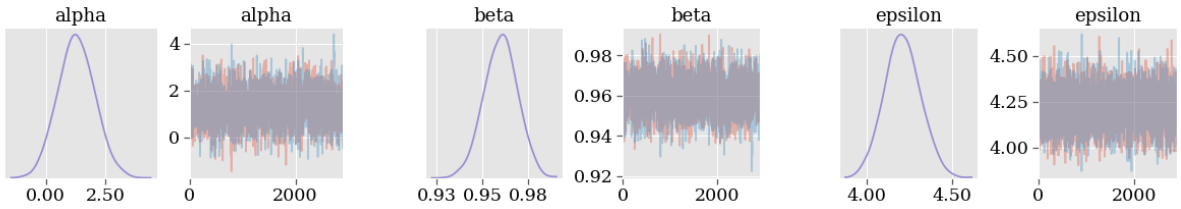


Figura 3.10: Traços dos parâmetros α, β and ϵ, respectivamente, utilizando velocidades da API como preditores para o Sensor 1

Como as variáveis previstas são modeladas a partir de uma distribuição normal, um valor médio de velocidade equivalente, fornecido pela reta de regressão, e o seu desvio-padrão, gerado por ϵ, devem ser obtidos como resultado. A partir das verificações preditivas posteriores, obtemos os intervalos de máxima probabilidade para as variáveis previstas. Da distribuição normal, coletamos o intervalo correspondente a um desvio padrão, que corresponde aos limites de probabilidade de aproximadamente 68% da distribuição normal, e o valor médio previsto. Este procedimento é realizado para todas as observações e é utilizado para calcular as veloci-

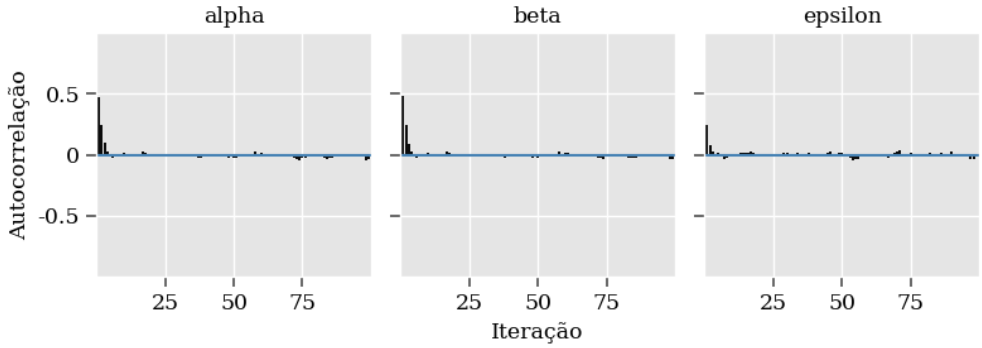


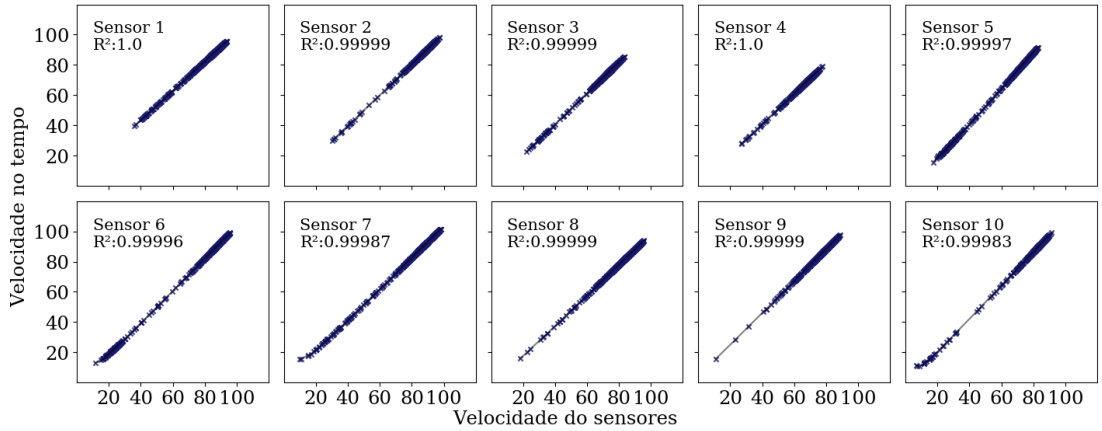
Figura 3.11: Autocorrelação para os parâmetros α, β and ϵ, respectivamente, utilizando velocidades da API como preditores para o Sensor 1

dades equivalentes a partir das Equações 2.4 e 2.5. A vantagem desta abordagem consiste no fato de que ela permite a extração de desvios-padrão dos dados observados sem a necessidade de calculá-los dentro dos períodos de agregação, que raramente são fornecidos pelos sistemas de coleta de dados de tráfego.

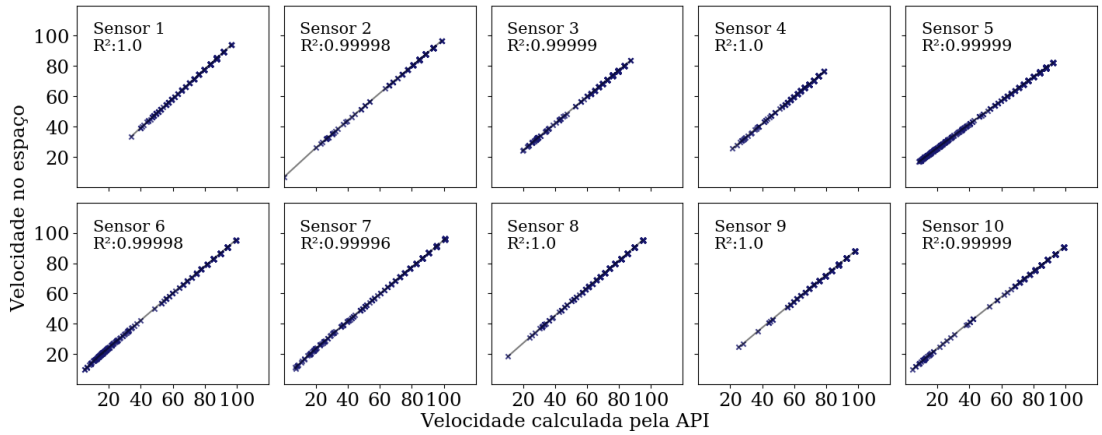
Após estimar as velocidades no tempo ou no espaço a partir das velocidades observadas, calculou-se o coeficiente de correlação Pearson, cujos resultados são mostrados na Figura 3.12. Pode-se notar que todos os valores de R^2 resultaram acima de 99%. Isto indica que o modelo linear *bayesiano* proposto é um método válido para correlacionar as velocidades de sensores e da API, uma vez que ele captura não apenas os valores médios equivalentes entre velocidades no tempo e no espaço, mas também suas variâncias.

3.4 Considerações finais

Este Capítulo descreveu o processo de obtenção e tratamento de um banco de dados históricos de sensores contendo dados de fluxos e velocidades. Adicionalmente, foi proposta uma regressão linear *bayesiana* para compatibilização entre velocidades de sensores e velocidades calculadas a partir de tempos viagem de uma API de tráfego, que gerou valores médios equivalentes entre velocidades no tempo e no espaço e também suas variâncias. Como próxima etapa, os intervalos gerados pela regressão linear *bayesiana* entre velocidades de API e de sensores serão usados para filtrar o banco de dados históricos a fim de gerar conjuntos de fluxos que serão usados para previsão.



(a) Velocidades no tempo estimadas a partir da relação de Wardrop



(b) Velocidades no espaço estimadas a partir da relação de Rakha

Figura 3.12: Velocidades estimadas convertidas pelas relações de Wardrop e Rakha, calculadas a partir de velocidades observadas

Capítulo 4

Previsão de fluxo de tráfego

Este capítulo propõe um método estatístico para previsão de fluxo de tráfego para as seções de rodovias selecionadas no capítulo 3, a partir de tempos de viagem coletados pela API e de um banco de dados históricos de sensores. Para cada previsão, foram utilizados como dados de entrada no modelo a velocidade calculada a partir do tempo de viagem da API, o dia da semana, a hora do dia e o horizonte de previsão desejado. Este procedimento retornou grupos de fluxos, utilizados como dados de entrada para um modelo hierárquico *bayesiano*, que retornou distribuições posteriores representativas das condições segundo as quais foram gerados, a partir das quais foram propostas medidas de avaliação de acurácia e dos intervalos previstos.

4.1 Agrupamento dos dados de fluxo

Estimar valores futuros requer a obtenção de dados históricos, de modo a estruturar conjuntos de dados representativos das condições segundo as quais foram gerados. Uma prática comum em trabalhos de previsão de tráfego veicular consiste em agrupar dados típicos de tráfego em função do dia da semana e horário de ocorrência. Além disso, conforme mostrado na Seção 3.3.1, velocidades observadas pela API podem ser relacionadas a intervalos de velocidades de sensores por meio de uma regressão linear *bayesiana*. A partir desses intervalos de probabilidade, é possível supor que, dada uma velocidade calculada pela API, exista um conjunto de de velocidades equivalentes de sensores que represente condições de ocorrência equivalentes.

Com essas premissas em mente, estruturou-se um procedimento de agregação de dados históricos de fluxo de tráfego que respondesse à seguinte pergunta: "Dada uma velocidade V, em uma seção de rodovia S, registrada às H horas de um dia da semana D, qual será o possível valor para o fluxo veicular Q, naquela seção, daqui a M minutos?". A ideia consiste em, a

partir de uma velocidade calculada em tempo real de um tempo de viagem fornecido pela API, agregar todos os fluxos históricos que contemplem as velocidades do sensor equivalentes à velocidade de entrada da API, observados nos mesmos dia da semana e horário de ocorrência, para um dado horizonte de previsão, a fim de obter uma distribuição de fluxo de tráfego para aquela seção da via.

Como os dados de sensores referentes ao mesmo período da coleta pela API (de 1/7/2019 a 14/7/2019) estavam agregados em intervalos de 15 minutos, foram estabelecidos horizontes de previsão de curto prazo de 0, 15, 30 e 60 minutos. Assim, foram gerados mais de 27800 subconjuntos de dados segundo os critérios de agregação mencionados, considerando todos os dez sensores utilizados nesta pesquisa. Diferentemente da etapa anterior, em que se utilizou um intervalo de 68% a fim de obter o desvio padrão para validação pelas Equações 2.4 e 2.5, optou-se por utilizar nesta etapa o intervalo de probabilidade de 95% para definição dos limites inferior e superior de velocidades de sensores. Como o intuito não é obter um valor para o desvio padrão, mas sim um intervalo de velocidades equivalentes, este valor foi escolhido por contemplar a maior parte dos dados da uma distribuição normal resultante do modelo de regressão linear *bayesiana* proposto na Seção 3.3.1. As etapas de agregação dos dados de fluxo foram realizadas na seguinte sequência:

1. Com a velocidade calculada pela API (V_{API}), definiram-se as velocidades máximas e mínimas equivalentes de sensores para o intervalo de máxima densidade de probabilidade de 95%;

2. Segundo as velocidades existentes nesse intervalo, filtrou-se o banco de dados históricos de sensores;

3. Desse subconjunto, foram selecionados todos os registros que ocorreram no mesmo horário H e dia da semana D em que foi observada a V_{API};

4. Desse novo subconjunto, foram registrados todos os valores no formato *datetime*;

5. A cada registro *datetime*, foram somados diferentes horizontes M de previsão;

6. A partir do conjunto de valores *datetime* registrados para cada horizonte de previsão, filtraram-se todos os fluxos Q existentes no banco de dados históricos.

A Figura 4.1 apresenta a visualização do conjunto de dados históricos para as diferentes etapas do processo de filtragem descritos anteriormente. Neste exemplo, foi escolhido o Sensor 3, localizado na rodovia SP270, km 34, sentido Leste, cuja velocidade calculada pela API foi de 72.8 km/h e ocorreu às 6 horas da manhã de uma segunda-feira. O intervalo de velocidades do sensor equivalentes à velocidade calculada pela API, nesse caso, compreendeu valores entre 62.3 e 80.1 km/h, para o intervalo de credibilidade de 95% gerado pela regressão linear *bayesiana*.

	Data	Dia	Hora	Velocidade	Volume	Datetime
10	2013-02-28	3	23:05:00	86.0	264	1362103500
200	2013-02-28	3	07:15:00	75.0	1908	1362046500
300	2013-02-27	2	22:55:00	81.0	330	1362016500
400	2013-02-27	2	14:35:00	83.0	1050	1361986500
700	2013-02-26	1	13:35:00	76.0	1104	1361896500
800	2013-02-26	1	05:15:00	86.0	792	1361866500
900	2013-02-25	0	20:55:00	81.0	438	1361836500
1000	2013-02-25	0	12:35:00	81.0	1338	1361806500
1200	2013-02-24	6	19:55:00	75.0	1290	1361746500

(a) Dados na coluna *Velocidade*, contendo velocidades históricas de sensores, foram filtrados segundo os intervalos gerados pela regressão linear *bayesiana* entre velocidades de sensores e da API. A primeira coluna refere-se ao identificador do registro no banco de dados.

	Data	Dia	Hora	Velocidade	Volume	Datetime
19811	2012-12-17	0	06:00:00	75.0	1104	1355731200
33059	2013-07-29	0	06:00:00	79.0	900	1375088400
35075	2013-07-22	0	06:00:00	80.0	888	1374483600
39107	2013-07-08	0	06:00:00	79.0	786	1373274000
41123	2013-07-01	0	06:00:00	72.0	1038	1372669200
65591	2013-06-24	0	06:00:00	80.0	972	1372064400

(b) Dados filtrados conforme dia da semana e horário de observação. Deste processo, são selecionados os valores da coluna *Datetime*

	Data	Dia	Hora	Velocidade	Volume	Datetime
19808	2012-12-17	0	06:15:00	73.0	1266	1355732100
33056	2013-07-29	0	06:15:00	77.0	1116	1375089300
35072	2013-07-22	0	06:15:00	80.0	1122	1374484500
39104	2013-07-08	0	06:15:00	78.0	768	1373274900
41120	2013-07-01	0	06:15:00	72.0	1290	1372670100
65588	2013-06-24	0	06:15:00	83.0	1068	1372065300

(c) Aos datetimes filtrados anteriormente foram acrescidos, nesse exemplo, 900 segundos, equivalentes a um horizonte de previsão de 15 minutos. A partir desses valores, filtrou-se o banco de dados históricos novamente e foram agregados todos os fluxos existentes nesse subconjunto

	Datetime	Dia	Hora	Horizonte	Vsensor	Fluxo	Vapi	Velocidades	Fluxos
1616	08/07/2019 06:00:00	0	6	0	77.95	812	72.75	[75.0, 79.0, 80.0, 79.0, 72.0, 80.0]	[1104.0, 900.0, 888.0, 786.0, 1038.0, 972.0]
1617	08/07/2019 06:00:00	0	6	15	74.94	920	72.75	[73.0, 77.0, 80.0, 78.0, 72.0, 83.0]	[1266.0, 1116.0, 1122.0, 768.0, 1290.0, 1068.0]
1618	08/07/2019 06:00:00	0	6	30	75.99	956	72.75	[76.0, 78.0, 81.0, 78.0, 72.0, 83.0]	[1248.0, 1428.0, 1482.0, 996.0, 1350.0, 1404.0]
1619	08/07/2019 06:00:00	0	6	60	79	1014	72.75	[78.0, 75.0, 83.0, 82.0, 72.0, 77.0]	[1398.0, 1530.0, 1416.0, 1038.0, 1710.0, 1854.0]

(d) O banco de dados final com todos os conjuntos de fluxos agregados apresenta a seguinte configuração: A cada localidade, horário e dia de ocorrência, velocidade da API e horizonte de previsão, é gerado um conjunto distinto de fluxos históricos

Figura 4.1: Detalhamento do procedimento de filtragem do banco de dados históricos para geração de conjuntos de fluxos pra a velocidade de 72.8 km/h calculada pela API e compreendida entre os intervalos de 62.3 e 80.1 km/h gerados pela regressão linear *bayesiana*, observada às 6 horas da manhã do dia 1/7/2019 (segunda-feira), para o sensor localizado na rodovia SP270, km 34, sentido Leste

A visualização dos conjuntos de fluxos gerados após agregação do banco de dados consistiu na etapa preliminar para definição do modelo de previsão, a fim de obter indícios acerca da configuração aproximada dos grupos de fluxo. A partir dessas configurações, foi possível propor quais as distribuições de probabilidade seriam mais adequadas para representar a distribuição de verossimilhança de um modelo hierárquico *bayesiano*. Uma abordagem intuitiva de visualização consiste na geração das distribuições de frequências. Entretanto, como fluxos são variáveis contínuas e buscam-se distribuições de probabilidade que os representem, utilizou-se uma função de geração de distribuições pelo método KDE (*Kernel Density Estimation*), disponibilizado pela biblioteca Seaborn (WASKOM et al., 2018), um algoritmo de estimação de densidade de probabilidade, que busca aproximar a distribuição geradora dos dados. Para isso, foram utilizadas as configurações padrão disponibilizadas pela biblioteca.

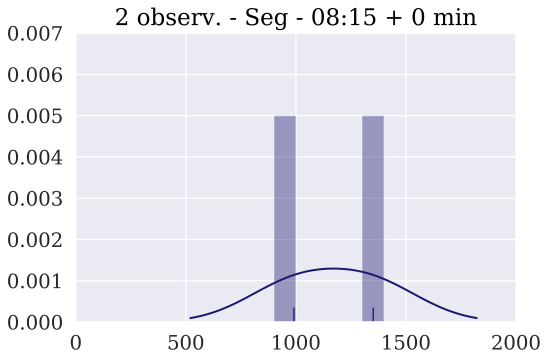
Na Figura 4.2 constam alguns exemplos dos conjuntos de fluxo obtidos. Os agrupamentos foram selecionados aleatoriamente, de forma a contemplar diferentes sensores e quantidades de dados, bem como velocidades distintas e para diferentes horizontes de previsão. Por inspeção, conjuntos com menor quantidade de dados, como os das Figuras 4.2a a 4.2d, apresentaram maior variabilidade nas formas das distribuições, podendo assumir ou não formas regulares. Já os conjuntos como os das Figuras 4.2e a 4.2h tenderam a apresentar, de modo geral, formas mais regulares, à medida em que se observou maior quantidade de dados.

4.2 Estruturação do modelo de previsão

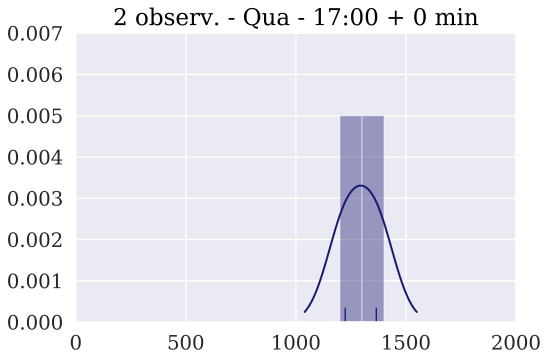
Nesta etapa, preocupou-se com a elaboração de um modelo de previsão que, além de poder fornecer valores pontuais de fluxos veiculares a partir de uma velocidade calculada pela API, e levando-se em conta outros fatores, como hora de ocorrência e dia da semana, também pudesse prever intervalos de credibilidade para esses parâmetros. A melhor forma de descrever esses processos é por meio da geração de distribuições de probabilidade. Conforme discutido no Capítulo 2, uma abordagem que se popularizou nos últimos anos por fornecer esses dois recursos é a estatística *bayesiana*. Sendo assim, as próximas seções descrevem as etapas para a estruturação de um modelo hierárquico *bayesiano* para previsão de fluxo de tráfego veicular.

4.2.1 Definição da distribuição de verossimilhança

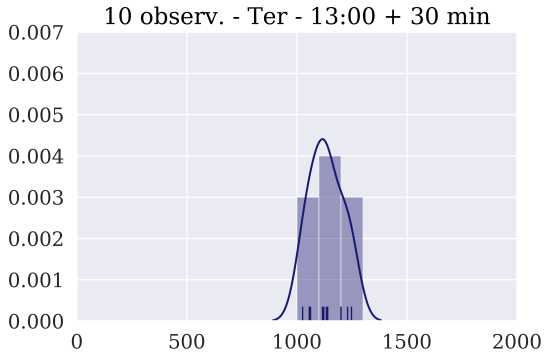
Para propor um modelo de previsão que represente os dados observados, é necessário descobrir qual a configuração aproximada de seus dados. Assim, a partir das inspeções visuais dos grupos de fluxos gerados, cujas amostras foram apresentadas na Figura 4.2, foram propostas algumas distribuições que pudessem representar os conjuntos obtidos. Inicialmente, supôs-se que as formas assumiriam configurações próximas à distribuição normal, descrita por uma média μ e um desvio padrão σ, já que tenderam, em sua maioria, a um valor central, e apresentaram dispersões variando em maior ou menor escala, frequentemente de forma simétrica ao redor da média. Entretanto, a distribuição normal não prevê com fidelidade as observações atípicas,



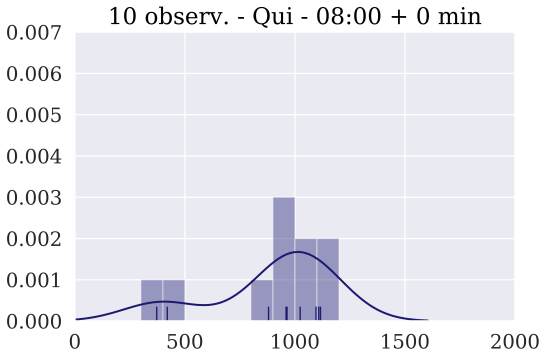
(a) Distribuição de fluxos para o sensor 2



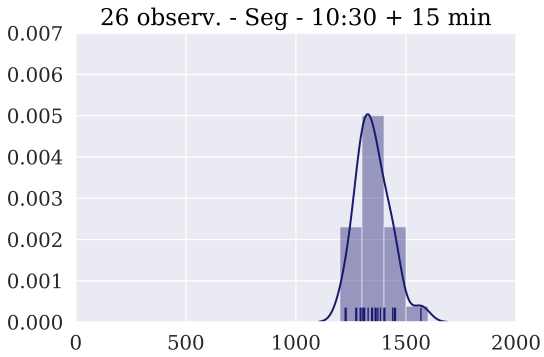
(b) Distribuição de fluxos para o sensor 2



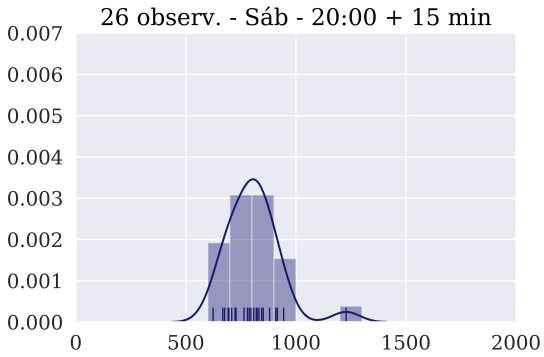
(c) Distribuição de fluxos para o sensor 3



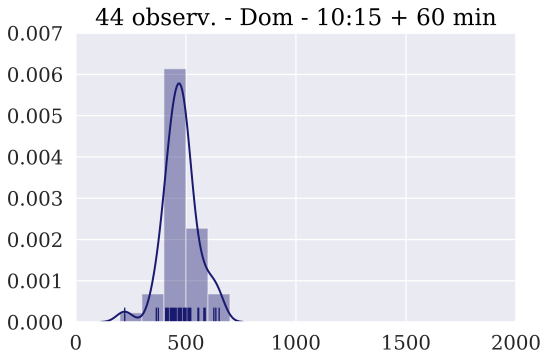
(d) Distribuição de fluxos para o sensor 9



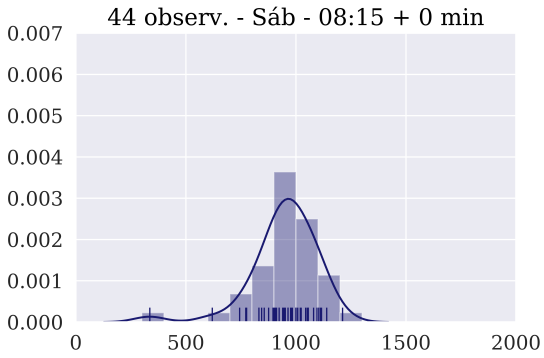
(e) Distribuição de fluxos para o sensor 8



(f) Distribuição de fluxos para o sensor 8



(g) Distribuição de fluxos para o sensor 1



(h) Distribuição de fluxos para o sensor 9

Figura 4.2: Distribuições de fluxos históricos agregados em função de diferentes velocidades observadas pela API, para diferentes dias da semana, horários, horizontes de previsão e locais de observação.

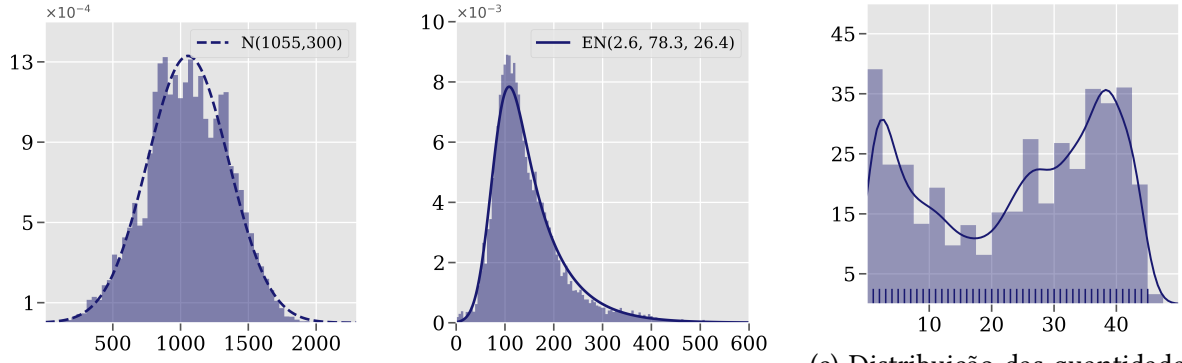
muito distantes dos valores centrais, percebidas com recorrência em vários conjuntos de fluxo, como no caso das Figuras 4.2d e 4.2h, já que observações situadas nos extremos tendem a deslocar a média em sua direção.

Uma forma de evitar este comportamento é a partir do uso da distribuição de Student, já que as suas caudas pesadas a tornam mais robusta à presença de *outliers* e, portanto, apta a incorporar maior variabilidade (MARTIN, 2016). Os três parâmetros que a descrevem são a média μ, que representa a tendência central, o parâmetro de escala σ, que atua de forma análoga ao desvio padrão da normal, e o parâmetro de normalidade ν, equivalente ao tamanho da amostra, que determina a forma da distribuição. Quanto menor este último parâmetro, mais pesadas tendem a ser as caudas, ou seja, mais variabilidade é incorporada, e à medida que este valor aumenta, a distribuição de Student tende a se aproximar da normal. Assim, com o intuito de estabelecer um modelo universal, adaptável a diversos tamanhos de banco de dados, a distribuição de Student foi escolhida para representar a verossimilhança dos fluxos no modelo *bayesiano*.

Após a definição da distribuição de verossimilhança, a próxima etapa para estruturar o modelo de previsão consistiu em buscar distribuições de probabilidade que pudessem descrever cada um dos três parâmetros da distribuição de Student. Em geral, nesta etapa se incorporam os conhecimentos já existentes a respeito desses parâmetros. Os valores para fluxos rodoviários encontrados na literatura tendem variar entre 0 a 2500 veículos por hora por faixa. Com isso em mente, uma alternativa seria utilizar, para representar as médias, uma distribuição uniforme cujos valores estivessem compreendidos nesse intervalo. Para os demais parâmetros, entretanto, dada a particularidade do método de geração dos conjuntos de dados proposto neste trabalho, não foi possível contar com evidências referenciadas na literatura para embasar a adoção de valores que os representassem. Por este motivo, realizou-se uma análise exploratória com todos os conjuntos de fluxos gerados.

Como a distribuição de Student é representada por um valor de tendência central, uma medida de escala e o parâmetro de normalidade, os grupos de fluxos foram analisados de forma a evidenciar essas medidas. Sendo assim, foram calculados os desvios-padrão amostrais e médias de cada conjunto de fluxo. Como ao parâmetro ν é comumente atribuído o tamanho da amostra, gerou-se o histograma dessa variável em função do tamanho dos conjuntos de dados gerados. As distribuições de frequências de cada uma dessas variáveis estão expostas na Figura 4.3.

Desse procedimento, foram observados padrões nos subconjuntos das médias e dos desvios-padrão. Como se pode observar na Figura 4.3a, o histograma das médias assumiu uma configuração aproximadamente normal. Foram calculados a média e o desvio-padrão do subconjunto das médias, que resultaram em 1055 e 300, respectivamente. A partir dessas estatísticas, traçou-se uma distribuição normal, cujo ajuste visual indicou aderência aos dados. Já para os desvios-padrão, procedeu-se ao ajuste de curvas pela função "*fit*"de ajuste de dados disponi-



(a) Distribuição das médias dos grupos de fluxo

(b) Distribuição dos desvios-padrão dos grupos de fluxo

(c) Distribuição das quantidades de observações para os grupos de fluxo

Figura 4.3: Histogramas das estatísticas geradas a partir de mais de 27000 grupos de fluxos. Tal análise serve como base para propor as distribuições *a priori* que serão incorporadas ao modelo *bayesiano* para previsão de fluxo veicular

bilizada pela biblioteca Scipy (VIRTANEN et al., 2020), na versão 1.3.1. Conforme se observa no histograma para os desvios-padrão na Figura 4.3b, a assimetria à direita conduziu a um melhor ajuste pela distribuição normal modificada exponencialmente, abreviada como EN. Neste caso, os parâmetros λ, μ e σ resultaram em 2.6, 78.3 e 26.4. Por fim, no caso do parâmetro de normalidade ν, não foi observado nenhum padrão que pudesse ser descrito por alguma distribuição de probabilidade. Portanto, a definição de uma distribuição para este parâmetro não será feita com base na análise exploratória dos dados de fluxo.

4.2.2 Definição das distribuições a *priori* e modelo *bayesiano*

A partir das análises dos grupos de fluxo, propôs-se o modelo representado na Figura 4.4. Segundo exemplo elaborado por Kruschke (2014), como o objetivo das previsões consiste em obter previsões para os parâmetros μ, σ e ν da verossimilhança, o modelo foi estruturado em mais de um nível hierárquico, de modo que foi necessário propor um nível adicional de distribuições a *priori* específicas para cada parâmetro. Além disso, como o objetivo principal deste trabalho é conferir intervalos de credibilidade para as estimativas de fluxo, com destaque para o parâmetro σ, que representa justamente a medida de variabilidade que buscamos fornecer às previsões, obter estimativas especificamente do parâmetro de escala da distribuição de Student é essencial para descrever os comportamentos de fluxo.

O modelo ilustrado na Figura 4.4 foi construído utilizando-se as distribuições para as médias e desvios-padrão mostrados na Figura 4.3. Observa-se que a distribuição de Student foi reparametrizada, pois o modelo centralizado, embora mais intuitivo, tornou a exploração do espaço paramétrico menos eficiente em relação ao modelo não-centralizado proposto. As distribuições Normal(μ_μ, σ_μ), Gama Inversa(0.5ν, $0.5\nu\sigma^2$) e Exponencial(ν) descrevem os parâmetros μ, σ e ν, respectivamente.

$$\mu_\mu \sim Normal(1055, 150), (0, \infty]$$

$$\sigma_\mu \sim Gama(1.3, 150), [10, 1000]$$

$$\mu \sim Normal(\mu_\mu, \sigma_\mu)$$

$$\nu_\nu \sim Exponencial(30)$$

$$\nu \sim Exponencial(\nu_\nu)$$

$$\sigma_\sigma \sim ExponencialNormal(78.3, 26.4, 2.6)$$

$$\sigma^2 \sim GamaInversa(\frac{\nu}{2}, \frac{\nu\sigma_\sigma^2}{2}), (0, \infty]$$

$$fluxo \sim StudentT(\mu, \sigma, \nu)$$

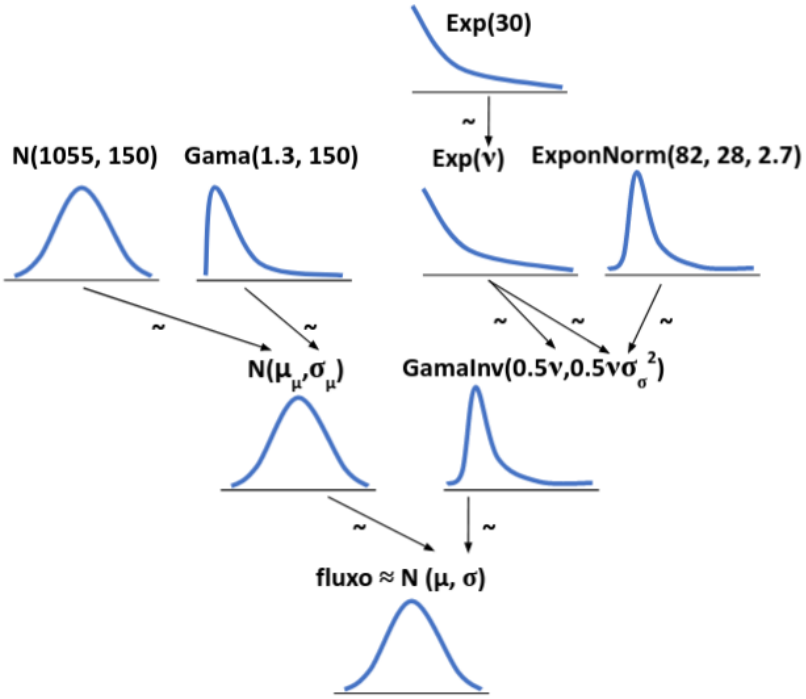


Figura 4.4: Modelo configurado com distribuições a *priori* para μ e σ definidas a partir de análise exploratória do conjunto dos dados de fluxo

Para definir o parâmetro μ da distribuição de Student, adotou-se a distribuição normal, cujos parâmetros μ_μ e σ_μ foram representados por uma distribuição normal, com média 1055 e desvio padrão de 150, e uma distribuição gama, com parâmetros alpha igual a 1.3 e beta igual a 150, já que, nessas configurações, contempla mais valores próximos ao desvio padrão de 300 encontrado pela análise exploratória, além de reduzir a possibilidade de ocorrência de valores a partir de 1000. Ainda assim, os valores para essa distribuição foram restritos ao intervalo [10, 1000], já que valores muito próximos de zero tenderam a levantar divergências durante a simulação das cadeias de Markov. Da mesma forma, optou-se por limitar a distribuição normal para a média apenas a valores positivos, já que, analiticamente, resultados negativos seriam passíveis de ocorrer e fluxos devem ser representados por valores positivos.

Preferiu-se limitar os parâmetros ν e ν_ν a distribuições exponenciais que concentrassem a maior parte de seus valores no intervalo entre 1 e 30. Essa escolha se deu em virtude de que

muitos dos conjuntos de fluxo agregaram observações afastadas da tendência central, mesmo para grupos com mais de 30 elementos. Supôs-se, assim, que seria mais representativo limitar os valores para ν a fim de garantir, por meio da priori, que os valores afastados da média fossem incorporados à posteriori. Assim, a distribuição escolhida para representar o parâmetro de normalidade foi a distribuição exponencial, com parâmetro de forma λ igual a 30.

Já o parâmetro σ foi definido por uma distribuição gama inversa, seguindo o formato reparametrizado da distribuição de Student, que o retorna em sua forma quadrática. Observa-se que seus parâmetros estão em função de ν, já definido anteriormente, e de σ_σ, definido pela distribuição exponencial normal, obtida na análise exploratória dos desvios-padrão dos conjuntos de fluxo, restrita também apenas aos valores positivos.

4.3 Obtenção das distribuições posteriores

Após definir os modelos de previsão, foram feitas as previsões com o uso das bibliotecas PyMC3 (SALVATIER; WIECKI; FONNESBECK, 2016), na versão 3.7, Arviz (KUMAR et al., 2019), na versão 0.5.1, e Theano (Theano Development Team, 2016), na versão 1.0.4. Para integração do espaço paramétrico, foi escolhido o método NUTS, por apresentar recursos computacionais mais eficientes para o cálculo de variáveis contínuas em relação aos métodos tradicionais de amostragem da distribuição posterior, como Metrópolis e Gibbs. Como método de inicialização das cadeias, utilizou-se a configuração automática fornecida pelo PyMC3, por otimizar a escolha do valor de inicialização no espaço paramétrico a fim de agilizar o processo de convergência, além de ser fortemente recomendada pelos desenvolvedores da plataforma.

O modelo foi configurado para rodar quatro cadeias em paralelo, quantidade recomendada por Vehtari et al. (2020) para garantir a geração de estatísticas confiáveis, cuja finalidade é de aferir a convergência das previsões. Foram geradas 1000 amostras da distribuição posterior para cada cadeia, para cada qual se realizaram 1000 amostras adicionais de *tuning*, o que totalizou 8000 iterações para cada grupo de fluxo. Outra configuração adotada foi definir o alvo de aceitação para 0.90, já que neste nível, os espaços paramétricos mais problemáticos tenderam a acusar nenhuma ou poucas divergências por parte do algoritmo de integração.

Para fins de verificação, foram conferidos visualmente os traços e os diagramas de autocorrelação de uma amostra de previsões de 1500 conjuntos de dados, a fim de buscar erros de convergência em quaisquer dos conjuntos obtidos. Além disso, foram feitas verificações preditivas posteriores desses mesmos conjuntos, para observar a aderência das previsões aos dados.

Embora a análise visual das cadeias seja recomendada para análise de convergência, ela seria inviável em um banco de dados tão extenso, e portanto, a análise do tamanho efetivo da amostra (ESS) e da estatística de Gelman e Rubin foi realizada para tornar esse processo mais eficiente. Assim, para cada previsão, verificou-se se essas estatísticas corresponderam a va-

lores de referência maiores que 400 e menores que 1.01, respectivamente, para cada um dos parâmetros μ, σ e ν. Além disso, adotou-se, por convenção, que previsões com até 8 divergências ou, em média, duas divergências por cadeia, ocorridas durante o processo de amostragem pelas cadeias de Markov usando o método NUTS, fossem consideradas como falsos-positivos, pois divergências em excesso poderiam sugerir problemas na integração do espaço posterior e, nesse caso, poderiam enviesar as distribuições posteriores. Assim, para os casos de desconformidade elencados, novas previsões foram feitas, elevando-se o *tuning* e o alvo de aceitação para 1500 e 0.95, respectivamente, em uma primeira rodada, e para 2500 e 0.99, para previsões com divergências existentes. Em ambos os casos, manteve-se o número de 1000 amostras por cadeia.

4.4 Análise das distribuições posteriores

Após verificar a convergência, foram realizadas verificações preditivas posteriores, com objetivo de realizar 1000 combinações de valores amostrados das distribuições dos parâmetros μ, σ e ν e simular diferentes distribuições Student. Esse procedimento teve objetivo de verificar se a sobreposição de diversas distribuições amostradas pela posterior tenderam a configurar uma distribuição de Student média. Por fim, para definir a distribuição representativa de cada conjunto de fluxo, foram utilizadas as médias das distribuições dos parâmetros μ, σ e ν.

Iniciou-se a análise pela comparação entre valores observados e previstos, por meio de erros absolutos e erros médios percentuais. Por convenção, quando há necessidade de adotar um valor representativo da distribuição, utiliza-se a sua média para representar o valor estimado. Para tanto, utilizou-se a média μ como expectativa da distribuição de Student, segundo a Equação 4.1, para comparar com o fluxo observado, registrado pelo sensor. A Equação 4.2 descreve o erro absoluto, que consiste em selecionar o valor previsto e compará-lo com o valor observado, e o erro relativo, descrito na Equação 4.3, foi escolhido para análise de acurácia por ser o mais difundido para verificação da qualidade dos modelos de previsão.

$$E[fluxo] = \mu \tag{4.1}$$

$$EA = |\mu - fluxo_{observado}| x100\% \tag{4.2}$$

$$ER = \frac{EA}{fluxo_{observado}} \tag{4.3}$$

Como a vantagem da inferência *bayesiana* consiste justamente em gerar intervalos de probabilidade, ou seja, fornecer indícios de quais os possíveis valores que aquele parâmetro pode apresentar, não faz sentido, entretanto, ater-se apenas a valores centrais extraídos das distribuições posteriores e suas métricas de acurácia. Por este motivo, a segunda abordagem proposta para análise das previsões consistiu justamente na análise desses intervalos, a partir

das métricas definidas nas Equações 2.10 e 2.11. Por fim, foi realizada análise exploratória das distribuições dos parâmetros e das medidas de erro relativo e absoluto, de forma geral para todas as previsões, e também divididas por sensor e horizontes de previsão.

4.5 Considerações finais

Neste capítulo foi apresentado um método para estimação de fluxo de tráfego veicular de rodovias a partir de um modelo hierárquico *bayesiano*. Estruturou-se um banco de dados que contemplasse diferentes condições de tráfego em função de variáveis temporais e de velocidade da API, a partir do qual se propôs uma distribuição de probabilidade que pudesse descrever todos esses conjuntos de forma robusta. Para definição do modelo, foi necessário explorar o próprio banco de dados para obter indícios de distribuições a *priori* que pudessem descrever o comportamento de seus parâmetros. Para avaliar a qualidade das previsões, foram propostas métricas de convergência, bem como análises visuais dos traços, gráficos de autocorrelação e verificações posteriores preditivas. Por fim, foram propostas duas formas de interpretar e analisar os modelos gerados, uma delas com base na análise tradicional por acurácia, e outra com base na análise de intervalos de credibilidade gerados pelas distribuições. A partir desses resultados, buscou-se fornecer não apenas previsões pontuais para o fluxo de tráfego, mas também intervalos de previsão que descrevessem a incerteza acerca valor obtido, de modo a se obter uma compreensão melhor acerca dos fenômenos de tráfego.

Capítulo 5

Resultados

Neste capítulo são apresentados os resultados das previsões de fluxos para o modelo *bayesiano* proposto no Capítulo 4. Inicialmente, é descrito o processo de interpretação de um único conjunto de dados, selecionado aleatoriamente. Em seguida, são apontados os resultados das previsões conduzidas para todos os conjuntos de dados, de forma a extrapolar a interpretação no nível individual. Também são mostradas distribuições de erros relativos e absolutos e demais análises exploratórias do conjunto de dados gerado. Por fim, são apresentadas as implicações práticas das distribuições posteriores geradas e suas diferentes formas de utilização.

5.1 Análise de convergência das cadeias

A fim de visualizar os resultados das previsões geradas pelo modelo hierárquico *bayesiano*, são apresentados os traços das cadeias geradas para os parâmetros de um conjunto de dados selecionado aleatoriamente. Na Figura 5.1 constam as cadeias referentes à observação para a localidade do sensor da rodovia SP280, km 29.5, sentido leste, cuja velocidade observada pela API foi de 93.8 km/h, ocorrida às 14 horas de um sábado, para o horizonte de previsão de 0 minutos. Na Figura 5.1, constam os traços referentes aos parâmetros do modelo.

Observa-se do lado esquerdo que as quatro cadeias tendem a concordar entre si e, de modo geral, todos os parâmetros apresentam traços bem misturados, conforme mostra o lado direito da Figura 5.1, o que indica a exploração adequada do espaço paramétrico. Os parâmetros designados como *mu_mu_2*, *sigma_mu_2*, *nu_2*, *sigma_2*, referem-se, respectivamente, aos parâmetros μ_μ, σ_μ, ν_ν, σ_σ definidos na Seção 4.2.2, enquanto μ_1, σ_1 e ν_1 referem-se a μ, σ e ν.

Os quatro primeiros parâmetros formam o segundo nível do modelo hierárquico, e pode-se ver que seus traços posteriores concordam com a forma das prioris propostas. O mesmo ocorre para os três últimos parâmetros, a partir do qual são retirados os valores de tendência central para definir a distribuição de Student.

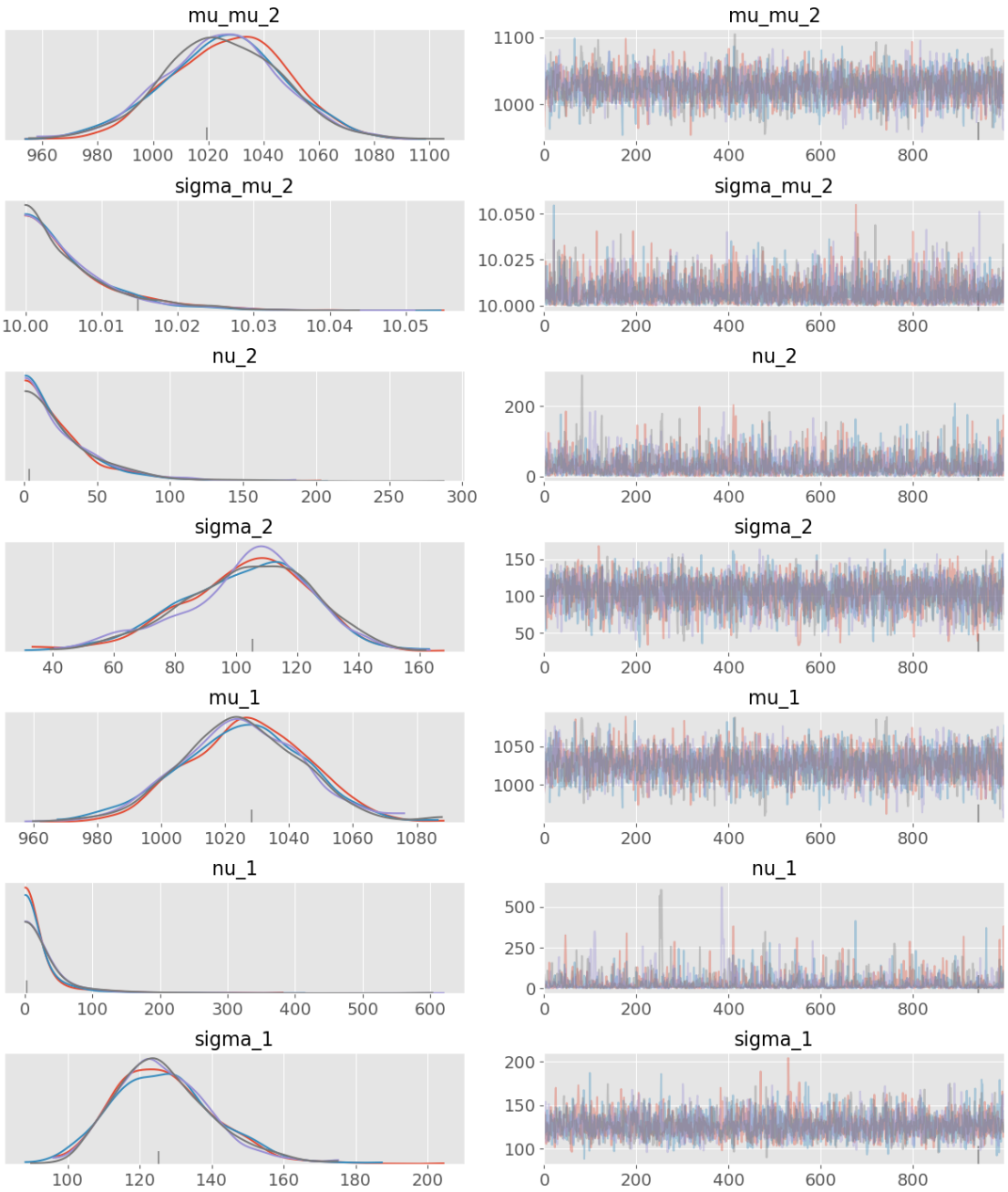


Figura 5.1: Traços das quatro cadeias de Markov gerados para os parâmetros μ, σ e ν, coincidindo entre si, o que indica que a amostragem ocorreu de forma uniforme pelo espaço paramétrico

Para análise de convergência, além da verificação dos traços, são utilizados os gráficos de autocorrelação, expostos na Figura 5.2. Nota-se que eles apresentam valores mais altos para essa estatística apenas nas primeiras iterações, com tendência de estabilização imediata antes da centésima iteração. Em resumo, ambas as análises visuais dos traços e dos gráficos de auto-

correlação fornecem indícios que constatam a adequabilidade do modelo aos dados e, portanto, é possível inferir que há convergência das cadeias.

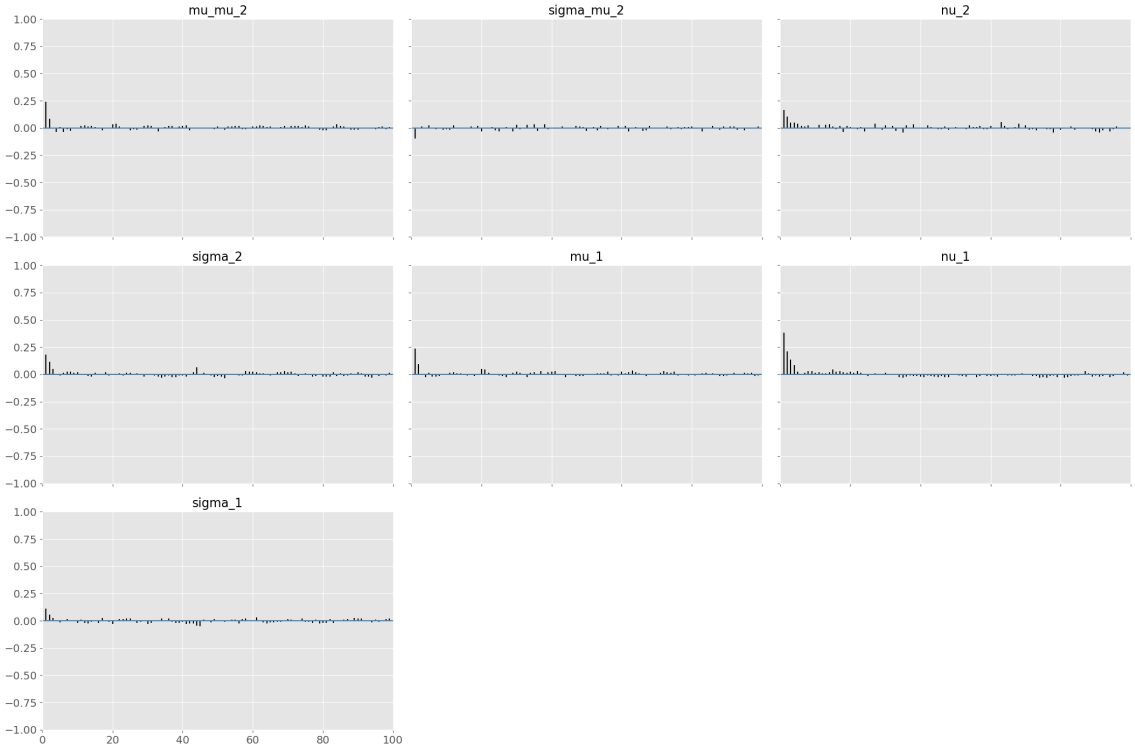
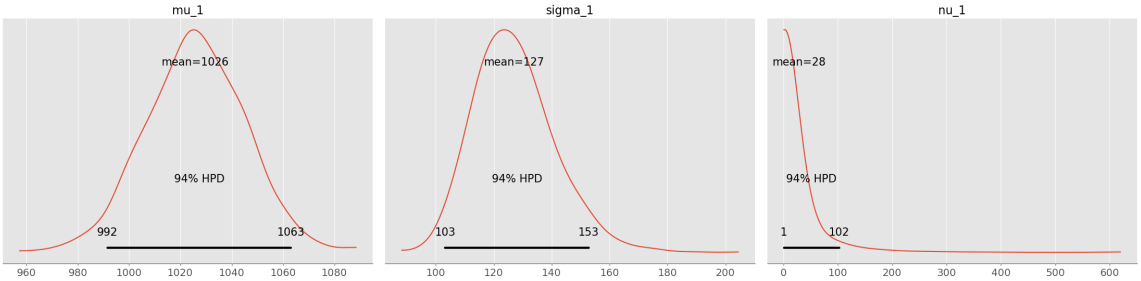


Figura 5.2: Os gráficos de autocorrelação indicam que as cadeias tendem a se estabilizar bem antes da centésima iteração

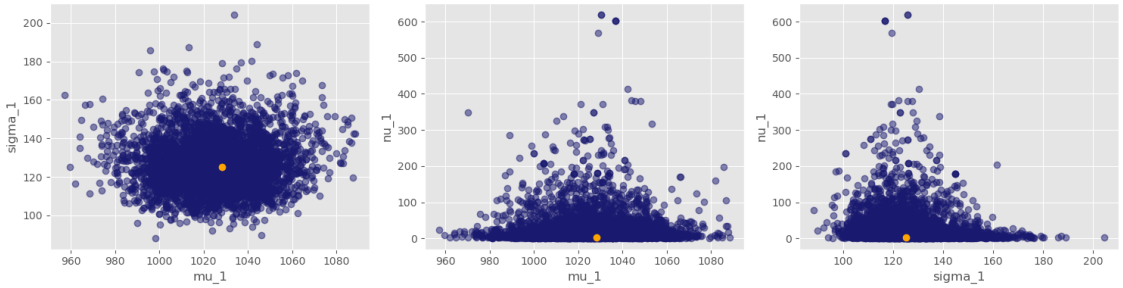
Por corresponder ao objeto de interesse das previsões, a Figura 5.3 contém as distribuições posteriores para os parâmetros μ, σ e ν apenas, acompanhadas de seus intervalos a 94% de credibilidade, e embora se espere um valor de 95%, ressalta-se que esses intervalos foram fornecidos automaticamente pelo pacote computacional Arviz, utilizado pela biblioteca PyMC3, e portanto, supõe-se que a não utilização do intervalo de 95% esteja relacionada a alguma especificidade do pacote computacional. Na Figura 5.3a, observa-se que μ corresponde à uma distribuição normal, cujo desvio padrão corresponde à cerca de 50 veículos por hora por faixa. Isso indica certa variabilidade acerca desse parâmetro e, ao mesmo tempo, o grau de certeza conferido a ele por meio da interação entre a distribuição normal proposta a *priori* e os dados. Já o parâmetro σ é descrito por uma distribuição de forma aproximada à da gama que o gerou, porém, observa-se que houve deslocamento dos valores mais próximos a zero, contemplados em sua priori, que variava originalmente de 10 a 1000, a valores próximos a 120 veículos por hora por faixa, variando de 103 a 153. Finalmente, a distribuição para ν corresponde à distribuição exponencial com uma gama extensa de valores, porém, em função de sua massa se concentrar próxima a zero, seu valor médio, que é utilizado para definir a distribuição de Student representativa dos dados, costuma ficar próximo a 30. Observa-se adicionalmente que, por restrição imposta pela priori, o valor mínimo para ν manteve-se acima de um.

Na Figura 5.3b, constam os espaços paramétricos com as combinações dois a dois entre os

parâmetros μ, σ e ν, em que cada ponto azul corresponde a uma amostra da posterior. Essa forma de visualização permite conferir se existem regiões de curvaturas muito complexas, o que poderia levar a dificuldades na convergência das cadeias de Markov. Neste caso, observa-se que os espaços posteriores estão bem distribuídos e apresentam curvaturas suaves. O ponto laranja, destacado ao centro de cada gráfico, representa uma divergência encontrada pelo algoritmo. Nesse caso, como ele se encontra em uma região bastante densa para todos os parâmetros, infere-se que se trata de um falso positivo, de forma que sua ocorrência não apresenta indicativo de falhas de convergência.



(a) Distribuições posteriores com seus respectivos intervalos a 95% de probabilidade

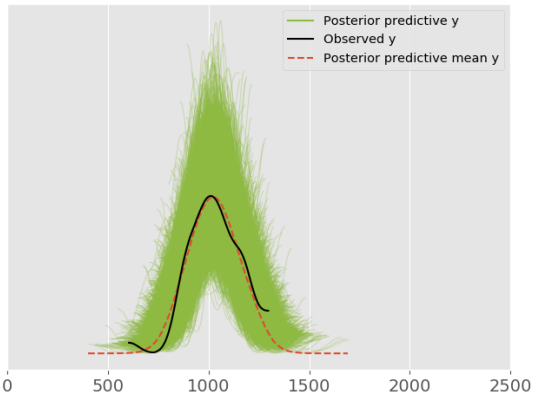


(b) Espaços posteriores dos parâmetros combinados dois a dois, a fim de identificar curvaturas complexas. O ponto em laranja representa uma divergência, possivelmente um caso de falso positivo apontado pelo algoritmo de previsão.

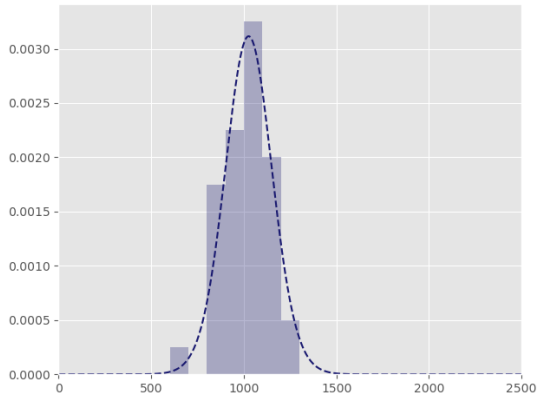
Figura 5.3: Distribuições posteriores para os parâmetros μ, σ e ν para a distribuição de Student do fluxo previsto

As últimas verificações são apresentadas na Figura 5.4a. Alternativamente às análises de convergência apresentadas, a Figura 5.4 contém previsões preditivas posteriores. Cada linha verde corresponde a uma simulação da distribuição de Student, gerada a partir de combinações de valores retirados aleatoriamente das distribuições posteriores dos parâmetros μ, σ e ν. A linha tracejada em laranja corresponde à média dessas previsões, e a preta, aos dados observados. A partir das previsões preditivas posteriores, é possível notar que houve boa aderência aos dados, o que indica coerência com o modelo proposto. Finalmente, utilizando-se as médias das distribuições posteriores para μ, σ e ν, a Figura 5.4b apresenta a distribuição posterior como resultado final deste exemplo em questão, cuja curva tracejada em destaque corresponde à distribuição prevista para os fluxos, e os dados são representados no histograma ao fundo para fins ilustrativos.

A análise visual apresentada até aqui é um modo simples de verificar a convergência das



(a) Conjunto de previsões preditivas posteriores, em verde, feitas a partir de amostras aleatórias de valores das distribuições geradas para μ, σ e ν. Alinha laranja representa a média das previsões preditivas posteriores e a linha preta, o histograma dos dados observados.



(b) Distribuição de Student prevista para os fluxos, ajustada com os valores médios das distribuições posteriores dos parâmetros μ, σ e ν. A previsão foi feita com velocidade observada pela API às 14 horas para o horizonte de previsão de zero minutos do sensor 9.

Figura 5.4: As previsões preditivas posteriores auxiliam na verificação da aderência das previsões aos dados. Como resultado final, a distribuição de Student é ajustada pelos valores médios de μ, σ e ν

Rodada	Amostras	Tuning	Alvo	Conjuntos	Porcentagem
1	1000	1000	0.9	27261	**98.0**
2	1000	1500	0.95	474	**1.70**
3	1000	2500	0.99	54	**0.20**
4	1000	5000	0.99	10	**0.04**
5	2500	5000	0.99	16	**0.06**

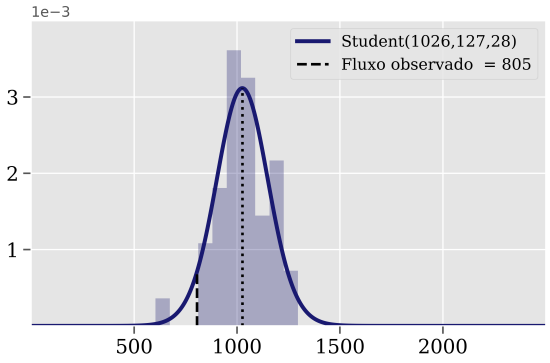
Tabela 5.1: Quantidades de conjuntos de dados rodados com diferentes configurações

cadeias em inferência *bayesiana*, embora seja pouco prática no escopo deste trabalho, que abrange mais de 27000 conjuntos de dados. Nesse sentido, a principal ferramenta de análise consiste na verificação das estatísticas de Gelman e Rubin e dos tamanhos efetivos das amostras. A Tabela 5.1 resume a quantidade de conjuntos de dados que não convergiram na primeira rodada de previsões, seguindo as configurações inicias pré-determinadas, e os resultados das iterações seguintes em novas configurações com cadeias mais longas e alvos mais altos.

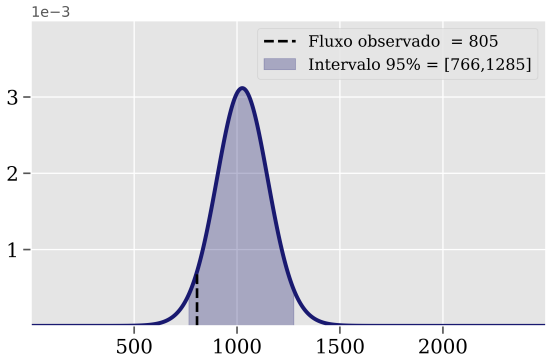
Não se observa, nos dados previstos, um padrão definido que esclareça a diferença de comportamentos de convergência entre conjuntos. Embora visualmente os dados pareçam condizentes com a verossimilhança proposta, é possível que seu tratamento algébrico leve o modelo a identificar multimodalidade dos dados. Nesse caso, é comum ocorrerem problemas na convergência das cadeias e o modelo de fato tende a não funcionar. Outro fator é que os espaços posteriores podem apresentar regiões de alta curvatura que restringem as transições a regiões muito densas do espaco posterior. Quando isso acontece, a amostra gera representantes com alta autocorrelação e não explora adequadamente o espaço paramétrico, o que torna as amostras não representativas para fins de previsão.

5.2 Implicações práticas da distribuição posterior

As principais aplicações decorrentes do método proposto neste trabalho são discutidas nesta Seção. A Figura 5.5a mostra, sobre o histograma dos dados históricos observados para o mesmo conjunto utilizado na análise da Seção 5.1, a distribuição de fluxos prevista, com destaque para a moda da distribuição, que nesse caso corresponde também à média, e para o fluxo observado. A partir desses dois valores, se considerarmos como valor previsto a média de 1026 veículos por hora por faixa e o fluxo observado de 805 veículos por hora por faixa, é possível calcular o erro absoluto de 221 veículos por hora por faixa e erro relativo de 27%. Este erro relativo, embora pontual, pode ser considerado alto em comparação com os erros médios obtidos por outros métodos de previsão correntes, conforme apresentado na Tabela 2.1, mas ainda assim, observa-se que o fluxo observado é contemplado pela distribuição posterior, o que indica que sua ocorrência ainda é possível, embora seja menos provável.



(a) Distribuição final com destaque para o fluxo previsto e o fluxo observado



(b) Distribuição final com o destaque para o intervalo de 95% de credibilidade

Figura 5.5: Aplicações da distribuição posterior. No primeiro caso, é possível calcular medidas de acurácia, a partir da média da distribuição como valor previsto e o fluxo observado. No segundo caso, é possível verificar que o valor previsto se enquadra no intervalo de 95% de probabilidade

Já a Figura 5.5b apresenta a estimação de fluxo de tráfego sob outra ótica: em vez de focar na acurácia, busca-se o intervalo de credibilidade no qual o valor observado está inserido. Neste caso, o valor estimado fica no limite do intervalo de credibilidade de 90.7%. Consequentemente, o valor observado se manteve dentro de um intervalo de 95% de credibilidade, considerado como valor de referência em análises estatísticas.

Entretanto, é de interesse do pesquisador definir seus próprios intervalos de referência e critérios de corte. Na Figura 5.6, observa-se que, para o sensor da SP270, no quilômetro 34 e sentido Leste, com velocidade observada pela API de 72.8 km/h, às 6 horas de uma segunda-feira, e para um horizonte de previsão de 15 minutos, o fluxo médio tende a 1838 veículos/hora/faixa. Considerando, hipoteticamente, que fluxos da ordem de 2000 veículos/hora/faixa sejam críticos para esta região, a probabilidade de valores maiores ocorrerem, segundo a distribuição obtida para essas condições, é de 32%. Nesse caso, intervalos bicaudais a 95% de credibili-

dade não são interessantes, já que o que se busca é a probabilidade da ocorrência de um valor extremo.

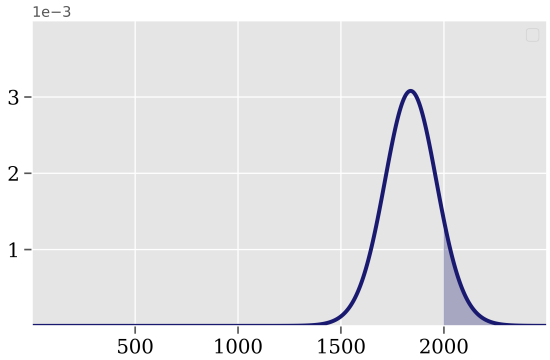


Figura 5.6: Utilização da posterior sem pretensão de obter valor específico. O pesquisador pode definir valores aceitáveis e analisar probabilidades dos eventos possíveis

Outra vantagem do método proposto é que, mesmo para um conjunto de dados com apenas uma observação, é possível obter uma distribuição posterior. A Figura 5.7 mostra dois exemplares de previsões para esses conjuntos. No caso específico dos conjuntos com apenas uma observação, a medida de escala da distribuição de Student é regularizada puramente pela distribuição a *priori*, já que não existe variação que possa ser inferida pelos dados. Porém, os desvios podem apresentar variações maiores, como no caso da Figura 5.7b, em que a distribuição abrange todos os valores possíveis de fluxo.

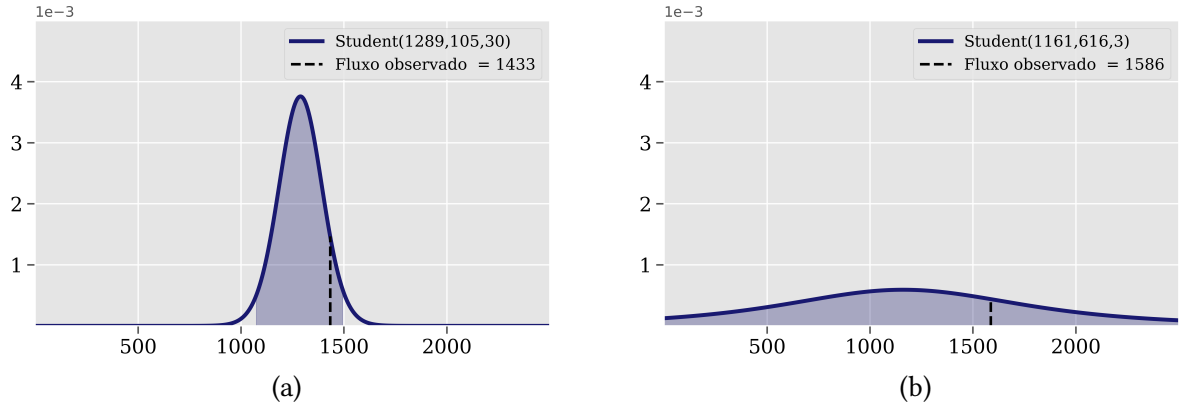
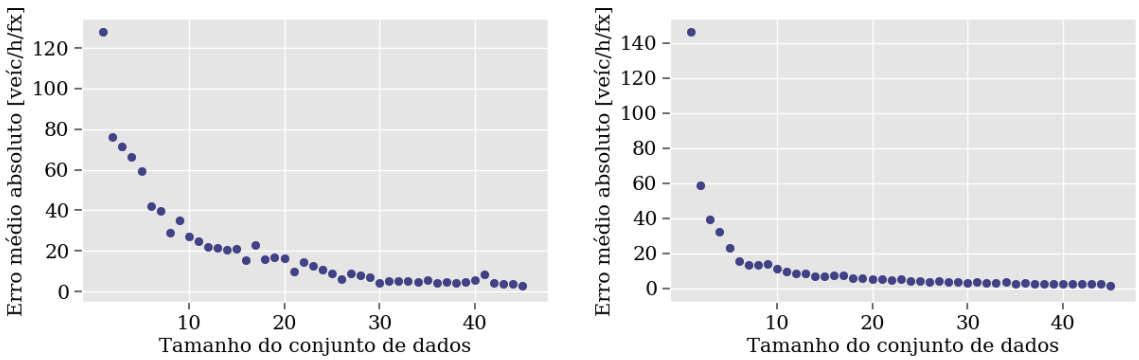


Figura 5.7: Distribuições posteriores geradas a partir de conjuntos de dados com apenas uma observação

A análise individual de cada uma dessas distribuições fornece subsídios para inferir as condições dos fluxos em uma seção de rodovia pré determinada e sob condições específicas de dia da semana e horário, em tempo real ou para horizontes próximos. Entretanto, é relevante observar como essas previsões se comportam de forma geral, e não apenas em nível individual. Inicialmente, em função das diferenças observadas pela comparação entre previsões realizadas com apenas uma observação, conforme evidenciado na Figura 5.7, torna-se relevante mapear o efeito do tamanho do banco de dados nos erros de previsão.

A Figura 5.8 mostra as médias dos erros médios absolutos em função da quantidade de dados usados nas previsões. Esses erros foram calculados pela diferença entre valores previstos e valores calculados para média e desvio padrão dos conjuntos de fluxos observados. Na Figura 5.8a é possível ver as previsões geradas para μ a partir de conjuntos de dados com apenas uma observação apresentam desvios médios de 126 veículos por hora por faixa, caindo para 76 para grupos com duas observações e progressivamente quanto maior a quantidade de dados. O mesmo ocorre para os parâmetros σ na Figura 5.8b. Em ambos os casos, essas diferenças tendem a se reduzir gradativamente até que se estabilizam em uma diferença mínima a partir de conjuntos de dados com 30 observações.



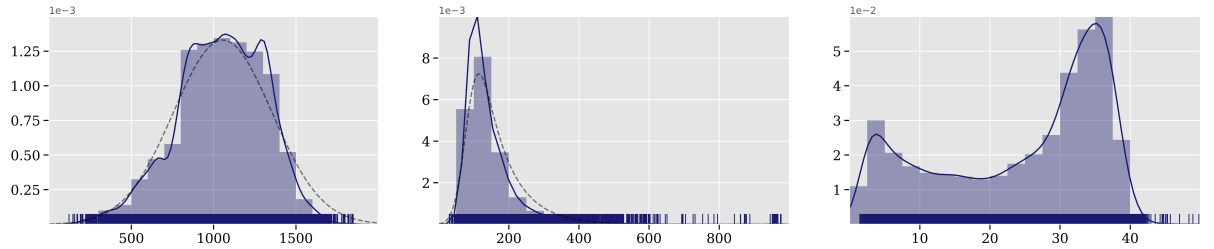
(a) Médias (b) Parâmetros de escala σ

Figura 5.8: Distribuição das médias das diferenças entre valores observados e previstos para os parâmetros μ e σ, agrupados conforme o tamanho do banco de dados

Dos parâmetros previstos, σ e ν apresentaram correlação negativa de 0.8. Tal correlação pode ser explicada pela própria natureza da distribuição de Student, já que tanto valores baixos para ν quanto valores altos para σ estão relacionados à alta variação no conjunto de dados. Sendo assim, é de se esperar que ambos os parâmetros se equilibrem no controle da variabilidade de forma conjunta. Por outro lado, ν e o tamanho dos conjuntos de dados apresentam coeficiente de correlação de 0.11. Isso leva a crer que mesmo grupos com maiores quantidades de dados podem apresentar maiores variâncias e gerar, consequentemente, valores menores para ν. Da mesma forma, conjuntos de dados menores não necessariamente implicam em maiores variâncias.

5.3 Distribuição dos parâmetros μ, σ e ν

Como forma de avaliar os parâmetros posteriores μ, σ e ν de forma geral, observam-se suas distribuições na Figura 5.9. As distribuições normal e exponencial gaussiana se replicam e se ajustam conforme os dados que as geraram. Isso pode se dever ao fato de as distribuições a *priori* terem sido geradas por análise exploratória do banco de dados e, de certa forma, isso indica que essas prioris são representativas dos dados e podem ser utilizadas para previsão de fluxos veiculares para a região segundo a qual foram geradas.



(a) Médias (b) Parâmetros de escala σ (c) Parâmetros de normalidade ν

Figura 5.9: Distribuição de frequências dos parâmetros μ, σ e ν das distribuições posteriores

Já o parâmetro ν não apresenta forma definida semelhante a nenhuma distribuição conhecida, embora sua distribuição também tenha sido compatível com o histograma referente ao tamanho dos conjuntos de dados de fluxo (Figura 4.3c). Essa semelhança visual pode indicar forte influência dos dados sobre as previsões, considerando que as prioris foram geradas com base nos dados.

5.4 Previsões por sensor

As Figuras 5.10 a 5.19 apresentam os resultados das previsões realizadas ao longo das duas semanas, para todos os dez sensores, com intervalos de credibilidade de 50%, 85% e 95% destacados em escala de cinza, para horizontes de previsão de 15 minutos. Como não se observou grande variabilidade entre os perfis dos diferentes horizontes de previsão, as demais previsões para horizontes de 0, 30 e 60 minutos constam no Apêndice A. É possível observar que, na maior parte dos casos, verificou-se proximidade entre valores observados e previstos. Algumas ressalvas, entretanto, devem ser sinalizadas. Por exemplo, o dia 9 de julho, terça-feira, foi um feriado regional, e como os dados não foram agregados de forma a considerar feriados, consequentemente esse padrão não foi captado pelo modelo. Todos os sensores, exceto o da Figura 5.16, apresentaram diferenças consideráveis aos sábados, com valores observados em geral maiores que os previstos. É possível que isso tenha ocorrido em função de mudanças no padrão de deslocamento nos fins de semana, já que os dados históricos estavam disponíveis apenas para o período de 2011 a 2014, enquanto os dados observados são de 2019. Já o sensor da Figura 5.18 apresentou uma diferença considerável entre os valores previstos e observados. Embora não se tenha verificado a ocorrência de nenhum evento atípico ao longo do período de coleta ou alterações na infraestrutura viária próxima ao sensor em questão, é possível que tenha ocorrido alguma atividade de manutenção, em que alguma pista tenha sido bloqueada. Por fim, é possível notar, especialmente no caso do sensor da Figura 5.13, que grande parte dos dados estão faltantes, o que indica que possivelmente houve mau funcionamento desse sensor durante as duas semanas de coleta. Falhas também podem ser observadas em outros sensores, por períodos mais curtos.

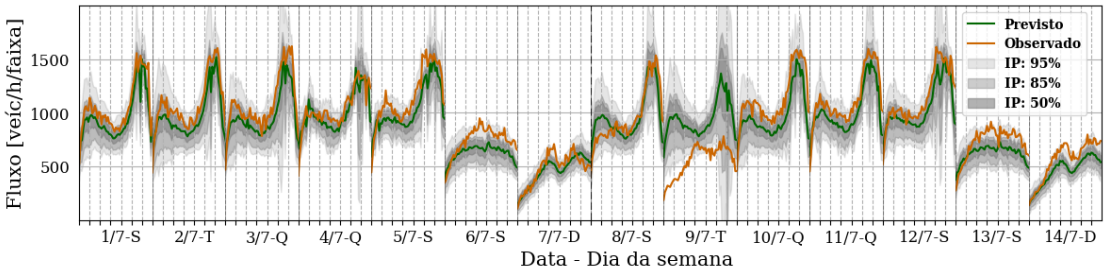


Figura 5.10: Previsões para rodovia SP-021, km 18.3, sentido externo

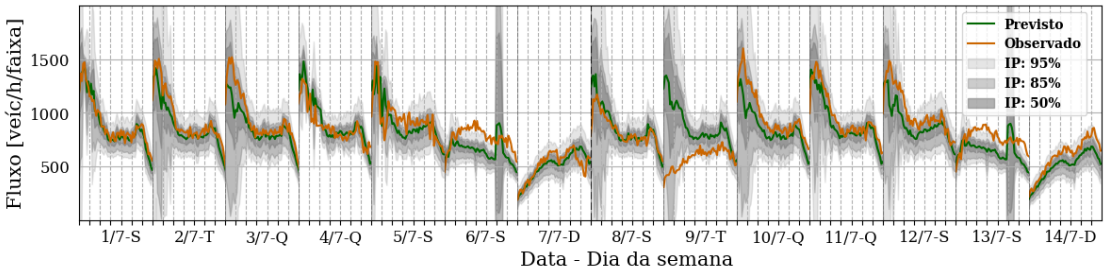


Figura 5.11: Previsões para rodovia SP-021, km 18.3, sentido interno

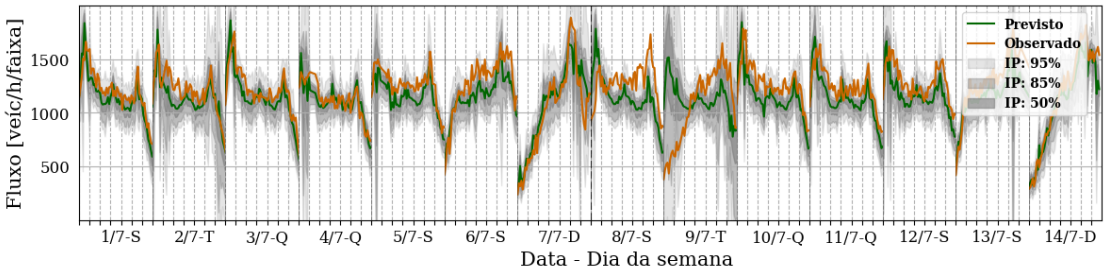


Figura 5.12: Previsões para rodovia SP-270, km 34, sentido leste

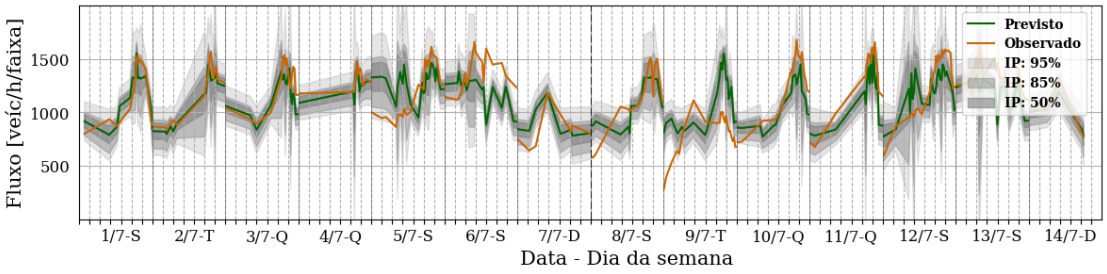


Figura 5.13: Previsões para rodovia SP-270, km 34, sentido oeste

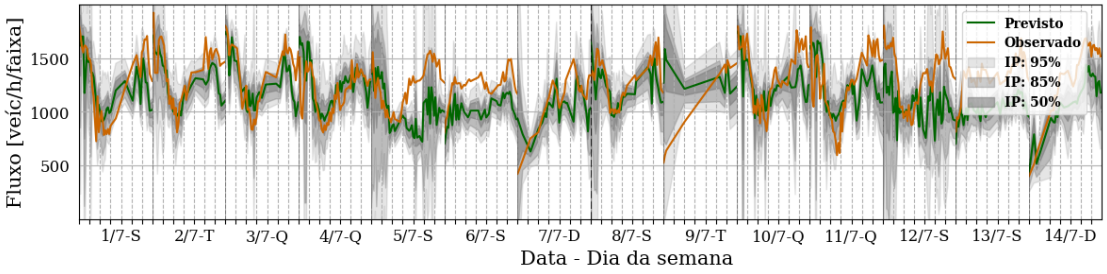


Figura 5.14: Previsões para rodovia SP-280, km 16.08, sentido leste

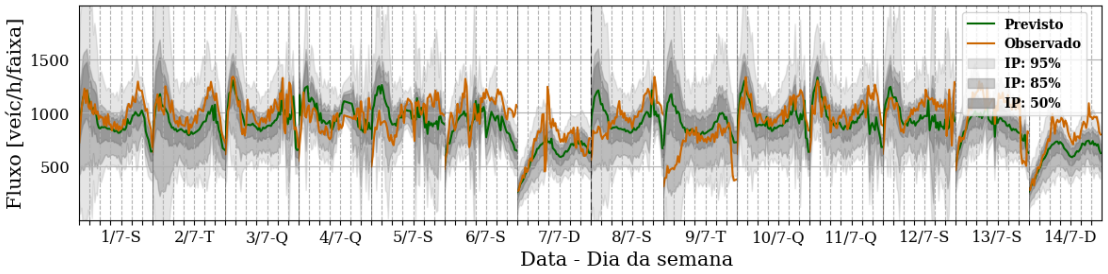


Figura 5.15: Previsões para rodovia SP-280, km 22.4, sentido oeste

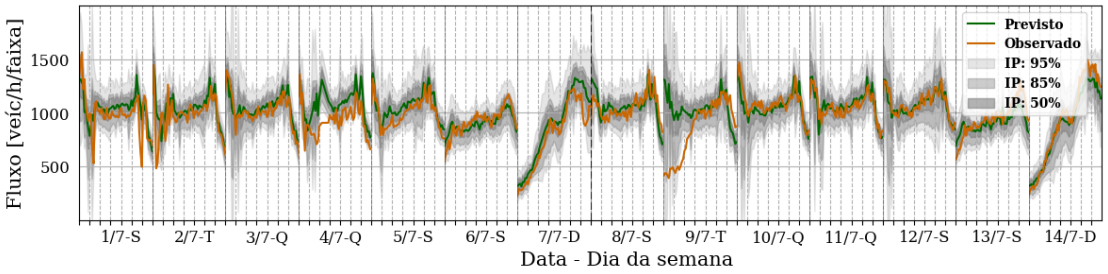


Figura 5.16: Previsões para rodovia SP-280, km 27, sentido leste

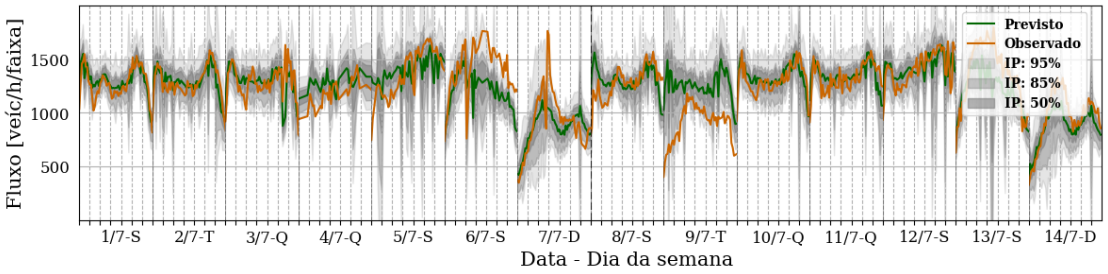


Figura 5.17: Previsões para rodovia SP-280, km 27, sentido oeste

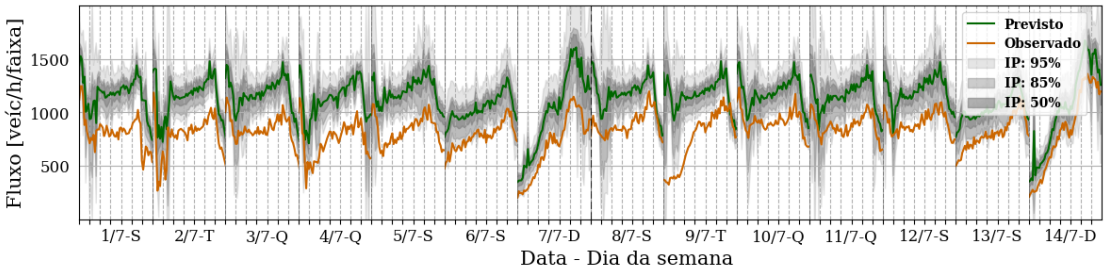


Figura 5.18: Previsões para rodovia SP-280, km 29.5, sentido leste

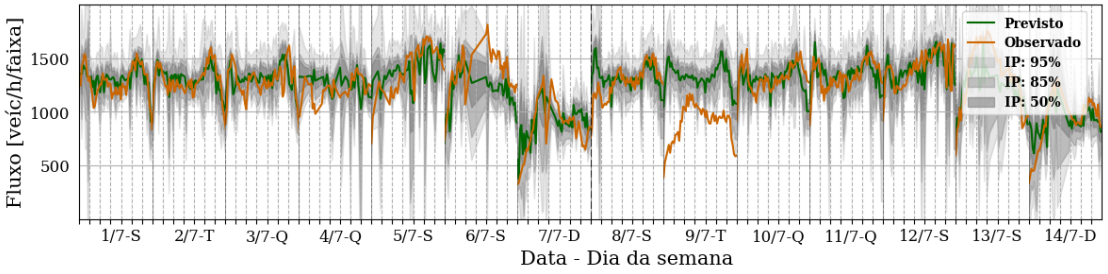


Figura 5.19: Previsões para rodovia SP-280, km 29.5, sentido oeste

Caso	Todos os Sensores		Remoção Sens. 4 e 9	
	PCIP	AMIP	PCIP	AMIP
Todo o período	82.6	693	88.7	705
Remoção do feriado	83.7	678	89.7	687
Remoção do feriado e fins de semana	85.1	714	92.2	727

Tabela 5.2: Probabilidade de cobertura do intervalo de previsão (PCIP, em %) e amplitudes médias dos intervalos de previsão (AMIP, em veíc./h/faixa), considerando as combinações entre cenários com a remoção de sensores com erros, dias atípicos e fins de semana

5.5 Intervalos de previsão

Para análise dos intervalos de previsão, foram utilizadas as métricas de probabilidade de cobertura do intervalo de previsão (PCIP), descrita pela Equação 2.10, que reflete a porcentagem dos fluxos observados que pertenceram aos intervalos previstos, e a amplitude média do intervalo de previsão (AMIP), dada pela Equação 2.11. Ambas as métricas foram calculadas em relação ao intervalo de credibilidade de 95% e estão apresentadas na Tabela 5.2. Em função dos efeitos do feriado e fins de semana sobre as previsões e da ocorrência de operações atípicas pelos sensores, a análise foi segregada em seis níveis, considerando desde as médias para todas as previsões, até os resultados filtrados para as condições ideais. Assim, à medida que se reduz o banco de dados apenas para as condições normais e típicas, ou seja, sem falhas nos sensores e em dias úteis de semana, é possível notar melhora nos valores do PCIP, de modo que no melhor dos casos, este valor atinge 92.2%, frente a 82.6% no caso em que todos os sensores são contabilizados por esta métrica.

Os valores para AMIP também tenderam a se reduzir à medida que os resultados foram filtrados para as condições ideais. Porém, para os casos em que os fins de semana foram removidos, os valores para esta métrica registraram aumento substancial. Isso se deve ao fato de que os fins de semana tenderam a apresentar intervalos de credibilidade menores, possivelmente em função de fluxos menores em comparação com dias úteis, ainda que os valores observados e previstos não tenham coincidido como na maior parte dos dias de semana. Desse modo, é possível inferir que os intervalos de credibilidade para dias de semana tenderam a ser mais amplos do que os de fins de semana, ou seja, apresentaram maior variabilidade.

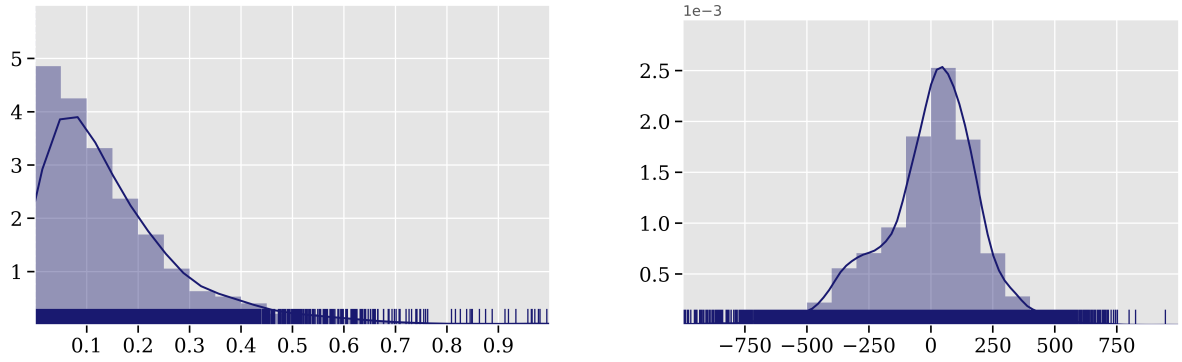
Idealmente, o valor para AMIP deve ser o menor possível, de modo que a variabilidade seja mínima para uma previsão. Já o PCIP deve atingir um valor aproximado de 95%, que corresponde ao intervalo de credibilidade considerado nessa análise. Ainda assim, esta métrica se manteve abaixo do esperado mesmo no melhor cenário. Uma alternativa para reduzir o intervalo de previsão e, consequentemente, melhorar esses indicadores, seria criar uma variável de agregação, no momento da geração dos grupos de fluxos, que considerasse feriados, eventos atípicos, ou mesmo vésperas de feriado, para os quais também podem ser propostas prioris específicas para essas condições.

Caso	Todos os Sensores		Remoção Sens. 4 e 9	
	MAPE	RMSE	MAPE	RMSE
Todo o período	17.6	210	14.3	186
Remoção do feriado	15.1	190	11.9	165
Remoção do feriado e fins de semana	14.2	189	10.8	160

Tabela 5.3: Erros médios percentuais (MAPE, em %) e erros quadráticos médios (RMSE, em veíc./h/faixa), considerando as combinações entre cenários com a remoção de sensores com erros, dias atípicos e fins de semana

5.6 Acurácias

Conforme destacado anteriormente, as métricas de acurácia não consistem no principal atrativo da inferência *bayesiana*, mas sim de seus intervalos de credibilidade conferidos pela distribuição posterior. Ainda assim, convém verificar o comportamento conjunto das métricas de erros relativos e de erros absolutos, cujas frequências estão distribuídas na Figura 5.20.



(a) Erro relativo (b) Erro absoluto

Figura 5.20: Distribuições de erros

Verifica-se que maior parte dos erros absolutos se encontram aproximadamente distribuídos segundo uma distribuição normal, de forma simétrica em relação à origem, atingindo geralmente valores da ordem de até 300 veículos/hora/faixa. Já os erros relativos ocorrem com mais frequência até o valor de aproximadamente 40%, atingindo média de 17.6%. Esse valor pode ser considerado alto em comparação com outros métodos propostos para previsão de fluxo veicular, conforme evidenciado na Tabela 2.1. Métodos estatísticos usualmente apresentam erros médios percentuais globais acima de 10%, enquanto os de aprendizado de máquina usualmente variam de 5% a 10%. Entretanto, se retirarmos o feriado das análises, o MAPE reduz para 15.1%, e excluindo-se fins de semana, para 14.2%. De modo análogo, se retirarmos os sensores 4 e 9, em que se evidenciou a ocorrência de anomalias na operação típica, os erros com todos os dados, retirando feriados e retirando fins de semana reduzem para 14.3%, 11.9% e 10.8%, respectivamente.

A fim de melhor visualizar os resultados obtidos, é conveniente apresentá-los segregados por sensor e horizontes de previsão, conforme Figuras 5.21 e 5.22, respectivamente. As três

primeiras colunas contêm as distribuições das médias dos parâmetros previstos. A primeira coluna apresenta o histograma dos fluxos previstos e a curva aproximada dos valores de fluxos observados. Exceto para os Sensores 4 e 9 que apresentaram condições de operação atípicas, os demais fluxos apresentaram distribuições condizentes com as observações, o que indica a eficácia do método para condições normais de funcionamento das rodovias. As três colunas seguintes se referem à distribuição de erros, que apresentaram uma distribuição normal para erros absolutos, aproximadamente exponencial para erros relativos, e assimetricamente positiva para as amplitudes. Por fim, a Figura 5.22 indica que não ocorrem diferenças relevantes para os diferentes horizontes de previsão, já que os resultados apresentaram formas semelhantes entre si.

5.7 Considerações finais

Neste capítulo foram apresentados e discutidos os resultados das previsões de fluxo de tráfego geradas a partir de um modelo hierárquico *bayesiano*. Inicialmente, foi avaliada a convergência das cadeias de Markov com um exemplo visual seguido da visualização dos espaços e distribuições posteriores dos parâmetros. Em seguida, foram mostrados exemplos práticos da utilização das distribuições posteriores, com base na análise de acurácia e de intervalos de credibilidade. Por fim, foram mostradas e discutidas as distribuições dos parâmetros μ, σ e ν e de erros relativos e absolutos.

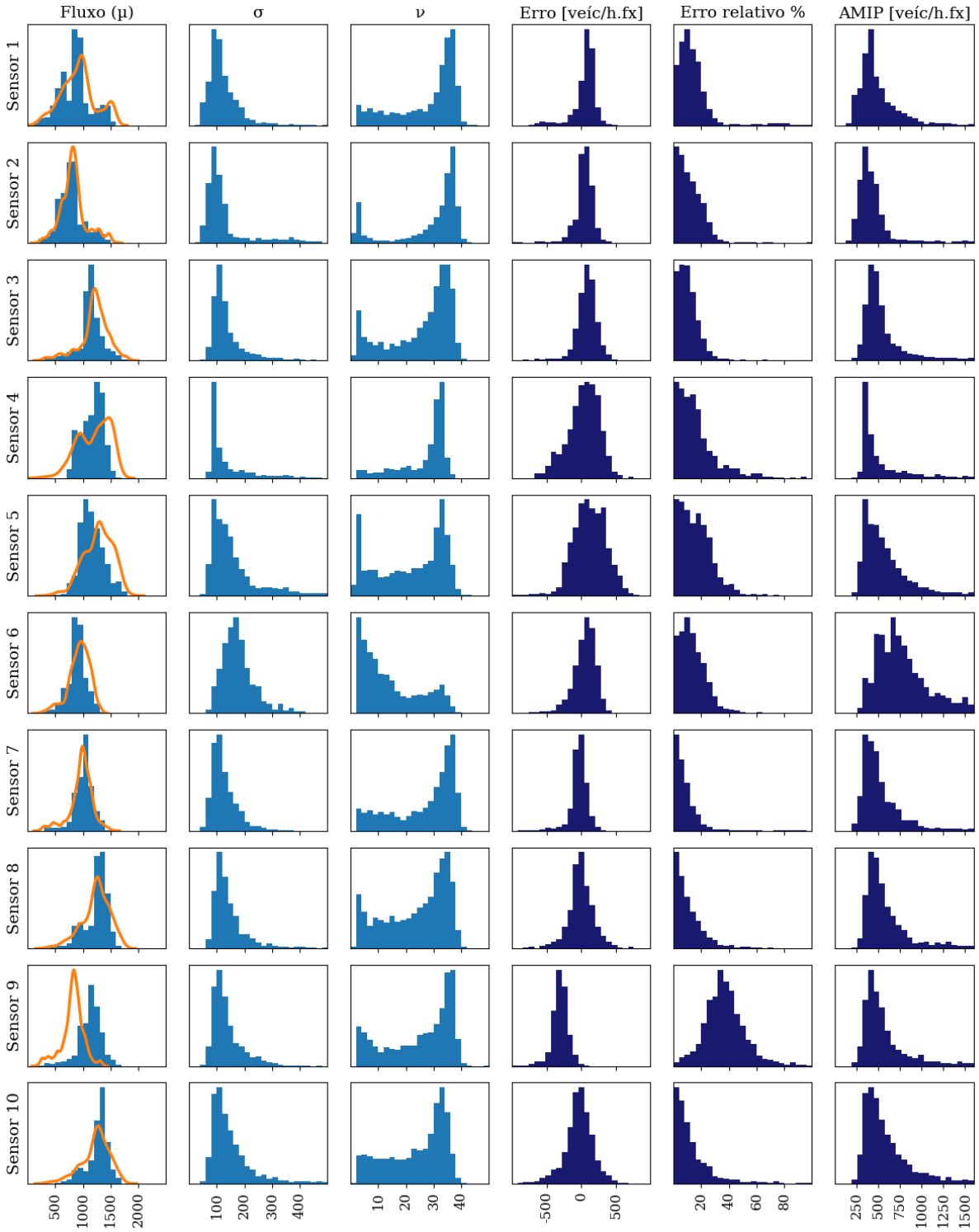


Figura 5.21: Distribuição de parâmetros previstos para cada sensor

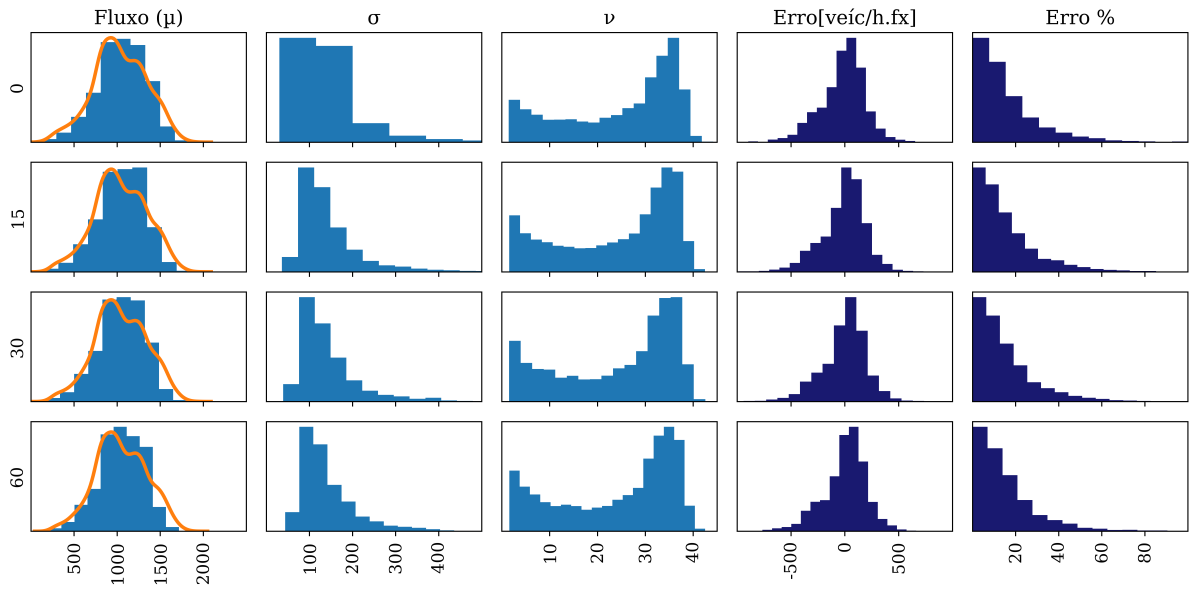


Figura 5.22: Distribuição de parâmetros previstos para cada horizonte de previsão

Capítulo

6 Conclusões e sugestões

Este trabalho propôs um método de estimação de fluxo de tráfego que forneceu intervalos de credibilidade para fluxos veiculares a partir de um modelo probabilístico, utilizando velocidades obtidas de uma fonte *online* como dados de entrada. A fim de cumprir este objetivo, os seguintes objetivos específicos foram atingidos:

1. Validou-se a aderência entre dados de velocidade *online* e dados de sensores;
2. Utilizaram-se dados históricos de fluxo veicular e de velocidade *online* para gerar distribuições de fluxo futuras a partir das velocidades observadas em tempo real;
3. Utilizou-se inferência *bayesiana* para prever fluxos viários, que resultaram em distribuições de probabilidade para os fluxos previstos.

Essas etapas tornaram possível responder às perguntas de pesquisa que motivaram o desenvolvimento deste trabalho, que consistiram em mostrar um modo de utilizar dados velocidade *online* para análises de tráfego e prever os intervalos de credibilidade de fluxos veiculares a partir de dados de velocidade. Nas seções seguintes são apresentadas as conclusões e principais contribuições deste estudo, bem como as recomendações para trabalhos futuros.

6.1 Conclusões

A maioria dos métodos de previsão de fluxo veicular não contempla análise de intervalos de credibilidade, dando enfoque majoritariamente na validação de seus modelos por medidas de acurácia entre valores previstos e observados. Na previsão de fluxo veicular, é pouco comum

a existência de abordagens que busquem intervalos de probabilidade que descrevam o comportamento dos parâmetros previstos. Nesse sentido, a principal contribuição deste trabalho consistiu em gerar previsões que contemplassem não apenas previsões pontuais, mas também intervalos de probabilidade para previsão de fluxo veicular.

Dados de velocidade *online* podem ser obtidos de maneira indireta a partir dos tempos de viagem fornecidos por API's de tráfego. No caso deste trabalho, esses dados se mostraram úteis quando aliados a dados de sensores, a fim de incorporar maior volume de informação à análise. Como dados fornecidos pela API têm custo por elemento requisitado, a estruturação de um banco de dados oriundos unicamente de fontes *online* poderia se tornar onerosa a um usuário final. Assim, unir ambas as fontes de informação potencializou a capacidade de análise no escopo deste trabalho.

As velocidades, grandezas físicas comuns a ambos dados históricos de sensores e da API não devem ser compatibilizadas de forma direta por possuírem naturezas distintas no tempo e no espaço, respectivamente. Assim, utilizando-se de uma abordagem baseada em dados, efetuou-se uma regressão linear *bayesiana* para relacionar ambas as fontes de informação. Dessa forma, a cada velocidade calculada pela API, conseguiu-se associar um intervalo de velocidades equivalentes de sensores.

A obtenção desses intervalos possibilitou agregar velocidades representativas de condições semelhantes segundo as quais foram geradas. Os subconjuntos gerados foram filtrados novamente de forma que condições referentes a dia da semana e horário do dia refinassem o nível de informação a eles atrelada. A projeção desses conjuntos em horizontes de previsão de curto prazo possibilitou gerar distribuições de fluxo que se configuraram como distribuições de forma aproximadamente normal, com desvios alocados simetricamente ao redor da média. No caso deste trabalho, optou-se por adotar uma distribuição de Student por permitir maior flexibilidade ao lidar com valores distantes da média.

A partir da análise exploratória de todos os conjuntos de fluxos, foi possível propor distribuições de parâmetros que pudessem ser incorporadas à análise *bayesiana*. Mesmo com um banco de dados extenso, observou-se a pequena quantidade de dados para alguns intervalos específicos, especialmente para velocidades mais baixas, que podem estar associadas a eventos atípicos e, por isso, não acumulam grande número de observações. Em comparação com os métodos usuais baseados em dados, um grande trunfo do método proposto foi possibilitar, por meio da incorporação da informação a *priori*, que as previsões pudessem ser realizadas apesar da restrição da quantidade de informações disponíveis. Este trabalho mostrou ainda que previsões realizadas com conjuntos contendo mais de 30 observações tenderam a ser influenciadas majoritariamente pelos dados, com reduzida influência das *prioris*. Isso indica que as previsões tendem a sofrer influência mais direta das prioris antes de os tamanhos dos conjuntos de dados atingirem esses patamares.

Embora genérico, o modelo é robusto para os sensores ou região de ocorrência a que se

referem, porque suas *prioris* puderam ser utilizadas para diferentes localidades, horários e condições de tráfego. Entretanto, é importante reforçar que essas distribuições foram geradas a partir de dados de tráfego de rodovias geometricamente similares, com ao menos duas faixas, localizadas na mesma região geográfica, além de apresentarem velocidades máximas regulamentadas entre 80 e 120 km/h. Por este motivo, não é possível afirmar que os valores das *prioris* podem ser estendidos para vias onde tais condições não forem observadas.

A utilização da distribuição de Student para descrever fluxos se mostrou bem sucedida para os objetivos a que se propôs, de modo que 98% dos conjuntos de dados convergiram logo na primeira configuração inicial proposta. Apesar de os outros 2% dos conjuntos de dados terem convergido nas rodadas seguintes sob novas configurações, é possível que alguns dos conjuntos de dados tenham falhado em convergir para a distribuição de Student por não apresentarem padrões capturáveis pelo modelo, como uma possível bimodalidade. Isso reduz a robustez do modelo e sugere que atenção deve ser dada caso se observe a multimodalidade.

Esse modelo é indicado para obter referenciais acerca da incerteza, ou da variabilidade para a previsão de fluxo veicular, em detrimento unicamente da acurácia. Nesse sentido, não se faz necessária a utilização de um valor específico da amostra, e nem cabe ao modelo classificar um valor previsto como certo ou errado, mas sim, garantir que as previsões estejam dentro de um intervalo provável. Sendo assim, não é pretensão deste trabalho comparar o modelo proposto com outros métodos a fim de averiguar sua superioridade ou inferioridade. Desse modo, sugere-se utilizá-lo tanto individualmente quanto como recurso adicional a outros métodos de previsão, a fim de garantir que valores previstos por outros métodos se encontrem dentro de um intervalo de valores provável.

Por fim, pode se concluir que a principal contribuição deste trabalho foi a criação de um método de agregação de fluxo de dados, aliando um banco de dados de fluxos e velocidades históricos a dados *online* de velocidade, que permitiu incorporar variabilidade e incerteza e não necessitou de grandes quantidades de dados, podendo ser estendida para regiões locais que apresentem condições semelhantes.

6.2 Recomendações para trabalhos futuros

O método proposto nesse trabalho evidenciou a necessidade de estabelecer distribuições a *priori* informativas para o modelo. Reconhecendo sua relevância, sugere-se buscar *prioris* mais específicas pela incorporação de características a serem exploradas por outros métodos nos conjuntos de fluxos. Com novas *prioris*, é possível comparar futuramente os novos métodos por meio do fator *bayesiano* ou outras técnicas de comparação de distribuições a *posteriori* contempladas na estatística *bayesiana*, a fim de obter uma avaliação quantitativa da qualidade entre modelos.

Por outro lado, além da implementação de outros modelos *bayseianos*, sugere-se associar o

método proposto a outros métodos correntes de previsão. Dessa forma, é possível validar os valores previstos por esses modelos considerando os intervalos de credibilidade previstos por inferência *bayesiana*, além de incorporar a incerteza a essas previsões.

Fatores exógenos, como trancamento de pista e incidentes de trânsito e alterações climáticas, como chuvas intensas, podem ser considerados como variáveis de entrada adicionais ao modelo de previsão. Para tanto, indica-se gerar bancos de dados específicos para as ocorrências que buscam prever, o que consiste por si só como tarefa custosa e de difícil execução. Tal procedimento tem o potencial de tornar o modelo mais robusto, embora mais complexo.

É possível que parte dos conjuntos cujas previsões não renderam resultados confiáveis sejam afetados por padrões não representados pela distribuição de verossimilhança proposta, como multimodalidade. Nesse caso, outra abordagem poderia ser proposta, como uma mistura gaussiana, por exemplo, a fim de capturar esse comportamento, o que pode aumentar o nível de complexidade do modelo, mas ao mesmo tempo troná-lo mais robusto.

Por fim, recomenda-se implementar o procedimento de atualização da posteriori quando da adição de novos dados observados. A inferência *bayesiana* possui o recurso de incorporar novas informações às previsões passadas sem necessariamente ter de incluir todos os dados novamente. Isso confere praticidade em atividades de monitoramento de rodovias, por exemplo, que contam com fluxo contínuo de geração de dados e podem se utilizar desses intervalos de fluxo para complementar o monitoramento em tempo real.

Referências

ABDOLLAHI, M.; KHALEGHI, T.; YANG, K. An integrated feature learning approach using deep learning for travel time prediction. *Expert Systems with Applications*, 2020. Elsevier, v. 139, p. 112864, 2020.

AHMED, M. S.; COOK, A. R. Analysis of freeway traffic time-series data by using box-jenkins techniques. *Transportation Research Record*, 1979. Transportation Research Board, n. 722, p. 1–9, 1979.

AMBÜHL, L.; MENENDEZ, M. Data fusion algorithm for macroscopic fundamental diagram estimation. *Transportation Research Part C: Emerging Technologies*, 2016. Elsevier, v. 71, p. 184–197, 2016.

AREM, B. V. et al. Recent advances and applications in the field of short-term traffic forecasting. *International Journal of Forecasting*, 1997. Elsevier, v. 13, n. 1, p. 1–12, 1997.

BACHMANN, C. et al. A comparative assessment of multi-sensor data fusion techniques for freeway traffic speed estimation using microsimulation modeling. *Transportation research part C: emerging technologies*, 2013. Elsevier, v. 26, p. 33–48, 2013.

BETANCOURT, M. A conceptual introduction to hamiltonian monte carlo. *arXiv preprint arXiv:1701.02434*, 2017. 2017.

BOQUET, G. et al. A variational autoencoder solution for road traffic forecasting systems: Missing data imputation, dimension reduction, model selection and anomaly detection. *Transportation Research Part C: Emerging Technologies*, 2020. Elsevier, v. 115, p. 102622, 2020.

CACERES, N. et al. Traffic flow estimation models using cellular phone data. *IEEE Transactions on Intelligent Transportation Systems*, 2012. IEEE, v. 13, n. 3, p. 1430–1441, 2012.

CAREY, M.; BOWERS, M. A review of properties of flow–density functions. *Transport Reviews*, 2012. Taylor & Francis, v. 32, n. 1, p. 49–73, 2012.

CHANDRA, S. R.; AL-DEEK, H. Predictions of freeway traffic speeds and volumes using vector autoregressive models. *Journal of Intelligent Transportation Systems*, 2009. Taylor & Francis, v. 13, n. 2, p. 53–72, 2009.

CHEN, C. et al. Short-time traffic flow prediction with arima-garch model. In: IEEE. *2011 IEEE Intelligent Vehicles Symposium (IV)*. [S.l.], 2011. p. 607–612.

CUNHA, A. L. B. N. d. *Avaliação do impacto da medida de desempenho no equivalente veicular de caminhões*. Tese (Doutorado) — Universidade de São Paulo, 2007.

DAI, X. et al. Deeptrend 2.0: A light-weighted multi-scale traffic prediction model using detrending. *Transportation Research Part C: Emerging Technologies*, 2019. Elsevier, v. 103, p. 142–157, 2019.

DJENOURI, Y.; ZIMEK, A.; CHIARANDINI, M. Outlier detection in urban traffic flow distributions. In: IEEE. *2018 IEEE International Conference on Data Mining (ICDM)*. [S.l.], 2018. p. 935–940.

DO, L. N. et al. An effective spatial-temporal attention based neural network for traffic flow prediction. *Transportation research part C: emerging technologies*, 2019. Elsevier, v. 108, p. 12–28, 2019.

DUAN, P. et al. Estimation of link travel time distribution with limited traffic detectors. *IEEE Transactions on Intelligent Transportation Systems*, 2019. IEEE, v. 21, n. 9, p. 3730–3743, 2019.

ERMAGUN, A.; LEVINSON, D. Spatiotemporal traffic forecasting: review and proposed directions. *Transport Reviews*, 2018. Taylor & Francis, v. 38, n. 6, p. 786–814, 2018.

FEI, X.; LU, C.-C.; LIU, K. A bayesian dynamic linear model approach for real-time short-term freeway travel time prediction. *Transportation Research Part C: Emerging Technologies*, 2011. Elsevier, v. 19, n. 6, p. 1306–1318, 2011.

GABRY, J. et al. Visualization in bayesian workflow. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 2019. Wiley Online Library, v. 182, n. 2, p. 389–402, 2019.

GARBER, N. J.; HOEL, L. A. *Traffic & highway engineering*. [S.l.]: Cengage Learning, 2008.

GELMAN, A. et al. Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian analysis*, 2006. International Society for Bayesian Analysis, v. 1, n. 3, p. 515–534, 2006.

GELMAN, A.; SIMPSON, D.; BETANCOURT, M. The prior can often only be understood in the context of the likelihood. *Entropy*, 2017. Multidisciplinary Digital Publishing Institute, v. 19, n. 10, p. 555, 2017.

GHADER, S.; DARZI, A.; ZHANG, L. Modeling effects of travel time reliability on mode choice using cumulative prospect theory. *Transportation Research Part C: Emerging Technologies*, 2019. Elsevier, v. 108, p. 245–254, 2019.

GHOSH, B.; BASU, B.; O'MAHONY, M. Multivariate short-term traffic flow forecasting using time-series analysis. *IEEE transactions on intelligent transportation systems*, 2009. IEEE, v. 10, n. 2, p. 246–254, 2009.

GHOSH, B.; BASU, B.; O'MAHONY, M. Bayesian time-series model for short-term traffic flow forecasting. *Journal of transportation engineering*, 2007. American Society of Civil Engineers, v. 133, n. 3, p. 180–189, 2007.

GHOSH, B.; BASU, B.; O'MAHONY, M. Random process model for urban traffic flow using a wavelet-bayesian hierarchical technique. *Computer-Aided Civil and Infrastructure Engineering*, 2010. Wiley Online Library, v. 25, n. 8, p. 613–624, 2010.

GOOGLE. 2021. `https://cloud.google.com/maps-platform/`. Visitado em 26.4.2021.

GOOGLE. 2021. `https://developers.google.com/maps/documentation/distance-matrix/overview`. Visitado em 26.4.2021.

GU, Z. et al. A big data approach for clustering and calibration of link fundamental diagrams for large-scale network simulation applications. *Transportation Research Part C: Emerging Technologies*, 2018. Elsevier, v. 94, p. 151–171, 2018.

GUESSOUS, Y. et al. Estimating travel time distribution under different traffic conditions. *Transportation Research Procedia*, 2014. Elsevier, v. 3, p. 339–348, 2014.

GUO, F. et al. Predictor fusion for short-term traffic forecasting. *Transp. Res. C, Emerg. Technol.*, 2018. v. 92, p. 90–100, 2018.

GUO, J.; HUANG, W.; WILLIAMS, B. M. Adaptive kalman filter approach for stochastic short-term traffic flow rate prediction and uncertainty quantification. *Transportation Research Part C: Emerging Technologies*, 2014. Elsevier, v. 43, p. 50–64, 2014.

HABTEMICHAEL, F. G.; CETIN, M. Short-term traffic flow rate forecasting based on identifying similar traffic patterns. *Transportation research Part C: emerging technologies*, 2016. Elsevier, v. 66, p. 61–78, 2016.

HE, S. et al. Freeway multisensor data fusion approach integrating data from cellphone probes and fixed sensors. *Journal of Sensors*, 2016. Hindawi, v. 2016, 2016.

HINSBERGEN, C. P. V. et al. Localized extended kalman filter for scalable real-time traffic state estimation. *IEEE transactions on intelligent transportation systems*, 2011. IEEE, v. 13, n. 1, p. 385–394, 2011.

KARLAFTIS, M. G.; VLAHOGIANNI, E. I. Statistical methods versus neural networks in transportation research: Differences, similarities and some insights. *Transportation Research Part C: Emerging Technologies*, 2011. Elsevier, v. 19, n. 3, p. 387–399, 2011.

KIDANDO, E. et al. Bayesian nonparametric model for estimating multistate travel time distribution. *Journal of Advanced Transportation*, 2017. Hindawi, v. 2017, 2017.

KRUSCHKE, J. K. *Doing Bayesian data analysis: A tutorial with R*. [S.l.: s.n.], 2014.

KUMAR, R. et al. Arviz a unified library for exploratory analysis of Bayesian models in Python. *Journal of Open Source Software*, 2019. The Open Journal, v. 4, n. 33, p. 1143, 2019. Disponível em: <https://doi.org/10.21105/joss.01143>.

KUMAR, S. V. Traffic flow prediction using kalman filtering technique. *Procedia Engineering*, 2017. Elsevier, v. 187, p. 582–587, 2017.

LANA, I. et al. Road traffic forecasting: recent advances and new challenges. *IEEE Intelligent Transportation Systems Magazine*, 2018. IEEE, v. 10, n. 2, p. 93–109, 2018.

LEDOUX, C. An urban traffic flow model integrating neural networks. *Transportation Research Part C: Emerging Technologies*, 1997. Elsevier, v. 5, n. 5, p. 287–300, 1997.

LEDUC, G. Road traffic data: Collection methods and applications. *Working Papers on Energy, Transport and Climate Change*, 2008. European Commission Joint Research Centre Institute for Prospective Technological Studies, v. 1, n. 55, p. 1–55, 2008.

LI, J.-Q. et al. Estimating queue length under connected vehicle technology: Using probe vehicle, loop detector, and fused data. *Transportation research record*, 2013. SAGE Publications Sage CA: Los Angeles, CA, v. 2356, n. 1, p. 17–22, 2013.

LI, L.; CHEN, X. M. Vehicle headway modeling and its inferences in macroscopic/microscopic traffic flow theory: A survey. *Transportation Research Part C: Emerging Technologies*, 2017. Elsevier, v. 76, p. 170–188, 2017.

LI, W. et al. Short-term traffic state prediction from latent structures: Accuracy vs. efficiency. *Transportation Research Part C: Emerging Technologies*, 2020. Elsevier, v. 111, p. 72–90, 2020.

LIN, L. et al. Quantifying uncertainty in short-term traffic prediction and its application to optimal staffing plan development. *Transportation Research Part C: Emerging Technologies*, 2018. Elsevier, v. 92, p. 323–348, 2018.

LIPPI, M.; BERTINI, M.; FRASCONI, P. Short-term traffic flow forecasting: An experimental comparison of time-series analysis and supervised learning. *IEEE Transactions on Intelligent Transportation Systems*, 2013. IEEE, v. 14, n. 2, p. 871–882, 2013.

LV, Y. et al. Traffic flow prediction with big data: a deep learning approach. *IEEE Transactions on Intelligent Transportation Systems*, 2014. IEEE, v. 16, n. 2, p. 865–873, 2014.

MARTIN, O. *Bayesian Analysis with Python*. [S.l.]: Packt Publishing Ltd, 2016.

MARTÍNEZ-DÍAZ, M.; PÉREZ, I. A simple algorithm for the estimation of road traffic space mean speeds from data available to most management centres. *Transportation Research Part B: Methodological*, 2015. Elsevier, v. 75, p. 19–35, 2015.

MCELREATH, R. *Statistical rethinking: A Bayesian course with examples in R and Stan*. [S.l.]: CRC press, 2020.

MEI, Y. et al. A bayesian approach for estimating vehicle queue lengths at signalized intersections using probe vehicle data. *Transportation Research Part C: Emerging Technologies*, 2019. Elsevier, v. 109, p. 233–249, 2019.

MORGUL, E. F. et al. Virtual sensors: Web-based real-time data collection methodology for transportation operation performance analysis. *Transportation Research Record*, 2014. SAGE Publications Sage CA: Los Angeles, CA, v. 2442, n. 1, p. 106–116, 2014.

MORI, U. et al. A review of travel time estimation and forecasting for advanced traveller information systems. *Transportmetrica A: Transport Science*, 2015. Taylor & Francis, v. 11, n. 2, p. 119–157, 2015.

O'HAGAN, A. et al. *Uncertain judgements: eliciting experts' probabilities*. [S.l.]: John Wiley & Sons, 2006.

OKUTANI, I.; STEPHANEDES, Y. J. Dynamic prediction of traffic volume through kalman filtering theory. *Transportation Research Part B: Methodological*, 1984. Elsevier, v. 18, n. 1, p. 1–11, 1984.

OLIVEIRA, G. M. de; CUNHA, A. L. A calibration method structured on bayesian inference of the hcm speed-flow relationship for freeways and multilane highways and a temporal analysis of traffic behavior. *arXiv preprint arXiv:1908.10852*, 2019. 2019.

POLSON, N. G.; SOKOLOV, V. O. Deep learning for short-term traffic flow prediction. *Transportation Research Part C: Emerging Technologies*, 2017. Elsevier, v. 79, p. 1–17, 2017.

QIAN, Z. S. et al. Modeling heterogeneous traffic flow: A pragmatic approach. *Transportation Research Part B: Methodological*, 2017. Elsevier, v. 99, p. 183–204, 2017.

QUEEN, C. M.; ALBERS, C. J. Intervention and causality: forecasting traffic flows using a dynamic bayesian network. *Journal of the American Statistical Association*, 2009. Taylor & Francis, v. 104, n. 486, p. 669–681, 2009.

RAKHA, H.; ZHANG, W. Estimating traffic stream space mean speed and reliability from dual-and single-loop detectors. *Transportation Research Record*, 2005. SAGE Publications Sage CA: Los Angeles, CA, v. 1925, n. 1, p. 38–47, 2005.

RIBEIRO, E. R. *Análise exploratória de método utilizando Wavelet para detecção de padrões e anomalias em dados históricos do tráfego veicular*. Tese (Doutorado) — Universidade de São Paulo, 2017.

ROESS, R. P.; PRASSAS, E. S.; MCSHANE, W. R. *Traffic engineering*. [S.l.]: Pearson/Prentice Hall, 2010.

RYU, U. et al. Construction of traffic state vector using mutual information for short-term traffic flow prediction. *Transportation Research Part C: Emerging Technologies*, 2018. Elsevier, v. 96, p. 55–71, 2018.

SAFFARI, E.; YILDIRIMOGLU, M.; HICKMAN, M. A methodology for identifying critical links and estimating macroscopic fundamental diagram in large-scale urban networks. *Transportation Research Part C: Emerging Technologies*, 2020. Elsevier, v. 119, p. 102743, 2020.

SALVATIER, J.; WIECKI, T. V.; FONNESBECK, C. Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, 2016. PeerJ Inc., v. 2, p. e55, 2016.

SCHOOT, R. V. D. et al. Analyzing small data sets using bayesian estimation: The case of posttraumatic stress symptoms following mechanical ventilation in burn survivors. *European Journal of Psychotraumatology*, 2015. Taylor & Francis, v. 6, n. 1, p. 25216, 2015.

SCHOOT, R. van de; DEPAOLI, S. Bayesian analyses: Where to start and what to report. *The European Health Psychologist*, 2014. v. 16, n. 2, p. 75–84, 2014.

SEKUŁA, P. et al. Estimating historical hourly traffic volumes via machine learning and vehicle probe data: A maryland case study. *Transportation Research Part C: Emerging Technologies*, 2018. Elsevier, v. 97, p. 147–158, 2018.

SEO, T. et al. Traffic state estimation on highway: A comprehensive survey. *Annual reviews in control*, 2017. Elsevier, v. 43, p. 128–151, 2017.

SHAHRIARI, S. et al. Ensemble of arima: combining parametric and bootstrapping technique for traffic flow prediction. *Transportmetrica A: Transport Science*, 2020. Taylor & Francis, v. 16, n. 3, p. 1552–1573, 2020.

SIMPSON, D. et al. Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical science*, 2017. Institute of Mathematical Statistics, v. 32, n. 1, p. 1–28, 2017.

SMITH, B. L.; DEMETSKY, M. J. Short-term traffic flow prediction models-a comparison of neural network and nonparametric regression approaches. In: IEEE. *Proceedings of IEEE international conference on systems, man and cybernetics*. [S.l.], 1994. v. 2, p. 1706–1709.

SORIGUERA, F.; ROBUSTÉ, F. Estimation of traffic stream space mean speed from time aggregations of double loop detector data. *Transportation research part C: emerging technologies*, 2011. Elsevier, v. 19, n. 1, p. 115–129, 2011.

TAN, M.-C. et al. An aggregation approach to short-term traffic flow prediction. *IEEE Transactions on Intelligent Transportation Systems*, 2009. IEEE, v. 10, n. 1, p. 60–69, 2009.

TEBALDI, C.; WEST, M.; KARR, A. F. Statistical analyses of freeway traffic flows. *Journal of Forecasting*, 2002. Wiley Online Library, v. 21, n. 1, p. 39–68, 2002.

Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, 2016. abs/1605.02688, 2016. Disponível em: <http://arxiv.org/abs/1605.02688>.

VEHTARI, A. et al. Rank-normalization, folding, and localization: An improved $\hat{R}$ for assessing convergence of MCMC. *Bayesian Analysis*, 2020. International Society for Bayesian Analysis, 2020.

VIRTANEN, P. et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 2020. v. 17, p. 261–272, 2020.

VLAHOGIANNI, E. I.; KARLAFTIS, M. G.; GOLIAS, J. C. Optimized and meta-optimized neural networks for short-term traffic flow prediction: A genetic approach. *Transportation Research Part C: Emerging Technologies*, 2005. Elsevier, v. 13, n. 3, p. 211–234, 2005.

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*, 2014. Elsevier, v. 43, p. 3–19, 2014.

WAGENINGEN-KESSELS, F. van et al. Genealogy of traffic flow models. *EURO Journal on Transportation and Logistics*, 2015. Springer, v. 4, n. 4, p. 445–473, 2015.

WANG, C. et al. A novel approach to estimate freeway traffic state: parallel computing and improved kalman filter. *IEEE Intelligent Transportation Systems Magazine*, 2018. IEEE, v. 10, n. 2, p. 180–193, 2018.

WANG, C.; YE, Z. Traffic flow forecasting based on a hybrid model. *Journal of Intelligent Transportation Systems*, 2016. Taylor & Francis, v. 20, n. 5, p. 428–437, 2016.

WANG, F.; XU, Y. Estimating o–d travel time matrix by google maps api: implementation, advantages, and implications. *Annals of GIS*, 2011. Taylor & Francis, v. 17, n. 4, p. 199–209, 2011.

WANG, J.; DENG, W.; GUO, Y. New bayesian combination method for short-term traffic flow forecasting. *Transportation Research Part C: Emerging Technologies*, 2014. Elsevier, v. 43, p. 79–94, 2014.

WANG, Y.; PAPAGEORGIOU, M. Real-time freeway traffic state estimation based on extended kalman filter: a general approach. *Transportation Research Part B: Methodological*, 2005. Elsevier, v. 39, n. 2, p. 141–167, 2005.

WARDROP, J. G. Some theoretical aspects of road traffic research. *Proceedings of the institution of civil engineers*, 1952. Thomas Telford-ICE Virtual Library, v. 1, n. 3, p. 325–362, 1952.

WASKOM, M. et al. *mwaskom/seaborn: v0.9.0 (July 2018)*. Zenodo, jul. 2018. Disponível em: <https://doi.org/10.5281/zenodo.1313201>.

WILLIAMS, B. M. Multivariate vehicular traffic flow prediction: evaluation of arimax modeling. *Transportation Research Record*, 2001. SAGE Publications Sage CA: Los Angeles, CA, v. 1776, n. 1, p. 194–200, 2001.

WILLIAMS, B. M.; DURVASULA, P. K.; BROWN, D. E. Urban freeway traffic flow prediction: application of seasonal autoregressive integrated moving average and exponential smoothing models. *Transportation Research Record*, 1998. SAGE Publications Sage CA: Los Angeles, CA, v. 1644, n. 1, p. 132–141, 1998.

WILLIAMS, B. M.; HOEL, L. A. Modeling and forecasting vehicular traffic flow as a seasonal arima process: Theoretical basis and empirical results. *Journal of transportation engineering*, 2003. American Society of Civil Engineers, v. 129, n. 6, p. 664–672, 2003.

WOLSHON, B.; PANDE, A. et al. *Traffic engineering handbook*. [S.l.]: John Wiley & Sons, 2016.

WU, Y. et al. A hybrid deep learning based traffic flow prediction method and its understanding. *Transportation Research Part C: Emerging Technologies*, 2018. Elsevier, v. 90, p. 166–180, 2018.

XU, C.; LI, Z.; WANG, W. Short-term traffic flow prediction using a methodology based on autoregressive integrated moving average and genetic programming. *Transport*, 2016. Taylor & Francis, v. 31, n. 3, p. 343–358, 2016.

XU, D.-w. et al. Real-time road traffic state prediction based on arima and kalman filter. *Frontiers of Information Technology & Electronic Engineering*, 2017. Springer, v. 18, n. 2, p. 287–302, 2017.

YAO, R.; ZHANG, W.; ZHANG, L. Hybrid methods for short-term traffic flow prediction based on arima-garch model and wavelet neural network. *Journal of Transportation Engineering, Part A: Systems*, 2020. American Society of Civil Engineers, v. 146, n. 8, p. 04020086, 2020.

YIN, H. et al. Urban traffic flow prediction using a fuzzy-neural approach. *Transportation Research Part C: Emerging Technologies*, 2002. Elsevier, v. 10, n. 2, p. 85–98, 2002.

ZHANG, Y.; ZHANG, Y.; HAGHANI, A. A hybrid short-term traffic flow forecasting method based on spectral analysis and statistical volatility model. *Transportation Research Part C: Emerging Technologies*, 2014. Elsevier, v. 43, p. 65–78, 2014.

ZHENG, W.; LEE, D.-H.; SHI, Q. Short-term freeway traffic flow prediction: Bayesian combined neural network approach. *Journal of transportation engineering*, 2006. American Society of Civil Engineers, v. 132, n. 2, p. 114–121, 2006.

ZHU, Z. et al. Integrating probabilistic tensor factorization with bayesian supervised learning for dynamic ridesharing pattern analysis. *Transportation Research Part C: Emerging Technologies*, 2021. Elsevier, v. 124, p. 102916, 2021.

ZHU, Z. et al. A generalized bayesian traffic model. *Transportation research part C: emerging technologies*, 2019. Elsevier, v. 108, p. 182–206, 2019.

Apêndices

APÊNDICE

A Previsões

Gráficos das previsões ao longo das duas semanas separados por sensor, contendo horários das 6 às 20 horas, para horizontes de previsão de 0, 30 e 60 minutos, respectivamente.

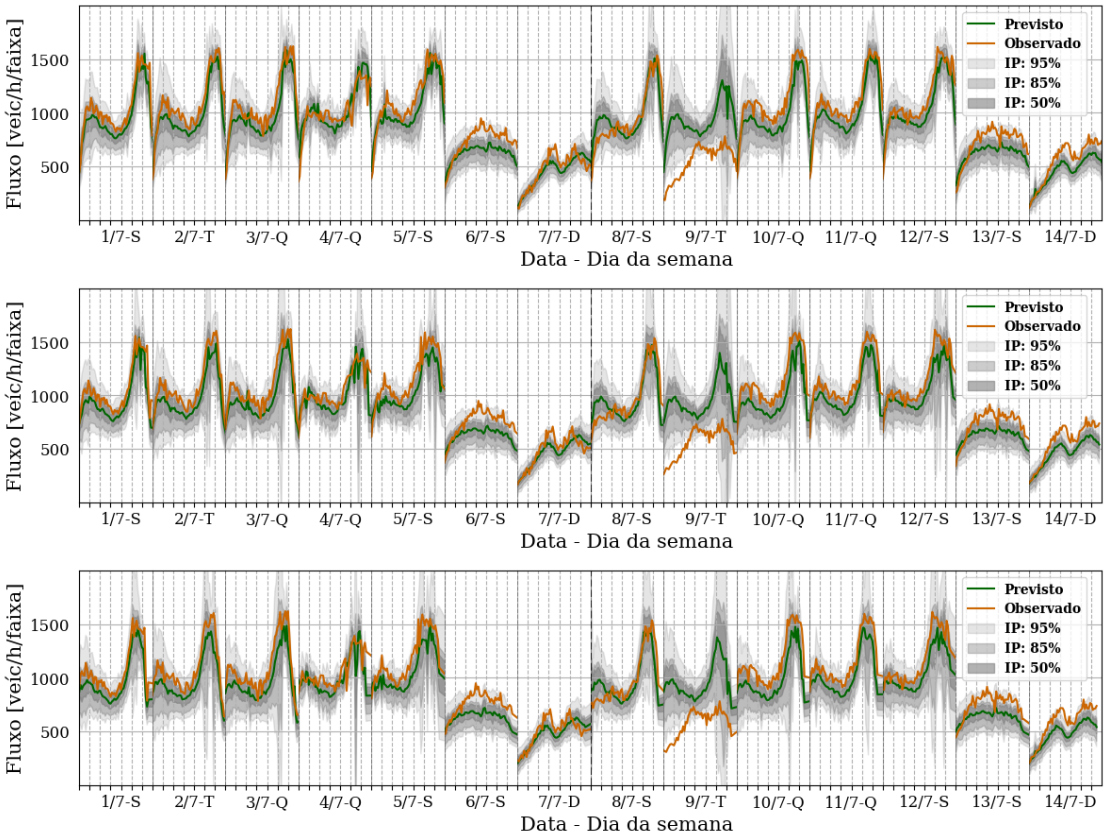


Figura A.1: Previsões para Sensor 1 na SP-021, km 18.3, sentido externo

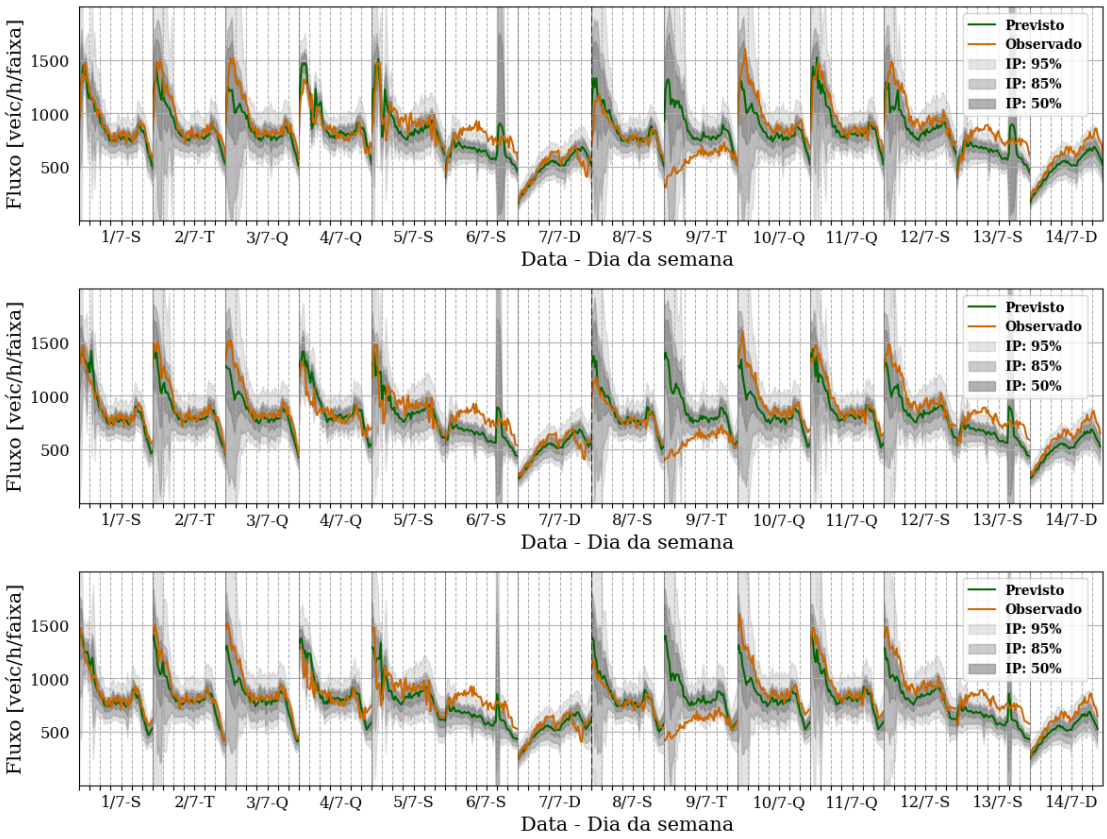


Figura A.2: Previsões para Sensor 2 na SP-021, km 18.3, sentido interno

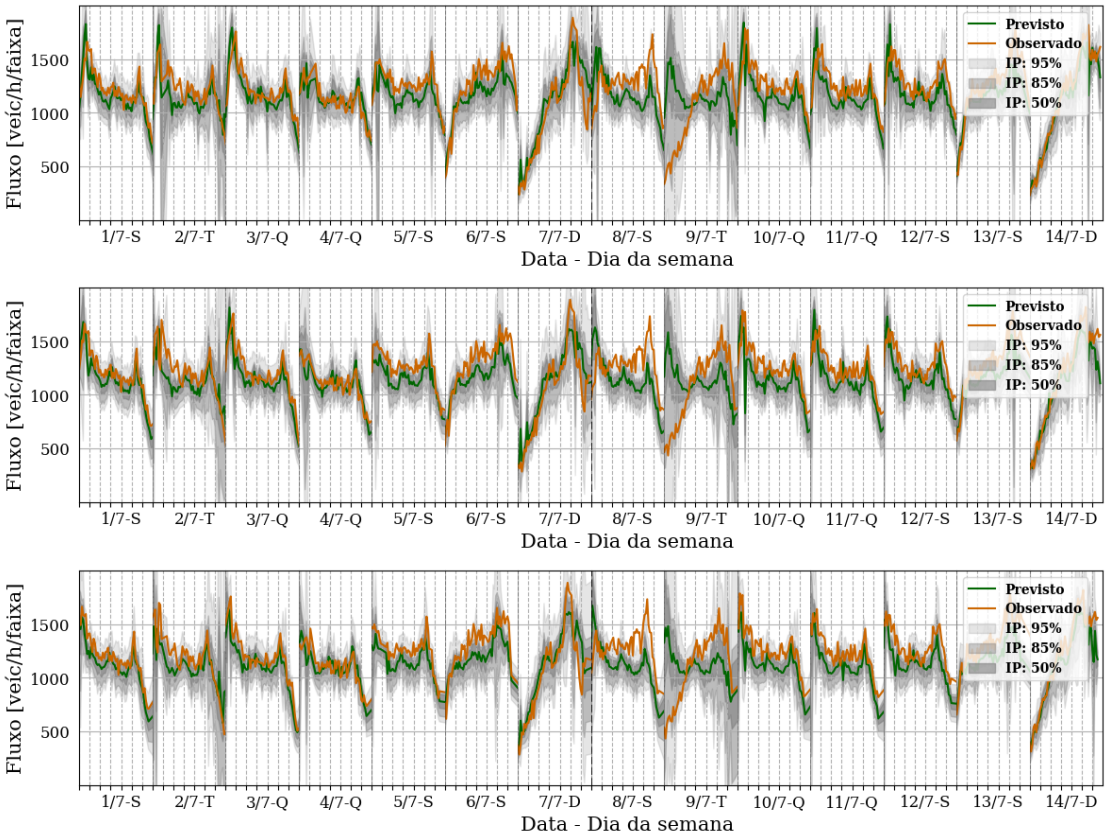


Figura A.3: Previsões para Sensor 3 na SP-270, km 34, sentido leste

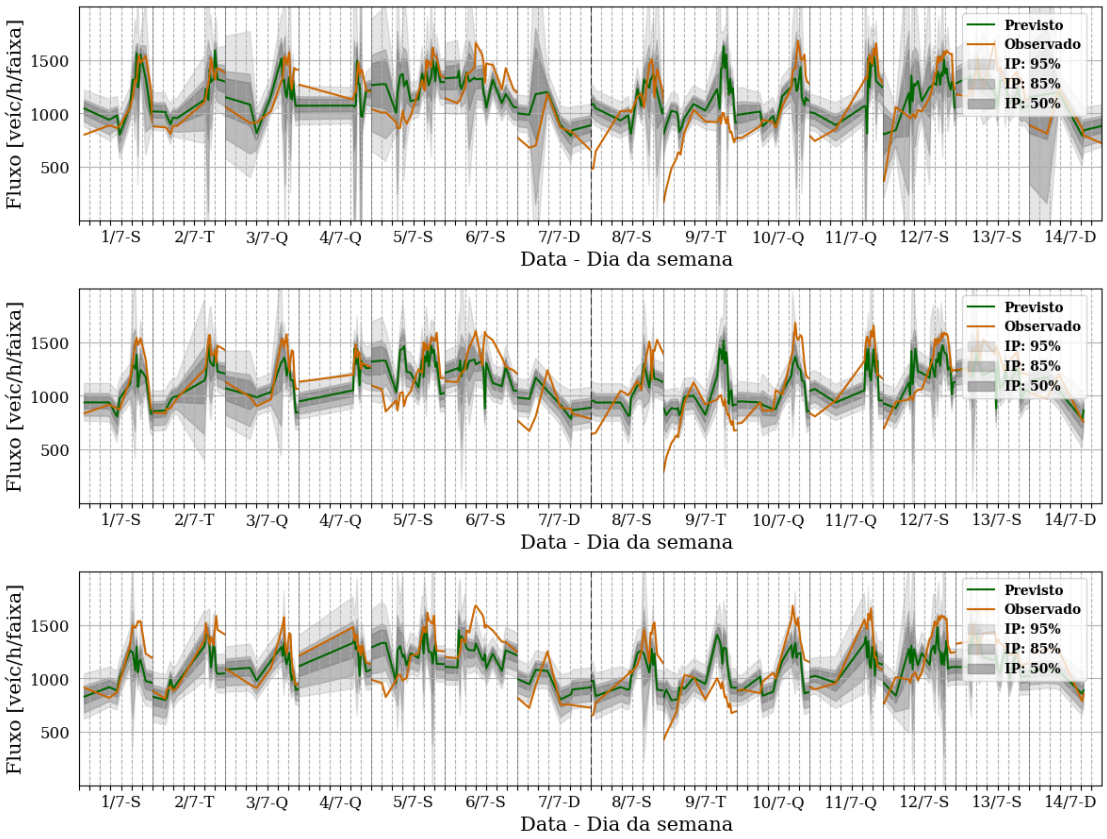


Figura A.4: Previsões para Sensor 4 na SP-270, km 34, sentido oeste

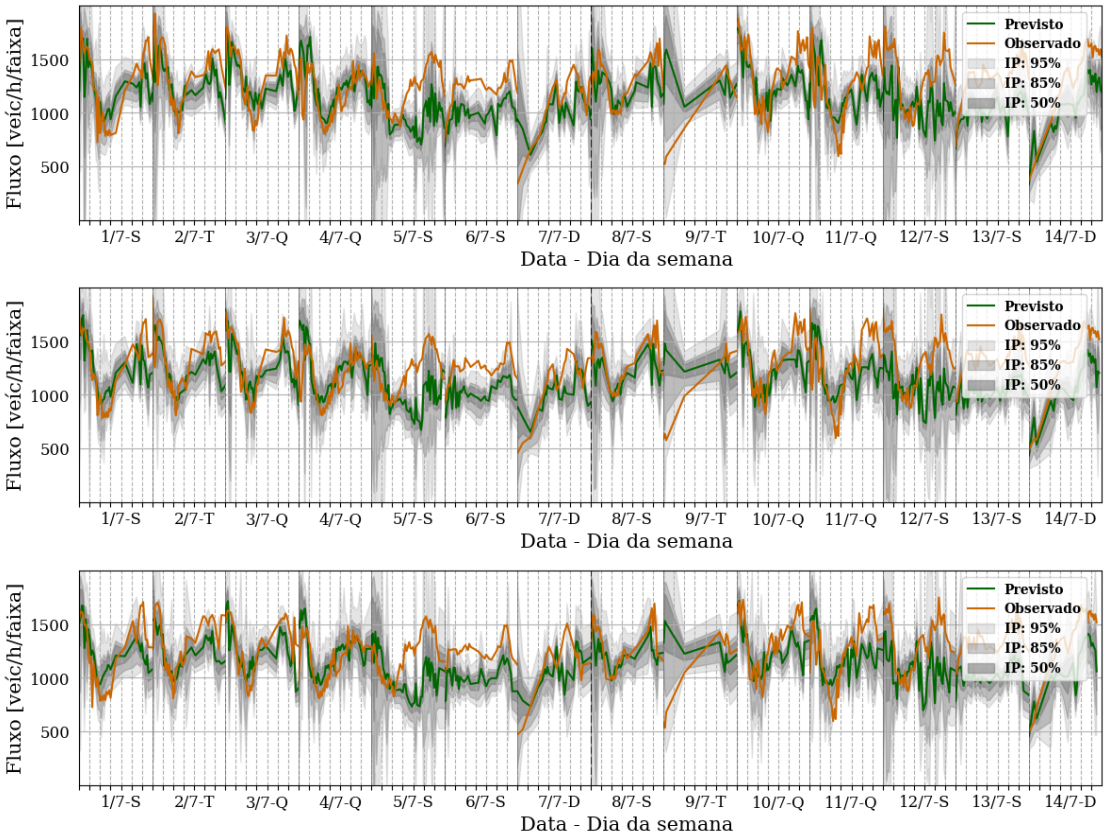


Figura A.5: Previsões para Sensor 5 na SP-280, km 16.08, sentido leste

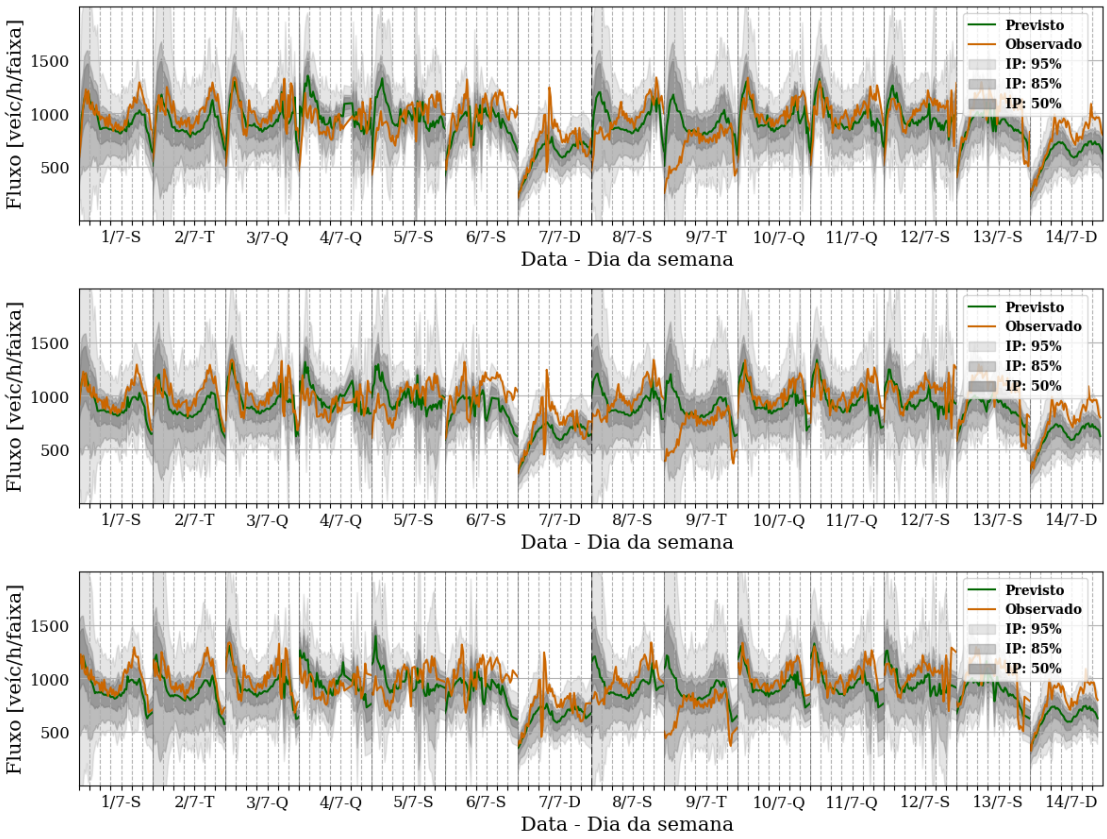


Figura A.6: Previsões para Sensor 6 na SP-280, km 22.04, sentido oeste

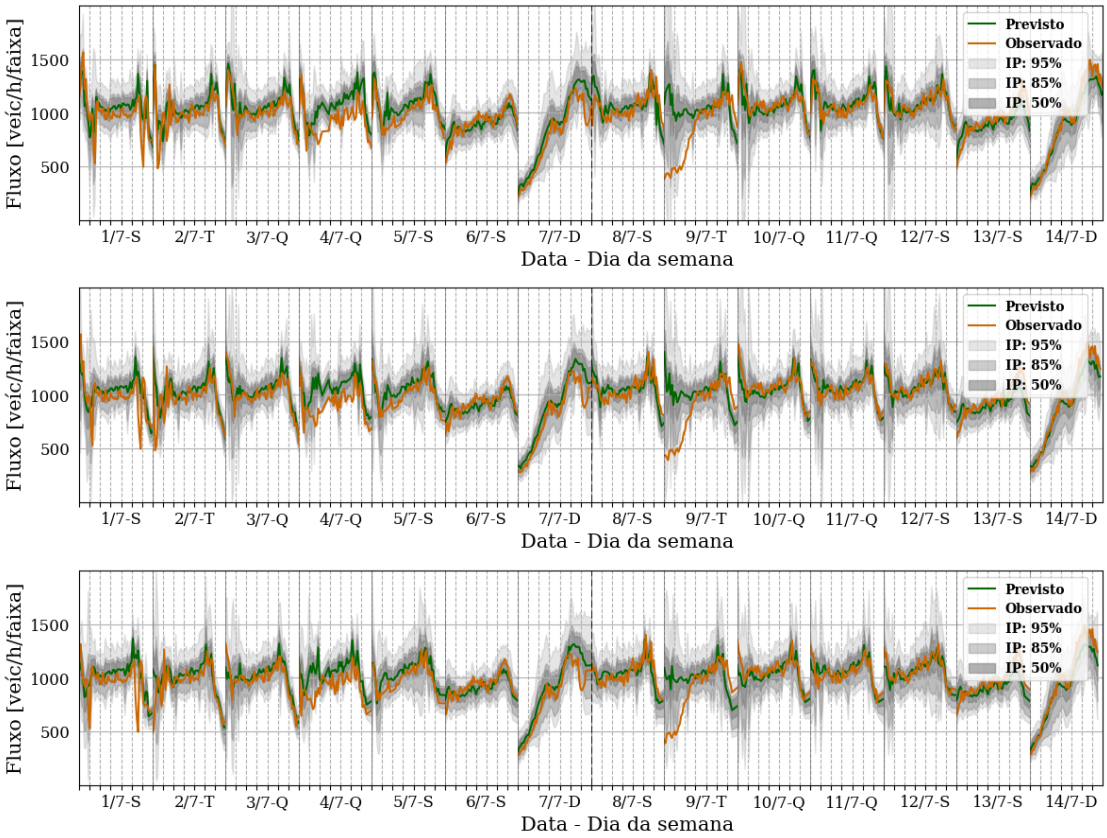


Figura A.7: Previsões para Sensor 7 na SP-280, km 27, sentido leste

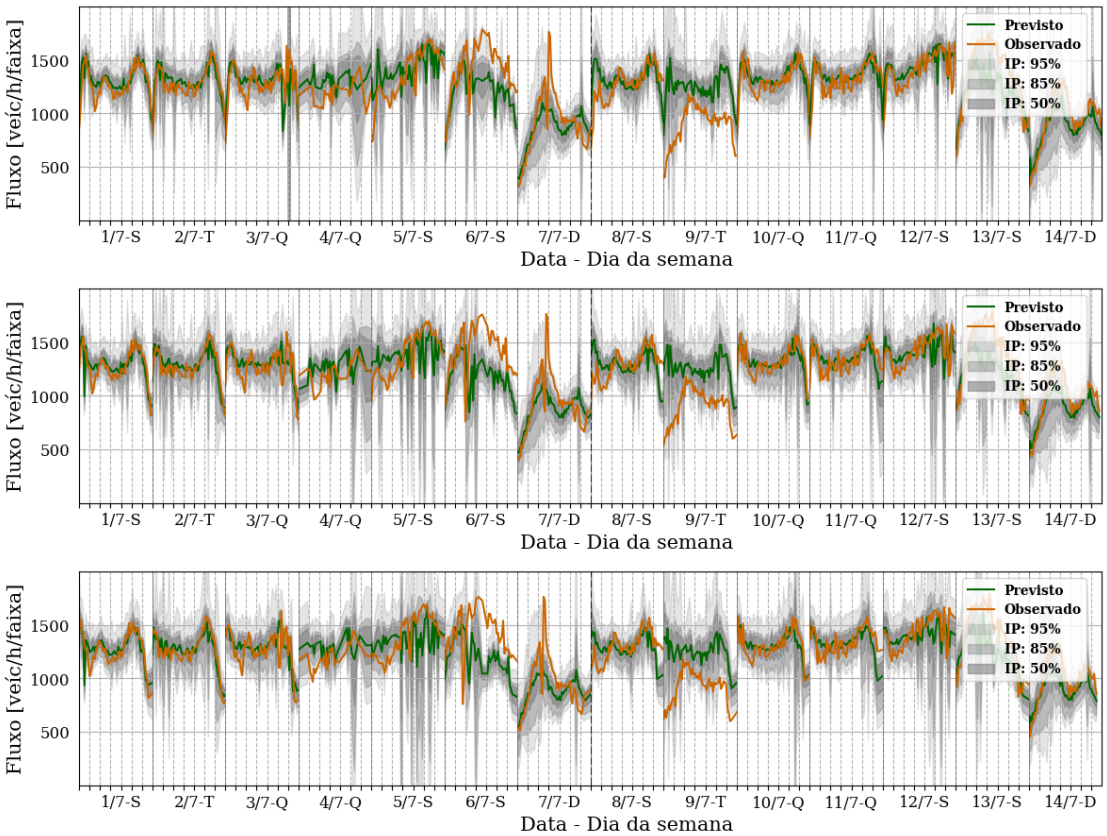


Figura A.8: Previsões para Sensor 8 na SP-280, 27, sentido oeste

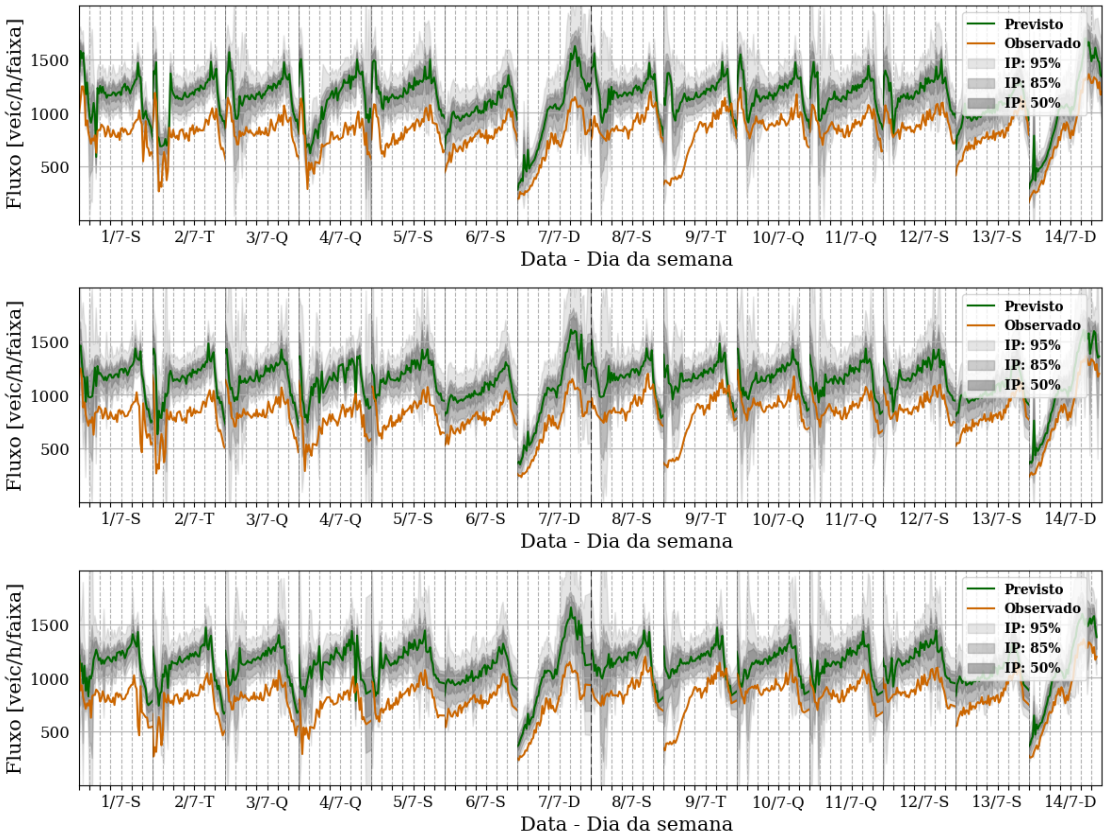


Figura A.9: Previsões para Sensor 9 na SP-280, 29.5, sentido leste

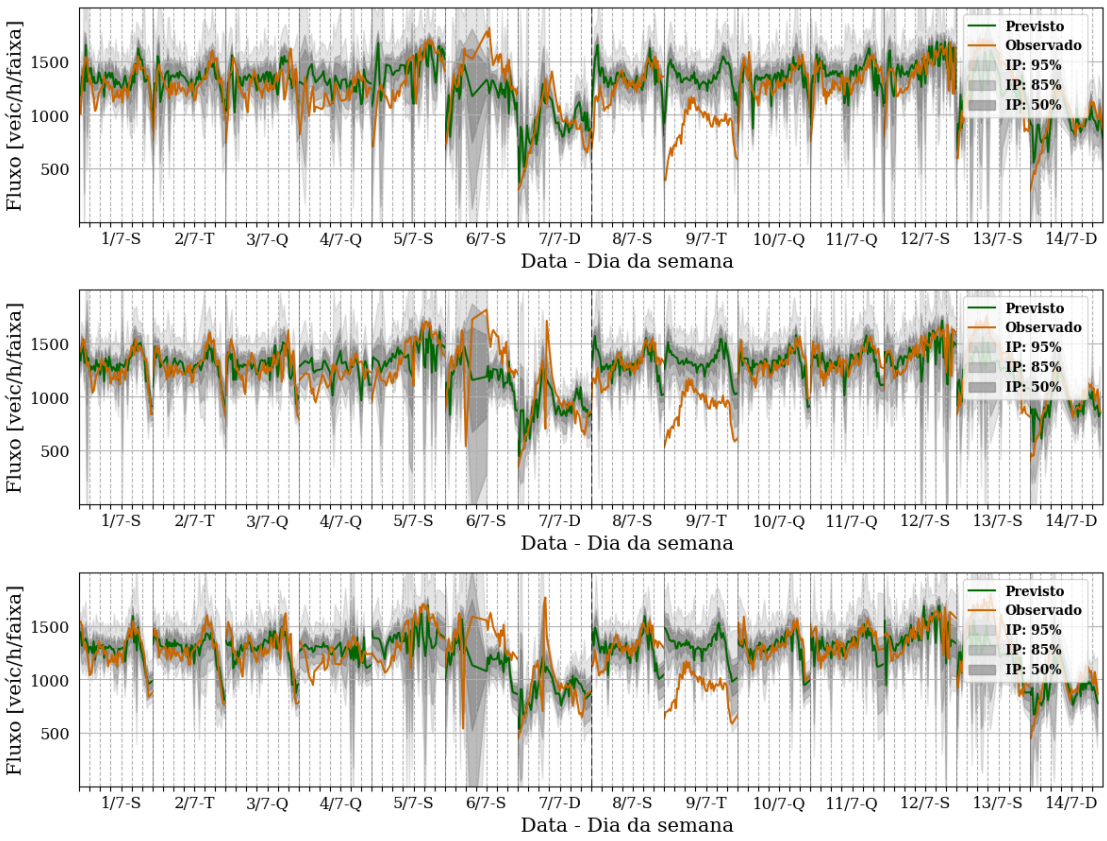


Figura A.10: Previsões para Sensor 10 na SP-280, 29.5, sentido oeste