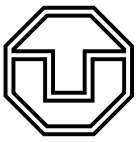

TECHNISCHE
UNIVERSITÄT
DRESDEN

Velocity Fluctuations and Extreme Events in Microscopic Traffic Data

Bachelor-Arbeit
zur Erlangung des Hochschulgrades
Bachelor of Science
im Bachelor-Studiengang Physik

vorgelegt von

Moritz Piepel

Institut für Theoretische Physik
Fakultät Physik
Bereich Mathematik und Naturwissenschaften
Technische Universität Dresden
2022

Submitted on September 12, 2022

1. Reviewer: Prof. Dr. rer. nat. Marc Timme, MA[1]
2. Reviewer: Dr. rer. nat. Alexander Keimer[2]

Supervision: Dr.-Ing. Angelika Hirrle[3] and Dr. rer. nat. Malte Schröder[1]

[1]Chair for Network Dynamics, Institute for Theoretical Physics, Faculty of Physics & Center for Advancing Electronics Dresden (cfaed), Technische Universität Dresden (TUD)

[2]Institute of Transportation Studies, University of California, Berkeley

[3]Chair of Traffic Process Automation, Institute of Traffic Telematics, "Friedrich List" Faculty of Transport and Traffic Sciences, Technische Universität Dresden (TUD)

Abstract

English:
Vehicle velocity distributions are of utmost relevance for the efficiency, safety, and sustainability of road traffic. Yet, due to technical limitations, they are often empirically analyzed using spatiotemporal averages. Here, we instead study a novel set of microscopic traffic data from Dresden comprising 346 million data points with a resolution of one vehicle from 145 detector sites with a particular focus on extreme events and distribution tails. By fitting q-exponential and Generalized Extreme Value distributions to the right flank of the empirical velocity distributions, we establish that their tails universally exhibit a power-law behavior with similar decay exponents. We also find that q-exponentials are best suitable to model the vast extent to which speed limit violations in the data occur. Furthermore, combining velocity and time headway distributions, we obtain estimates for free flow velocities that always exceed average velocities and sometimes even significantly exceed speed limits. Likewise, congestion effects are found to play a very minor, almost negligible role in traffic flow at the detector sites. These results provide insights into the current state of traffic in Dresden, hinting toward potentially necessary policy amendments regarding road design, speed limits, and speeding prosecution. They also reveal the potentials and limitations of the data set at hand and thereby lay the groundwork for further, more detailed traffic analyses.

Abstract

German:
Geschwindigkeitsverteilungen von Fahrzeugen sind von größter Bedeutung für die Effizienz, Sicherheit und Nachhaltigkeit des Straßenverkehrs. Aufgrund technischer Beschränkungen werden sie jedoch häufig empirisch anhand von raum-zeitlichen Mittelwerten analysiert. Hier untersuchen wir stattdessen einen neuartigen Satz mikroskopischer Verkehrsdaten aus Dresden, der 346 Millionen Datenpunkte mit einer Auflösung von einem Fahrzeug von 145 Detektorstandorten umfasst. Ein besonders Augenmerk legen wir auf Extremereignisse und Tails von Verteilungen. Durch den Fit von q-Exponential- und verallgemeinerten Extremwertverteilungen an die rechte Flanke der empirischen Geschwindigkeitsverteilungen stellen wir fest, dass ihre Tails universell ein Potenzgesetzverhalten mit ähnlichen Exponenten aufweisen. Wir stellen außerdem fest, dass q-Exponentiale am geeignetsten sind, um das Ausmaß der Geschwindigkeitsüberschreitungen in den Daten zu modellieren. Darüber hinaus erhalten wir durch die Kombination von Geschwindigkeits- und Bruttozeitlückenverteilungen Schätzungen für Freiflussgeschwindigkeiten, die stets über den Durchschnittsgeschwindigkeiten liegen und teilweise sogar die Geschwindigkeitsbegrenzungen deutlich überschreiten. Ebenso wird festgestellt, dass Staueffekte eine sehr geringe, fast vernachlässigbare Rolle im Verkehrsfluss an den Detektorstandorten spielen. Die Ergebnisse geben einen Einblick in die aktuelle Verkehrssituation in Dresden und weisen auf möglicherweise notwendige politische Änderungen in Verkehrsplanung hinsichtlich der Straßengestaltung, der Geschwindigkeitsbegrenzung und der Verfolgung von Geschwindigkeitsüberschreitungen hin. Sie zeigen aber auch die Möglichkeiten und Grenzen des vorliegenden Datensatzes auf und legen damit den Grundstein für weitere, detailliertere Verkehrsanalysen.

Contents

1 Introduction

> Extreme-Value Statistics can be regarded as the art of extrapolation.
>
> *Axel Bücher and Chen Zhou* [BZ21]

For years, traffic volumes in Germany have kept increasing. In 2019, the number of kilometers driven in road passenger traffic was 29% higher than in 1991 [Umw22]. Most of this can be attributed to cars [Ver22]. Until 2030, mobility demand is projected to keep growing [Kri16]. This, along with the increasing controversies about the allocation of urban space and the need for a sustainable transformation of the traffic system brought about by climate change, poses a huge challenge for urban and in particular traffic planning [Ver22].

In order to inform inclusive, sustainable, and efficient political decisions but also to improve traffic control, traffic data availability is key. Usually, this data is recorded in the form of spatiotemporal averages. In 2020, the traffic detector system in Dresden was upgraded to be able to resolve individual vehicles and store this microscopic traffic data for scientific use. Since then, 346 million vehicle movements have been recorded which will be analyzed in this thesis for the first time.

A main advantage of the microscopic data is that it paints a realistic picture of the traffic system whereas in mesoscopic or macroscopic data particularly extreme events are suppressed due to averaging. Meanwhile, extreme events are of high relevance for important applications such as road safety analyses, traffic stability, and accident avoidance. Thus, this novel data set allows us to explore the potential of extreme value theory that is well established e.g. in finance [Roc10] but has rarely found applications in traffic science.

This thesis serves to provide an overview of the data set at hand, examing multiple aspects with connections to mathematics, physics, and traffic science. It builds on the work by Paul Rathke, who first analyzed a fraction of the data and examined the potential of extreme value theory in [Rat20].

Chapter 2 lays down the mathematical and statistical groundwork necessary for the following analyses. After introducing and defining basic terms, it motivates the limit distributions of partial maxima and peaks over threshold and compares the two approaches. Furthermore, aspects from the field of traffic science like the hydrodynamic relation, the quantities measured by traffic detectors, and the fundamental diagram or flow–density diagram are discussed.

Chapter 3 then outlines how the data is recorded, processed, and saved. It describes the specifications of the measured quantities and the database structure. It also sets out the measurement period, data gaps, the total traffic volume, and its distribution across different detectors, the modal share, detector locations, average velocities per detector and traffic volume, and average velocities per time of day, day of the week, and week of the year, serving as a basis for i.i.d. assumptions that will later become necessary for extreme value theory analyses.

In Chapter 4, we examine car velocity distributions. This is split into two parts: Firstly, we consider the full distributions and divide them into four distinct sections. Using exponential and q-exponential distributions, we characterize the right flanks of these distributions. We also employ dimensionality reduction algorithms to determine principal components of the distributions and identify clusters. Secondly, we turn towards extreme value theory to describe the right tails of the distributions. After establishing criteria restricting the data to subsets that can be assumed as independent and identically distributed, we verify the fit methodology and determine the necessary parameters, namely the block size and quantile. Lastly, we classify the distributions found in the traffic data according to the extreme value distribution types.

Chapter 5 focuses on traffic science-related questions and presents empirical probability distributions for detector occupation times and time gaps. Based on that, we discuss distributions of traffic flows and traffic densities. Combining flow and density distributions, we construct characteristic flow–density, velocity–flow, and velocity–density diagrams, allowing us to draw conclusions about the state of traffic in Dresden and the role of congestion effects. In addition, we simulate the prior averaging conducted by many detectors and discuss its influence on the distributions and characteristic diagrams, thereby comparing the potential of microscopic versus mesoscopic or macroscopic traffic data. We also develop and put to use a method to estimate free-flow velocities.

Applying the previously established probabilistic models, Chapter 6 compares real speed limit violations to those predicted by the different velocity probability distributions. Furthermore, building on the work of the previous chapter, we use average and free-flow velocities and the corresponding speed limits to identify potential mismatches between road design and the speed

limits.

Finally, in Chapter 7 we summarize our results and discuss their implications and limitations before providing an outlook for potential future works and open questions in Chapter 8.

At this point, I would also like to thank Marc Timme for allowing me to write this thesis with his group in spite of this tight schedule and Alexander Keimer for agreeing to support and review this thesis when we approached him out of the blue. Without the constant support and supervision of Malte Schröder who answered my questions regarding math and physics and Angelika Hirrle who brought in her expertise in traffic science, this work would not have been the same. I am also thankful to Lina Tölle, Christoph Röllig, Infinity Rage Cage, Henrik Wolf, and my parents for the advice, support, and motivation they gave me.

2 Theoretical Background

2.1 Mathematics and Statistics

2.1.1 Elements of Statistics and Limit Distributions of Partial Sums

A **random variable** X is a function $X : \Omega \to E$, where Ω is the set of possible outcomes (the sample space) of a random experiment and E is a measurable space which for the purpose of this thesis can be assumed as the real numbers $\mathbb{R}$ [Col01, p. 20]. If the image of X, $X(\Omega) \subset E$, is countable, we call the random variable discrete, otherwise, we call it continuous. The distribution of discrete random variables are described with **probability mass functions** (PMF) whereas continuous random variables are described with **probability density functions** (PDF). **Cumulative distribution functions** (CDF) on the other hand apply to both cases and return the probability that the outcome is less than or equal to the given value.

A set of random variables is **independent and identically distributed** (i.i.d.) if each of the random variables has the same cumulative distribution function (that is $P(X_i \leq x) = P(X_j \leq x)$ for all i, j and x) and all random variables are mutually independent, e.g. random variable X_i does not influence random variable X_j [Geo15, p. 71]. Independence is equivalent to $P(X_i \leq x_i \wedge X_j \leq x_j) = P(X_i \leq x_i) \cdot P(X_j \leq x_j)$.

Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of i.i.d. random variables where each of the random variables has a well-defined first moment μ and a well-defined second central moment σ^2 (such a sequence is called a random process [cf. Col01, p. 25]). The **central limit theorem** is then concerned with the behavior of the partial sums of this sequence, $\Sigma_{i=1}^{n} X_i = n \cdot \bar{X}_n$ [HF10, p. 3], , where $\bar{X}_n$ indicates the mean of all X_n. It can be proven [cf. Geo15, pp. 152 sqq.] [cf. Düm16, pp. 219–220] that a linear normalization[1] of these partial sums weakly converges to the normal

[1] It is immediately apparent that the partial sums can diverge unless normalized.

distribution[2]:

$$\frac{\Sigma_{i=1}^{n} X_i - n \cdot \mu}{\sqrt{n}} = \sqrt{n}\left(\bar{X}_n - \mu\right) =: Y_n \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\right) \quad \text{for} \quad n \to \infty \tag{2.1}$$

This has to be understood in the following sense: In the limit $n \to \infty$ (or, in practice, for sufficiently high n), assume we repeatedly draw the random variable Y_n. Then, the realizations of Y_n, y_n, will be distributed according to the normal distribution with mean 0 and variance σ^2. Note that we did not make any assumptions about the distributions of the individual X_i apart from them being i.i.d.

In this sense, the normal distribution is the asymptotic distribution of the (normalized) partial sums. For a finite number of random variables, the normal distribution is a good approximation for the distribution close to its peak – however, for extreme values corresponding to the tails of the distribution, this does not necessarily hold.

2.1.2 Limit Distributions of Partial Maxima

As opposed to classical statistics, which is mainly concerned with central moments of distributions, extreme value theory tries to describe the so-called tails of distributions. For many distributions with thin tails, this does not yield meaningful insights as extreme events are very rare – e.g. for the normal distribution, 99% of values fall into the 3-σ-interval – but for distributions with a relatively high probability for extreme events, this is a different story.

Now, we are thus interested in the behavior of the **partial maxima** of a sequence of random variables, that is $\max(X_1, X_2, \ldots, X_n)$. How are these partial maxima[3] distributed for $n \to \infty$?

Thus, a normalization is necessary in order to stabilize location and scale of distribution when $n \to \infty$ [Col01, p. 46]. In analogy to how we normalized the partial sums in Equation 2.1, we choose a linear ansatz using two real sequences, $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ (whose structure and values are not relevant at this point; we only assume that they exist): $\frac{1}{a_n}(\max(X_1, X_2, \ldots, X_n) -$

[2]Weak convergence or convergence in distribution $X_n \xrightarrow{d} X$ is equivalent to $\lim_{n\to\infty} F_n(x) = F(x)$ for all x with F continuous at x [Col01, p. 26].

[3]It is sufficient to develop a theory for partial maxima as $\min(X_1, X_2, \ldots, X_n) = -\max(-X_1, -X_2, \ldots, -X_n)$, such that a theory dealing with maxima can also be applied to minima [Col01, p. 53].

b_n). We can thus rewrite the cumulative distribution function [cf. HF10, p. 4]:

$$\begin{aligned} & P\left(\frac{\max(X_1, X_2, \ldots, X_n) - b_n}{a_n} < x\right) \\ &= P(X_1 < a_n x + b_n \wedge X_2 < a_n x + b_n \wedge \ldots \wedge X_n < a_n x + b_n) \\ &= F(a_n x + b_n)^n \end{aligned} \tag{2.2}$$

The Khintchine theorem provides that in the limit $n \to \infty$, the sequences $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ can be replaced with constants a and b. Now, we have reason to assume that this normalization could yield a non-degenerate limit distribution function $G(x)$, if the limit exists:

$$\lim_{n\to\infty} P\left(\frac{\max(X_1, X_2, \ldots, X_n) - b_n}{a_n} < x\right) = \lim_{n\to\infty} F(a_n x + b_n)^n = G(x) \tag{2.3}$$

This is essentially the extreme value analog to the central limit theorem [Col01, p. 46]. In theory, this would already be a sufficient basis for an in-depth analysis. However, in practice, this result is not very helpful yet as the underlying cumulative distribution function $F(x)$ is not known. Obviously, it could be estimated using standard statistical methods, but this would not be very useful as small deviations in $F(x)$ would lead to huge deviations in $F(x)^n$ for a large n [cf. Col01, pp. 45 sq.].

We now want to discuss the limiting forms of Equation 2.3 – this is what Fisher and Tippett first did in 1928 in [FT28], when they formulated their **Fisher–Tippett theorem**[4]. It states that if $G(x)$ (as in Equation 2.3) exists, that is if $\lim_{n\to\infty} F(a_n x+b_n)^n$ converges weakly (which might very well not be the case), it belongs to one of exactly three types – Gumbel, Fréchet and Weibull – which can be expressed using the the cumulative distribution function of the **Generalized Extreme Value (GEV) Distribution** [cf. HF10, p. 6]:

$$G(x) = \exp\left(-(1+\gamma x)^{-1/\gamma}\right), \quad 1+\gamma x > 0, \quad \gamma \in \mathbb{R} \cup \{-\infty, +\infty\} \tag{2.4}$$

It should be noted that this result holds irrespective of the underlying distribution function $F(x)$ – as is the case for the central limit theorem. It is merely the type of the GEV distribution that depends on the underlying distribution function $F(x)$. Distribution functions which give rise to one particular type of GEV distribution are said to lay in that GEV distributions

[4] In literature, it is also referred to as the Fisher–Tippett–Gnedenko theorem to acknowledge Boris Vladimirovich Gnedenkos contributions in 1943.

domain of attraction [LR83, pp. 15 sq.]. However, it should be noted that there are some non-degenerate and degenerate examples of distributions that do not fall into any of the GEV distributions' domains of attraction, e.g. outlined in [Löe08, pp. 25 sq.] and [exercises 1.13 and 1.18 in HF10, pp. 35 sq.].

How can we draw any meaningful insights when the sequences $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ are unknown? Well, if Equation 2.3 holds, we can rewrite it using $x = (x^* - b_n)/a_n$, where G^* has different parameters but also clearly belongs to one of the GEV distribution types which can be thought of as equivalence classes [Col01, p. 48]:

$$\begin{aligned} &P((\max(X_1, X_2, \ldots, X_n) - b_n)/a_n < (x^* - b_n)/a_n) \\ =\,&P(\max(X_1, X_2, \ldots, X_n) < x^*) \\ =\,&G((x^* - b_n)/a_n) = G^*(x^*) \quad \text{for} \quad n \to \infty \end{aligned} \tag{2.5}$$

As the parameters are not known a priori and result from an estimation based on the data, this is not an issue. Thus, the sequences are irrelevant for further analysis.

The GEV distributions have the property of **max-stability** [cf. Haa84]:

$$\frac{\max(X_1, X_2, \ldots, X_n) - b_n}{a_n} \xrightarrow{d} X \quad \text{for} \quad n \to \infty \tag{2.6}$$

Here, all X follow a GEV distribution. In essence, max-stability describes distributions for which the process of taking maxima generates the same distribution (up to normalization). It is intuitively understandable (and easy to show) that GEV distributions fulfil Equation 2.6 – in fact, the class of max-stable distributions and the class of GEV distributions are equivalent [cf. Col01, p. 50]. This is what the proof of the Fisher–Tippet theorem, outlined in [LR83, pp. 8 sqq.], relies on.

From now on, we write $G_\gamma(x)$ instead of $G(x)$ to indicate that the behavior of the limit distribution qualitatively depends on the parameter γ, called the **extreme value index** [cf. HF10, p. 6].

As in the case of the central limit theorem, we would also obtain a location parameter μ and a scale parameter σ – however, these can be neglected as they do not fundamentally change the distributions which can easily be transformed using $x' = \frac{x-\mu}{\sigma}$ (with $-\infty < \mu < \infty$ and $\sigma > 0$). Distributions which satisfy $F_1(x) = F_2(ax + b)$ are said to be of the same type. Distributions

belonging to the same type exhibit the same qualitative behavior [LR83, p. 9].

When dealing with real-world data, it is generally preferable to fit the GEV distribution compared to fitting a Gumbel, Fréchet, or Weibull distribution as this would constitute an a priori choice of the extreme value index. Thus, its uncertainty could not be described by the model [cf. Col01, pp. 63 sq.].

Equation 2.4 leaves three cases that qualitatively differ for further analysis: $\gamma = 0$, $\gamma > 0$ and $\gamma < 0$.

The Gumbel Distribution (Type I): $\gamma = 0$

The case of $\gamma = 0$ has to be understood in the sense of taking the limit $\gamma \to 0$ (as $\lim_{n\to\infty} \left(1 + \frac{-x}{n}\right)^n = \exp(-x)$ with $n = -\frac{1}{\gamma}$).

The Gumbel distributions cumulative probability and probability density functions are then given as:

$$G_0(x) = \exp(-\exp(-x)), \quad -\infty < x < \infty \tag{2.7}$$

$$g_0(x) = \exp(-(x + \exp(-x))), \quad -\infty < x < \infty \tag{2.8}$$

As $\exp(-\exp(-x)) \to 1$ for large x, the decay behavior is dominated by $\exp(-x)$ which is exponential, decaying faster than power-law – this is called a thin, sometimes also light, tail. The domain of attraction of the Gumbel distribution includes the normal, lognormal, gamma,

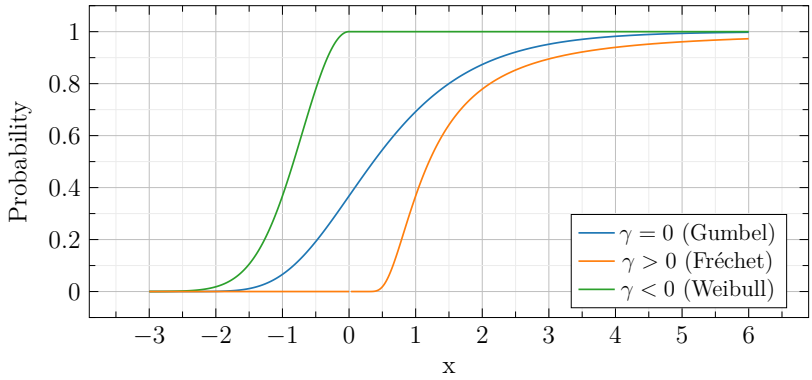

Figure 2.1: Cumulative distribution functions of the Gumbel, Fréchet and Weibull distributions.

and exponential distributions among others [LR83, pp. 19 sqq.]. Usually, their moments exist.

The Fréchet Distribution (Type II): $\gamma > 0$

The Fréchet distributions cumulative probability and probability density functions are given as[5]:

$$G'_\gamma(x) = G_\gamma\left(\frac{x-1}{\gamma}\right) = \begin{cases} 0, & x \leq 0 \\ \exp\left(-x^{-1/\gamma}\right), & x > 0 \end{cases} \tag{2.9}$$

$$g'_\gamma(x) = \begin{cases} 0, & x \leq 0 \\ \frac{1}{\gamma} \cdot x^{-(1+1/\gamma)} \cdot \exp\left(-x^{-1/\gamma}\right), & x > 0 \end{cases} \tag{2.10}$$

Often, the distribution is expressed using $\frac{1}{\gamma} = \alpha > 0$ as $G'_\gamma(x) = \exp(-x^{-\alpha})$.

As $\exp\left(-x^{-1/\gamma}\right) \to 1$ for large x, the decay behavior is dominated by $x^{-(1+1/\gamma)}$ which is polynomial – this is called a fat, sometimes also heavy or power-law, tail. The domain of attraction of the Fréchet distribution includes the Pareto and Cauchy distributions among others [LR83, pp. 19 sqq.]. The higher the extreme value index is, the more probability mass is centered in the tail, and thus the heavier the tail is.

Both the Fréchet distribution and the distributions in its domain of attraction are characterized by the fact that not all of their moments exist, that is $E[X^j] = \infty$ for $j \geq 1/\gamma$ [Woe10, p. 21].

The Fréchet distribution $G_\gamma(x)$ has a lower endpoint at $x = -1/\gamma$. It also has a positive skew.

The Weibull Distribution (Type III): $\gamma < 0$

The Weibull distribution's[6] cumulative probability and probability density functions are given as:

[5] $G'_\gamma(x)$ is only introduced to increase readability and highlight the qualitative behavior of the distribution.

[6] As the Weibull distribution also appears in other applications, this is technically the reverse-Weibull distribution – however, as we are only interested in the distribution with an upper bound we will simply refer to it as the Weibull distribution.

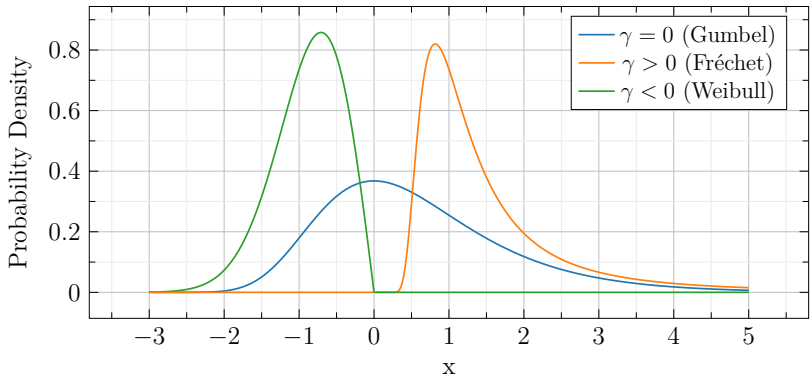

Figure 2.2: Probability density functions of the Gumbel, Fréchet and Weibull distributions.

$$G'_\gamma(x) = G_\gamma\left(-\frac{x+1}{\gamma}\right) = \begin{cases} \exp\left(-(-x)^{-1/\gamma}\right), & x < 0 \\ 1, & x \geq 0 \end{cases} \tag{2.11}$$

$$g'_\gamma(x) = \begin{cases} -\frac{1}{\gamma} \cdot (-x)^{-(1+1/\gamma)} \cdot \exp\left(-(-x)^{-1/\gamma}\right), & x < 0 \\ 0, & x \geq 0 \end{cases} \tag{2.12}$$

Often, the distribution is expressed using $-\frac{1}{\gamma} = \alpha > 0$ as $G'_\gamma(x) = \exp(-(-x)^\alpha)$.

The Weibull distribution $G_\gamma(x)$ has an upper endpoint at $x = -1/\gamma$ – essentially, it does not have a tail, often also referred to as a short (or finite) tail. The domain of attraction of the Weibull distribution includes distributions that have an upper endpoint like the uniform, beta or truncated exponential distribution [LR83, pp. 19 sqq.]. For short-tailed distribution, all moments exist [Woe10, p. 21]. It also has negative skew.

2.1.3 Block Maxima (BM)

In order to apply the GEV distribution to real-world data, it is necessary to identify maxima. This is done using the so-called **block maxima method**.

Let us look again at Equation 2.3 and identify $\max(X_1, X_2, \ldots, X_n) =: M_n$. When given a dataset of size n, it is clearly not very useful to take the maximum of the dataset as this would only yield one value – not enough to construct a distribution. Thus, the dataset is divided up into blocks of size k. This gives $m = \lceil N/k \rceil$ block maxima M_k to which the distribution can then be fitted.

Choosing a value for the block size k is rather complicated. On the one hand, the GEV

distribution is the limit distribution for $k \to \infty$. A too low value of k would thus introduce bias – in the extreme case of $k = 1$ it would simply reproduce the original distribution which is certainly not a GEV distribution. On the other hand, choosing a too high value of k would drastically reduce the number of block maxima to which the GEV distribution can be fitted, increasing variance. There is no general rule on how this tradeoff should be done – the block size can only cautiously be chosen on a per-case basis when dealing with real-world data [cf. Col01, p. 54].

The block maxima do depend on the order of the data. Thus, they will be different when shuffling the data before drawing block maxima. This can be advantageous when the data follows some weak time-dependent, e.g. seasonal trend: Then, the individual blocks might not follow the same distribution which can be mitigated by shuffling the data before drawing block maxima, removing the influence of the trend. When the data does not follow such a trend but an observation depends on the previous n observations, it is preferable not to shuffle before drawing block maxima as the dependence can be ignored as long as the block size k is larger than n as only the maximum value is selected, leaving aside all other (dependent or independent) values.

However, according to [Col01, p. 74], the block maxima method is not necessarily the most desirable option when the entire time series of the data is available. In the next section, we will discuss one of the alternatives.

2.1.4 Limit Distributions of Peaks over Threshold (POT)

Apart from the block maxima method, there is a more natural, intuitive way to define extreme values – that is by defining a threshold over which a value is then considered extreme. How could a distribution of these values be described statistically?

$$A = \{X_i : X_i > u,\, i \in (1, \ldots, n)\} \tag{2.13}$$

Again, $F : \mathbb{R} \to, x \mapsto F(x)$ is the cumulative distribution function of the X_i. The probability of X_i to exceed the threshold u by $y = x - u$ is then $F'(x) = 1 - F(x) = 1 - F(u + y)$ (with $y \geq 0$). In order to obtain the cumulative distribution function this needs to be normalized by the overall probability of x exceeding u, $F'(u) = 1 - F(u)$. Thus, the probability of X_i exceeding u (given that $X_i > u$ as denoted by the subscript) is:

$$P_{X_i>u}(X_i > x) = \frac{1 - F(x)}{1 - F(u)} \quad \text{for} \quad x \geq u \tag{2.14}$$

But, as with the GEV distribution, F is usually not known in real-world scenarios. Thus, we try to obtain it as the limiting distribution when the threshold $u \to \infty$, given (or on the condition) that $X > u$. As an ansatz, we choose the GEV distribution as we would only expect to see the block maxima for a large enough threshold u. We start with Equation 2.3 and Equation 2.4 [cf. Col01, pp. 75 sqq.], assuming sufficiently large n and using the Taylor expansion $-\log x \approx 1 - x$ in the last step:

$$\begin{aligned} F(x)^n = G(x) &= \exp\left(-(1+\gamma x)^{-1/\gamma}\right) \\ \Leftrightarrow\ n \log F(x) &= -(1+\gamma x)^{-1/\gamma} \\ \Leftrightarrow\ (1 - F(x)) &= \frac{1}{n}(1+\gamma x)^{-1/\gamma} \end{aligned} \tag{2.15}$$

Using this, we can solve Equation 2.14, renaming γ to the shape parameter ξ in order to avoid ambiguity:

$$P_{X_i>u}(X_i > u + y) = \frac{1 - F(u+y)}{1 - F(u)} = \frac{(1+\xi(u+y))^{-1/\xi}}{(1+\xi u)^{-1/\xi}} = \left(1 + \frac{\xi y}{1+\xi u}\right)^{-\frac{1}{\xi}} \tag{2.16}$$

Again understanding $\xi = 0$ as taking the limit $\xi \to 0$, this yields the **Generalized Pareto Distribution (GPD)** (absorbing $1/(1+\xi u)$ into ξ). Its cumulative distribution function and probability density function are:

$$G_\xi(y) = 1 - (1+\xi y)^{-\frac{1}{\xi}} \quad \text{for} \quad 0 \leq y \quad (\text{and } y < -1/\xi \text{ for } \xi < 0) \tag{2.17}$$

$$g_\xi(y) = (1+\xi y)^{-\frac{\xi+1}{\xi}} \quad \text{for} \quad 0 \leq y \quad (\text{and } y < -1/\xi \text{ for } \xi < 0) \tag{2.18}$$

This finding is called the **Pickands–Balkema–De Haan theorem** [Pic75], [Mak07, p. 116], [BH74].

As in the case of the GEV distribution, y can be substituted with $\frac{x-\mu}{\sigma}$ to include a location and scale parameter.

The most notable difference for the different cases of the shape parameter ξ is that for negative

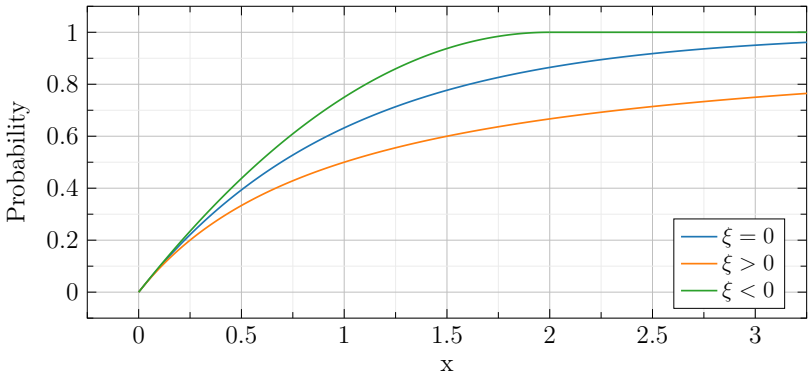

Figure 2.3: Cumulative distribution function of the Generalized Pareto Distribution for different cases of ξ.

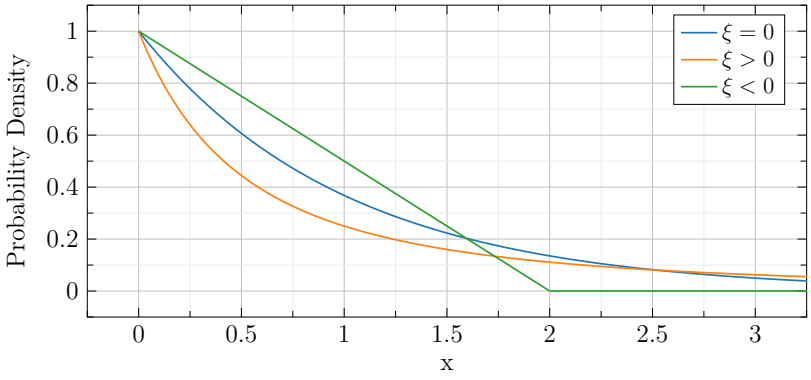

Figure 2.4: Probability density function of the Generalized Pareto Distribution for different cases of ξ.

values of ξ the distribution has a right endpoint at $x_+ = -1/\xi$. For $\xi = 0$, the GPD distribution is identical to the exponential distribution. Otherwise, the decay follows a power-law and the higher ξ is, the heavier the tail of the distribution is and the mass of the probability density function is centered towards higher values. Furthermore, for $\xi \geq 1$, the mean is infinite, otherwise, it is finite.

When the generalized Pareto distribution is a valid model for a given threshold u, it is also a reasonable model for every $u' > u$ [Col01, p. 83]. This can easily be shown analytically and is also intuitively understandable as a power-law distribution is scale-free.

2.1.5 BM vs. POT: Comparing the Two Approaches

The approach of block maxima and peaks over threshold share many similarities. This is to be expected as for many data sets, the block maxima and peaks over threshold have significant overlap, as can be seen in Figure 2.5. Additionally, for block maxima and peaks over threshold drawn from the same data set, their empirical cumulative distribution functions are identical in the asymptotical limit of large values. Conversely, the shape parameters of GEV and GPD distributions are identical if the block maxima and peaks over threshold were drawn from the same data set [HF10, pp. 89 sq.] [Rod17] – and due to max-stability, the shape parameters are independent of the threshold or the block size as long as they are large enough.

Another sense in which the approaches are similar is the tradeoff between bias and variance [cf. Col01, pp. 54, 78]. Choosing a low value for the threshold or block size respectively might violate the assumption of a GPD or a GEV model as the limiting distribution; choosing a high value leaves fewer values for the analysis variance and uncertainty of the fit.

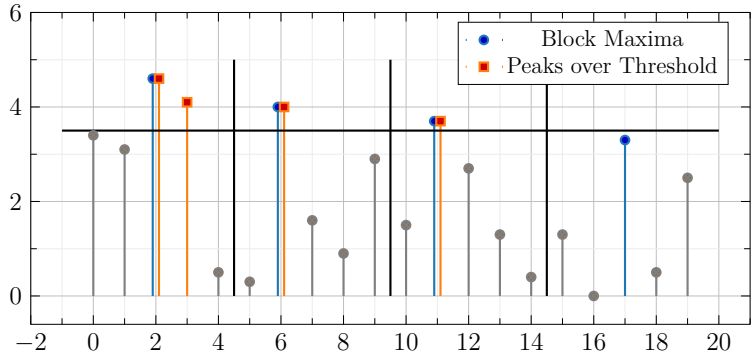

Figure 2.5: Block maxima (BM) and peaks over threshold (POT) for 20 random real numbers between zero and five with a block size of five and a threshold of 3.5.

In many data sets, extreme events are quite rare. Then, it is important to make efficient use of the available extreme events. In these cases, it is often preferable to employ the peaks over threshold approach as it selects all events considered extreme whereas the block maxima approach per definition disregards all events smaller than the maximum, even if they are only marginally smaller.

This can, however, also be an advantage. If the data is dependent, extreme events tend to cluster. These clusters then violate the i.i.d. assumption at the heart of extreme value theory. This is an issue, particularly for the peaks over threshold approach. The block maxima approach on the other hand circumvents this issue by only selecting one value from each cluster as long as the clusters are small enough compared to the block size. Thus, the independence criterion is less strict for the GEV distribution.

A way by which the two approaches can be synthesized is by considering not only the maximum from each block but selecting the r largest values. This is a generalization of the GEV model to r largest order statistics, allowing to extract more information from one block than just the maximum. [Col01, p. 66], albeit at the expense of complexity. This is particularly useful when the data is very limited. However, due to the amount of data available for the following analyses, this is not relevant here.

2.1.6 Maximum Likelihood Estimation

Often, it is desirable to describe real-world data with a statistical model. These statistical models are usually characterized by their parameters. How can this description be conducted in a rigorous manner?

The parameters, in this case, are known as the estimands – the quantities that need to be estimated. The rule or function defining the estimation process is called the estimator. The result, the concrete values for the different parameters or estimands, are the estimates [cf. Düm16, pp. 25 sqq.].

The approach that the maximum likelihood estimation (MLE) uses is simple – it basically answers the question: Given the observed data, which parameter configuration has the highest probability of reproducing this measurement [cf. Geo15, pp. 219 sqq.]? This parameter configuration is then taken as the estimate.

For this, the **likelihood function** for observed data $x_1, \ldots, x_n$ with sample size n and prob-

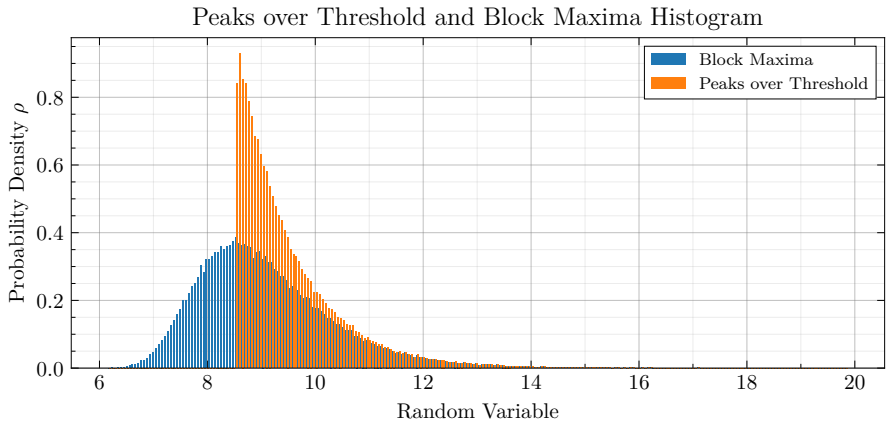

Figure 2.6: Histogram of block maxima and peaks over threshold drawn from the same data set (with arbitrary scaling of the x-axis).

ability density function $f(x)$ with parameter θ (which can generally be thought of as a vector of multiple scalar parameters) is introduced:

$$L(\theta) = \prod_{i=1}^{n} f_\theta\left(x_i\right). \tag{2.19}$$

Then, the likelihood function has to be maximized with respect to the parameter θ, yielding the MLE estimator $\hat{\theta}$. Often, it is easier to maximize the **log-likelihood function** – as the logarithm is strictly increasing, this yields the same maximum with respect to the parameter as the likelihood function.

$$\ell(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f_\theta\left(x_i\right). \tag{2.20}$$

Maximum likelihood estimation will become relevant in Chapter 4 where it is used to fit statistical models to observed velocity distributions. An advantage of the MSE model is that (under certain regularity conditions outlined in [Geo15, p. 229]) the estimator follows a (multivariate) normal distribution for sufficiently large sample size n [cf. Col01, pp. 31 sqq.], allowing us to learn about the uncertainties of the estimates. In the following, all uncertainties will be provided as standard errors, corresponding to a statistical significance of 68.3%. A comparison of different estimators for extreme value distributions can be found in [YKÖ21]. The likelihood functions for the GEV distribution can be found in [Woe10, p. 26].

2.2 Traffic Science

2.2.1 Hydrodynamic Relation and Continuity Equation

From a physics perspective, the number of vehicles in a particular volume is a conserved quantity – at least on short timescales, when accidents and the production and disposal of vehicles can be ignored. Certainly, this is true for an isolated system.

Let us consider a road section of length L with N vehicles traveling at velocity v. Then, one vehicle will take $T = L/v$ for this road section. Given that there are N vehicles on this road section, this allows for multiple configurations: Either the vehicles are packed very densely and travel quite slowly, or vice versa, or in an intermediary state. If we introduce a traffic density ρ (that is being integrated over a volume V) and a traffic flux density Φ (through a detector area A perpendicular to the direction of travel), this can be expressed as follows:

$$\begin{aligned} N &= \int_V dV\, \rho = \int_T dt \int_A d\vec{A} \cdot \vec{\Phi} \\ &= L \cdot \rho = \frac{L}{v} \cdot \Phi \\ &\Leftrightarrow \Phi = v \cdot \rho \end{aligned} \tag{2.21}$$

This if often referred to as the **hydrodynamic relation** [TK10, p. 17].

A change in the number of vehicles on a given road section can be written as a closed surface integral of the traffic flux density $\vec{\Phi}$[7] – if the number of vehicles in a given volume changes, vehicles must either be flowing in or out of that volume. Using the divergence theorem, this yields the continuity equation:

$$\begin{aligned} &\frac{dN}{dt} = -\oint_{\partial V} d\vec{S} \cdot \vec{\Phi} = -\int_V dV\, \vec{\nabla} \cdot \vec{\Phi} = \int_V dV\, \frac{\partial \rho}{\partial t} \\ &\Rightarrow \frac{\partial \rho}{\partial t} + \vec{\nabla} \cdot \vec{\Phi} = 0 \end{aligned} \tag{2.22}$$

As we can usually assume streets as one-dimensional systems, the traffic flow density is identical to the traffic flux (or, equivalently, traffic flow) and the continuity equation simplifies to $\frac{\partial \rho}{\partial t} + \frac{\partial \Phi}{\partial x} = 0$. Clearly, this is only true when no other sinks or sources like ramps are present.

[7] This is using the sign convention of positive flux density flowing from inside to outside.

Here, the traffic flow and traffic density are extensive properties that scale with the system (when the system size is the number of vehicles) that is being considered while the velocity is an intensive property that is independent of system size.

2.2.2 Traffic Quantities and Detector Data

Real-world detectors usually have no means of directly measuring the traffic flow or the traffic density. Particularly the traffic density is difficult to measure since most detectors are fixed in place and thus only allow for an indirect estimate of the traffic density which is a spatial quantity. Many detectors instead record the occupation time t_{occ}, the net time gap between two vehicles t_{gap} and the individual vehicle velocity v [TK10, pp. 14 sq.]. In this context, the traffic flow can intuitively be understood as the rate of vehicles passing the detector:

$$\Phi = \frac{n}{\Sigma_{i=0}^{n}(t_{\text{occ, i}} + t_{\text{gap, i}})} \quad , \qquad [\Phi] = \text{h}^{-1} \tag{2.23}$$

From this, using Equation 2.21 and the average velocity $v = \Sigma_{i=0}^{n} v_i / n$, we can calculate the traffic density [cf. Sch04, pp. 6 sq.]:

$$\rho = \frac{\Phi}{v} = \frac{n}{\Sigma_{i=0}^{n}(t_{\text{occ, i}} + t_{\text{gap, i}})} \cdot \frac{n}{\Sigma_{i=0}^{n} v_i} \quad , \qquad [\rho] = \text{km}^{-1} \tag{2.24}$$

Qua definitionem, the slope of the secant in a flow–density diagram, Φ/ρ, is the average traffic velocity (analogous to the phase velocity $v_p = \omega/k$). The slope of a curve in the flow–density diagram, $d\Phi/d\rho$, on the other hand, corresponds to the propagation velocity of perturbations and jam fronts [TK10, p. 79] (analogous to the group velocity $v_g = d\omega/dk$). As some models describe traffic using partial differential equations that are effectively wave equations, this propagation velocity can also be understood as the velocity of density waves traveling through traffic [Sch04, pp. 75, 104]. In the following, in order to distinguish the two, we will refer to $\Phi/\rho = v$ as the traffic velocity and to $d\Phi/d\rho = u$ as the propagation velocity.

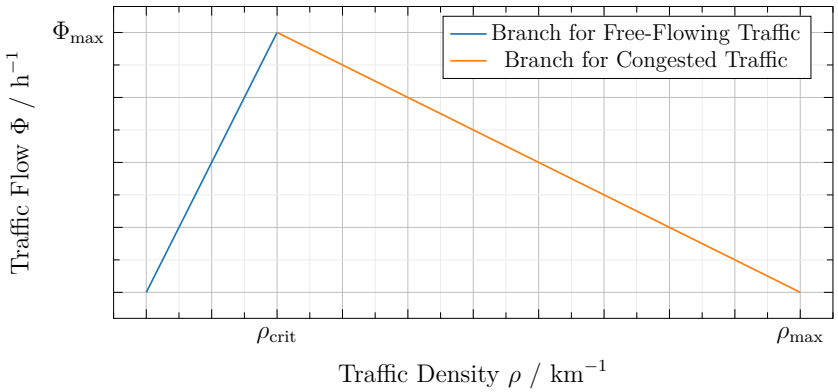

Figure 2.7: Schematic fundamental diagram (with a triangular shape from the Lighthill-Whitham-Richards (LWR) model class) [cf. TK10, pp. 82 sqq.].

2.2.3 The Fundamental Diagram

There are various models predicting the shape of the flow–density diagram. This predicted, theoretical shape is then called the fundamental diagram [cf. TK10, p. 32]. An example of an often used, simple, and easy-to-implement class of models are the Lighthill-Whitham-Richards (LWR) models, depicted in Figure 2.7. There are a few characteristic properties that most fundamental diagrams share and that can also often be observed in real-world flow–density diagrams: For a low traffic density and a low traffic flow, they follow a linear relation. This is the free-flowing, stable regime or region in which vehicles barely influence each other. Thus, the traffic velocity and the propagation velocity are nearly identical. A change in density, e.g. induced by a vehicle accelerating or decelerating, is only experienced locally by an observer moving with traffic. This is, however, only true up to a critical density ρ_{crit} when interactions between vehicles begin to dominate. As random perturbations like a breaking car set off chain reactions that often grow in amplitude and are no longer only experienced locally, increasing the traffic density even further will decrease the traffic flow. Thus, the critical density ρ_{crit} corresponds to a maximum traffic flow Φ_{max} (which represents the capacity of the road section). Then, traffic velocity and propagation velocity are no longer identical, in fact, the propagation velocity is even negative, indicating that perturbations propagate opposite to the direction of traffic flow. This reflects the fact that drivers are usually only reacting to events and perturbations in front of them, not behind them. This is the congested, unstable regime or region. Some models also predict a meta-stable region that allows for free-flowing traffic beyond the critical density, but once perturbations become too large, the traffic breaks down and becomes congested[8]. In this case, the fundamental diagram is no longer a function but rather a curve in flow–density space. Eventually, when vehicles are packed so densely that they cannot move anymore (with a density ρ_{max}), the traffic stops flowing altogether and the

[8]These models then allow for hysteretical effects, where the state of the system depends on its history or path, leading to loops in the flow–density diagram.

traffic flow vanishes.

In addition to the flow–density diagram, depending on the situation, it is sometimes also useful to consider the velocity–flow or the velocity–density diagrams shown in Figure 2.8.

The so-called free speed v_0 is the traffic velocity in the limit $\rho \to 0$ – this is the speed that drivers would choose if no other vehicles or obstacles were present on the street given physical limitations [TK10, p. 28]. Sometimes, the speed limit is also considered as an upper bound for the free-flow velocity, however, as we will see, the free-flow velocity often exceeds the speed limit.

Numerous different models are predicting the shape of fundamental diagrams and, more generally, traffic flow dynamics, some of which are discussed in [Sch04] and [TK10].

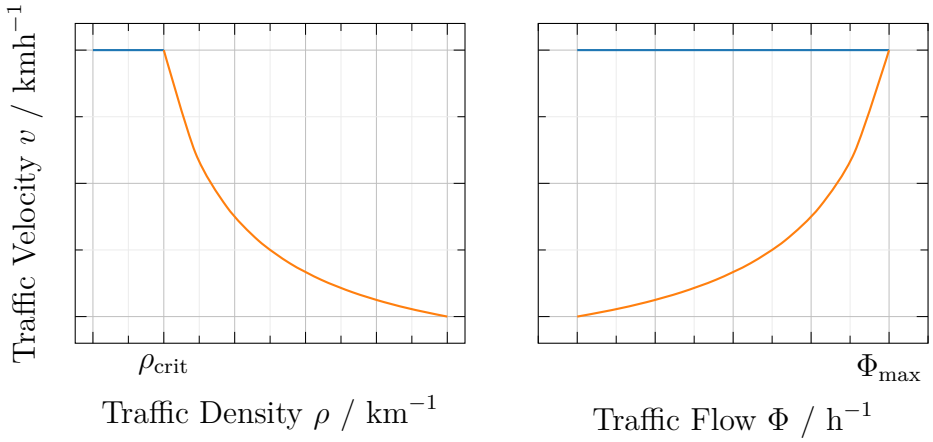

Figure 2.8: Velocity–density and velocity–flow diagrams for the fundamental diagram shown in Figure 2.7. Mathematically, the velocity–flow relation is no longer a function (as it is not right-unique) since there are two different configurations for each traffic flow: a high–velocity, low–density, and a low–velocity, high–density configuration.

3 Microscopic Traffic Data from Dresden

3.1 Data Source and Structure

All the data analyzed in this thesis is taken from the traffic management system VAMOS2[1] in which the TU Dresden is involved through its Chair of Process Traffic Automation.

The VAMOS2 system comprises data from 145 detectors in 42 distinct locations all over Dresden. When there are multiple detectors in one location, each of them monitors one of the traffic lanes. It creates a data point for each vehicle that is recorded. The data spans roughly two years, from July 27, 2020, to April 26, 2022, totaling 501 days of data recordings. Due to technical issues, no data was recorded in October and November 2021 and from January 11 to February 24, 2022. During that time, about 346 million vehicle movements were recorded in total.

The data from the VAMOS2 system is accessed through a redundant copy of the actual traffic data which is stored in an SQL database containing two tables, `quelle_pzs` and `pzs_archiv`[2]. According to the operator, the data is classified according to the TLS standard described in [Str12].

Column name	Type	Description
`s_idz`	int4	Unique identifier for each detector.
`pzs_id`	varchar	Unique identifier for each detector location.
`spur`	int2	Lane that is being monitored by a given detector at a given location.
`standort`	varchar	Location of a given detector.

Table 3.1: Colums of `quelle_pzs`.

[1] Further information about VAMOS2 can be found here: https://tu-dresden.de/bu/verkehr/vis/vpa/forschung/Individualverkehr/verkehrsmanagementsystem-vamos-dresden.

[2] PZS is a German abbreviation and refers to ”Pegelzählstelle”.

Column name	Type	Description
`zeit`	timestamp	Time when the vehicle is entered into the VAMOS2 database.
`id`	varchar(50)	Corresponds to `pzs_id` in `quelle_pzs` and specifies the detector.
`gap`	int4	Time between the detection of the back of the last vehicle and the front of the current vehicle in ms.
`lane`	int2	Number of the detector lane (numbered from one to four).
`class`	varchar	Vehicle class according to Table 3.3.
`length`	float4	Length of the vehicle in m (rounded to one decimal place).
`class_id`	int2	Identifier of the vehicle class according to Table 3.3.
`velocity`	int4	Velocity of the vehicle in m/s (rounded to zero decimal places).
`timestamp`	timestamp(0)	Time when the vehicle is recorded by the detector in s (rounded to zero decimal places).
`occupation`	int4	Occupation of the detector in ms.

Table 3.2: Colums of `pzs_data`; specification according to [Str12, pp. 312 sqq.].

The detectors use pairs of induction loops and measure the velocity and length of the vehicles [for induction loops cf. TK10, pp. 13 sq.]. The detectors themselves are specified according to [Str12, pp. 51, 137]. Whenever a metal vehicle moves over the induction loops, the inductance changes which can be measured by evaluation electronics. Depending on the vehicle, the magnitude of the inductance change and the inductance profile over time is different and can be used to deduct the type of vehicle on the induction loop. By measuring the time gap between the actuation of the first and the second induction loop (not to be confused with the time gap between vehicles), the velocity of the vehicle is calculated. Then, using the velocity and the occupation time of one of the loops, the vehicle length is calculated.

There are two types of relevant induction loop setups according to [Str12]. The first one uses loops of length 2.5 m which are spaced 1.5 m apart (Type I) while the other ones uses loops of length 1.0 m which are spaced 1.5 m apart (Type II). Thus, the effective length of the setup used to measure the velocity is 4 m for type I and 2.5 m for type II. In Dresden, both detector types are found and the specific type for one detector can be deducted from consistency checks.

Due to technical limitations, it is impossible to increase the accuracy of the velocity beyond the resolution of the data of 1 km/h [cf. Str12, p. 313]. When the velocity cannot be measured, the detectors report a speed of 255 km/h. The occupation is then calculated from the velocity and the length. When the occupation cannot be calculated, the detectors report an occupation

of 65 535 ms. When the time gap between two vehicles cannot be measured (e.g. because it is the first vehicle being measured), the detectors report a gap of 65 535 ms. The time headway (essentially the gross time gap) between two vehicles is the sum of the (net) time gap and the occupation. When the measured value exceeds the datatype range, the detectors report the value for the error code minus one.

Vehicle Class	Class ID	Description
nk Kfz	6	Classification not possible (nicht klassifizierbare Fahrzeuge)
Krad	10	Motorcycles (Motorräder)
Pkw	7	Car (Pkw)
Lfw	11	Delivery vans up to 3.5 t (Lieferwagen bis 3.5 t)
PkwA	2	Car with trailer (Pkw mit Anhänger)
Lkw	3	Truck (Lkw)
LkwA	8	Truck with trailer (Lkw mit Anhänger)
Sattel-Kfz	9	Articulated vehicle (Sattelkraftfahrzeuge)
Bus	5	Bus

Table 3.3: 8+1 classification of different vehicle types according to [Str12, p. 152].

3.2 Traffic Parameters and Traffic Changes over Time

Even though the vehicle volume per detector spans over an entire order of magnitude as seen in Figure 3.1, the detectors with the lowest vehicle volume still detected vehicles well into the million range over the entire time of data collection. This allows us to conduct statistical analysis even of the detectors with a comparatively low vehicle volume, and it also confirms the fact that all detectors are placed on roads with a lot of traffic. As the detectors are primarily used for traffic control, covering those roads with the most demand is crucial.

Since cars make up almost the entire traffic volume as seen in Figure 3.2, we will focus on analyzing car traffic for the rest of this thesis. The second largest group are vans and delivery trucks, while all other categories can basically be neglected.

Figure 3.3 clearly shows that – as one would expect in a city – most detectors are placed on streets with a speed limit of either 30 km/h or 50 km/h, with detectors 0831 and 0841 being the only exceptions. Yet, it is already apparent that the individual standard deviation is rather high, allowing for velocities above the speed limit, and the 95% quantile corresponds to a velocity well exceeding the speed limit.

As time-dependencies vanish when averaging over a lot of data, the dependency of the volume and average velocities on time need to be known a priori in order to exclude data that does not fulfill necessary requirements. Figure 3.4 and Figure 3.5 display total traffic volume and average velocity per hour of the day, day of the week, and week of the year.

Traditionally, in traffic science, one distinguishes between Monday, Tuesday through Thursday, Friday, Saturday and Sunday [cf. TK10, p. 25]. However, here, it seems like both traffic volume and average velocity are fairly constant from Monday through Friday with only Saturday and Sunday being exceptions. The same can be said about traffic from 7.00 am to 6.00 pm. When studying the change of traffic over the course of the year, the school holidays (e. g. in calendar weeks 40 and 52, 53) immediately stand out. Additionally, total traffic volume reduces towards the end of the year, most likely due to pandemic restrictions and work-from-home orders.[3]

[3]Germany saw a so-called "lockdown light" starting from November 2, 2020. A full lockdown was then imposed on December 16, 2020. It went out of effect on June 11, 2021. Even though there was no full lockdown the following winter, a work-from-home order was passed on November 18, 2021, lasting until April 3, 2022. It should be noted, however, that many employers and employees proceeded with extra caution, often exceeding the legal requirements. Source: https://de.wikipedia.org/wiki/COVID-19-Pandemie_in_Deutschland.

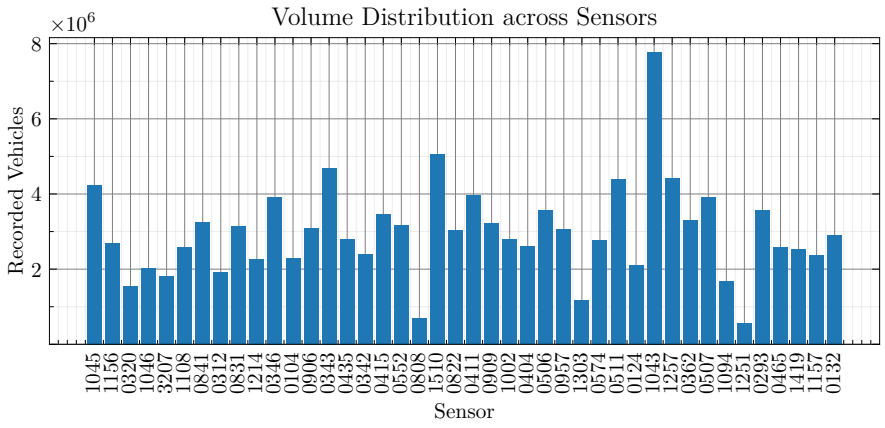

Figure 3.1: Share of the total traffic volume for each detector (from July 2020 to January 2021; different lanes have been combined).

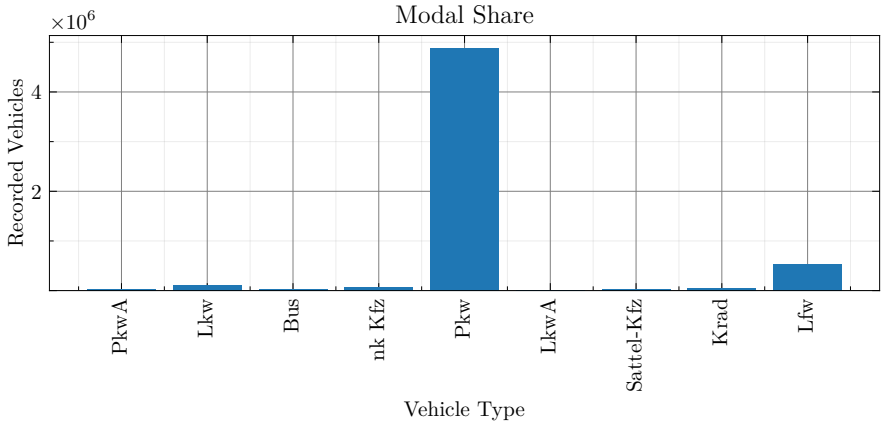

Figure 3.2: Share of the total traffic volume for all detectors by vehicle type (for Oct–05–2020 to Oct–11–2020).

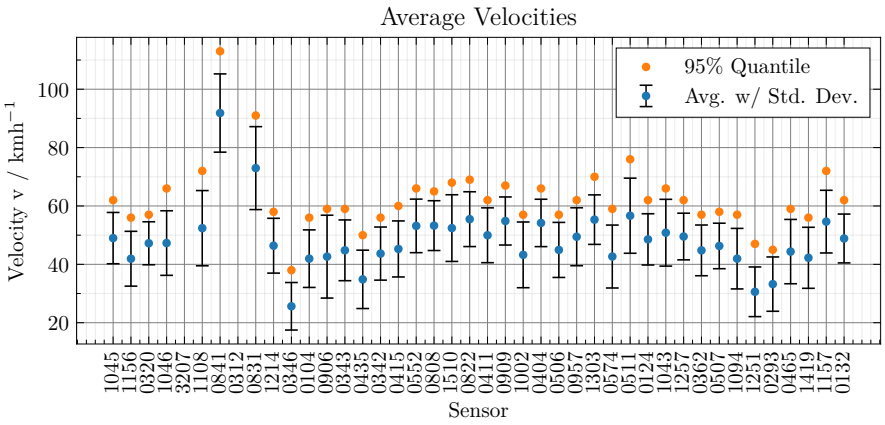

Figure 3.3: Average velocities (including standard deviations) and 95% quantile for all detectors (different lanes have been combined). As the detectors 3207 and 0312 report faulty data (with velocities up to 1000 km/h), the data for these two detectors is not plotted.

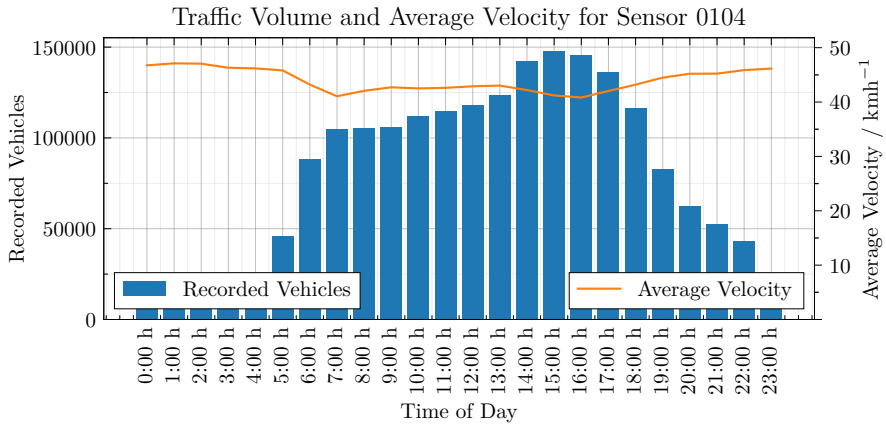

Figure 3.4: Traffic volume and average velocity per hour of the day. Between 07.00 am and 06.00 pm, both traffic volume and average velocity are approximately constant.

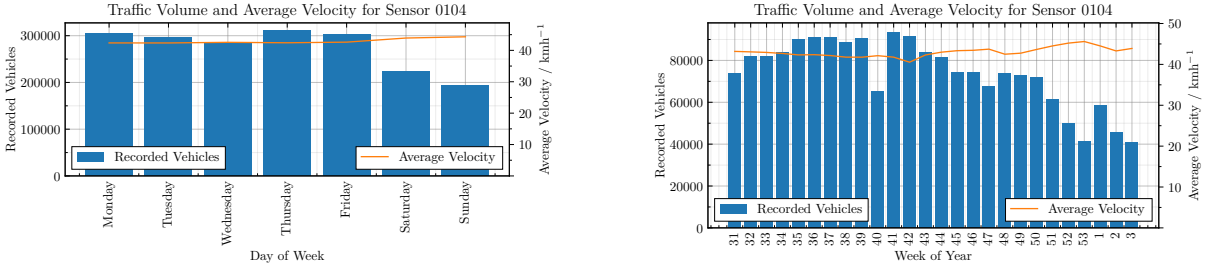

Figure 3.5: Traffic volume and average velocity per day of the week and per week of the year. From Monday through Friday, both traffic volume and average velocity are approximately constant. Throughout the year, traffic volume and average velocity do change, but the relative changes are much less than throughout individual days.

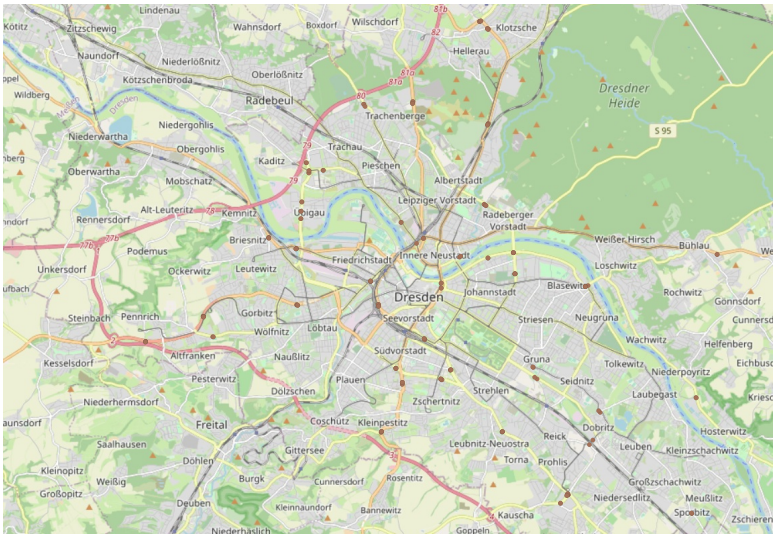

Figure 3.6: Detector locations indicated by brown dots (screenshot from [Ope17]).

4 Analysis of Vehicle Velocities

4.1 Car Velocity Distributions

4.1.1 Empirical Velocity Distributions

Usually, traffic detectors record traffic over a fixed time interval and average the measurements. In Dresden, for example, these time intervals used for traffic control span 60 s, 112 s, 5 min or 15 min.[1] When the quantity of concern is the velocity, these averaged velocities fulfill the conditions of the central limit theorem (CLT) and follow a normal distribution, e.g. as seen in Figure 4.1. Thus, in traffic science, vehicles velocities, e.g. for simulation purposes, are often assumed to follow a normal distribution, e.g as discussed in [AA14] or [Sch10]. However, the fact that the average velocities follow a normal distribution does not imply that the individual velocities are normally distributed and the empirical findings show that this is in fact not true in Dresden.

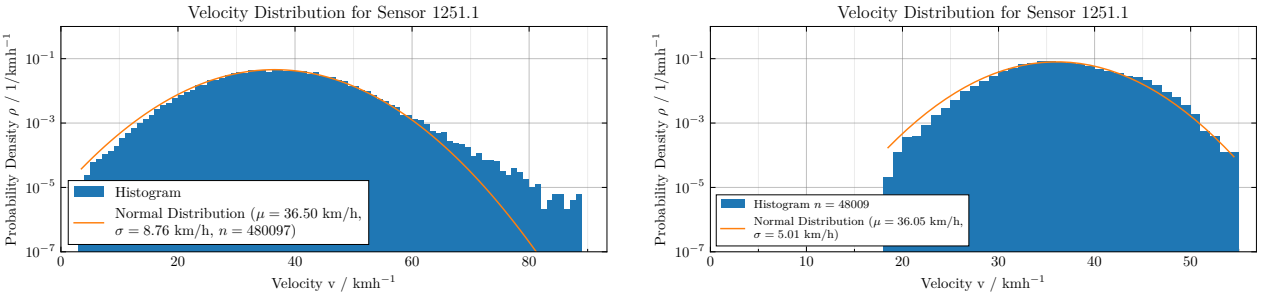

Figure 4.1: The velocities recorded at detector 1251.1 approximately follow a normal distribution (left: prior to averaging; right: after averaging over $k = 10$ vehicles). (Averaging over a number of vehicles rather than a time interval is preferable according to [TK10, p. 15].)

[1] These numbers were given to the author by employees of the Chair of Traffic Process Automation who regularly work with the traffic control system, they could not be independently confirmed as they are not published.

Generally speaking, the empirical velocity distributions can be split into four sections:

- **left flank:** Low velocities do not follow a clear trend. Rather, they heavily depend on detector placement. For some detectors, the velocities follow a uniform distribution. For others, the distribution shows a secondary maximum, possibly corresponding to a detector placement following a traffic light, thus not leaving cars enough time to fully accelerate after the traffic light has turned green. We also find detectors where very low velocities are not recorded whatsoever, for example when the detector is on a high-speed road without obstacles that never sees congestion effects.
- **peak:** Close to the peak, typically plus or minus one standard deviation, the distribution approximately follows a normal distribution, indicated by the parabolic shape in the logarithmic plot.
- **right flank:** The right flank typically follows a distribution with an exponential-like behavior, corresponding to a linear decline in a logarithmic plot. With an exponential decay, high velocities are much more pronounced compared to a normal distribution as visible in Figure 4.2. This will be further analyzed in the next section, Subsection 4.1.2.
- **extreme values:** However, for extreme velocities, even the exponential distribution underestimates the probabilities. Distributions that can describe these velocities will be further discussed in Section 4.2.

4.1.2 Right Flank Velocity Model Estimation

Now, we want to establish a simple model for velocities corresponding to the right flanks of the empirical velocity distributions, that is velocities exceeding $\mu + \sigma$ (where μ and σ have been calculated from the data directly). Figure 4.2 shows the relative frequency of different exponential decay parameters λ that have been fitted to the values exceeding $\bar{v} + \sigma_v$ using maximum likelihood estimation. As the full velocity distributions are not very skewed and display a Gaussian behavior towards their peaks, it is reasonable to assume that 50% of the total probability belongs to the interval $[0, \mu]$ and 68.3% to the interval $[\mu - \sigma, \mu + \sigma]$, leaving $1 - (0.5 + 0.683/2) = 0.1585 = 15.85\%$ to the values exceeding $\bar{v} + \sigma_v$. This is confirmed by the actual velocity distributions which have $(16.452 \pm 0.019)\%$ of the probability mass concentrated in this interval.

The spread of the exponential decay parameter is fairly small as shown in Figure 4.3 and it does not correlate with any other detector characteristics like total traffic volume, car volume, average velocity, velocity standard deviation, lane, etc. pp. This suggests that it is a somewhat universal property across all detectors studied here. On average, the exponential decay

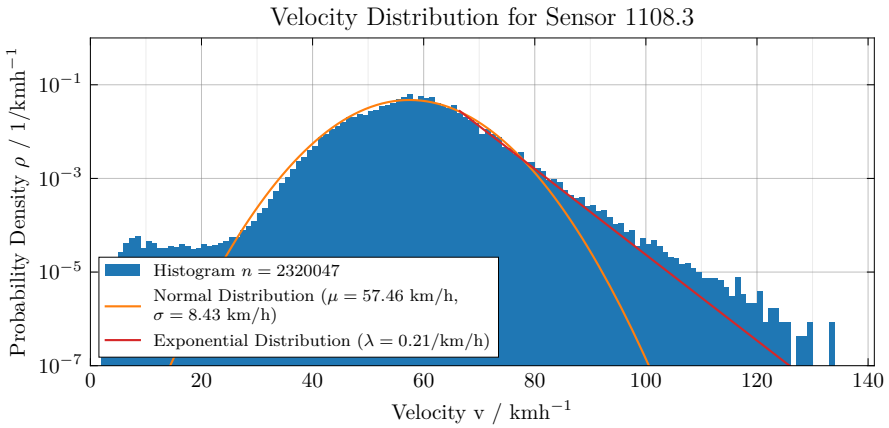

Figure 4.2: The right side decay of the velocities recorded at detector 1108.3 approximately follows an exponential distribution.

parameter is $\bar{\lambda} = (0.240 \pm 0.065)\,\mathrm{kmh}^{-1}$.

Albeit that given its simplicity the model explains the behavior of the velocities quite well, it underestimates higher velocities as previously mentioned and as one would need to expect. Closer analysis reveals that the exponential decay parameter is indeed time-dependent, as can be seen in Figure 4.4. It fluctuates on rather long time scales compared to the local traffic equilibrium, influenced e.g. by the time of day and seasonal events. In fact, for any given point in time (or, rather, time interval) the exponential distribution seems to be a better model than for the entire, cumulated, and time-independent velocity distribution. The fact that the right flank velocity distributions seem to originate from exponential distributions in superposition points towards superstatistics as an alternative approach.

Superstatistics have been used in various context; the most relevant here being train [BB07] and airplane delays [Mit+21]. It predicts the emergence of so-called q-exponentials when individual exponential distributions are in superposition:

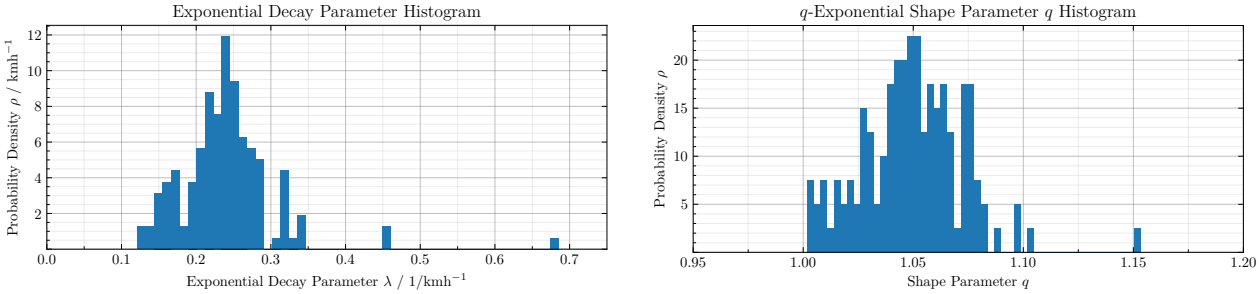

Figure 4.3: Histograms of exponential decay parameters λ (left) and q-exponential shape parameter q (right).

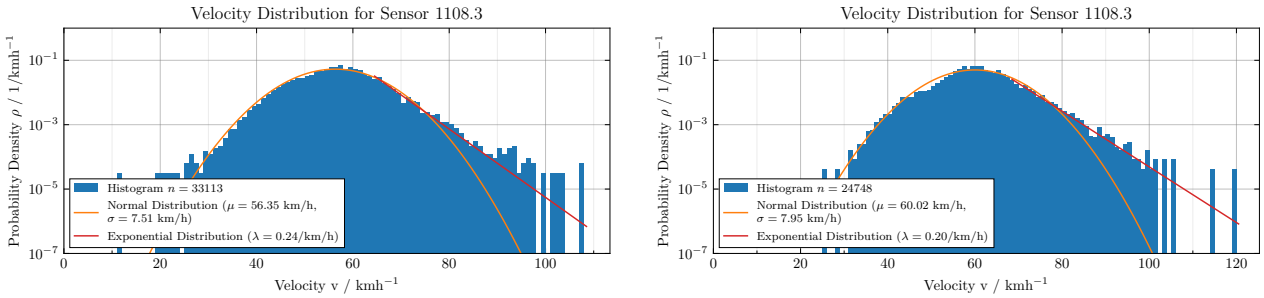

Figure 4.4: The exponential decay parameter is time-dependent (left: October 1–7, 2021, right: April 1–7, 2022).

$$\exp_q(x) = (1 + (1-q)x)^{1/(1-q)}, \quad q \in \mathbb{R}, \quad 0 < q < 2 \tag{4.1}$$

Here, q is a parameter that describes the deviation from a usual exponential distribution, where $q = 1$ corresponds to the exponential distribution. As outlined before, the deviations from an exponential decay arise from the fluctuating exponential decay parameter. q can then be determined from the distribution that the exponential decay parameter follows over time; a simple model for which along with more mathematical details can be found in [BB07]. The methodology used here to perform the fits is inspired by [Mit+21] and relies on maximum likelihood estimation. Figure 4.5 shows that the q-exponential clearly provides a much better fit than a normal exponential distribution, e.g. displayed in Figure 4.2. Yet, even though the effect is real, the deviations from the exponential distribution are rather small over all detectors, indicated by q being strictly greater but close to 1, visible in Figure 4.3. Again, as the spread of the q-parameters is fairly small, it is reasonable to make a universality assumption with $\bar{q} = (1.0498 \pm 0.0233)$. As Equation 4.1 asymptotically behaves like $x^{(1/(1-q))}$ for $x \to \infty$, this corresponds to a tail behavior of $x^{-20.08}$.

4.1.3 Principal Component Analysis and Histogram Clusters

Now, we want to attempt to cluster the total velocity distributions for cars (leaving aside restrictions for i.i.d. assumptions which will be introduced in Section 4.2), characterized by their normalized histograms which can be expressed as 254-dimensional vectors where each component corresponds to the probability of one bin of 1 km/h. First, we conduct a principal component analysis, seen in Figure 4.6. This is done using SCIKIT-LEARN [Ped+11]. As only two principal components explain the vast majority of the total variance, we attempt a visual representation of all histograms using dimensionality reduction. Principal component analysis itself does not yield visually useful results, thus we resort to UMAP [MHM18] as a nonlinear

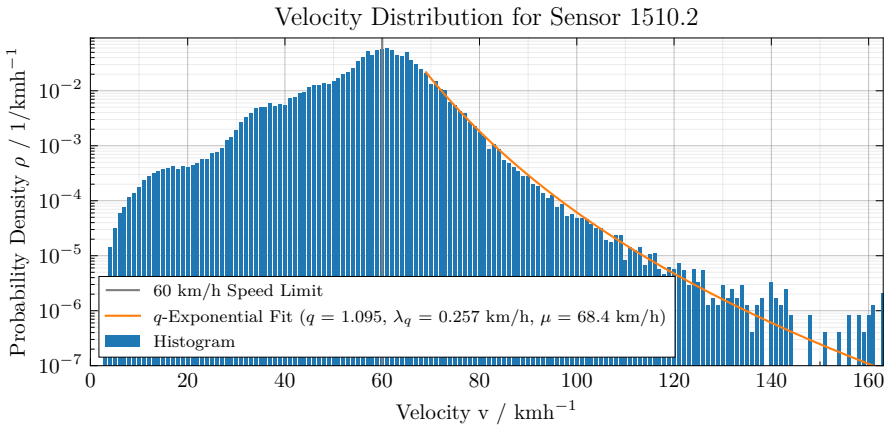

Figure 4.5: q-Exponential distribution fit to the right flank of the empirical velocity distribution for detector 1510.2.

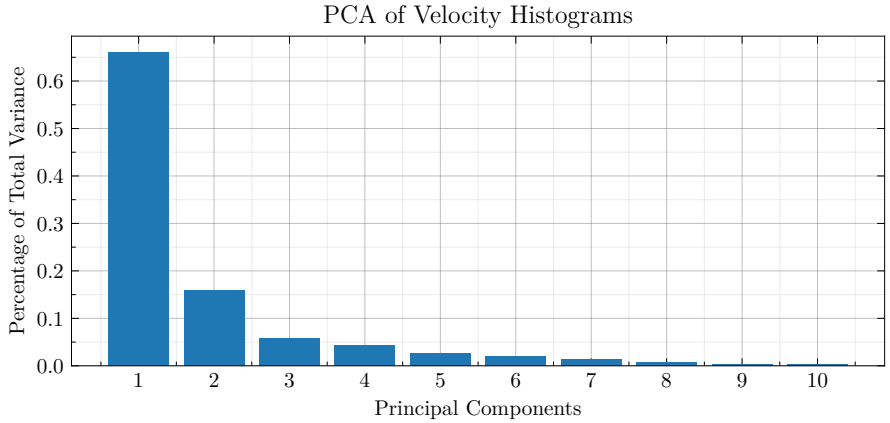

Figure 4.6: Percentage of explained total variance by the number of principal components.

dimensionality reduction algorithm, visualized in Figure 4.7.

As can be seen in Figure 4.7 (left), the clusters in the UMAP plot are highly correlated with average velocities for that detector. Interestingly, this does not translate to speed limits: While 100 km/h– and 70 km/h–zones cluster nicely in the two-dimensional representation, the positioning of detectors in 50 km/h– and 60 km/h–zones cannot really be distinguished; they mix seemingly randomly, cf. Figure 4.8. This hints at the fact that there is a lot of variance among these detectors, with "slow" detectors where vehicles barely reach the speed limit and "fast" detectors where vehicles regularly exceed the speed limit.

The number of four clusters in Figure 4.7 (right) has been arbitrarily chosen to obtain clusters that are visually separable and equisized. Similarly, clusters 1 and 4 and clusters 2 and 3 merge when only two clusters are chosen. This is necessary as KMeans does not establish the number of clusters but rather assigns each vector to one of a predetermined number of clusters. While the 2d-UMAP projection is clearly highly correlated with the average velocity and so are the clusters, this cannot explain the stark divide between clusters 1/4 and 2/3. The color scale is linear, i.e. vectors with almost identical average velocities are found both in clusters 1/4 and 2/3. As it seems unlikely that the strong separation of the clusters is merely a statistical artifact, this should be studied further.

When the histograms are normalized to the same velocity expectation value, the clusters are explained by the width of the distributions with high and low-velocity standard deviations appearing together respectively. This has to be expected: When the first moment of the distribution is eliminated and cannot explain the remaining total variance, only higher moments can contribute with the second central moment contributing most. Once the variance is taken into account by normalizing the distributions to the same standard deviation, the histograms appear uniformly distributed in the 2d-plain and clusters disappear. This is a behavior that

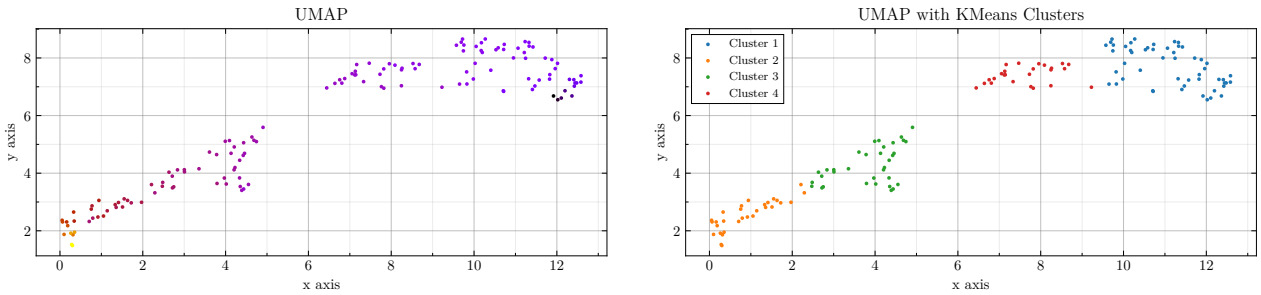

Figure 4.7: UMAP projection of all histograms. On the left, the color indicates the average velocity for that detector where black is the lowest and yellow is the highest. On the right, the histograms have been assigned to four total clusters using the KMeans algorithm [Llo82].

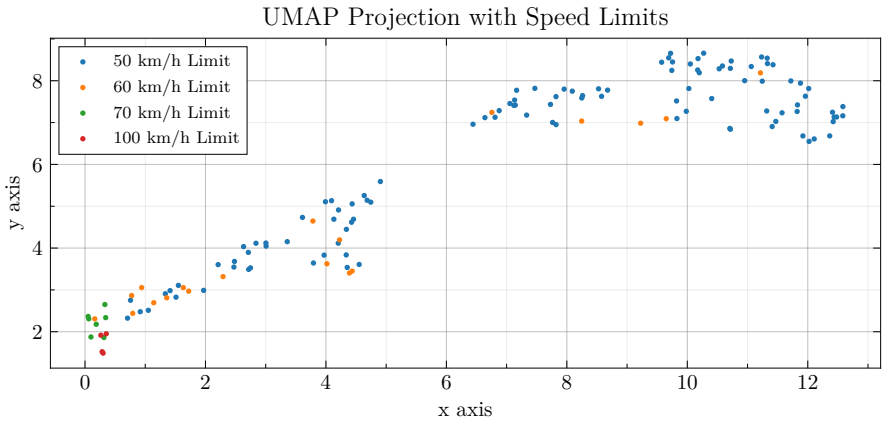

Figure 4.8: UMAP projection of all histograms. The color indicates the speed limit which clearly does not coincide with clusters.

one would expect to see from probability distributions, yet it establishes that there is no other underlying structure that differentiates the detectors.

4.2 Properties of Extreme Velocities

4.2.1 Statistical Assumptions

When applying statistical analysis tools to real-world data, it is often – e. g. for extreme value theory – important that the data is i.i.d. Thus, based on our prior findings, we establish selection criteria that try to ensure that the data is i.i.d. Subsequently, only the data fulfilling these i.i.d. criteria are analyzed. As examining the full velocity distributions conditional on variables like time, type of vehicle, etc. would not be feasible, we use the total traffic volume and the mean velocity as proxies for the distribution.

- **Vehicle type:** Clearly, different vehicle types will see different velocity distributions. For example, trucks might not actually be capable of reaching high velocities in an in-city context; busses are usually driven by specially trained personnel that might be less prone to exceed speed limits; motorcycles might be encouraged to drive fast as they are less likely to be detected by speed cameras. Thus, for the rest of this thesis, unless explicitly mentioned, we will focus on cars only.

- **Time of day:** As the driving behavior of one driver is limited by how many vehicles

are on the road and how fast they are driving which depends on the time of the day, the velocity distribution for one car also depends on the time of the day. Even though a finer resolution would be possible, the most obvious choice is to differentiate between day and night traffic. As a high vehicle volume is desirable for a meaningful statistical analysis, we restrict the data we consider to 7.00 am to 6.00 pm. During this time, as established in Section 3.2, both traffic volume and average velocity are approximately constant.

- **Day of week:** Since the traffic volume drops significantly on weekends and the average velocity adjusts accordingly (cf. Section 3.2), we only assume traffic from Monday through Friday as i.i.d. This deviates from traffic science literature [cf. TK10, p. 25], which usually considers Monday, Tuesday through Thursday, and Friday separately.

- **Week of year:** Over the year, total traffic volume per week varies by more than a factor of two, both through seasonal and pandemic-induced effects. Yet, the traffic volume, even over e.g. the Christmas season stays so high that the average velocity is only barely affected. Thus, we do not make any restrictions based on the week of the year.

- **Lane:** Under many circumstances, traffic on two lanes going into the same direction can be assumed as approximately i.i.d. However, as we study an in-city scenario where many of the detectors are placed close to intersections and traffic lights in order to be able to control traffic flow and thus turning lanes play a role, we choose not to consider separate lanes as i.i.d. and analyze each lane individually.

- **Velocity:** The detectors report faulty measurements as a velocity of 255 km/h and velocities over or equal to 254 km/h as 254 km/h. Thus, for the following statistical analyses, we only include velocities less than 254 km/h as these correspond to actual velocity measurements.

It quickly becomes apparent that these criteria are by no means strict. Yet, they are necessary to proceed further with statistical analysis. This is supported by another argument, applying to extreme value theory: Imagine a set of random numbers that are i.i.d. Now, we add random numbers drawn from different probability distributions which all have a compact support whose upper bound is lower than the upper bound of the distribution of the i.i.d. random numbers or which have tails that drop off so quickly that in a finite set of random numbers drawn from those non-i.i.d. distributions one will find virtually none exceeding a certain threshold. Then, the set of i.i.d. and non-i.i.d. random variables is no longer i.i.d. Yet, when extracting extreme

values, either through the block maxima or the peaks over threshold method, they will be the same as if they had been drawn from the i.i.d. set – given that the block size or the threshold respectively are high enough.

This is helpful as in traffic, the velocity distributions of individual vehicles become dependent when traffic densities are high. In those situations, however, velocities are low, meaning they will ”automatically” be excluded from the analysis when studying high velocities through the block maxima or peaks over threshold method. The same applies to other perturbations like traffic lights, public transport, and bikes or pedestrians. In essence: High velocities require free-flowing traffic which suggests that the velocities of different vehicles are independent.

When high velocities are dependent and thus cluster, e.g. because of road racing, the block maxima method is preferable over the peaks over threshold method. This is because the block maxima method is blind towards clusters in the sense that it still only selects the maximum value of the cluster whereas the peaks over threshold method would select the entire cluster. This is particularly relevant in high-flow situations where clusters occur.

It should also be noted that velocity distributions of different vehicles are certainly not identical but instead depend on the driver. Still, it is reasonable to assume that there is a typical velocity distribution compared to which deviations based on individual driving behavior are small so that under the previously established criteria the individual velocity distributions can approximately be assumed as identical.

4.2.2 Fit Method Verification

We use SciPy's `stats.rv_continuous.fit` for `genextreme` and `genpareto` [Vir20] to analyze the extreme velocities. As outlined in Subsection 2.1.6, the fitting should ideally be conducted using maximum likelihood estimation, which has been implemented in the relevant SciPy modules. In order to validate our results, we first have to verify the fit method.

Thus, we generate random numbers following GPD and GEV distributions with a fixed shape parameter and then fit the corresponding distribution in order to confirm that we can recover the shape parameter from the data. Then, we draw block maxima and peaks over threshold from the data and refit the distributions. This also yields the same shape parameter as predicated by the mathematical theory. Even when drawing block maxima from the GPD data and peaks over threshold from the GEV data and then fitting GEV and GPD distributions, we obtain the shape parameter used to generate the GEV and GPD data in the first place with

sufficient precision. Here, it should be noted that SciPy uses an opposite sign convention for the GEV shape parameter γ compared to this thesis.

Then, we draw block maxima and peaks over threshold from sets of random numbers drawn from a standard Gaussian distribution and a standard exponential distribution. In both cases, this yields, as expected, $\gamma \approx \xi \approx 0$. However, both for the GEV and the GPD distribution, the shape parameter is always negative, slowly converging towards zero for $k \to \infty$ and $q \to 1$ respectively.

Finally, we apply our fit method to real-world traffic data. Both fit methods are consistent in the sense that they reproduce the same fit parameters for identical data. As the block maxima depend on the order of the data, the GEV shape parameter γ varies slightly when shuffling the data prior to drawing block maxima.

It should be noted that the GEV fit method runs in local minima occasionally, thus aborting the fit algorithm and missing the global minimum that corresponds to the actual shape parameter. This behavior becomes less apparent when increasing the block size. However, even for lower block sizes, it can be avoided by repeating the fitting process multiple times and then removing outliers. The GPD fit method does not display this kind of behavior.

In theory, maximum likelihood estimation would allow to extract precise values for the uncertainties of the obtained parameters. However, this is not implemented in SciPy's `stats.rv_continuous.fit`. As the data sets we are dealing with are fairly large, it would be necessary to implement a custom, runtime-efficient MLE algorithm in order to obtain analytical uncertainties, which is outside the scope of this thesis.

Thus, alternatively, we resort to a bootstrapping approach in order to obtain uncertainties on the 1-σ-level, corresponding to approximately 68% certainty. The fit method then works as follows:

1. Split up the data into five data subsets, each containing 40% of the data (and thus overlap).
2. Select five block sizes (80%, 90%, 100%, 125%, and 150% of the initially chosen block size plus corresponding quantiles).[2]

[2] In theory, the shape parameters for different block sizes should be identical as long as all block sizes are large enough. Yet, in finite data sets, the shape parameter will vary as the block maxima that are selected naturally fluctuate. These fluctuations allow us to obtain a better estimate of the actual shape parameter and its uncertainty.

3. Shuffle the subsets of the data and draw block maxima and peaks over threshold according to the different block sizes and quantiles.
4. Fit the corresponding distribution five times for each subset and each block size or quantile and remove potential outliers that correspond to unphysical local minima in the fit algorithm.
5. Calculate the average and standard deviation of the results. Obviously, for the location and scale parameter, only the initial block size or quantile is considered.

4.2.3 Block Size and Quantile Selection

Even though in traffic science, averages and other quantities are usually taken over a time interval, we choose the blocks from which we extract the maxima to consist of a fixed number of vehicles. As outlined in [TK10, p. 15], this is statistically preferable. The usual convention in the traffic sciences merely stems from technical limitations and data availability.

In order to maintain comparability, we ensure that we fit the GEV and GPD distributions to the same number of data points by fixing the block size k and then choosing the threshold corresponding to the quantile $1 - 1/k$. Thus, for a block size of e.g. 100, we obtain exactly one percent of the data, both for the GEV and the GPD distribution.

Figure 4.9 shows how the shape parameter changes depending on the block size and the quantile respectively. It is immediately apparent that particularly for the GEV distribution, the shape parameter quickly converges to a fixed value, and it does so for a block size well below 100. Thus, we (somewhat arbitrarily) select 100 as the block size, supported by the fact that a), as previously explained, multiple block sizes are considered for the fitting and b) a much higher block size would induce variance caused by a reduced number of available data points.

The traffic data used here is discretized as only integer velocities are recorded. Meanwhile, the peak over threshold velocities usually fall into a fairly small interval. Thus, the number of peaks over a given threshold determined by a certain quantile as a function of the quantile is a discontinuous function of the quantile when $q \to 1$.[3] Consequently, the GPD shape parameter oscillates as a function of the quantile as compared to a slow and steady converging behavior of the GEV shape parameter.

[3] Effectively, the threshold can only be x or $x + 1$ and as the peaks follow a power-law type decay, the relative abundance of x compared to $x + 1$ is much higher (and so on). As we are then fitting the distribution to significantly less data, we obtain a different shape parameter.

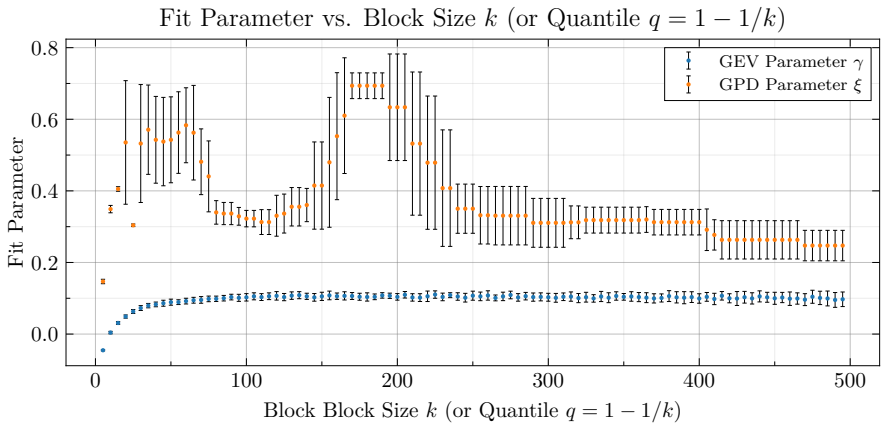

Figure 4.9: Shape parameter as a function of the block size k and corresponding quantile $q = 1 - 1/k$ (for detector 0465.1). While ξ is distorted by rounding effects, γ shows a converging behavior with variance and block size increasing together.

4.2.4 Classification of Extreme Value Distribution Types

As Figure 4.10 clearly shows, the shape parameters for all detectors cluster around the shape parameter mean, meaning that their extreme values approximately (qualitatively, apart from scale and location) follow the same distributions.

One thing to note is that all GPD shape parameters are significantly higher than the GEV shape parameters, even though theory predicts that they should be identical. This cannot be fully explained. Yet, it seems reasonable that this is due to the discrete nature of the velocities caused by the rounding to integer values. This is in contrast to the results in Subsection 4.2.2, where the fit methods would reliably reproduce the same shape parameters. However, here we are dealing with discrete as opposed to continuous random variables. Even for Gaussian, GEV, or GPD data, when rounded to integer values, the GPD shape parameters obtained through maximum likelihood estimation are systematically shifted to higher values. This effect does not occur for the GEV distribution, most likely because block maxima span over a larger interval than peaks over threshold, i.e. the GEV distributions support is larger, cf. e.g. Figure 2.6. This might reduce the relative influence of rounding. The shape parameters for the GEV distribution are also generally speaking more plausible as the right tail of the velocity log-histograms is almost linear for most detectors, corresponding to an exponential decay and thus a shape parameter of zero or close to zero.

This essentially invalidates the estimates for the GPD distribution, even though visually they seem to provide a reasonable fit. This is unfortunate as the peaks over threshold approach would in principle allow to 'stitch' together a full probability distribution for all velocities,

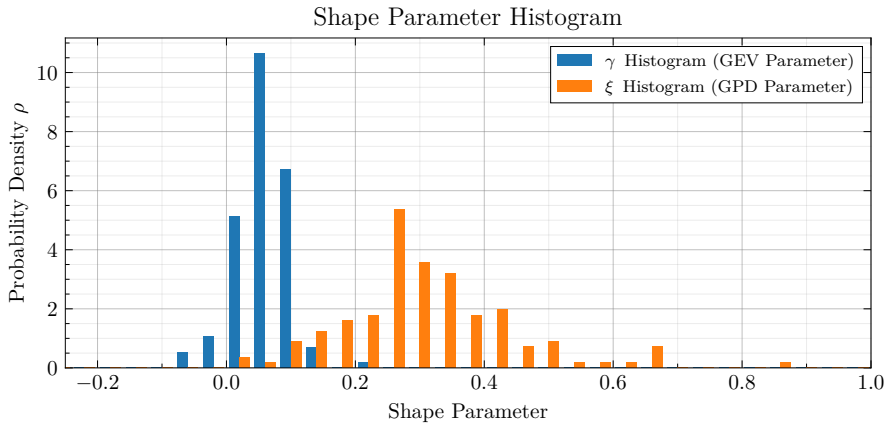

Figure 4.10: Histogram of GEV and GPD shape parameters for all detectors.

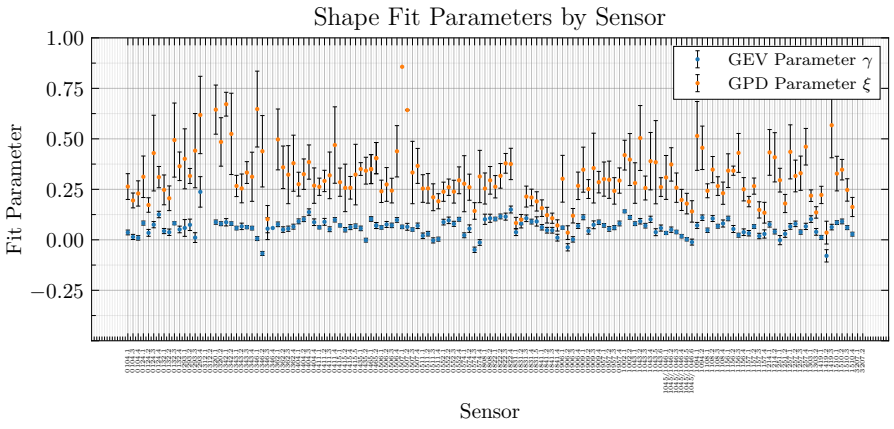

Figure 4.11: GEV and GPD shape parameters including uncertainty for all detectors.

where a Gaussian or exponential distribution would describe the part of the distribution closer to the mean and the GEV distribution would describe the extreme events exceeding some threshold.

This alone leaves the GEV distribution as the more suitable option to model extreme velocities. As a side benefit, this comes with the increased ability of the block maxima method to be applied to dependent data.

The GEV shape parameter is uncorrelated to the average velocity, the 0.95-quantile velocity, the velocity standard deviation, the total traffic volume, or the total car traffic volume. This suggests that the extreme values for all detectors do qualitatively behave similarly and that the tail behavior is a somewhat universal property of car velocity distributions.

Including the 1-σ-uncertainty of the GEV shape parameter, 88.7% of all detectors belong to the Fréchet class of distributions. 2.8% belong to the Weibull class of distributions. The rest cannot clearly be assigned to one of the Gumbel, Fréchet, or Weibull classes given their uncertainties. Thus, the extreme velocities of most detectors follow a power-law type decay behavior implied by the Fréchet class. However, this behavior is not very pronounced as the shape parameter is relatively close to zero. Averaged over all detectors, the shape parameter with one standard error is $\bar{\gamma} = (0.0612 \pm 0.0014)$. As established in Subsubsection 2.1.2, the tails asymptotically behave like $x^{-(1+1/\bar{\gamma})} \approx x^{-17.34}$ for $x \to \infty$.

As $1/\bar{\gamma} > 16$, the first sixteen moments exist (as outlined in Subsubsection 2.1.2) – this is clearly always true for finite data sets but not necessarily for the underlying distributions. Thus, even for the theoretical distributions, calculations of the mean, variance, etc. yield meaningful results.

All Fréchet distributions have no right endpoint, "though it is unreasonable on physical grounds to believe that [extreme events] are truly without limit" [Col01, p. 66], which is certainly also true for car velocities. This is, however, not an issue as the probability distributions fall off so quickly that in any real-world finite set of random variables there will basically never be a velocity that is truly unphysical, but points to the fact that the model needs to be applied diligently to incredibly high velocities.

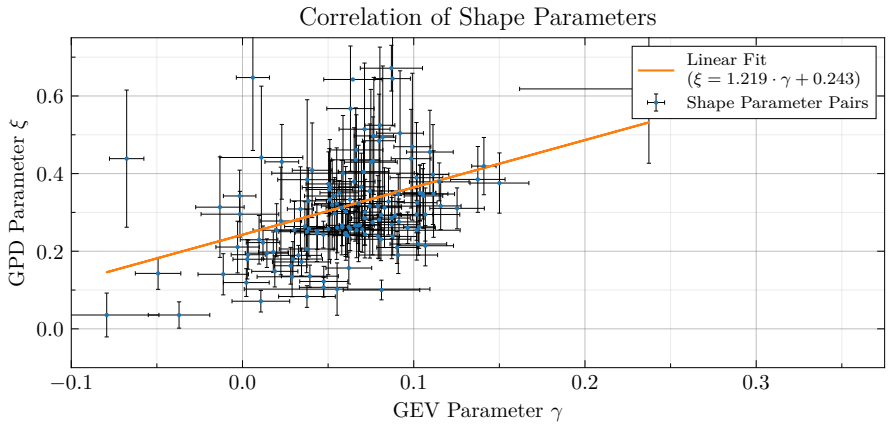

Figure 4.12: Correlation of GEV shape parameter γ and GPD shape parameter ξ. The linear trend line indicates a systematic shift of ξ to higher values, most likely caused by rounding. Other than that, the correlation is very loose, underlining that the fit method for the GPD shape parameter is invalid.

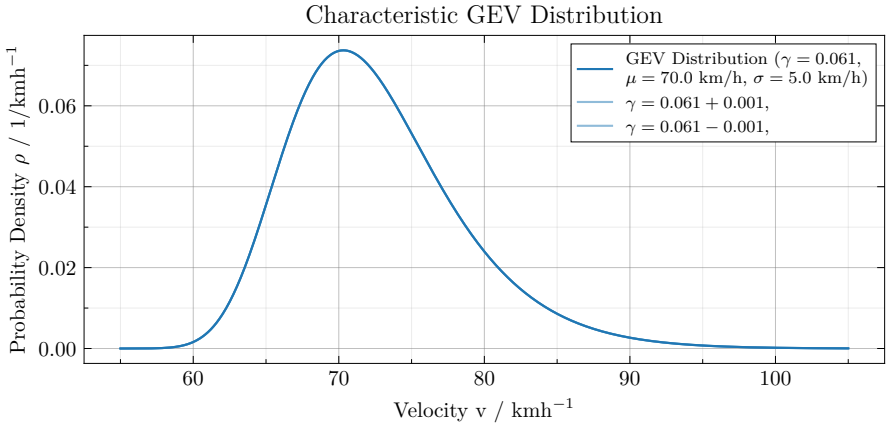

Figure 4.13: Characteristic GEV distribution for car velocities for $k = 100$ and $\gamma = (0.0612 \pm 0.0014)$. The uncertainty of the shape parameter is so small that it can only barely be seen in the plot. (The location and shape parameters have been arbitrarily chosen for this representation.)

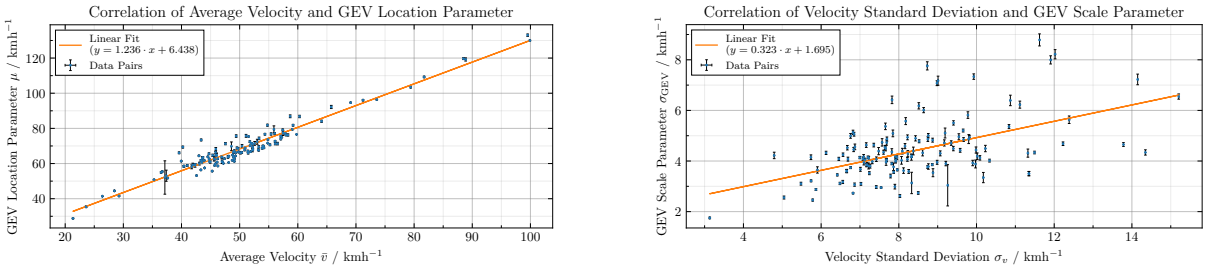

Figure 4.14: Relation between velocity mean and standard deviation and GEV parameters. While the GEV location parameter is essentially a linear function of the average velocity and can thus be estimated from the average velocity, the same statement cannot be made about the GEV scale parameter and the velocity standard deviation, which seems to sensitively depend on the underlying distribution of the velocities.

5 Traffic Analysis

5.1 Flow and Density Distributions

We compute traffic flows and densities as outlined in Subsection 2.2.2. For that, we need the time headways or gross time gaps between vehicles, the sum of the net time gaps displayed in Figure 5.2 and the occupation times displayed in Figure 5.1.

It is immediately apparent that the histograms for both quantities are discretized, albeit on different scales. This hints at the fact that the data resolution might not be as high as suggested by the data type used to store the data in the database. The discretization of the time gaps could be caused by instruments with insufficient precision within the detectors. The discretization of the occupation time could be both the cause or the consequence of the discretization of the velocity measurements, however, we cannot say as this is not discussed in [Str12] and we do not have any further knowledge of the inner workings of the detectors.

As the detectors are of fixed length l, the velocity is the inverse occupation time (modulo detector length). Thus, their probability distributions are related by $g(t_{\text{occ}}) = l/t_{\text{occ}}^2 \cdot f(l/t_{\text{occ}})$, where f is the corresponding velocity probability density function.

If the time gap between two vehicles is sufficiently large, they should be effectively independent of each other and pass the detector at a constant rate. Then, the number of vehicles per time interval should follow a Poisson distribution. Consequently, the time gap between two vehicles[1] should follow an exponential distribution. This corresponds to a linear behavior in the logarithmic plot which can be observed for sufficiently high time gaps, e.g. in Figure 5.2. This behavior does not extend to low time gaps, because then vehicles are no longer independent of each other, for example, due to safety distance requirements (and physical limitations).

[1] Here, differentiating between gross and net time gaps does not make a big difference as occupation times are constant for a given velocity and one order of magnitude smaller than net time gaps, thus effectively shifting the distribution of gross time gaps compared to the distribution of net time gaps.

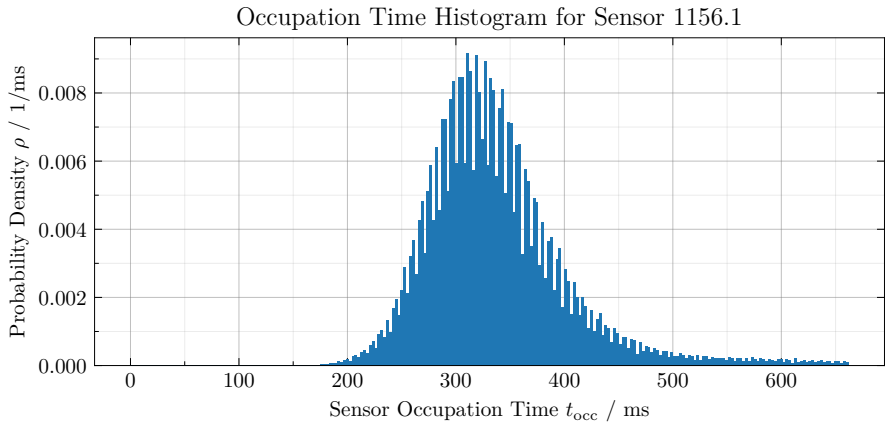

Figure 5.1: Histogram of detectors occupation times vehicles for detector 1156.1.

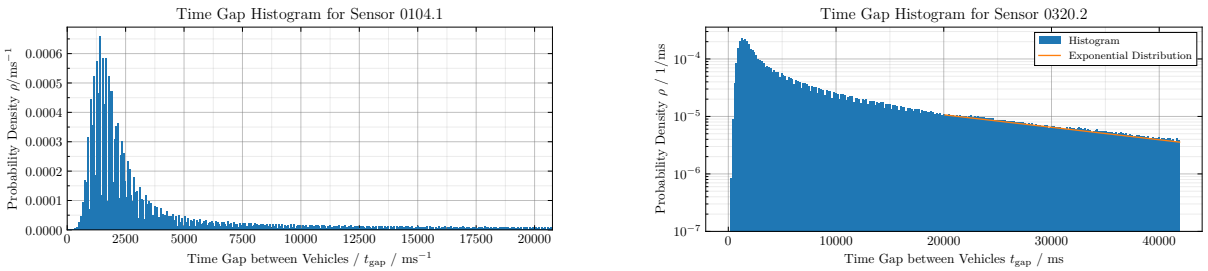

Figure 5.2: Histogram of time gaps between vehicles: linear scale for detector 0104.1 (left) and logarithmic scale for detector 0320.2 (right).

Traffic flow and traffic density are macroscopic properties that can only really be assigned meaning after averaging over a time interval or a number of vehicles. As outlined in Subsection 4.2.3, it is preferable to average over a fixed number of vehicles rather than over a time interval. Yet, in the process of averaging, a lot of information over the system state is lost. Thus, depending on the intended application, it can be beneficial to average over fewer vehicles. In the limit, this would only be one vehicle, where its time gap to the preceding vehicle and its velocity would determine its microscopic traffic flow and density.

Figure 5.3 shows the effect of averaging over $k = 1$ vs. $k = 10$ vehicles on the traffic flow. When averaging over more vehicles, the distribution becomes more and more concentrated, converging towards one value for $k \to \infty$.

Figure 5.4 presents two types of traffic flow histograms. All detectors fall into one of the groups or a combination of both. However, most detectors follow a type-II-like behavior, with flow peaking well below $1000\,\mathrm{h}^{-1}$ and then decreasing monotonously.

It should be noted that the traffic flow never actually reaches zero. Due to technical limitations, the maximum time gap that the detector system records is 65 534 ms, about one minute, corresponding to a (minimum) microscopic traffic flow of $55\,\mathrm{h}^{-1}$ reported by the system, while actual traffic flows can be arbitrarily close to zero.

Traffic densities, shown in Figure 5.5, then exhibit a similar behavior as they are obtained from traffic flows by dividing by the corresponding velocities.

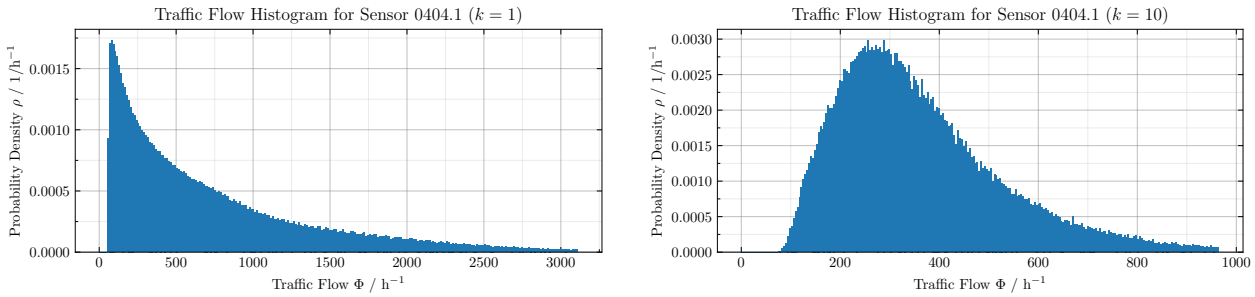

Figure 5.3: Influence of averaging on the distribution of traffic flows ($k = 1$ on the left; $k = 10$ on the right).

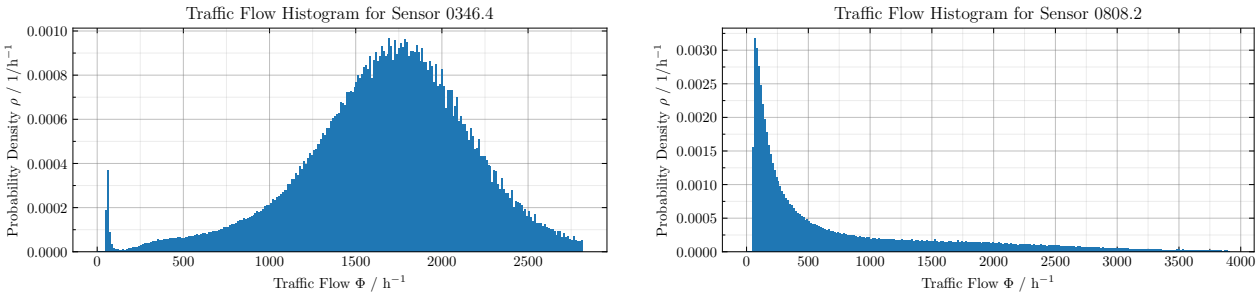

Figure 5.4: Histograms of traffic flow ($k = 1$): type I (left) and type II (right).

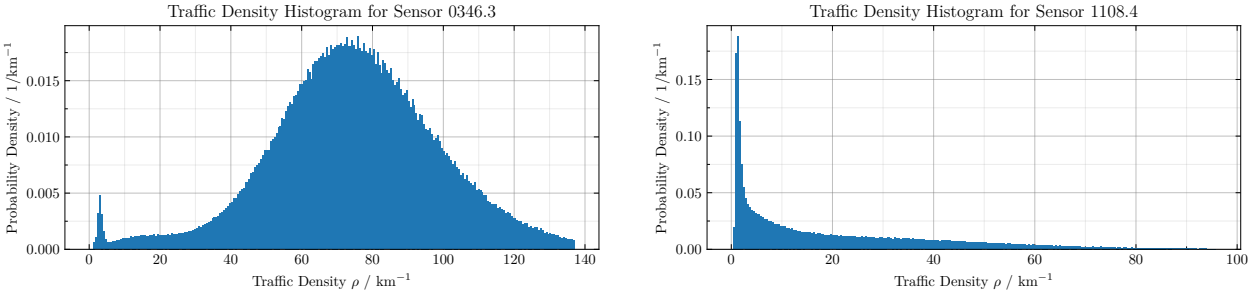

Figure 5.5: Histogram of traffic densities ($k = 1$): type I (left) and type II (right).

5.2 Characteristic Diagrams and Classification

Characteristic diagrams are then obtained from the data by creating a 2-dimensional histogram of the relevant quantities. In particular, the flow–density diagram is of interest here as it allows us to easily differentiate between congested and free-flowing states of traffic.

Figure 5.6 shows such a flow–density diagram. The average flow is obtained by averaging over all flows corresponding to a certain density. Towards higher densities, one finds regions for which flow–density combinations have not been recorded. These correspond to linear functions whose slope, representing the velocity, are non-integer – values that cannot be recorded by

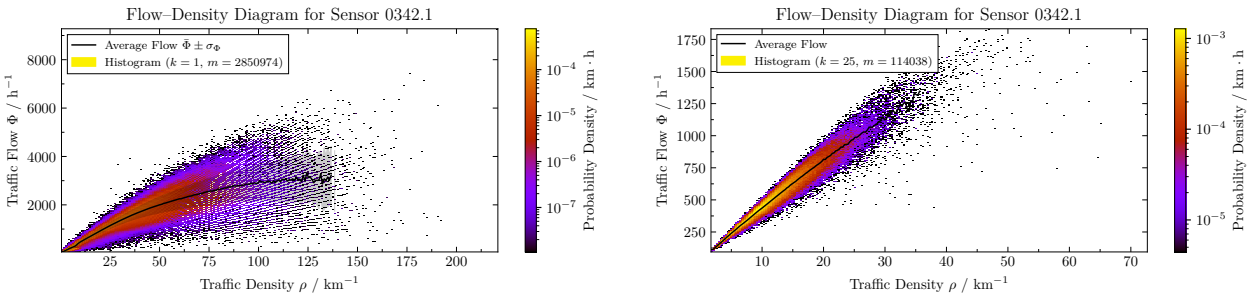

Figure 5.6: Flow–density diagrams for detector 0342.1: without prior averaging ($k = 1$, left) and with averaging ($k = 25$, right).

the system in use in Dresden. As one would expect, the average velocity $v = \Phi/\rho$, decreases with increasing density. The (theoretical) propagation velocity $u = d\Phi/d\rho$, however, decreases faster, almost reaching zero. Yet, it does not become negative which would correspond to backward propagation, indicating a traffic jam. Even though the propagation velocity does almost reach zero, the vast majority of all traffic measurements are far from that, showing that the road system usually operates far from its peak capacity.

In fact, this finding can be generalized to all detectors found in Dresden – congestion virtually never occurs at the detector sites. For most detectors, mean velocities stay more or less constant over the typical density ranges, and (theoretical) propagation velocities stay well above zero. This is unfortunate as it does not allow for analyzing congestion effects, but is clearly good news for everybody driving on Dresden roads.

Usually, traffic detectors are set up to average over several vehicles or a certain time interval before computing traffic quantities. This reduces complexity and allows to draw real-time conclusions about the state of the system. Though, a lot of information is lost in the process. Thus, here, we have opted to forego the averaging over consecutive vehicles and only average in the end to infer the flow–density relationship. Figure 5.6 illustrates the difference that this makes: While the density–flow diagram with averaging does not show any congestion effects whatsoever and suggests an almost constant traffic velocity, the other, 'microscopic' one shows clear signs of these effects.

Even though the velocity–flow diagram is most intuitive in some sense and is most suited for identifying congested states of traffic, it is not necessarily the most natural way to present the data. That would be the velocity–flow diagram that is shown in Figure 5.7. This representation only relies on the velocity and time headways which can be directly measured as opposed to the density. Depending on the context and application, this can be preferable as the density estimates are biased as can also be seen in Figure 5.7 – low densities and low velocities do not occur in conjunction as the density is calculated as $\rho = \Phi/v$ while the flow has a lower

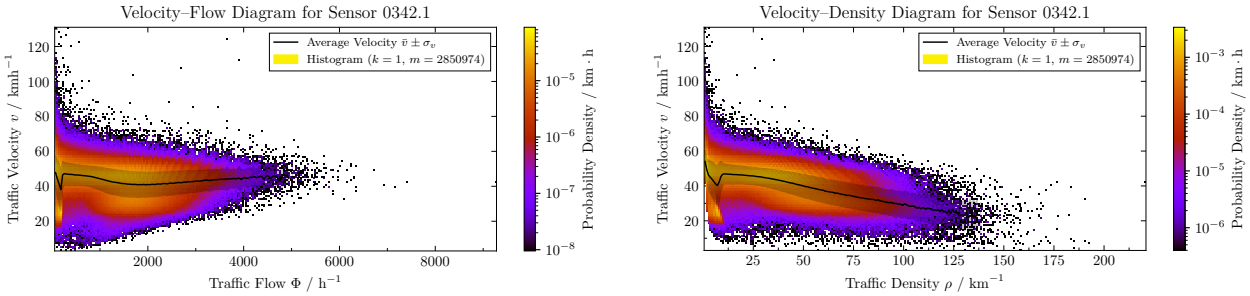

Figure 5.7: Velocity–flow (left) and velocity–density diagrams (right) for detector 0342.1.

bound.

The velocity–flow diagram can be broken down even further: Figure 5.8 shows the velocity distributions conditional on a given traffic flow. Essentially, these are cross-sections of the velocity–flow diagram in Figure 5.7 parallel to the flow axis. As one would expect, they exhibit similar behavior as the entire empirical velocity distributions studied in Subsection 4.1.1. Qualitatively, the distributions for different traffic flows do not differ significantly, apart from the fact that the mean tends to increase with decreasing traffic flow (even though here, for detector 0342.1, the average velocity is relatively constant with a small dip towards intermediate traffic flows as can be seen in Figure 5.7).

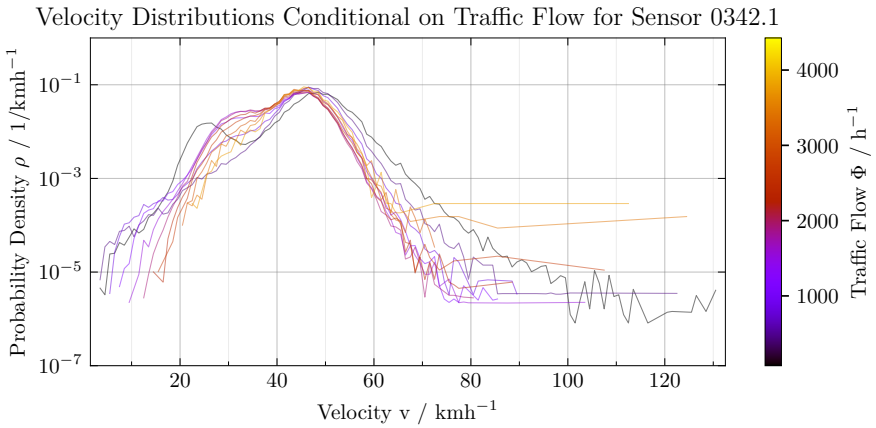

Figure 5.8: Velocity distribution for detector 0342.1 conditional on the traffic flow. It should be noted that these empirical distributions are in fact histograms with bin size 1 km/h which are only plotted as continuous functions for visual representation purposes.

5.3 Estimation of Free-Flow Velocity

For traffic planning and road design, it is highly relevant to know the free-flow velocity v_0. This is the velocity that drivers would choose in absence of obstacles or perturbations like other cars [TK10, pp. 28 sqq.]. E. g., if one were to design a road like a highway but with a speed limit of 30 km/h, drivers would likely drive faster than that – this constitutes a mismatch between the speed limit and road design.

A good proxy for the free-flow velocity is the velocity that drivers choose in the limit of a vanishing traffic density, that is $\rho \to 0$ [Str12, pp. 28 sqq.]. This leaves two at least in principle equivalent options: Firstly, one can determine the point where the velocity axis is being intercepted in the velocity–density diagram. Secondly, one can determine the slope of the flow–density diagram in the limit $\rho \to 0$.[2]

In reality, however, both approaches are flawed. The detectors employed in Dresden are stationary and cannot resolve traffic spatially. Thus, traffic densities can only be deduced from traffic flows and velocities. Then, however, densities and velocities are no longer independent: As the time gap has an upper bound, the traffic flow has a lower bound. As $\rho = \Phi/v$, low densities do per definitionem correspond to high velocities. Clearly, this cannot be used for proper free-flow velocity estimation.

The same issue arises when trying to determine the free-flow velocity from the flow–density

[2] As the flow–density relation is linear in the origin, we have $\lim_{\rho\to 0} v = \lim_{\rho\to 0} \Phi/\rho = \lim_{\rho\to 0} d\Phi/d\rho$.

diagram, restricting both densities and flows to low values.

This motivates an approach relying only on measured velocities and time gaps. When time gaps are long, drivers are also approximately free to choose their velocity. In principle, ceteris paribus, we could thus find the average velocity for $t_{\text{gap}} \to \infty$ (which also implies $\Phi \to 0$). Figure 5.9 (left) illustrates this behavior: The longer the time gaps between vehicles, the faster vehicles drive, converging towards the free-flow velocity. This is, however, an idealized image that would likely apply to a highway but in cities, the assumption of all other things being equal does not necessarily hold – in fact, for extreme gaps for some detectors under certain conditions, the average velocity even decreases again as the condition $t_{\text{gap}} \to \infty$ introduces bias (right part of Figure 5.9). This can be explained by the fact that only very few vehicles actually experience these time gaps, for example, because they only occur at night when drivers might be inclined to drive more carefully or because they correspond to vehicles that were previously stopped by traffic lights.

For some detectors, the average velocity even shows sharp peaks or troughs for specific time gaps. These can be attributed to traffic control measures like traffic lights with a fixed period, where vehicles with a fixed time gap are e.g. those that just accelerated in order to avoid a red light or those that had to come to a full stop and wait for a full light cycle.

In order to mitigate this effect, we do not simply select the highest rolling average velocity over all time gaps or the average velocity for the maximum time gap. Rather, we compute the rolling average velocities for all time gaps and then select the highest quintile of these velocities. Selecting the highest quintile is clearly a somewhat arbitrary choice, yet it is a reasonable trade-off between considering too many velocities which do not all correspond to free-flow traffic states and considering too few velocities and thus introducing bias. This is supported by the fact that selecting the highest quartile or the highest decile yields the same free-flow velocity. Often, the highest quintile of average velocities corresponds to the highest quintile of time gaps, yet for some detectors, it corresponds to time gaps as low as 10 s. Even though this is a fairly small time gap, it is reasonable to assume that these vehicles have chosen

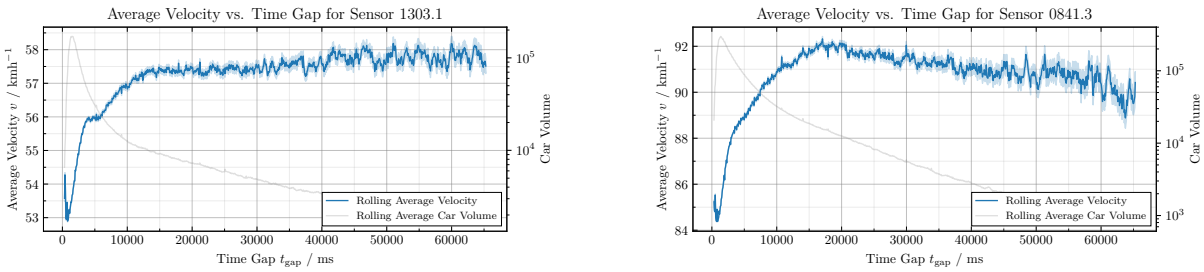

Figure 5.9: Average velocity as a function of the time gap to the previous vehicle.

their velocity freely if the average velocity does not significantly increase for higher time gaps. In order to obtain the free-flow velocity, we then average over all velocities corresponding to a time gap whose average velocity lies in the highest quintile of the average velocities. Usually, this incorporates 5%–20% of all recorded velocities.

Using this approach, we obtain the free-flow velocities found in Figure 5.10 as a function of the average velocity. Clearly, the free-flow velocities are higher than the average velocities. This is per construction but also per definitionem. If, as a thought experiment, we think of a driver driving at its free-flow velocity and then add other vehicles, the presence of these other cars can only really force the driver to slow down, not accelerate.

Figure 5.10 also shows that the free-flow velocity and average velocity converge towards each other for higher average velocities. We have no means to establish a causal relationship, but two hypotheses could explain this type of behavior:

1. Drivers prefer to exceed the average velocities that are usually driven on that road by a fixed amount, e.g. deeming an increased velocity in the range from 5–10 km/h to be acceptable, regardless of the speed limit. Then, in a 30 km/h-zone the relative difference will be much greater than in a 70 km/h–zone.
2. Drivers have a free-flow velocity that is higher than the speed limit by a fixed amount. When the average velocity is higher, they are already closer to their desired free-flow velocity, leaving less room to drive faster even with low time gaps.

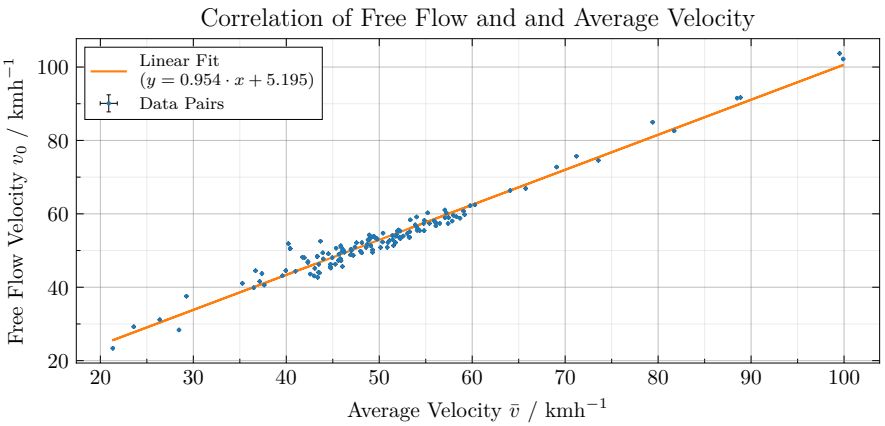

Figure 5.10: Correlation of free-flow velocity and average velocity.

6 Applications

6.1 Speed Limit Violations

Now, we want to use the previously established statistical models in order to model speed limits violations and compare them to actual speed limits violations founds in the data. The speed limits were extracted from OpenStreetMap [Ope17]. The penalties in form of fines, demerit point and months of driving bans that we use here are according to the German penalty catalogue regulation[1]. In theory, even exceeding the speed limit by 1 km/h can lead to a fine. In practive, however, a tolerance of 3% of the driven velocity or at least 3 km/h is subtracted in favor of the driver.[2] Here, we take an even more conservative approach: As the exponential and q-exponential distributions are only valid models for velocities exceeding $\mu_v + \sigma_v$, we disregard any speed limits violations that are below $\mu_v + \sigma_v$. This is justified as for many detectors, the mean velocity surpasses the speed limit – in these cases, a strict enforcement of speed limit violations would lead to more than half of all drivers receiving a ticket which seems unreasonable.

According to local newspapers, the city of Dresden raised 2.7 million Euros in 122.800 speed

Potential...	**Fines / €**	**Demerit Points**	**Driving Bans / Months**
Obs. Violations	876×10^6	1340×10^3	522×10^3
Normal Dist.	780×10^6	633×10^3	151×10^3
Exponential Dist.	814×10^6	1038×10^3	348×10^3
q-Exponential Dist.	840×10^6	1425×10^3	543×10^3

Table 6.1: Annual speed limit violations for all detectors and predictions using normal, exponential, and q-exponential distributions. All penalties are potential in the sense that they are not actually issued but rather correspond to the speed limit violations found in the data.

[1] Anlage 1 zu §1 Abs. 1 Verordnung über die Erteilung einer Verwarnung, Regelsätze für Geldbußen und die Anordnung eines Fahrverbotes wegen Ordnungswidrigkeiten im Straßenverkehr (BKatV), retrieved on July 20, 2022 from https://www.gesetze-im-internet.de/bkatv_2013/.

[2] cf. https://www.bussgeld-info.de/ab-wann-blitzt-ein-blitzer/, retrieved on July 20, 2022.

limit violations in 2021[3], 2.2 million Euros in 79.233 speed limit violations in 2020[4] and 3.67 million Euros in 143.285 speed limit violations in 2019[5].

Table 6.1 shows the potential penalties for the speed limit violations found in the data. Most strikingly, the actual potential fines exceed those actually collected by a factor of more than 300. Clearly, this figure needs to be taken with a grain of salt: If these penalties were issued, drivers would quickly adapt their behavior. Given that there are only 27 speed cameras in Dresden[6], it is not surprising that there are more speed limit violations in the data than those that are sanctioned. Yet, the factor of 300 suggests that drivers only momentarily reduce their speeds at speed camera sites and engage in excessive speeding that does not have any consequences elsewhere.

While fines already apply to minor speed limit violations, demerit points and driving bans only apply to significant exceedances of more than 20 km/h. As the vast majority of all vehicles are centered around the velocity mean where the normal distribution is a decent model, the normal distribution is capable of describing the total fines reasonably well. However, this is not true for demerit points or driving bans which are dramatically underestimated by the normal distribution. While the exponential distribution is more capable for all penalties, its performance is exceeded by the q-expoential distribution which has the lowest relative deviation from the observed speed limit violations while slightly overestimating high velocities and thus potential demerit points and potential driving bans.

This finding is not only true for the total penalties per year found in Table 6.1, but also extends to the predictions for the individual detectors and their respective velocity distributions as can be seen in Figure 6.1. However, deviations between observed and precited penalties are much greater than for the total penalties as fluctuations do not average out over all detectors. While most detectors see potential fines per day in the range of 10000€ – 30000€, one detector even records car velocities that would amount to 50000€ in fines per day. Figure 6.2 shows the results for demerit points and months of driving bans. The potential demerit points for each detector are also best modeled by the q-exponential distribution, albeit the tendency of the q-exponential to overestimate high velocities is clearly visible here. The deviations between model and observations can be significantly reduced by fine-tuning the fit to each distribution which would, however, be much more work than an umbrella-type fit method

[3] cf. https://www.zeit.de/news/2022-02/15/blitzer-bussgelder-spuelen-millionen-in-kassen-der-grossstaedte, retrieved on August 22, 2022.

[4] https://www.tag24.de/dresden/politik-wirtschaft/dresden-verliert-million-bei-blitzer-einnahmen-hier-kommt-der-naechste-hin-1795531, retrieved on August 22, 2022.

[5] cf. https://www.tag24.de/dresden/crime/rekordeinnahmen-hier-stehen-die-fiesesten-blitzer-in-dresden-rathaus-kasse-gefuellt-falschparker-1400661, retrieved on August 22, 2022.

[6] cf. https://www.blitzer.de/ort/blitzer-in-Dresden-Sachsen, retrieved on September 6, 2022.

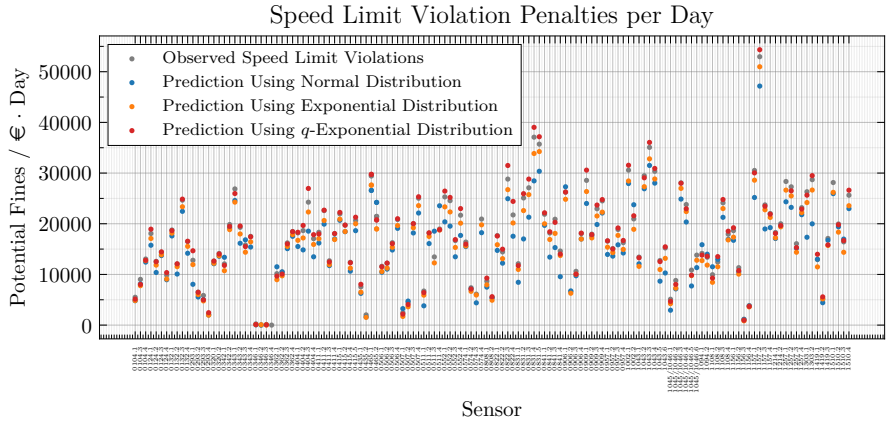

Figure 6.1: Potential fines per day per detector found in the data and predicted by the respective normal, exponential, and q-exponential distributions.

employed here. The months of driving bans estimation exhibits the same behavior, even though the overestimation by the q-exponential is even more pronounced.

The potential amount of driving bans underlines the remarkable extent to which speeding occurs: For some detectors, more than 330 potential months of driving bans are accrued per 10000 vehicles. If these were enforced, this would lead to a new traffic equilibrium as all drivers would end up being banned from driving after some time, assuming that it is not always the same drivers who are speeding.

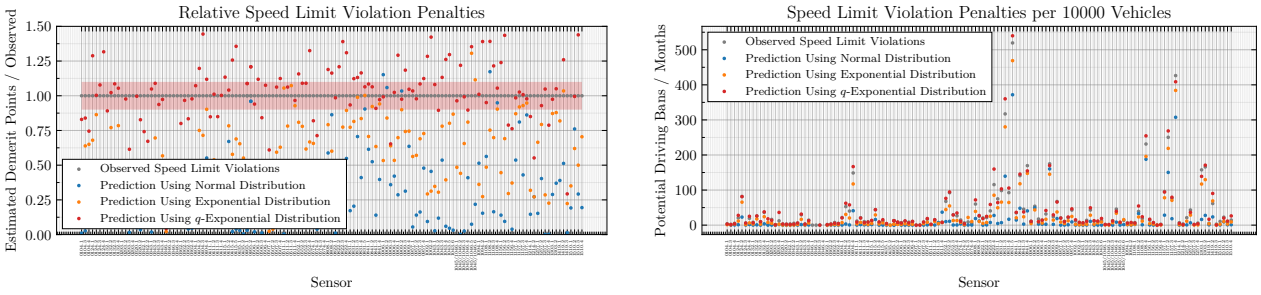

Figure 6.2: Left: Estimated demerit points relative to observed potential demerit points. The red shaded area are the observed demerit points ±10 percentage points. Right: Potential months of driving bans per 10000 vehicles.

6.2 Speed Limit and Road Design Mismatches

From a traffic planning perspective, it is intuitive that good traffic design comprises two elements:

- an average velocity that (approximately) matches the free-flow velocity and
- a free-flow velocity that (approximately) matches (or at least does not surpass) the speed limit.

If the first condition does not hold, it indicates that vehicles are facing constraints caused by the traffic volume – because otherwise, they would choose to increase their velocity. This can be an early indication that road capacities are being stretched.

If the second condition does not hold, this constitutes a mismatch between the speed limit and the road design. This can mean two things: Either, the speed limit should be changed, because currently drivers are not allowed to go as fast as the road design would suggest or because the road design does not allow them to go as fast as would be legal. Or, the road design should be changed, incentivizing drivers to drive faster or slower to match the speed limit. E.g., on a highway, even if the speed limit was only 30 km/h, drivers would be inclined to go faster, and, vice versa, in a crowded suburban area with winding roads and a lot of street parking, even without a speed limit, drivers would not be able to go much faster than 30 km/h.

Figure 6.3 shows that traffic is in a good situation on most detectors sites, generally speaking. Yet, there are some clear outliers in both directions: For detector 0293.1, the average velocity sits at about 75% of the speed limit while the free-flow velocity sits at about 85%. For detector 0511.1, the average velocity is approximately 115% while the free-flow velocity is roughly 125% of the speed limit.

In all of these cases, a closer analysis from a traffic science perspective, taking into account the specific context of each detector, could reveal further insights and room for improvement. Furthermore, the speed limits obtained from OpenStreetMap are static and reflect the status quo on August 1, 2022. Thus, temporary effects like construction sites or changes in the speed limit could not be taken into account.

Even though there are some outliers, Figure 6.4 exhibits that most free-flow velocities fall under the speed limit which is also true for the average free-flow velocities, except for the 70 km/h–limit. It should be taken into account, however, that as there are only a few detectors

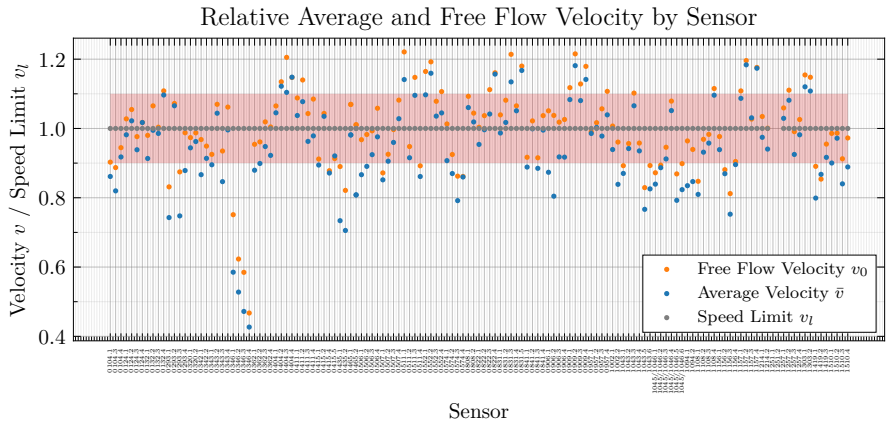

Figure 6.3: Free-flow and average velocities normed to the speed limit. The red shaded area is the speed limit ±10 percentage points that are deemed to be acceptable. As the speed limit for detector 1251 is time-dependent, it is not shown here.

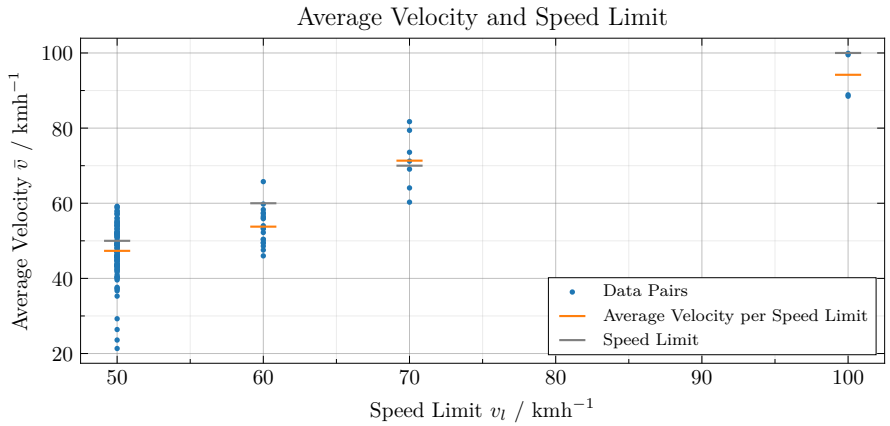

Figure 6.4: Free-flow velocity and average free-flow velocity per speed limit.

in 70 km/h–zones the average free-flow velocity for these detectors does not carry a lot of significance.

7 Results and Discussion

In this thesis, we have analyzed an extensive set of microscopic traffic data from Dresden. First, we focused on the properties of velocity distributions. Then, we turned our attention to various aspects of traffic science like flow and density distributions, characteristic diagrams, and free-flow velocities.

Analyzing all traffic detectors, we have found that car velocity distributions can be split into four distinct parts: the peak, left and right flanks, and extreme values. The left flank is highly dependent on the detector location while the peak behaves approximately Gaussian. For fixed, short-time intervals, the right flank is best described by an exponential distribution. The exponential distributions decay parameter is a fairly universal property across all detectors with $\bar{\lambda} = (0.240 \pm 0.065)\,\mathrm{kmh}^{-1}$. Integrating over all times, these individual exponential distributions overlap and their superpositions lead to the emergence of q-exponentials with power-law tails. Again, the q-exponential distributions shape parameter is similar across all detectors with $\bar{q} = (1.0498 \pm 0.0233)$. The fluctuations of the exponential decay parameters are relatively small, being reflected in q being close to 1 which is consistent as $q = 1$ would imply a normal exponential distribution without fluctuations in the decay parameter.

In order to introduce a model for extreme values, we first had to make assumptions regarding the vehicle type, the hours of the day, the days of the week, the weeks of the year, lanes, and velocities that restricted the data to a subset that can be assumed as independent and identically distributed. Notably, deviating from the usual hypothesis in traffic science that differentiates between Mondays, Tuesdays to Thursdays, and Fridays, we found no relevant differences between different weekdays. We have also developed a methodology that can consistently and accurately fit the Generalized Extreme Value distribution parameters and that has been verified using mock and real data. To conduct the fits, a block size of $k = 100$ has turned out to provide a good trade-off between bias and variance. Subsequently, we determined that the distributions for extreme car velocities of 88.7% of all detectors belong to the Fréchet class of the extreme value distribution types with a shape parameter of $\bar{\gamma} = (0.0612 \pm 0.0014)$. Thus, we have concluded that the tails of the distributions do in fact exhibit a power-law behavior

that asymptotically scales like $x^{-(1+1/\bar{\gamma})} \approx x^{-17.34}$. Given the relatively large uncertainty of $\bar{q}$, this is indeed compatible with the results for the q-exponentials and increases the significance of the result. It should be noted, however, that q-exponentials underestimate the tails of the distributions, likely because they are fitted to much more data, most of which is not considered to be extreme. q-exponentials are also of interest as they provide a possible explanation for the emergence o power-law tails. Considering the rather high power-law exponent, we have also shown that on average the first 16 moments of the underlying theoretical distributions exist. Furthermore, we have found that the GEV location parameter (for block size $k = 100$) is very well described as a linear function of the average velocity, while the GEV scale parameter and the velocity standard deviation are only loosely, albeit positively correlated. We have not been able to properly fit the Generalized Pareto Distribution, most likely due to the discrete nature of the data.

This thesis starts with the quote "Extreme-Value Statistics can be regarded as the art of extrapolation" [BZ21]. In conclusion, we can say that the data suggests that there is actually no need to extrapolate. Given the high power-law exponent of -17.34, even an observation period that is 20 times as long would on average only result in the maximum velocity being observed increase by 1 km/h. This underlines the value of the data set at hand: Essentially all the information about the underlying velocity distributions is already engraved in the data and more data would not necessarily enable more insights, albeit that non-integer velocities would be useful, e.g. for better fitting results.

All these findings hinge on the fact that the data is microscopic. As we have demonstrated, prior averaging and thus simulating mesoscopic or macroscopic traffic data yields Gaussian distributions for the velocities.

Applying our learnings about velocity distributions, we have determined that q-exponentials are best suitable to model speed limit violation fines, demerit points, and driving bans. Astonishingly, the amount of annual potential fines found in the data exceeds those collected by the city of Dresden by a factor of more than 300. We have also illustrated that for some detectors, immediately issuing and enforcing driving bans would lead to the road being empty after less than a month as all drivers would have been banned, exemplifying the extent to which speeding occurs.

Employing principal component analysis as well as non-linear dimensionality reduction, we have confirmed that average velocities are in fact the factor explaining most of the variance between detector velocity distributions and explain the distributions projection onto the 2d-plain best. The 2d-projection has been found to be comprised of two clusters whose origins

remain unclear. Lastly, the 2d-projection indicates that speed limits are only very loosely correlated with the actual velocity distributions. In particular, is it not possible to differentiate between 50 km/h and 60 km/h limits based merely on the empirical velocity distributions.

One caveat, however, that all these results entail is that we have not been able to verify the detector measurements independently and cannot conclude about statistical fluctuations, systematic distortions, or potential measurement errors producing truly unphysical velocity measurements.

Shifting towards aspects from traffic science, we have confirmed that for large enough time headways, the time headways follow an exponential distribution whereas the number of vehicles per time interval follows a Poisson distribution. Additionally, we have provided empirical velocity distributions for occupation times, time headways, traffic flows, and traffic densities.

We have also found that congestion has virtually no effect on traffic in Dresden. When averaging over a fixed number of vehicles before constructing the flow–density diagrams (i.e. simulating mesoscopic or macroscopic data), the traffic velocity as indicated by the flow–density diagram essentially becomes constant and independent from the traffic density. Directly plotting the microscopic traffic data, however, has revealed that the propagation velocity does in fact decrease with increasing traffic density, albeit that the propagation velocity does not dip below zero and that this only occurs at a small fraction of all points in time analyzed here. These findings suggest that the Dresden road system operates far from its capacity.

Using time headways as a proxy to identify free-flow traffic states, we have established a methodology to estimate free-flow velocities. The free-flow velocities we have obtained are consistently higher than average velocities as one would expect, but free-flow and average velocities converge towards each other for higher velocities. Adding speed limits to the picture, we have identified roads where we suspect a mismatch between road design and speed limit, indicated by large deviations between free-flow velocities and speed limits. Even though this effect is suppressed by the traffic volume decreasing average velocities such that they are in line with speed limits, here, traffic planning officials should consider changing either speed limits or the road design. Again, as already discussed on the basis of the velocity distributions, real traffic behavior is barely different between 50 km/h– and 60 km/h–zones.

In general, it should be noted that these findings can only be generalized to some extent: While the number of detectors and their locations suggest some universality, the detector locations have been chosen to support traffic control and not for representative scientific studies.

8 Outlook

In this thesis, we have explored the potential of the extensive microscopic traffic data set from Dresden. Our results illustrate the value of the data set. Yet, due to time constraints, this thesis was only able to scratch the surface of the data set and its potential is much greater. Thus, here, we provide an overview of open questions and potential starting points for future analyses.

Vehicle Velocity Distributions:
In the previous chapter, most distribution fits were based on prior assumptions which were then visually validated. To improve the results, hypothesis testing and statistical fits could be employed. Furthermore, most uncertainties were obtained through bootstrapping approaches. This could be refined by analytical uncertainties which would require a new implementation of maximum likelihood estimation supporting uncertainties that is capable of processing the required amounts of data in a reasonable amount of time.

So far, we did not have access to individual vehicle velocity distributions. If these were available, they could be assigned to the domains of attraction of the GEV distribution types in order to validate or correct our results [cf. LR83, p. 16]. As mean velocity fluctuations over time are relatively small, we have implicitly assumed the stochastic process to be stationary. If these fluctuations were taken into account, methods for extreme value theory describing non-stationary processes could be used, potentially improving the results [Col01, p. 93].

Moreover, we have so far restricted our analyses to cars as these constitute the majority of all vehicles. It might be of interest, however, to also analyze other vehicle types, e.g. delivery trucks, to determine how their velocity distributions differ. If it was possible to gather more data, emphasis should be placed on roads with higher speed limits as these are naturally underrepresented in the city data set. The origin of the two velocity histogram clusters found in the 2d-projection could be investigated further. Lastly, the correlations of consecutive vehicles could be considered.

Detector Location:
So far, detector location has more or less been neglected. Still, correlations with traffic lights, intersections, road type, city center versus outskirts, arterial roads versus smaller roads, and road direction (into or out of the city) seem probable. Future work could look at how these factors influence velocity distributions and in particular the tails of the distributions.

Time:
The time dependency of the data has also not been considered yet. Taking this data into account would allow us to compare morning commute, evening commute, and night velocity distributions. On top of that, the data also encompasses time-dependent speed limits. This would allow us to investigate if and how drivers react to speed limits that are not constant. Coincidentally, the entire Covid pandemic and the various stay-at-home orders and other countermeasures fall into the data measurement time frame, allowing us to explore their effect on motorized mobility in Dresden. The same is true for gas price spikes that occurred in the aftermath of the Russian invasion of Ukraine, almost doubling the cost of gasoline and diesel.

Network Structure:
If the detector locations were to be modeled as a network structure, this would allow us to model the flow of traffic not only at one detector site but rather through the entire network. For example, flow-splitting at individual nodes could be statistically described, allowing for predictions of the flow of traffic through the entire network and identifying potential bottlenecks or other inefficiencies and improvements to routing. Furthermore, this could help to establish temporal correlations between events at different nodes. Potentially, the statistical routing could then be crosschecked with commuting data or real-time traffic data from Google Maps or other traffic data providers.

Traffic Models:
As touched on in Subsection 2.2.3, there is a large number of macroscopic models in traffic science predicting the shape of the fundamental diagram based on a fixed number of parameters. These parameters are usually not known a priori. However, as the flow and density profiles at the detector sites have been observed, the parameters of these macroscopic models could be fitted to match the density at the detector given appropriate time-dependent boundary conditions. In particular, the detector sites 1303, 1241, 0841, and 0831 could be of interest as they are located quite far away from perturbation-inducing obstacles like traffic lights, intersections, etc. Here, 0831 might be the best candidate as its flow–density diagram exhibits congestion effects and not only the free-flow regime.

In addition to the macroscopic models, there are also microscopic car-following models describing the behavior of individual vehicles, e.g. the Intelligent Driver Model (IDM) [Alb+22] or the BandoTFL model [GK22]. These require two detectors without many obstacles between them to fit a model. If it was possible to identify individual car trajectories in the traffic data of detectors 1043 and 1045/1046 (Fetscherstraße) or 0362 and 0343 (Washingtonstraße), a microscopic model could be calibrated to the first detector to compare its predictions with the observations at the second detector.

The microscopic car following model parameters or the data itself could also help answer the question: Do drivers going faster keep less safety distance?

9 List of References

[AS72] Milton Abramowitz and Irene Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. 10th printing. National Bureau of Standards, Dec. 1972. ISBN: 0-486-61272-4.

[AA14] Sherif M. Abuelenin and Adel Y. Abul-Magd. "Empirical study of traffic velocity distribution and its effect on VANETs connectivity". In: *2014 International Conference on Connected Vehicles and Expo (ICCVE)*. 2014, pp. 391–395. DOI: 10.1109/ICCVE.2014.7297577.

[Alb+22] Saleh Albeaik et al. "Limitations and improvements of the intelligent driver model (IDM)". In: *SIAM Journal on Applied Dynamical Systems* 21.3 (2022), pp. 1862–1892. DOI: 10.1137/21M1406477.

[BH74] A. A. Balkema and Laurens de Haan. "Residual Life Time at Great Age". In: *The Annals of Probability* 2.5 (1974), pp. 792–804. DOI: 10.1214/aop/1176996548.

[Bas11] Bojan Basrak. "Fisher-Tippett Theorem". In: *International Encyclopedia of Statistical Science*. Ed. by Miodrag Lovric. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 525–526. ISBN: 978-3-642-04898-2. DOI: 10.1007/978-3-642-04898-2_254.

[BB07] Keith Briggs and Christian Beck. "Modelling train delays with q-exponential functions". In: *Physica A: Statistical Mechanics and its Applications* 378.2 (2007), pp. 498–504. ISSN: 0378-4371. DOI: 10.1016/j.physa.2006.11.084.

[Bro+08] Ilja Nikolajewitsch Bronstein et al. *Taschenbuch der Mathematik*. 7., vollst. überarb. u. erg. Aufl. Verlag Harri Deutsch, July 2008. ISBN: 3-8171-2017-6.

[BZ21] Axel Bücher and Chen Zhou. "A Horse Race between the Block Maxima Method and the PeakoverThreshold Approach". In: *Statistical Science* 36.3 (2021), pp. 360–378. DOI: 10.1214/20-STS795.

[CP12] Lin Cheng and Sooksan Panichpapiboon. "Effects of intervehicle spacing distributions on connectivity of VANET: a case study from measured highway traffic". In: *IEEE Communications Magazine* 50.10 (2012), pp. 90–97. DOI: 10.1109/MCOM.2012.6316781.

[Col01] Stuart Coles. *An introduction to statistical modeling of extreme values*. Springer Series in Statistics. London: Springer-Verlag, 2001. ISBN: 1-85233-459-2. DOI: 10.1007/978-1-4471-3675-0.

[Düm16] Lutz Dümbgen. *Einführung in die Statistik*. 1. Aufl. 2016. Basel: Birkhäuser, 2016. ISBN: 9783034800044. DOI: 10.1007/978-3-0348-0004-4.

[FH15] Ana Ferreira and Laurens de Haan. "On the block maxima method in extreme value theory: PWM estimators". In: *The Annals of Statistics* 43.1 (2015), pp. 276–298. DOI: 10.1214/14-AOS1280.

[FT28] R. A. Fisher and L. H. C. Tippett. "Limiting forms of the frequency distribution of the largest or smallest member of a sample". In: *Mathematical Proceedings of the Cambridge Philosophical Society* 24.2 (1928), pp. 180–190. DOI: 10.1017/S0305004100015681.

[Geo15] Hans-Otto Georgii. *Stochastik: Einführung in die Wahrscheinlichkeitstheorie und Statistik*. De Gruyter, 2015. ISBN: 9783110359701. DOI: 10.1515/9783110359701.

[GD08] Nikolas Geroliminis and Carlos F. Daganzo. "Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings". In: *Transportation Research Part B: Methodological* 42.9 (2008), pp. 759–770. ISSN: 0191-2615. DOI: 10.1016/j.trb.2008.02.002.

[GK22] Xiaoqian Gong and Alexander Keimer. "On the well-posedness of the "Bando–Follow the Leader" car following model and a time-delayed version". In: (2022). DOI: 10.13140/RG.2.2.22507.62246.

[Gum58] E. J. Gumbel. *Statistics of Extremes*. Columbia University Press, 1958. ISBN: 9780231891318. DOI: 10.7312/gumb92958.

[Haa84] Laurens de Haan. "A Spectral Representation for Max-stable Processes". In: *The Annals of Probability* 12.4 (1984), pp. 1194–1204. DOI: 10.1214/aop/1176993148.

[HF10] Laurens de Haan and Ana Ferreira. *Extreme Value Theory: An Introduction*. 1st Edition. Springer, 2010. ISBN: 144192020X. DOI: 10.1007/0-387-34471-3.

[Har+20] Charles R Harris et al. "Array programming with NumPy". In: *Nature* 585.7825 (2020), pp. 357–362. DOI: 10.1038/s41586-020-2649-2.

[HLR13] Chao Huang, Jin-Guan Lin, and Yan-Yan Ren. "Testing for the shape parameter of generalized extreme value distribution based on the L_q-likelihood ratio statistic". In: *Metrika* 76 (July 2013). DOI: 10.1007/s00184-012-0409-5.

[Hun07] J. D. Hunter. "Matplotlib: A 2D graphics environment". In: *Computing in Science and Engineering* 9.3 (2007), pp. 90–95. DOI: 10.1109/MCSE.2007.55.

[Kri16] Wolfram Krick. “Die Verkehrsverflechtungsprognose 2030 und deren zugrundeliegende regionale Strukturdaten”. In: (2016).

[LR83] Georg Leadbetter M. R. Lindgren and Holger Rootzén. *Extremes and Related Properties of Random Sequences and Processes*. Springer New York, 1983. ISBN: 9781461254492. DOI: 10.1007/978-1-4612-5449-2.

[Llo82] Stuart Lloyd. “Least squares quantization in PCM”. In: *IEEE transactions on information theory* 28.2 (1982), pp. 129–137.

[Lod+19] Allister Loder et al. “Understanding traffic capacity of urban networks”. In: *Scientific reports* 9.1 (2019), pp. 1–10. DOI: 10.1038/s41598-019-51539-5.

[Löe08] Matthias Löewe. “Extremwerttheorie”. In: (2008).

[Mak06] Mikhail Makarov. “Extreme value theory and high quantile convergence”. In: *Journal of Operational Risk* 1.2 (2006). DOI: 10.21314/JOP.2006.009.

[Mak07] Mikhail Makarov. “Applications of exact extreme value theorem”. In: *Journal of Operational Risk* 2 (2007), pp. 115–120. DOI: 10.21314/JOP.2007.024.

[MHM18] Leland McInnes, John Healy, and James Melville. “Umap: Uniform manifold approximation and projection for dimension reduction”. In: *arXiv:1802.03426* (2018).

[McN99] Alexander J. McNeil. “Extreme value theory for risk managers”. In: *Departement Mathematik ETH Zentrum* 12.5 (1999), pp. 217–37.

[Mit+21] Evangelos Mitsokapas et al. “Statistical characterization of airplane delays”. In: *Scientific Reports* 11.1 (2021), pp. 1–11. DOI: 10.1038/s41598-021-87279-8.

[New04] Mark Newman. “Power Laws, Pareto Distributions and Zipf’s Law”. In: *Contemporary Physics - CONTEMP PHYS* 46 (Dec. 2004). DOI: 10.1080/00107510500052444.

[Ope17] OpenStreetMap contributors. *Planet dump retrieved from https://planet.osm.org*. https://www.openstreetmap.org. 2017.

[Ped+11] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[Pic75] James Pickands. “Statistical Inference Using Extreme Order Statistics”. In: *The Annals of Statistics* 3.1 (1975), pp. 119–131. ISSN: 00905364. DOI: 10.1214/aos/1176343003.

[Pir10] Mir Nabi Pirouzi Fard. “Probability plots and order statistics of the standard extreme value distribution”. In: *Computational Statistics* 25.2 (2010), pp. 257–267. DOI: 10.1007/s00180-009-0174-8.

[Rat20] Paul Rathke. “Potentialanalyse der Extremwerttheorie zur Anwendung im Strassenverkehr”. In: (2020).

[Roc10] Marco Rocco. “Extreme Value Theory for Finance: A Survey”. In: *Bank of Italy Occasional Paper* 99 (2010). DOI: 10.2139/ssrn.1998740.

[Rod17] Gabriel Rodrguez. “Extreme value theory: An application to the Peruvian stock market returns”. In: *Revista de Métodos Cuantitativos para la Economa y la Empresa* 23 (2017), pp. 48–74.

[Sch04] Andreas Schadschneider. *Physik des Strassenverkehrs*. 2004. URL: http://www.thp.uni-koeln.de/~as/Mypage/PSfiles/verkehr.pdf.

[Sch+18] Benjamin Schäfer et al. “Non-Gaussian power grid frequency fluctuations characterized by Lévy-stable laws and superstatistics”. In: *Nature Energy* 3 (Feb. 2018). DOI: 10.1038/s41560-017-0058-z.

[Sch10] Hagen Schüller. “Modelle zur Beschreibung des Geschwindigkeitsverhaltens auf Stadtstrassen und dessen Auswirkungen auf die Verkehrssicherheit auf Grundlage der Strassengestaltung”. In: (2010).

[Smi85] Richard L. Smith. “Maximum Likelihood Estimation in a Class of Nonregular Cases”. In: *Biometrika* 72.1 (1985), pp. 67–90. DOI: 10.2307/2336336.

[Str12] Bundesanstalt für Strassenwesen. *Technische Lieferbedingungen für Streckenstationen (TLS)*. Bundesministerium für Verkehr, Bau und Stadtentwicklung, 2012. URL: https://www.bast.de/DE/Verkehrstechnik/Fachthemen/v5-tls/tls-streckenstationen.html.

[TK10] Martin Treiber and Anne Kesting. *Verkehrsdynamik und -simulation: Daten, Modelle und Anwendungen der Verkehrsflussdynamik*. Springer Berlin Heidelberg, 2010. ISBN: 978-3-642-05227-9. DOI: 10.1007/978-3-642-05228-6.

[Umw22] Umweltbundesamt. “Fahrleistungen, Verkehrsleistung und Modal Split”. In: *Umweltzustand und Trends* (2022). URL: https://www.umweltbundesamt.de/daten/verkehr/fahrleistungen-verkehrsaufwand-modal-split.

[VD09] Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009. ISBN: 1441412697.

[Ver22] Bundesministerium für Verkehr und digitale Infrastruktur. “Verkehr in Zahlen 2021/2022”. In: (2022). URL: https://www.bmvi.de/SharedDocs/DE/Publikationen/G/verkehr-in-zahlen-2021-2022-pdf.pdf.

[Vir20] Pauli et al. Virtanen. “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python”. In: *Nature Methods* 17 (2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

[Woe10] Birgit Woeste. “Eine Anwendung der Block Maxima Methode im Risikomanagement”. Diplomarbeit. Westfälische Wilhelms-Universität Münster, 2010.

[YKÖ21] Asuman Yilmaz, Mahmut Kara, and Onur Özdemir. “Comparison of different estimation methods for extreme value distribution”. In: *Journal of Applied Statistics* 48.13-15 (2021), pp. 2259–2284. DOI: 10.1080/02664763.2021.1940109.

A Appendix

All programming necessary for this thesis were done in Python 3.10 [VD09]. Additionally, calculations were aided by NumPy [Har+20]; all statistical analysis was done through SciPy [Vir20]. The plots were created using Matplotlib [Hun07].

Due to the complexity and extend of the data set analyzed in this thesis, only a very small fraction of all diagrams and plots created during research did find their way into this thesis. The entire set of more than 6000 plots as well as the Python code and the full size plots can be requested from the author at: moritz.piepel@gmail.com.

All map material was obtained from OpenStreetMap [Ope17] on August 1, 2022. The map data also contains the speed limits used in the previous sections.

The microscopic traffic data was provided by Straßen- und Tiefbauamt (STA) of Landeshauptstadt Dresden. The author thanks them for their support.

List of Figures

List of Tables