Freeway On-Ramp Bottleneck Activation, Capacity, and the Fundamental Relationship

DISSERTATION

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Seoungbum Kim

Graduate Program in Civil Engineering

The Ohio State University

2013

Dissertation Committee:

Professor Benjamin A. Coifman, Advisor

Professor Mark R. McCord

Professor Rabi G. Mishalani
Abstract

Almost all freeway delay arises from bottlenecks of one form or another, e.g., incidents, merging, lane drops, and weaving. A bottleneck is said to be active when it is restricting flow. Conventionally an active freeway bottleneck is modeled as if it occurs at a discrete point in space, with queued traffic upstream of the bottleneck location and unqueued traffic downstream. As will be shown herein, this conventional point bottleneck model appears to be too simplistic to capture all of the traffic flow dynamics in the vicinity of bottlenecks. Empirical studies of bottlenecks are encumbered with the challenge of simultaneously estimating capacity and detecting the instant when the arriving demand exceeds the bottleneck capacity. We believe this fact may have helped to obscure a poor fit between the point bottleneck model and actual bottleneck effects that occur over an extended distance. To date a comprehensive understanding of the subtle but important factors that contribute to the bottleneck mechanism remains elusive.

This dissertation examines a merging bottleneck while revisiting commonly held assumptions and uncovering systematic biases that likely have distorted our understanding of bottleneck formation, bottleneck capacity, and even empirical studies of the fundamental relationship (FR) of macroscopic traffic flow. This simulation-based study incorporates microscopic driver behavior with macroscopic traffic flow theory and seeks to provide better insight into those bottleneck features. The simulation extends a conventional car following model to also include a driver relaxation factor for the
vehicles that enter or are immediately behind an entering vehicle (termed "affected vehicles"). Rather than instantaneously changing speed, headway or velocity after an entrance maneuver, these affected vehicles initially take a shorter headway and spacing for the given speed than allowed by the car following model and then gradually approach the model's speed-spacing relationship over many seconds.

The simulation results show that the queue initially forms downstream of the on-ramp due to the driver relaxation. The downstream end of the queue then grows further downstream of this formation location, and later recedes upstream back to the on-ramp. The spatial nature of these findings is clearly in contradiction with a point bottleneck model but is consistent with some empirical studies. According to the simulation results, it takes several minutes between the time of the initial bottleneck activation and the time the queue reaches the on-ramp. Simulating conventional detector measurements over the period, we show that flow is higher than the underlying FR would predict (termed “supersaturated flow”) in any sample containing an affected vehicle with unsustainably short headway. If one does not already know capacity (as is typically the case in an empirical study) then this systematic bias in flow due to the affected vehicles is not readily apparent in the detector measurements: During the initial queue formation speeds remain close to free speed and the supersaturated states can exceed the bottleneck capacity.

We speculate that the driver relaxation mechanism is common and that many empirical bottleneck capacity studies have erroneously mistaken several minutes of the supersaturated flows to be the bottleneck capacity, when in fact these unsustainably high
flows simply reflect the fact that the system is starting to store vehicles further
downstream during the earliest portion of bottleneck activation. Instead of flow
eventually dropping "from capacity", we see flow drop "to capacity" from supersaturation.
This dissertation investigates this issue in detail and examines how the supersaturated
states can distort an empirically observed FR.

While this study is based on a simulation work, the major findings provide clues
to better interpret past empirical observations that have not been fully explained, yielding
new insights into merge bottleneck mechanisms, capacity, and the FR. It is hoped that the
work developed herein will help shape future empirical studies to test the validity of the
new interpretations and quantify the impacts from the previously overlooked phenomena.
Dedication

This document is dedicated to my family and most especially my wife
Acknowledgments

I would like to thank my advisor, Professor Benjamin Coifman, whose support and encouragement made this dissertation possible. I would like to thank Professor Mark McCord and Professor Rabi Mishalani, who provided useful background to fulfill this research.

I would like to thank my lovely wife, who endured the long distances required for my studies and helped me through the hard times, and finally my family, who supported my study at the Ohio State University.
Vita

2004..........................................................B.S. Civil Engineering,
The Hanyang University

2008..........................................................M.S. Civil Engineering,
The Ohio State University

2004 to present...........................................Graduate Research Assistant,
The Ohio State University

Fields of Study

Major Field: Civil Engineering
Table of Contents

Abstract.................................................................................................................. ii

Dedication .............................................................................................................. v

Acknowledgments................................................................................................. vi

Table of Contents.................................................................................................. viii

List of Tables ......................................................................................................... xi

List of Figures ........................................................................................................ xii

Chapter 1  Introduction .......................................................................................... 1

Chapter 2  Literature review ................................................................................. 8

  2.1 Freeway merge bottleneck studies ............................................................. 8

  2.2 Merging behavior ........................................................................................ 11
    2.2.1 Previous studies in merging sections .................................................. 12
    2.2.2 Relaxation process .............................................................................. 15

Chapter 3  Methodology ....................................................................................... 18

  3.1 Terminology ................................................................................................ 18

  3.2 Car following models .................................................................................. 20

  3.3 Relaxation model ....................................................................................... 23
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.4</td>
<td>Catching-up and joining the queue</td>
<td>30</td>
</tr>
<tr>
<td>Chapter 4</td>
<td>Numerical analysis</td>
<td>33</td>
</tr>
<tr>
<td>4.1</td>
<td>Introduction of simulation</td>
<td>33</td>
</tr>
<tr>
<td>4.2</td>
<td>Queue formation near the on-ramp</td>
<td>35</td>
</tr>
<tr>
<td>4.3</td>
<td>Defining the loading and settling periods</td>
<td>40</td>
</tr>
<tr>
<td>4.4</td>
<td>Alternative scenarios</td>
<td>42</td>
</tr>
<tr>
<td>4.4.1</td>
<td>Cases with constant headway ramp arrivals</td>
<td>42</td>
</tr>
<tr>
<td>4.4.2</td>
<td>Cases with non-evenly spaced (NES) arrivals</td>
<td>44</td>
</tr>
<tr>
<td>4.4.3</td>
<td>Model calibration and other models</td>
<td>48</td>
</tr>
<tr>
<td>4.5</td>
<td>Summary</td>
<td>55</td>
</tr>
<tr>
<td>Chapter 5</td>
<td>Unaccounted for impacts of relaxation on empirical studies</td>
<td>57</td>
</tr>
<tr>
<td>5.1</td>
<td>Bottleneck capacity</td>
<td>58</td>
</tr>
<tr>
<td>5.2</td>
<td>The fundamental relationship</td>
<td>63</td>
</tr>
<tr>
<td>5.3</td>
<td>Summary</td>
<td>69</td>
</tr>
<tr>
<td>Chapter 6</td>
<td>Conclusions and future work</td>
<td>71</td>
</tr>
<tr>
<td>6.1</td>
<td>Conclusions</td>
<td>71</td>
</tr>
<tr>
<td>6.2</td>
<td>Limitations and Future works</td>
<td>74</td>
</tr>
<tr>
<td>References</td>
<td></td>
<td>79</td>
</tr>
<tr>
<td>Appendix A: Simulation results without the driver relaxation process</td>
<td>86</td>
<td></td>
</tr>
</tbody>
</table>
Appendix B: Simulation results with 20 mph of wave speed ............................................. 88
Appendix C: Simulation results using different car following models ............................... 94
Appendix D: Implementation of Laval and Leclercq's relaxation model ......................... 98
List of Tables

Table 1: Combined demand from mainline and ramp flow ........................................ 34
List of Figures

Figure 1: Newell’s simplified car following model (a) the underlying triangular flow-density relationship, (b) speed-spacing relationship associated with b, (c) example of car following vehicle trajectories. (adapted from Newell, 2002) ............................................. 22

Figure 2: Relaxation from an entering vehicle (vehicle j) and a vehicle (vehicle i+1) immediately behind, as shown (a) in the time-space plane, and in detail (b) at $t_0$, (c) at $t_1$, and (d) at $t_2$. ................................................................................................................................. 26

Figure 3: Car following after relaxation process (a) predicted trajectories after relaxation process (b) the assumed underlying triangular $q_kFR$. ................................................................. 29

Figure 4: Two-step processes to calculate vehicle trajectories when catching-up the queue for vehicle i+2, (a) step 1- derive $\{i+2\}$ from car following and $[i+2]$ for unqueue, (b) step 2- upon catching-up at $t_j$, take partial vehicle trajectories from $[i+2]$ and $\{i+2\}$ to find the final trajectory i+2. ............................................................................................................. 32

Figure 5: (a) Time-space plot of mainline speed (mph) with constant headway ramp entrances with mainline demand = 2,080 vph and ramp demand = 360 vph. Diagonal line shows last vehicle past before ramp turns on, the collection of points show the u-end and d-end. Along with the corresponding (b) flow (vph), (c) calculated density (vpm), and (d)
rescaled cumulative arrivals after the ramp turns on (zero values omitted for clarity). The dark area in (d) shows the region with speeds below 50 mph and the jagged solid line is the boundary between the loading and settling periods.

Figure 6: Time-space plots of mainline speed (mph) with constant headway ramp entrances, corresponding to the 9 demand combinations from Table 1. The mainline demand in row 1 is 1,960 vph, in row 2 is 2,080 vph, and in row 3 is at RCap, 2,200 vph. The ramp demand increases from left to right. Traffic flows from bottom to top, with the ramp at mile zero. The ramp turns on at t=0, after the mainline has loaded.

Figure 7: (a) Time-space plot of mainline speed (mph) with NES ramp entrances with mainline demand = 2,080 vph and ramp demand = 360 vph. Diagonal line shows last vehicle past before ramp turns on, the collection of points show the u-end and d-end. Along with the corresponding (b) flow (vph), (c) calculated density (vpm), and (d) rescaled cumulative arrivals after the ramp turns on (zero values omitted for clarity). The dark area in (d) shows the region with speeds below 50 mph and the jagged solid line is the boundary between the loading and settling periods.

Figure 8: Time-space plots of mainline speed (mph) with NES ramp entrances, corresponding to the 9 demand combinations in Table 1. Compare to Figure 6.

Figure 9: d-end and u-end associated with Figure 8b and d over an extended period.

Figure 10: Figure 8f re-plotted over a larger time range.
Figure 11: u-ends and d-ends from three dv values (3 mph, 1 mph, -1 mph) simulated from (a) constant headway ramp arrivals, (b) NES ramp arrivals. Mainline and ramp demands are 2,080 vph and 360 vph respectively and dcc is 2 ft/sec². ........................................ 49

Figure 12: u-ends and d-ends from three dcc values (1 ft/sec², 2 ft/sec², 3 ft/sec²) simulated from (a) constant headway ramp arrivals, (b) NES ramp arrivals. Mainline and ramp demands are 2,080 vph and 360 vph respectively and dv is 1 mph......................... 50

Figure 13: (a) u-ends and d-ends from two wave speed values (20 mph, 12 mph) simulated from deterministic arrivals, (b) u-end and d-end from two wave speed values (20 mph, 12 mph) simulated from NES ramp arrivals. Mainline and ramp demands are 2,080 vph and 360 vph respectively. ............................................................... 51

Figure 14: Time-space plots of mainline speed (mph) with constant headway (a-c) and NES (d-f) ramp entrances, corresponding to the middle row Table 1 using the model from Laval and Leclercq (2008). ................................................................. 53

Figure 15: Contour plot of flow corresponding to Figure 14c and f (compare to Figure 5b and Figure 7b). ................................................................. 54

Figure 16: Time series detector measurements at three locations downstream of the on-ramp (by column). The first row shows the cumulative arrival curve after subtracting a background flow equal to RCap, the second row shows q, and the bottom row shows v,
all with a common time axis. Each plot shows one curve for the constant headway arrivals and another for the NES ramp arrivals. .......................................................... 59

Figure 17: q-k relationships at three locations downstream of the on-ramp with constant headway arrivals, by row, when the on-ramp has constant headway arrivals (a)-(c): 30 sec sampling period; (d)-(f): 60 sec sampling period. The underlying triangular qkFR is shown with dashed lines in each plot................................. 65

Figure 18: q-k relationships at three locations downstream of the on-ramp with NES ramp arrivals, by row, when the on-ramp has constant headway arrivals (a)-(c): 30 sec sampling period; (d)-(f): 60 sec sampling period. The underlying triangular qkFR is shown with dashed lines in each plot.......................................................... 67

Figure 19: Repeating the scenario from Figure 5a-c using Newell’s car following model without the driver relaxation process......................................................... 87

Figure 20: Figure 5 with 20mph wave speed.................................................... 89

Figure 21: Figure 6 with 20mph wave speed.................................................... 90

Figure 22: Figure 7 with 20mph wave speed.................................................... 91

Figure 23: Figure 8 with 20mph wave speed.................................................... 92

Figure 24: Figure 9 with 20mph wave speed.................................................... 92
Figure 25: Figure 10 with 20mph wave speed............................................................... 93

Figure 26: Repeating the scenario from Figure 5a-c using the car following model
developed by Gazis et al (1959) and including our driver relaxation process from Section
3.3........................................................................................................................................... 95

Figure 27: Repeating the scenario from Figure 5a-c using the car following model
developed by Ozaki (1993) and including our driver relaxation process from Section 3.3.
............................................................................................................................................... 96

Figure 28: u-end and d-end associated with Figure 20 over an extended period ............ 97
Chapter 1    Introduction

Virtually all freeway delay arises from a bottleneck of one form or another. Many different freeway features can cause bottlenecks, e.g., grades (Newell, 1998; Laval, 2006), lane drops (Banks, 1991, Bertini and Leal, 2005; Coifman and Kim, 2011), merge areas (Banks, 1991, Persaud et al, 1998; Cassidy and Bertini, 1999; Bertini and Cassidy, 2002; Bertini and Malik, 2004; Cassidy and Rudjanakanoknad, 2005), and weaving sections (Banks, 1989, 1990, and 1991). No matter what the source of the bottleneck may be, the bottleneck is said to be active when it is restricting flow, i.e., demand exceeding capacity, yielding distinct traffic conditions upstream and downstream of the active bottleneck.

Most contemporary bottleneck studies employ the point bottleneck model, wherein the bottleneck process is assumed to occur over a negligible distance along the roadway (Daganzo, 1997; Zhang and Levinson, 2004). In this case an active bottleneck is defined as a point on the network with queuing upstream and unqueued conditions downstream (see, e.g., Bertini and Leal, 2005). Typically the bottleneck capacity (BCap) is defined as the highest sustained throughput, and in most empirical bottleneck studies BCap is observed immediately prior to activation. Many researchers have observed a capacity drop where discharge flow drops immediately after the bottleneck becomes active (e.g., Banks, 1990; Hall and Agyemang-Duah, 1991; Persaud et al, 1998; Cassidy

---

1 Ultimately we will show that this definition is too vague and we would rephrase it to be that an active bottleneck is defined as a section of the network with queuing upstream and strictly unqueued conditions downstream.
and Bertini, 1999; Zhang and Levinson, 2004; Chung et al, 2007; Duret et al, 2010; Leclercq et al, 2011). Several studies have stressed the importance of measuring BCap downstream of the bottleneck to prevent inclusion of demand in excess of BCap observed upstream of the growing queue and doing so without any intervening ramps to ensure that the entire throughput is measured (e.g., Hurdle and Datta, 1983; Hall and Agyemang-Duah, 1991, Cassidy and Bertini, 1999).

There is growing evidence to suggest that the point bottleneck model does not accurately capture the entire bottleneck process. For example, Cassidy and Bertini (1999) and Daamen et al (2010) observed empirically that the initial queue forms about 1 km downstream of the merge point at a merge bottleneck, suggesting a spatial component not included in the point bottleneck model. The prevailing use of the point bottleneck model persists in part because typical noise in empirical traffic data has precluded defining a precise boundary between queued and unqueued traffic states.

Ultimately, empirical bottleneck studies are encumbered with the difficult challenge of simultaneously measuring BCap, identifying the instant that the bottleneck becomes active (i.e., starts restricting flow), and establishing where the bottleneck actually forms. There are several different techniques commonly used to determine when a bottleneck is active:

[1] Some studies look for a speed drop upstream of the bottleneck, indicative of queuing (e.g., Banks, 1990; Hall and Agyemang-Duah, 1991).
Some look for a positive correlation between flow and occupancy, which is indicative of the unqueued regime (sometimes called the free flow regime) of the flow-occupancy relationship (e.g., Hall and Agyemang-Duah, 1991). Both [1] and [2] have latency, requiring the queue to grow back to the detection location before the queuing can be detected.

Cassidy and Bertini, (1999) used rescaled cumulative arrival curves to construct a queuing diagram and measure accumulation between detector stations (thus identifying any queuing that might not reach the stations) and verified that the locally observed conditions at the stations were consistent with [1] and [2].

This dissertation examines a merge bottleneck to provide better insight into bottleneck mechanisms. The analysis incorporates microscopic driving behavior with macroscopic traffic flow theory to show that an on-ramp bottleneck's activation may occur several minutes earlier than conventional bottleneck models would detect, and that unsustainably high flows after the true activation time could easily be mistaken for BCap, leading to an overestimate of capacity. In the present case these discrepancies arise due to driver relaxation, whereby a driver will accept a short headway for some time (often 20 sec or more, e.g., Smith, 1985) so that they can enter a lane that is constrained by downstream conditions and then will gradually increase the short headway to their preferred headway (e.g., Newman, 1963; Cohen, 2004; Wang and Coifman, 2008; Leclercq et al, 2007; Xuan and Coifman, 2012). Likewise, the driver immediately behind an entrance will go through the relaxation process in response to their newly shortened
headway. Average headway is the reciprocal of flow, q, so as the drivers undergoing the relaxation process travel downstream they gradually increase their headway, and thus, the corresponding q should drop\(^2\).

Most bottleneck studies do not account for driver relaxation, and this dissertation seeks to demonstrate that driver relaxation is an important factor that can confound the results of empirical studies if it is not accounted for. We investigate the systematic impact of driver relaxation at an on-ramp bottleneck and see how it affects: bottleneck activation, queue formation, and traffic states in the vicinity of the active bottleneck. Based on the findings, we argue that if mainline demand is close to BCap and drivers are perpetually entering the freeway from an on-ramp, then the maximum sustainable throughput should drop as a function of distance downstream of the on-ramp due to driver relaxation. Thus, the demand will first exceed the diminishing sustainable throughput at some location downstream of the on-ramp, and the queue first forms at this instant and location.

Although throughput becomes more constrained as drivers relax (or gradually increase their headway), traffic downstream of the on-ramp should be traveling at or near free speed\(^3\), \(v_f\), even after this relaxation starts to limit throughput. The simulations presented herein show that this relaxation process can extend at least 1.5 mi downstream of the on-ramp, much further beyond the ramp than most empirical studies contemplate. The initial

---

\(^2\) By extension, if throughout this segment speed remains close to free speed, \(v_f\), the fundamental equation, \(q=kv\), dictates that decreasing \(q\) should correspond to decreasing density.

\(^3\) The "free speed" is often called "free flow speed" in the literature, in direct reference to the "free flow regime" of the fundamental relationship, FR. We avoid the "free flow" terminology in this document due to several ambiguities, including, (1) this work uses "unqueued" instead of "free flow" to refer to the left hand regime of the FR in part to emphasize that the queued regime arises strictly within a queue, (2) in the case of a parabolic FR there is a range of speeds seen in the unqueued (or free flow) regime, so the "free speed" makes clear that we are specifying the truly unconstrained speed, and (3) the term, "free flow speed" is poorly worded since it is referring strictly to a speed and not a flow.
period of activation is characterized by very minor accumulations downstream of the on-ramp that are below the sensitivity of [1]-[3]. As vehicles accumulate in this downstream section the queue slowly grows upstream, eventually reaching the on-ramp. Shortly thereafter, the downstream accumulations dissipate, and the queue moves almost entirely upstream of the on-ramp. A detailed discussion of these impacts will be presented in Chapter 4. Needless to say, this view implicitly assumes that the on-ramp bottleneck process occurs over an extended distance and should not be modeled as a single point bottleneck.

Although there is a general need to account for driver relaxation, due to the inherent limitations of conventional traffic data, only a few empirical studies have explicitly considered driver relaxation at on-ramps. None of the prior work found in the literature review (Chapter 2) considered the relaxation process in the context of bottleneck process. Cohen (2004) employed relaxation phenomenon to provide better prediction of vehicle trajectories. Leclercq et al (2007) investigated the evolution of traffic state within a queue. Daamen at al (2010) found evidence of driver relaxation at an on-ramp bottleneck, but only undertook a detailed study of the vehicles in the merge area while the relaxation process extended beyond the downstream end of their study segments.

While this study is based on a simulation study, the major findings provide clues to better interpret past empirical observations that have not been fully explained, yielding new insights into merge bottleneck mechanisms, capacity, and the fundamental relationship (FR). The three major contributions of this dissertation are: (1) Demonstrate
that once an on-ramp bottleneck becomes active the queue on the mainline initially forms downstream of the on-ramp due to driver relaxation process, consistent with the empirical results of Cassiday and Bertini (1999). (2) Demonstrate that the transient, unsustainable high flow downstream of the on-ramp, after the true activation, is likely to be mistaken for sustainable capacity flow prior to activation in conventional studies and then investigate the implications in the context of so-called capacity drop. (3) Demonstrate how the relaxation model can perturb an empirically observed FR. The present work should not be viewed as a complete model of the very complicated bottleneck process. Rather, these results are intended to highlight the impacts of what we believe to be an important factor that has previously gone largely overlooked. Ultimately the findings emphasize the need for microscopic empirical data collected at the appropriate locations. It is hoped that the work developed herein will help shape future empirical studies to test the validity of the new theories and quantify the impacts from the previously overlooked phenomena.

The remainder of this dissertation is organized as follows. Chapter 2 reviews previous studies of freeway bottleneck operation as well as the relevant studies of car following and driver relaxation. Chapter 3 presents the details of the car following and relaxation models used in this study to simulate vehicle trajectories. Chapter 4 uses simulation to investigate the systematic impact (i.e., on bottleneck activation and queue formation) of driver relaxation at an on-ramp bottleneck on a one-lane freeway under various demands. Chapter 5 discusses the implications of the analysis from Chapter 4 in the context of capacity and the fundamental relationship. The dissertation closes in
Chapter 6 with the conclusions and a discussion of future research directions branching off of this work.
Chapter 2  Literature review

This dissertation examines driver relaxation to provide better interpretation on past empirical observations that have not been fully explained, yielding new insights into merge bottleneck mechanisms: bottleneck activation and queue formation, capacity, and the FR. Section 2.1 provides a detailed review of the relevant literature on bottlenecks in general and the merge bottleneck process in particular. Section 2.2 discusses driving behaviors with two focal points, first, behavior near the merge section, and second, the driver relaxation process in general.

2.1 Freeway merge bottleneck studies

Merge bottlenecks are the focus of this dissertation. The merge bottleneck process has been extensively studied in the context of bottleneck activation, capacity drop, the FR, and other aspects. Several important points in this area have already been discussed in Chapter 1, such as the techniques commonly used to determine if a bottleneck is active ([1]-[3] in Chapter 1), and the widespread use of the point bottleneck model. As noted earlier, there is emerging empirical evidence showing that the bottleneck process spans a much larger distance along the roadway than previously thought, e.g., both Cassidy and Bertini (1999) and Daamen et al. (2010) both found the initial queue formed approximately 1 km downstream of the merge point in different merge bottlenecks. The
point bottleneck model does a poor job capturing the spatial dynamics over these extended distances.

In fact bottleneck activation is closely related with another phenomenon, the so-called capacity drop observed by many researchers; they have empirically observed a capacity drop where discharge flow drops immediately after the bottleneck becomes active. This interpretation has been widely accepted. If it is true that capacity drops immediately after a bottleneck becomes active, then an operating agency should strive to prevent congestion from forming in the first place to maintain the highest possible throughput. There have been numerous studies that have empirically observed capacity drop at different locations, the magnitude of the reported drop in flow from the observed maximum throughput varies from 1% to 18% (Banks, 1991; Cassidy and Bertini, 1999; Hall and Agyemang-Duah, 1991; Hall and Hall, 1990; Persaud and Hurdle, 1991; Zhang and Levinson, 2004, Bertini and Malik, 2004, Chung et al, 2007). Various traffic controls have been developed to limit demand in an effort to prevent capacity drop, most notably ramp metering (e.g., Pinnel et al, 1967; Athol and Bullen, 1973; Banks, 1991; Persaud et al, 1998; Cassidy and Rudjanakanoknad, 2005).

As noted in the Introduction (and will be discussed in greater detail in subsequent chapters), this dissertation research has yielded a new interpretation of the empirical data, suggesting the commonly observed capacity drop phenomena is actually the supersaturated flow dropping to capacity flow. If the capacity drop is at least in part actually a drop to capacity, then ramp metering is likely to reduce throughput below
sustainable capacity in an effort to prevent queuing. In which case, more vehicles could potentially be served if they are waiting in the queue.

Cassidy and Rudjanakanoknad (2005) and Chung et al (2007) have offered some of the most extensive diagnostics of the capacity drop phenomenon found in the literature. For the subject bottlenecks, they found that a surge in demand triggered an increased number of lane change maneuvers in the mainline. In the context of Coifman et al (2006), one could interpret an increase in lane change maneuvers as reducing the available mainline capacity. If we adopt the term apparent-point-bottleneck, APB, to specify the location where one would place the point bottleneck, Coifman and Kim (2011) argues that there are subtle features downstream of the APB due to entering vehicles (both to lane change maneuvers and on-ramps) that also serve to limit throughput. Although entering vehicles consume capacity that would otherwise be available at the APB, an on-ramp consumes the capacity it diverts and can potentially replenish the demand lost to lane change maneuvers occurring between the APB and the ramp (e.g., Coifman and Kim found that throughput actually increased up to a mile downstream of an APB due to vehicles entering from on-ramps).

Finally, ignoring the spatial issues with the point bottleneck model, another problem arises due to ambiguities in the definition. An active point bottleneck is commonly defined to be a point on the network with queuing upstream and strictly unqueued conditions downstream (see, e.g., Bertini and Leal, 2005). Unfortunately, the

---

4 For example, if the surge in flow right before activation is unsustainable, then ramp metering could inadvertently pull the available demand below BCap from Chapter 1. In the strictest definition we should not even say “sustainable capacity” but instead, “BCap” here. We use “sustainable” only to make a clear distinction in case one overestimated BCap by measuring it during the loading period rather than “prior to true activation” or “downstream of all queuing activity”.

10
boundary between queued and unqueued conditions is vague and not well defined. As will be discussed in Chapter 4, it is possible that during the first few minutes of queue formation after the bottleneck becomes active, the lightly-queued traffic state may erroneously be taken to be unqueued conditions and the active bottleneck would go undetected until queuing grew more pronounced.

2.2 Merging behavior

This research simulates a merge bottleneck to understand the details of how the bottleneck formation. The simulation partly depends on the microscopic behavior of the individual drivers, particularly in the context of the merging behavior. Merging behavior has been extensively studied, though the majority of that work has been in the contest of other applications. Most of the studies on merging behavior studies have focused on how entering vehicles accelerate on the on-ramp (e.g., Wattleworth et al, 1967) and under what circumstance the driver decides to actually merge (e.g., Kou and Machemehl, 1997). Merging models are also used to measure the merging capacity (e.g., Worrall et al, 1969; Worrall and Bullen, 1970; Rorbech and Bullen, 1976), optimize access control (e.g., Beegala et al, 2005), design geometric configuration (e.g., Drew, 1967), and calculate the level of service (LOS) of the merge section (e.g., Drew et al, 1968). This literature is reviewed in Section 2.2.1.

Nonetheless, the merging behavior will impact throughput and capacity, and this impact has been largely overlooked in the traffic flow theory literature, with only a few studies explicitly considering the impacts it in the context of the bottleneck mechanism. In fact most of the cutting edge traffic flow theory literature still employs the point
bottleneck model with the implicit assumption that the merging behavior occurs instantaneously. Only a few studies actually consider the driving behavior after a merge, i.e., the relaxation process. Section 2.2.2 reviews the literature related to the relaxation process.

2.2.1 Previous studies in merging sections

One of the simplest merging sections is where mainline and on-ramp flows come together. When demand is high the two streams compete for the available capacity. From the freeway operation point of view it is important to understand merging behavior and characteristics for traffic management applications such as access control that seeks to smooth the mainline traffic stream. Sarvi et al (2002) considered the merging behavior from the perspective of the merging driver, it consists of three stages: (1) acceleration-deceleration stage at the on-ramp, (2) gap acceptance stage, and (3) car-following after the merge. Most of the literature on the merging area has focused on first two stages.

In the 1960's and 1970's there were several field studies to investigate merging behavior. Worrall et al (1967) discussed empirically observed gap acceptance and gap rejection by drivers. Wattleworth et al (1967) used photographic-based data collection to investigate various merging characteristics such as the speed of the merging vehicles at the ramp nose and at the merge point, acceleration in the on-ramp, and the distribution of the merging point. Taylor and Carter (1970) also used photographic based data collection to study gap acceptance and rejection characteristics to find, “the most psychologically effective,” gap size. Drew (1967) focused on gap acceptance characteristics for stopped vehicles who chose a gap after rejecting several gaps, and contrasted their performance
against that of *moving vehicles* who chose the first gap without stopping because there was a sufficient gap on the mainline as they approached. Drew compared the distributions of the critical gap between the stopped vehicles and the moving vehicles. Drew argued that gap acceptance behavior is dependent upon, “ramp geometrics and ramp controls.” These early studies often focus on the characteristics of the merging vehicle such as gap acceptance and speed in the acceleration lane rather than the impact on the mainline.

More recently, researchers have started to employ mathematical models to account for vehicle dynamics on the on-ramp. For example, Kou and Machemehl (1997) developed a model for vehicle acceleration-deceleration with consideration of variability in space in the merging location. They investigated whether merging positions are correlated with the time gap or the relative speed between merging and mainline vehicle. Sarvi et al (2002) developed another model that incorporates acceleration and deceleration for drivers on the ramp by modifying a stimulus-response based car-following models (e.g., Herman and Rothery, 1963). Since the study focused on a merging section under congested conditions, every entrance is forced to merge by the end of the merge section. The modified model was applied in the acceleration lane until the merging maneuver is complete and then the original, unmodified stimulus-response based car-following model was applied in the mainline. In their model the entering vehicle adjusted to the new lane instantaneously, without any relaxation process.

There are also a few studies that consider cooperation between mainline traffic and merging traffic in an effort to model the merging behavior in more realistic way. Another approach uses *cooperation models* that consider the merging vehicle’s behavior
in the acceleration lane. These models are like the *acceleration-deceleration models* from Kou and Machemehl (1997) and Sarvi et al (2002); however, the major distinction is the fact that the cooperation models also consider the reaction of the mainline traffic. Kita (1999) examines the interaction between the mainline vehicles in the outside lane and merging vehicles from the ramp and model the cooperation using a game theory. Wang and Mahmassani (2005) developed a cooperation model that implements the cooperative interaction. They employed conventional gap selection models, acceleration models, and gap acceptance models for merging behavior. However, they also used a cooperation model applied to both the mainline and merging vehicles to account for speed and spacing adjustment between the two groups. At the end of the framework, a merging vehicle either successfully enters the mainline and starts a car following mode or fails to enter because there was not a sufficient gap to merge. In the latter situation, the merging vehicle waits for the next available gaps to merge. Choudhury et al (2007) developed a similar model that accounts for cooperation between merging and mainline vehicles but they further considered a forced merging when a merging vehicle fails to find appropriate gap in the mainline. However, as noted by Laval and Leclercq (2008), applying a car following model at the end of merging would “predict speeds far below empirical observations,” (the overreaction effect) because the car following rule could compel a sudden deceleration by the mainline vehicles in order to adjust the truncated spacing created by the merge. Of course, whatever overreaction effect induced by cooperation is far less than the conventional practice of abruptly applying a car following rule at the end of the merge. In a cooperative merge the mainline traffic adjusts their speed over time
and space to create an extra large gap to accommodate a merging vehicle ahead. However, there is still the overreaction effect, e.g., Daamen et al (2010), found that after most merging maneuvers the spacing (or time headway) ahead of the merging vehicle is not large enough that they can comfortably begin car following right away. This finding applied whether it was a normal merge, cooperative merge, or forced merge and it applied not only to the merging vehicle but also to the vehicle immediately behind the merge. In fact the empirical evidence from Daamen et al suggests that there is a relaxation process whereby the merging vehicle and a vehicle behind it tolerate a truncated spacing for a little while after a merge and then gradually increase the short spacing to reach their preferred speed-spacing relationship. This relaxation process will be discussed in Section 2.2.2.

2.2.2 Relaxation process

When mainline traffic is near capacity or queued, a vehicle can only enter the lane by accepting a short gap, i.e., accepting a headway that is shorter than the driver would normally sustain at the given speed. Typically the driver of the entering vehicle gradually increases the short headway over a period of time (often 20 sec or more, e.g., Smith, 1985) to their preferred headway (e.g., Newman, 1963; Cohen, 2004; Wang and Coifman, 2008; Leclercq et al, 2007; Xuan and Coifman, 2012). Obviously this accommodation also occurs over some distance. Likewise, the driver immediately behind an entrance will go through the relaxation process in response to their newly shortened headway. Although this general phenomenon is both intuitive and widely accepted, most bottleneck studies do not account for driver relaxation (e.g., as evidenced by the continued reliance on
simple point bottleneck models that lack a spatial component to the bottleneck mechanism).

There are several studies that observed driver relaxation empirically based on individual vehicle trajectories. Both Newman (1963) and Smith (1985) used time series photos to measure relationships between vehicle trajectories. Both studies found evidence of the driver relaxation process, e.g., Newman shows some merging vehicles taking more than 30 sec for the relaxation process. Smith found that the relaxation process often takes over 20 sec. Skabardonis (2002) asserted that various simulation software under-predicted speed and flow in weaving and lane drop segments because of the overreaction effect in the car-following models (as discussed above, in Section 2.2.1). To this end, Cohen (2004) demonstrated that applying different sensitivity values into the existing FRESIM model can account for the relaxation process and can yield better consistency with field data than the procedure without the relaxation process. Leclercq et al (2007) studied an on-ramp that was subject to queuing from a bottleneck far downstream and found impacts from driver relaxation on the spatial evolution of traffic states (Laval and Leclercq, 2008, subsequently developed a model of these observations). Daamen at al (2010) also found evidence of driver relaxation at an on-ramp bottleneck by looking at distribution of difference in time headway between two locations downstream of the merge and speculated that the initial queue formation downstream of the merge could be due to the relaxation process.
Unlike the earlier studies, this dissertation developed a relaxation process to investigate the impacts of the driver relaxation on bottleneck process (i.e., bottleneck activation, capacity, FR) in the vicinity of a merging bottleneck rather than within a queue. As noted above, Laval and Leclercq (2008) developed a model incorporating driver relaxation for vehicles that enter a queue far upstream of the actual bottleneck, but they never contemplated the impacts on bottleneck process, capacity, or empirical FR. In the present work we develop a model for vehicles that enter the traffic stream at the bottleneck itself. Our model has parameters that account for variability of the relaxation process associated with initial speed of vehicle from the on-ramp and its acceleration, which makes the model distinct from the model developed by Laval and Leclercq (2008). However, we also implement Laval and Leclercq’s model and find consistent macroscopic trends, though their model induces some unusual microscopic effects, e.g., their model triggers the relaxation process at the end of a merge no matter what the merging vehicle’s speed is.
Chapter 3  Methodology

This dissertation incorporates a microscopic behavior, driver relaxation, with macroscopic traffic flow theory, FR, to derive emergent macroscopic behavior and provide insight into empirically observed macroscopic phenomena. This chapter provides the background of the microscopic models that will be used in subsequent chapters to study the macroscopic behavior.

The chapter begins in Section 3.1 by reviewing terminology necessary for the subsequent discussion. The simulation model is based upon a commonly used car following model with the speed-spacing relationship derived from an underlying FR, as discussed in Section 3.2. However, this work goes beyond the conventional car following model by also including driver relaxation for affected vehicles (in this case those vehicle entering from a ramp and the mainline vehicle immediately behind an entering vehicle), as discussed in Section 3.3. Finally, Section 3.4 discusses the integration of the two behaviors to calculate the resulting vehicle trajectories for all vehicles.

3.1 Terminology

Before proceeding, it is important to define several key terms. The macroscopic traffic state (flow, q, density, k, and space mean speed, v) is commonly assumed to fall on some fundamental relationship, FR, that may vary over time and space. However, perturbations can cause the traffic state to deviate from the underlying FR, e.g., the shock
due to the arrival of a queue from downstream. The FR is commonly characterized in terms of a bivariate relationship between two of the three parameters. In our discussion, we will refer to the flow-density curve, qkFR, one of the three commonly used bivariate realizations of the FR. We assume a triangular qkFR (e.g., as found in Munjal and Pipes, 1971; Hall et al, 1986; Banks, 1989), which has several key parameters: the wave speed, w, corresponding to the slope of the queued regime (i.e., the right hand side of the qkFR), the free speed, \( v_f \), corresponding to the slope of the unqueued regime (i.e., the left hand side of the qkFR), and capacity. Unfortunately, capacity means several different things in the context of freeway flow. On the one hand, there is the maximum throughput that an infinitesimally short segment of road can accommodate if provided sufficient demand from upstream and no queuing downstream. This parameter corresponds to q at the apex of the qkFR. We call this parameter the *roadway capacity*, RCap, since it characterizes the particular point along the roadway (in contrast to the emergent BCap that arises from a series of points along the roadway). Although RCap exists at all locations, at most locations one should rarely see q that high (see, e.g., Figure 6 in Hall et al, 1992). On the other hand, in Chapter 1 we spoke strictly of BCap, the maximum sustainable throughput past a bottleneck. In a point bottleneck model BCap would typically be the smallest RCap over many successive infinitesimally short segments, and this minimum RCap would occur at the assumed point bottleneck location. In an extended bottleneck model we use BCap as shorthand to capture all of the factors that contribute to the bottleneck capacity, (e.g., lane change maneuvers, inhomogeneous vehicles, etc.).

---

5 In each case the third parameter can be calculated from the fundamental equation.
3.2 Car following models

There has been extensive research into car-following behavior, leading to many different models of this microscopic behavior. One of the earliest car-following models was proposed by Herrey and Herrey (1945) and was developed such that a driver keeps the minimum safe distance. In the 1950's and 60's researchers at the General Motors Research Laboratories developed a series of car following models (e.g., Chandler et al., 1958; Herman et al., 1958; Herman and Potts, 1959; Gazis et al., 1959; Gazis et al., 1961) that have become known collectively as the General Motors car following model, or GM model for short (Charkroborty and Kikuchi, 1999). The GM model is based on a stimulus-response framework where a following vehicle’s response (e.g., acceleration) depends on the stimulus (e.g., relative speed) from its leader. A generalized form of the GM model is presented in Equation 1.

\[
a_n(t + \Delta t) = c \frac{v_n(t + \Delta t)^m}{[x_{n-1}(t) - x_n(t)]^l}[v_{n-1}(t) - v_n(t)]
\]

[Eq. 1]

where,

\(a_n(t)\): acceleration/deceleration of the vehicle \(n\) at time \(t\)

\(v_n(t)\): speed of the vehicle \(n\) at time \(t\)

\(x_n(t)\): distance of the vehicle \(n\) at time \(t\)

\(\Delta t\): reaction time

\(c, m,\) and \(l\): parameters

In Equation 1, \(a_n(t + dt)\) is the response, \(v_{n-1}(t) - v_n(t)\) is the stimulus, and the rest of the right hand side of the equation is the sensitivity, which is a function of follower’s speed (nominator) and spacing (denominator). Integrating Equation 1 with respect to time,
t, yields a function between speed and spacing. Thus, the implicit assumption of the GM model is that the driver attempts to maintain their vehicle operation on some underlying speed-spacing relationship. Whenever the lead vehicle's behavior causes the vehicle operation to diverge from the speed-spacing relationship, after some reaction time the following driver will then change speed in an attempt to bring the operating point back to the speed-spacing relationship.

The GM model remains widely used today and has been thoroughly calibrated (i.e., \( c, m, l \) in Equation 1) based on field data by many researchers (e.g., May and Keller, 1967; Heyes and Ashworth, 1972; Ceder and May, 1976; Aron, 1988; Ozaki, 1993). In fact many subsequent car-following models turn out to be a subset of the GM model. One recent example is the car-following model developed by Newell in 2002.

Newell (2002) began with the fact that when the macroscopic flow-density relationship is assumed to be triangular (Figure 1a), the corresponding speed-spacing relationship is piecewise linear (Figure 1b). Whenever speed is below \( v_f \), in this model a driver is in car following mode. If the spacing is \( S_{crit} = v_f / R_{Cap} \) or larger, the driver travels at \( v_f \). Otherwise at lower speeds, to predict the follower’s vehicle trajectory Newell’s car following model essentially assumes that the follower replicates the leader’s trajectory, shifted by \( \frac{1}{w \times k_{jam}} \) in time and \( \frac{1}{k_{jam}} \) in space, as illustrated in Figure 1c. The parameters, \( w \) and \( k_{jam} \) come from some assumed triangular \( qkFR \), as shown in Figure 1a.
In fact, Newell’s model is a special case of the linear form of the GM model from Chandler et al (1958).⁶

![Diagram](image)

Figure 1: Newell’s simplified car following model (a) the underlying triangular flow-density relationship, (b) speed-spacing relationship associated with b, (c) example of car following vehicle trajectories. (adapted from Newell, 2002)

---

⁶ Note that since w is slope of the queued regime in a triangular qkFR, Newell’s car following model follows directly from the Lighthill, Whitham, and Richards’ macroscopic traffic flow theory (Lighthill and Whitham, 1955; Richards, 1956), with a triangular qkFR.
Newell's car-following model has been employed in several studies (e.g., Mauch and Cassidy, 2002; Brockfeld et al, 2004; Daganzo, 2005; Ossen et al, 2006; Duret et al, 2008; Wang and Coifman, 2008, Wang et al, 2011). Although the model has few parameters, it has proven to be very robust. For example, Ahn et al (2004) empirically showed a leader’s vehicle trajectory is generally the same as its follower “except for a translation in time and space”, while Coifman (2002) used the same shifting technique to accurately estimate vehicle trajectories and travel times over extended links within a queue. Given both the simplicity and robustness of Newell's car following model, this dissertation employs Newell’s model whenever car following occurs in the simulation (i.e., at a microscopic level during queued conditions) unless otherwise noted. Although most of the results presented herein use Newell's model, this research also examined other car following models that are subsets of the GM model to make sure the general results are not an artifact of the specific car following model used. The additional experiments will be discussed in Section 4.4.3.

3.3 Relaxation model

As noted in Chapter 1, when demand approaches or exceeds the available capacity, entering drivers will not be able to find gaps large enough to enter the freeway in a steady state. Instead, the empirical evidence shows that these drivers will accept a short headway for tens of seconds while they enter a lane that is constrained by downstream conditions and then will gradually increase the short headway to their preferred headway. In most cases the entering driver will also put the following driver in a similar non-steady state, and the following driver will also undertake the relaxation
process over many seconds too. As noted previously, both the entering and following vehicles are considered to be affected by the relaxation process. The impacts of this relaxation process, particularly in the vicinity of merge bottlenecks, have largely been overlooked in the traffic flow theory literature, and even in the car following literature it has not received sufficient attention, as discussed in Section 2.2.2. Such study is hampered by the fact that there are clearly a large number of variables in the process (e.g., due to driver behavior or vehicle performance).

Conventional traffic flow theory largely ignores the impacts of the driver relaxation process. Our work only seeks to achieve a "good" first-order approximation in order to demonstrate the importance of accounting for driver relaxation in traffic flow theory. Obviously there is a need to eventually develop more robust models of the driver relaxation process; however, such a detailed development is impeded by a lack of empirical data and is beyond the scope of the present work.

This section presents the relaxation model developed in this study and its implementation. The foundation of our model comes from the empirical observations of Smith (1985) and Cohen (2004), who observed that the affected vehicles tend to keep their speed over an extended time. For example, Cohen observed that, “neither the lane changer nor the follower underwent any significant reduction in speed. This observation was independent of how short the time spacings were between lane changer and its leader and between follower and lane changer.” In other words, a follower undergoing relaxation process is insensitive to a short spacing ahead and will systematically deviate

---

7 In this dissertation the "relaxation process," refers to the entire adjustment period: from the time of the initial truncated headway to the instant of first achieving the desired headway.
from a conventional car following model that is strictly contingent on a predetermined speed-spacing relationship.

In our first-order approximation the affected vehicles will tolerate a truncated spacing and maintain their current speed provided their relative speeds are less than the vehicle speed of the current leader (thus, the relative spacing is increasing), otherwise the follower will decelerate until achieving this status. The relaxation process ceases for an affected driver when they first reach their preferred speed-spacing relationship, at which point they resume conventional car following. For each entrance maneuver, the relaxation process is modeled as if it only directly impacts the two affected vehicles. All non-affected vehicles are assumed to obey the steady state car following rules, with the following driver remaining on their preferred speed-spacing relationship throughout. Obviously whenever one of these non-affected drivers is following an affected vehicle, the non-affected driver will still be impacted indirectly, since the lead vehicle's trajectory includes impacts of driver relaxation that will be passed on to the follower through the conventional car following rules.

Figure 2 shows a hypothetical example on a one lane freeway with an on-ramp. The mainline vehicles are indexed sequentially from i-1 to i+1 while a vehicle entering from the ramp is labeled "x". In Figure 2a, vehicles i and i+1 are initially traveling at free speed, $v_f$, with spacing in front at time $t_0$ of $S_i(t_0)$ and $S_{i+1}(t_0)$, respectively, as shown in Figure 2b. These spacings are in excess of $S_{crit}$. Thus, neither driver is car following at $t_0$, and the resulting macroscopic traffic state is unqueued. At $t_1$, however, vehicle x enters the freeway from the on-ramp, immediately ahead of vehicle i+1. Both vehicle x and i+1,
are now below their preferred spacing for the given speed (determined from an underlying speed-spacing relationship, e.g., Figure 1b), as illustrated in Figure 2c. The drivers will change speed to correct their spacing and gradually approach their preferred speed-spacing relationship over time.

![Diagram](image)

Figure 2: Relaxation from an entering vehicle (vehicle j) and a vehicle (vehicle i+1) immediately behind, as shown (a) in the time-space plane, and in detail (b) at $t_0$, (c) at $t_1$, and (d) at $t_2$.

As noted above, whenever a following vehicle's spacing is shorter than preferred, in our study that vehicle will respond depending on how the relative spacing is changing over time. If the relative spacing is increasing because the lead vehicle is traveling faster, then the follower will maintain its current speed until achieving the preferred spacing for the speed and then will resume conventional car following. For example, vehicle x maintains a constant speed from $t_1$ to $t_2$ since as shown in Figure 2a, this vehicle's relative
spacing, \(S_x\), is increasing. In this case it turns out that vehicle \(x\) maintains this same speed until \(t_n\) as shown in Figure 2a. On the other hand, if the lead vehicle is traveling slower than the following vehicle, the follower will decelerate until they reach a speed such that the relative spacing is increasing, at which point they will then maintain that speed until they reach their preferred speed-spacing relationship. For example, starting at \(t_1\), vehicle \(i+1\) finds that the relative spacing ahead is decreasing over time, \(S_{i+1}(t_2) < S_{i+1}(t_1)\). At \(t_2\), the vehicle starts decelerating at a fixed rate, \(dcc, a_{i+1}(t_2) = -dcc\). If in the next time step the spacing continues to decrease, then the vehicle increases its deceleration rate by another \(dcc, a_{i+1}(t_3) = a_{i+1}(t_2) - dcc\). This deceleration is repeated at each time step until either the spacing starts increasing or the vehicle reaches its preferred speed-spacing relationship. In this example, the spacing of vehicle \(i+1\) starts to increase at \(t_m\). So the vehicle stops decelerating and stays at a constant speed, \(v_{i+1}(t_m)\), from \(t_m\), until reaching its preferred speed-spacing relationship at \(t_k\), and then the vehicle begins car following. Note that this deceleration process is based on an iterative process with a given time step. Thus, the way a vehicle decelerates depends on the magnitude of the time step as well as the various constants in the model. This model does not directly account for the driver reaction time, so it is likely that future research could incorporate this idea and improve the results further by modeling behavior in a more realistic way.

The portions of the trajectories subject to the relaxation process are shown with bold curves in Figure 2a. In this figure the trajectories are only shown to the point where the drivers cease the relaxation process, the subsequent car following behavior for
vehicles x and i+1 are not shown here but will be discussed shortly in the context of Figure 3.

Equation 2 presents the details of the process used to update the speed, position, and acceleration during the relaxation process for some vehicle k. This process is based on the classical equations of motion. Finally, note that in our simulations the vehicles enter from the on-ramp anticipating the fact that they are beginning the relaxation process and they enter the mainline at a speed that is a fixed amount, dv, slower than their new leader, with an initial acceleration of zero. So the case of Figure 2, \( v_k(t) = v_i(t) - dv \) and \( a_k(t) = 0 \).

\[
\begin{align*}
x_k(t + dt) &= x_k(t) + v_k(t)dt + \frac{1}{2}a_k(t)dt^2 \quad \text{[Eq. 2a]} \\
v_k(t + dt) &= v_k(t) + a_k(t)dt \quad \text{[Eq. 2b]} \\
a_k(t + dt) &= \begin{cases} 
0, & \text{if } S_k(t + dt) > S_k(t) \\
_{a_k(t)} - dcc, & \text{otherwise}
\end{cases} \quad \text{[Eq. 2c]}
\end{align*}
\]

where,

- \( x_k(t) \): the position of the \( k^{th} \) vehicle at time \( t \)
- \( v_k(t) \): the speed of the \( k^{th} \) vehicle at time \( t \)
- \( a_k(t) \): the acceleration of the \( k^{th} \) vehicle at time \( t \)
- \( S_k(t) \): the spacing from the \( k^{th} \) vehicle to its leader at time \( t \)
- \( dt \): time step of the simulation
- \( dcc \): unit rate of deceleration

After an affected vehicle finally achieves its preferred speed-spacing relationship the relaxation process ends, its status as an affected vehicle ceases, and in the simulation
the driver begins conventional car following, e.g., vehicle x at $t_n$ and vehicle $i+1$ at $t_k$ in Figure 2. As discussed in Section 3.2, this study employs Newell's lower order car following model to generate vehicle trajectories for all vehicles on the mainline, except those that are presently undertaking a relaxation process. According to the Newell's model, the following vehicle replicates the lead vehicle's trajectory, shifted in time and space. Thus the trajectories of vehicle x and $i+1$ after $t_n$ and $t_k$, respectively, in Figure 2 are generated by shifting their lead vehicle's trajectory by $\frac{1}{w \times k_{jam}}$ in time and $\frac{1}{k_{jam}}$ in space, as implemented in Figure 3a. As in Figure 2, the solid lines are trajectories that are not car following and those in bold are undertaking a relaxation process. Now, however, the dashed trajectories in Figure 3a come from the car following model.

Figure 3: Car following after relaxation process (a) predicted trajectories after relaxation process (b) the assumed underlying triangular $qkFR$. 

\[
\text{Distance} \quad \text{On-ramp} \quad \text{Distance} \quad \text{Flow} \\
\text{veh. $i+1$} \quad \text{veh. $i$} \quad \text{veh. $x$} \quad \text{RCap} \\
1/k_{jam} \quad 1/w \times k_{jam} \quad w \quad k_{jam} \quad \text{Density} \\
1/\text{veh. $i$} \quad \text{t}_n \quad \text{t}_k \\
\text{Trajectory undergoing car following} \quad \text{Trajectory undergoing relaxation}
\]
3.4 Mainline vehicles joining the queue

Although there are only a few affected vehicles that undergo the relaxation process, during queued conditions each affected vehicle defines a new "prototype" trajectory for all subsequent vehicles within the queue to follow. We simulate the mainline vehicles starting from a distance sufficiently upstream of the on-ramp to ensure that all of the mainline vehicles start in unqueued conditions, traveling at \( v_f \). Thus, the initial spacing is in excess of \( S_{crit} \), and instead of car following, these drivers initially choose their preferred speed. However, because the combined mainline and ramp demands exceed \( R_{Cap} \), most of these mainline vehicles will eventually reach a queue and catch up to their leader, traveling slower than \( v_f \). For example, in Figure 3a, for the period while \( v_{i+1}(t) < v_f \), there is a good chance that the next mainline vehicle upstream, \( i+2 \), will catch-up to vehicle \( i+1 \) and begin car following as soon as \( S_{k+2}(t) < S_{crit} \) (assuming vehicle \( i+2 \) is not already car following due to queued conditions). Although vehicle \( i+2 \) is never affected directly by the ramp vehicles and does not undertake a relaxation process, it too will be impacted by the entering vehicles because \( i+2 \) will have to move slower in response to the delays in its leader's trajectory (i.e., vehicle \( i+1 \)).

To account for these catching-up situations we employ a two-step process to calculate the mainline vehicle trajectories. Consider vehicle \( i+2 \) in Figure 4a. The first step starts by calculating two possible trajectories for vehicle \( i+2 \). The first trajectory is derived directly from the Newell’s car-following model assuming the driver is in car following, thus shifting trajectory \( i+1 \) by \( \frac{1}{w \times k_{jam}} \) in time and \( \frac{1}{k_{jam}} \) in space, denoted by “\( \{i+2\} \)” in the figure. The second possible trajectory is derived assuming vehicle \( i+2 \) can
travel at $v_i$, and is denoted by “[i+2]”. In general vehicle i+2 must follow the most restrictive of these two possible trajectories, i.e., the one that is furthest upstream. In our simulations there is no opportunity for a vehicle to leave the lane, so when a vehicle returns to $v_f$ after car following, it does so at $S_{crit}$. This state can still be captured by the car following model; so once a vehicle begins car following, our simulator will treat it as if it continues car following throughout the remainder of the simulation (unless another vehicle enters from the ramp immediately ahead of the given mainline vehicle, in which case the mainline driver will undergo a relaxation process, as per Section 3.3, and then will resume car following afterwards). In the hypothetical example shown in Figure 4a, vehicle i+2 catches up to its leader at $t_j$ and from then on the car following model becomes more restrictive. The final trajectory for vehicle i+2 is shown in Figure 4b with the portion prior to $t_j$ coming from [i+2] and after $t_j$ from {i+2}. Vehicle i+2 eventually returns to $v_f$, but it does so at $S_{crit}$, the threshold for car following.
Figure 4: Two-step processes to calculate vehicle trajectories when catching-up the queue for vehicle i+2, (a) step 1- derive \{i+2\} from car following and \([i+2]\) for unqueue, (b) step 2- upon catching-up at \(t_j\), take partial vehicle trajectories from \([i+2]\) and \{i+2\} to find the final trajectory i+2.
Chapter 4  Numerical analysis

This dissertation seeks to demonstrate the impacts of driver relaxation at an on-ramp bottleneck using simulation and this chapter discusses the process. Section 4.1 outlines the study site and conditions examined in this dissertation. Section 4.2 presents one of the scenarios in great detail. Section 4.3 discusses the implications of the previous section and defines several key terms based on these observations. Section 4.4 discusses the remaining scenarios. Finally, Section 4.5 closes with a brief summary.

4.1 Introduction of simulation

This study seeks to simulate traffic past an on-ramp using the model from Chapter 3. The model is applied to a one-lane freeway section, Figure 2, with $v_f = 60$ mph, $RCap = 2,200$ vph, and $w = -12$ mph. The one-lane section includes an on-ramp at mile 0 and the simulated mainline segment is long enough to ensure that no queuing reaches either end (thereby ensuring that queuing does not impact the demand process). All vehicles are assumed to be homogenous, with identical driving characteristics. The model was tested under nine combinations of mainline and ramp flow: $(1,960 \text{ vph}, 2,080 \text{ vph}, 2,200 \text{ vph}) \times (120 \text{ vph}, 240 \text{ vph}, 360 \text{ vph})$, as shown in Table 1. These specific values were chosen somewhat arbitrarily with following goals. All scenarios would have mainline demand at or below $RCap$, while most scenarios would have a combined
demand at or above RCap. We also wanted enough scenarios to show general trends without so many scenarios that analysis would become intractable for a reader.

Table 1: Combined demand from mainline and ramp flow

<table>
<thead>
<tr>
<th>Ramp flow</th>
<th>120vph</th>
<th>240vph</th>
<th>360vph</th>
</tr>
</thead>
<tbody>
<tr>
<td>1960vph</td>
<td>2080vph</td>
<td>2200vph</td>
<td>2320vph</td>
</tr>
<tr>
<td>2080vph</td>
<td>2200vph</td>
<td>2320vph</td>
<td>2440vph</td>
</tr>
<tr>
<td>2200vph</td>
<td>2320vph</td>
<td>2440vph</td>
<td>2560vph</td>
</tr>
</tbody>
</table>

Note that the largest mainline demand is equal to RCap, and that the combined demand is below RCap in one case, equal to RCap in two, and above RCap in the remaining six. For the sake of clarity (i.e., limiting extraneous noise) the mainline has strictly constant headways between arrivals in the presented results, a point we will revisit in Chapter 6. The ramp is evaluated both with strictly constant headway arrivals and separately with non-evenly spaced arrivals (NES). The latter come from a random number generator using the uniform distribution between 0 and 10,000 sec for the given number of ramp arrivals, as per Table 1. This process is equivalent to assuming exponential headways. Only one set of NES arrivals is generated for a given ramp flow in Table 1, and this set is then applied to all three mainline demands. Since each set of the NES arrivals is only one realization for a given set of conditions in mainline and ramp flow in Table 1, any subsequent result herein should not be considered to be representative of an average response for the given pairwise combination of ramp and mainline demands.

When vehicles enter from the ramp, in our simulation they do so at the midpoint between two mainline vehicles, thus, the two affected vehicles initially have the same
spacing. Due to a lack of empirical calibration data, we are forced to use a heuristic method to set \( dv \) and \( dcc \) for the affected vehicles. The majority of this chapter presents the results for \( dv = 1 \) mph and \( dcc = 2 \) ft/sec\(^2\), though this research considered other values for each parameter, as will be discussed in Section 4.4.3.

Each simulation includes 4,000 mainline vehicles. The on-ramp flow is held at zero until 100 sec after the first mainline vehicle passes the on-ramp, allowing the mainline to stabilize before any on-ramp vehicles enter. Then at \( t=0 \) the on-ramp abruptly begins flowing at the set rate. This simulation time step, \( dt \), is 0.2 sec.

### 4.2 Queue formation near the on-ramp

This section presents the simulation results from one of the nine demand scenarios with constant arrival headway in Table 1. In particular, the case with 2,080 vph of mainline demand and 360 vph of ramp demand. Figure 5a shows a gray scale plot of the resulting mainline speed in the time-space plane with constant headway ramp arrivals. Traffic flows from bottom to top. Each measurement is calculated from the individual vehicle trajectories using a moving average every 5 sec over a time window of 31.1 sec\(^8\) at every 0.1 mile. In this case \( v \) is the harmonic average of the individual vehicle speeds passing the given location. As shown in the color bar, the lighter the color the faster the speed, and the white region corresponds to \( v_f \). The plot also shows two points from each delayed trajectory, indicating the location where the vehicle first drops below \( v_f \) and then the location where the vehicle first returns to \( v_f \). Taken collectively, these two groups of points respectively define the envelope of the upstream end of the queue, \( u\text{-end} \), and

---

\(^8\) This unusual period is used simply to prevent aliasing in the flow at RCap due to samples with partial headways.
downstream end of the queue, \textit{d-end} (avoiding the more common, but ambiguous terms, "head" and "tail"). The straight line passing through the origin is the trajectory of the last mainline vehicle before the on-ramp starts flowing. Figure 5b-c show the corresponding q, and k, respectively, where q comes from the number of vehicles per sample and k is calculated via the fundamental equation, $k=q/v$.\textsuperscript{9}

\textsuperscript{9} The moving average is centered on the reported time, as a result the plots are non-causal, the impacts of an event start becoming evident 15.5 sec before the event; which is why q starts increasing shortly before the on-ramp starts flowing.
Figure 5: (a) Time-space plot of mainline speed (mph) with constant headway ramp entrances with mainline demand = 2,080 vph and ramp demand = 360 vph. Diagonal line shows last vehicle past before ramp turns on, the collection of points show the u-end and d-end. Along with the corresponding (b) flow (vph), (c) calculated density (vpm), and (d) rescaled cumulative arrivals after the ramp turns on (zero values omitted for clarity). The dark area in (d) shows the region with speeds below 50 mph and the jagged solid line is the boundary between the loading and settling periods.
Starting at t=0 the combined ramp and mainline demand exceeds RCap and the bottleneck becomes active, but because of the driver relaxation process the first several minutes of activation does not exhibit any clear indicators of queuing. Flows in excess of RCap are common during these first few minutes after activation and the labels are shown with a dark background in Figure 5b (some flows are 250 vph over RCap, i.e., 11% above RCap). Over this region q and k remain positively correlated thus, precluding timely detection via /2/ from Chapter 1. The shaded area in Figure 5d shows the region where speeds are below 50 mph. Speeds remain above 50 mph everywhere until 3.3 min after the bottleneck activates and the first location where speed drops below 50 mph is downstream of the ramp. It takes several minutes for the queue to grow upstream of the on-ramp, precluding timely detection via /1/.

If we did not know the underlying qkFR, there would be no indication of queuing until at least 200 sec after activation. By the strictest definition there is clear evidence of delays between the ramp and the d-end almost immediately after activation since v is less than v_f and the underlying qkFR is triangular\textsuperscript{10}, but the drop is only 5-10 mph. This small speed drop combined with the positive correlation of q and k would commonly be interpreted as being indicative of the unqueued regime of a parabolic qkFR\textsuperscript{11} and so it would probably be overlooked in an empirical study. As evident in Figure 5a, the u-end of the fixed queue does not even reach the ramp until about 3 min after the bottleneck becomes active. Prior to this point the u-end is characterized by a succession of moving

\textsuperscript{10} Coifman and Kim (2011) previously argued that any v below v_f may be indicative of a sample that includes queued conditions for a portion of the sample, and that appears to be the case in the current study as well.

\textsuperscript{11} The earliest qkFR computed by Greenshields (1935) was parabolic and this shape still remains dominant in some domains, so if empirical data exhibit a pattern indicative of a parabolic qkFR is not likely to arouse suspicion.
bottlenecks emanating from the ramp, each triggered by an entering vehicle. The moving queue from each ramp entrance propagates downstream and is caught up by several free flowing mainline vehicles, as evident by the points between 0 and 0.2 mi over this time period.¹² As the affected drivers undertake the relaxation process, the feasible throughput drops towards RCap. The moving bottlenecks start to coalesce and the fixed queue forms in a manner that is somewhat similar to Duret et al (2010), i.e., the queue from one entry cannot completely dissipate before the impacts of the next entry arrive.

Now consider rescaled cumulative arrivals. This dissertation uses the last mainline vehicle before the ramp starts flowing for reference (the diagonal line) and sets it to be the 0-th vehicle. The numbers in Figure 5d show the cumulative arrivals minus t*RCap after the passage of the 0-th vehicle every 0.1 mi at 60 sec intervals. The zero values are not shown so as to highlight the samples with accumulation; likewise, all values prior to the 0-th vehicle are not shown and when the rescaled cumulative arrivals drop below -99 it is denoted with "*" for brevity. Since these rescaled cumulative arrivals are calculated using a moving time frame, the columns exhibit the same slope, v_f. For example, take the fourth column starting at the ramp and moving downstream, we see that 18 vehicles in excess of RCap have passed 0.1 mi since the bottleneck activated, and this quantity drops to 12 veh at 0.5 mi. Taking the difference between these two values we see an accumulation of 6 vehicles between 0.1 and 0.5 mi. Given the fact that 153 vehicles passed over this time, even when using /3/ from Chapter 1 it would be easy to miss this accumulation of 6 vehicles. Downstream of the ramp the rescaled cumulative

¹² Technically the first vehicle behind an entrance is upstream of the ramp, which is why these fluctuations extend a small distance upstream of the ramp.
arrivals are strictly non-decreasing over time until reaching the black jagged line. This line denotes the first instance when the rescaled cumulative arrivals decrease at the given location, and thus, q drops below RCap for a short period before subsequently returning to RCap. The numbers on the right side of Figure 5d show the rescaled cumulative arrivals at the given location at the end of the time window. Except for 0.1 mi, there are no more vehicles stored downstream of the ramp. Also note that there was never any accumulation past the d-end, but one would have to go more than 1.8 mi downstream of the ramp to find this case over all times.

Without the driver relaxation of Section 3.3, the entire bottleneck collapses to a point bottleneck at the on-ramp (as illustrated in Figure 19 of Appendix A for the same scenario presented in Figure 5). The system stabilizes after a few seconds, with the d-end remaining within 0.1 mi of the on-ramp from the moment the bottleneck became active as drivers accelerate past the point bottleneck. Queuing is immediately evident upstream of the on-ramp, and the u-end slowly grows. In fact one could derive the same state diagram directly using Lighthill, Whitham, and Richards' macroscopic traffic flow theory (Lighthill and Whitham, 1955; Richards, 1956) with a triangular qkFR.

4.3 Defining the loading and settling periods

Formalizing the analysis from Section 4.2, the inclusion of driver relaxation leads to a number of important findings, as follows. We call the several minutes immediately after demand first exceeds RCap the loading period. During the loading period, upstream of the on-ramp there is little or no speed drop, and no evidence of queuing. Downstream of the on-ramp q is supersaturated and in excess of RCap due to driver relaxation;
however, $q$ and $k$ remain positively correlated while $v$ only drops slightly below free speed, $v_f$. These supersaturated conditions are actually the initial formation of the queue, storing the demand in excess of $R_{Cap}$. These vehicles must be delayed while awaiting their turn to pass the d-end, hence the slight drop in speed. Past the d-end $q$ never exceeded $R_{Cap}$. In the above example the fixed queue formed around 0.2 mi and then grew in both directions. The d-end eventually extended more than 1.8 mi downstream of the ramp due to the segment saturating and the drivers undertaking the relaxation process having to travel further before reaching $v_f$. The u-end took several minutes to reach the ramp, after which point, delays and queuing first become evident upstream of the on-ramp. The loading period ends shortly after u-end passes the on-ramp because the on-ramp vehicles enter directly into the queue at lower speeds than before and thus, the relaxation distance shrinks. These results are consistent with Cassidy and Bertini (1999) who found the initial queue formation 1 km downstream of an on-ramp bottleneck.

With the shorter relaxation distance, the storage downstream of the ramp collapses and the d-end recedes back to about 0.4 miles downstream of the on-ramp. We refer to this interval as the \textit{settling period}. During the settling period $q$ between the ramp and the d-end drops below $R_{Cap}$ for a few minutes while the excess vehicles that were stored further downstream dissipate at $R_{Cap}$, consuming capacity that would otherwise be available at the on-ramp.\textsuperscript{13} This dissipation manifests as an upstream moving disturbance, within which both flow and speed drop to their lowest values for the given location. The settling period ends when flow downstream of the ramp recovers to $R_{Cap}$.

\textsuperscript{13} After a period of $q$ above $R_{Cap}$, this drop below $R_{Cap}$ should not be surprising since the long-term average $q$ cannot exceed $R_{Cap}$.
After the settling period, the d-end stabilizes, as does the bottleneck process overall for this case with constant headway ramp arrivals, e.g., speeds within the queue are roughly constant after the settling period. Of course the u-end continues to grow upstream, storing the demand in excess of capacity.

4.4 Alternative scenarios

This section repeats the simulation for all nine scenarios listed in Table 1. Section 4.4.1 presents the results with constant headway ramp arrivals and Section 4.4.2 presents the results with NES ramp arrivals. Finally, Section 4.4.3 discusses model calibration and reexamines the results from Section 4.2 when other car following models are employed.

4.4.1 Cases with constant headway ramp arrivals

Figure 6 repeats the simulation from Figure 5a for all nine scenarios listed in Table 1, with constant headway ramp arrivals. Comparing these nine plots, it should be clear that the shape of the queue depends on the combination of demands from the ramp and the mainline. No fixed queue forms for the three cases where the combined demand remains at or below RCap (i.e., the scenarios depicted in Figure 6a, b, and d), one only sees evidence of moving bottlenecks that quickly dissipate after each entrance. These moving bottlenecks are similar to those seen during the earliest part of the loading period for the u-end in Section 4.2, except demand is not high enough for the individual disturbances to coalesce into a fixed queue.
Figure 6: Time-space plots of mainline speed (mph) with constant headway ramp entrances, corresponding to the 9 demand combinations from Table 1. The mainline demand in row 1 is 1,960 vph, in row 2 is 2,080 vph, and in row 3 is at RCap, 2,200 vph. The ramp demand increases from left to right. Traffic flows from bottom to top, with the ramp at mile zero. The ramp turns on at t=0, after the mainline has loaded.

In the remaining six plots a fixed queue forms, the darker shading shows reduced speeds downstream and upstream of the on-ramp. For the three queued cases with mainline demand below RCap, Figure 6c, e, and f, the first 10-30 sec after demand first exceeds capacity are seemingly indistinguishable from the moving bottlenecks of the three cases in which no fixed queue formed. For the next 150 to 300 sec the fixed queue remains exclusively downstream of the on-ramp, with the only sign of delay at the on-ramp being the moving bottlenecks emanating downstream from the entering vehicles. As
in Section 4.2, within the fixed queue the supersaturated $q$ is above $R_{\text{Cap}}$, but past the d-end, $q$ does not exceed $R_{\text{Cap}}$. Speeds remain above 50 mph during the loading period, making this queuing very difficult to detect empirically. For the three cases with mainline demand at $R_{\text{Cap}}$, Figure 6g, h, and i, the very first entering vehicle causes a fixed queue to propagate upstream of the ramp, rather than the downstream moving bottlenecks seen in the other six plots. As a result, there is virtually no loading period in these plots. Finally, in all nine plots the system stabilizes by the end of the first 1,300 sec.

4.4.2 Cases with non-evenly spaced (NES) arrivals

Figure 7 repeats the detailed presentation from Figure 5 only now using the NES arrivals. Compared to Figure 5, the u-end reaches the on-ramp a little later than the constant headway ramp arrivals and there are three surges of flow in excess of $R_{\text{Cap}}$ evident in Figure 7b because of the specific NES arrival pattern. However, the basic relationships first described in Figure 5 remain. There is no indication of queue downstream of the on-ramp for several minutes after the on-ramp turns on. During this initial period the speed downstream of the on-ramp is only 5 to 10 mph below $v_f$, flow exceeds $R_{\text{Cap}}$, flow and density are positively correlated, and although there is vehicle accumulation downstream of the on-ramp, the quantity is small compared to the flow.
Figure 7: (a) Time-space plot of mainline speed (mph) with NES ramp entrances with mainline demand = 2,080 vph and ramp demand = 360 vph. Diagonal line shows last vehicle past before ramp turns on, the collection of points show the u-end and d-end. Along with the corresponding (b) flow (vph), (c) calculated density (vpm), and (d) rescaled cumulative arrivals after the ramp turns on (zero values omitted for clarity). The dark area in (d) shows the region with speeds below 50 mph and the jagged solid line is the boundary between the loading and settling periods.
Figure 8 repeats the same nine scenarios from Figure 6 using NES times between individual ramp arrivals, though the ramp arrivals still have an average flow equal to the respective column in Table 1. The basic findings from Section 4.4.1 remain, but the NES ramp arrival introduces noise that permeates the entire bottleneck process. The most notable difference from Figure 6 is in plots b and d, where the combined demand is exactly RCap. In Figure 8 a standing queue forms in both cases, complete with a loading and settling period. The queue grows when the short term demand exceeds RCap, but when demand falls below RCap, the excess capacity can only be used if there is already a queue, resulting in a standing queue. Unlike plots c, and e-i, the u-end does not grow indefinitely; it stops growing after 0.5-2 miles and then fluctuates, as illustrated in Figure 9. Across the six cases with queuing in Figure 6, the duration of the loading and settling periods differ in Figure 8 due to the short-term ramp flow fluctuations (e.g., the loading period in Figure 8e is now shorter than in Figure 8f even though the latter has larger combined demand). Rather than stabilizing after the initial settling period, in Figure 8 the queues continue to cycle through smaller loading and settling periods in response to the fluctuating ramp demand. As result, one now sees upstream moving disturbances and the d-end fluctuating after the initial settling period, as illustrated in Figure 10, showing a larger portion of the data from Figure 8f. Compared to Figure 6, the moving bottlenecks are less pronounced during the early portion of the loading period in Figure 8c, e, and f.
Figure 8: Time-space plots of mainline speed (mph) with NES ramp entrances, corresponding to the 9 demand combinations in Table 1. Compare to Figure 6.

Figure 9: d-end and u-end associated with Figure 8b and d over an extended period.
Figure 10: Figure 8f re-plotted over a larger time range.

4.4.3 Model calibration and other models

There are two entering vehicle parameters in the relaxation model: dv and dcc. Lacking calibration data, we repeated the analysis in this section using several values for these parameters. Figure 11 and Figure 12 show u-end and d-end from three dv and dcc values respectively when mainline and ramp demands are 2,080 vph and 360 vph (compare to Figure 6f and Figure 8f).
Figure 11: u-ends and d-ends from three dv values (3 mph, 1 mph, -1 mph) simulated from (a) constant headway ramp arrivals, (b) NES ramp arrivals. Mainline and ramp demands are 2,080 vph and 360 vph respectively and dcc is 2 ft/sec².
Figure 12: u-ends and d-ends from three dcc values (1 ft/sec$^2$, 2 ft/sec$^2$, 3 ft/sec$^2$) simulated from (a) constant headway ramp arrivals, (b) NES ramp arrivals. Mainline and ramp demands are 2,080 vph and 360 vph respectively and dv is 1 mph.

Note that the thinnest line in Figure 12 and Figure 13 corresponds to u-end and d-end in Figure 5 and Figure 7. As can be seen, the relaxation distance increases as the magnitude of dv decreases (entering at higher speed- requiring greater response time) or dcc decreases (lower deceleration rate- less responsive). As a result, the d-end and u-end both move downstream, and the duration of the loading and settling period increases. The reverse is true for increased dcc or dv.

The magnitude of w used in this section is typical of empirically observed values (e.g., Coifman and Wang, 2005). We have evaluated the results using a range of w and here too, the general relationships presented in Section 4.2 hold over the entire range. As the magnitude of w increases, queuing shrinks and generally both the loading and settling
periods have a shorter duration due to quicker driver response as illustrated in Figure 13.

Appendix B repeats the results from Sections 4.2 to 4.4.2 for \( w = -20 \) mph.

![Figure 13](image)

Figure 13: (a) u-ends and d-ends from two wave speed values (20 mph, 12 mph) simulated from deterministic arrivals, (b) u-end and d-end from two wave speed values (20 mph, 12 mph) simulated from NES ramp arrivals. Mainline and ramp demands are 2,080 vph and 360 vph respectively.

Obviously the results presented herein should depend on the choice of the car following model. As discussed in Chapter 3, Newell’s car following model is a linear form of the more general GM car following model. For completeness, we repeated the analysis from Section 4.2 after replacing Newell’s car following model with other, non-linear GM car following models from Gazis et al (1959) and Ozaki (1993), while retaining the driver relaxation process from Section 3.3. More details of these results are presented in Appendix C. Although the shape of the queue changes slightly, the general trends remain unchanged with a clear loading and settling period characterized by initial
queue formation downstream of the on-ramp, high speeds and supersaturated flows during the loading period, q drops below RCap during the settling period and finally the traffic state stabilizes with the d-end receding upstream towards the on-ramp.

As noted earlier, this dissertation seeks to use the simplest model to illustrate the impacts of driver relaxation. The relaxation model developed in Section 3.3 addresses some of the limitations of the relaxation model from Laval and Leclercq (2008) discussed in Section 2.2. However, given the potential calibration issues with our model, we also implemented the analysis using the more complicated car following model from Laval and Leclercq (2008) that was developed to account for lane change maneuvers within a queue and incorporates driver relaxation (see Appendix D for more details about their model). This analysis using different models of driver relaxation provides valuable insight into the robustness of the findings in the context of the emergent properties of the bottleneck process (e.g., queue formation, capacity, shape of the qkFR).

Using Laval and Leclercq’s model, Figure 14a-c shows the results for mainline demand of 2,080 vph and constant headway ramp arrivals (corresponding the same flow parameters and arrival patters of Figure 6d-f). While Figure 14d-f show the corresponding results for NES ramp arrivals (corresponding the same flow parameters and arrival patters of Figure 8d-f). In addition, Figure 15 shows contour plots of flow corresponding to Figure 14c and f (corresponding the same flow parameters and arrival patters of Figure 5b and Figure 7b).
Figure 14: Time-space plots of mainline speed (mph) with constant headway (a-c) and NES (d-f) ramp entrances, corresponding to the middle row Table 1 using the model from Laval and Leclercq (2008).
Figure 15: Contour plot of flow corresponding to Figure 14c and f (compare to Figure 5b and Figure 7b).

The basic results were similar to those from our relaxation process, as follows.

Using Laval and Leclercq, when the mainline demand was below RCap and the combined demand exceeded RCap we saw supersaturated states with q above RCap downstream of the on-ramp (e.g., see labels with black background in Figure 15). Queues were not evident upstream of the on-ramp until several minutes after demand exceeded capacity (i.e., until after the loading period), in the interim, v remained at or close to $v_f$. 

54
throughout the study area (as with our relaxation model, the traffic states looked as if they came from the unqueued regime of a parabolic qkFR). The extent of the affected region differed slightly when using Laval and Leclercq's model- typically it only reached a mile downstream of the on-ramp during loading, but the d-end stayed further downstream during the settling period and beyond. The behavior with constant headway ramp arrivals once queuing set in from Laval and Leclercq was slightly different than our model. Our model results in a single loading period and single settling period before stabilizing at RCap. Laval and Leclercq's relaxation process leads to a damped oscillation that cycles through loading and settling several times before stabilizing at RCap. In any event, both models provide evidence suggesting that q exceeds RCap during the loading period and this supersaturated q lasts for several minutes with only subtle indications that the queue has started forming.

4.5 Summary

Chapter 4 discussed simulation results near a freeway on-ramp bottleneck using the relaxation model from Chapter 3. Totally nine combinations of mainline and on-ramp demands were considered, as listed in Table 1. For ramp arrivals, this study examined both constant headway arrivals and NES arrivals. Followings are major findings from the simulations.

(a) Given NES arrivals, we sometimes see a standing queue even when demand does not exceed Rcap but the queue only grows a small distance and then fluctuates, providing just enough storage to allow surges in demand to dissipate during lulls in demand.
(b) In both constant headway and NES cases where the combined demand is greater than RCap but the mainline demand is less than RCap, there is no sign of delay or queue upstream of the on-ramp over an extended period (150 sec - 350 sec of loading period) after the on-ramp turns on. In the meantime the queue initially forms further downstream and u-end propagates upstream. During the loading period, speed is still close to $v_f$, flow and density are positively correlated, and vehicle accumulation downstream of the on-ramp is small. If one does not account for driver relaxation, these features would likely prevent an empirical study from detecting the fact that the bottleneck is already active during the loading period (see, e.g., Appendix A for the expected form of the queue in the absence of driver relaxation).

(c) Shortly after the queue reaches the on-ramp the settling period starts. The d-end collapses and then stabilizes a short distance beyond the ramp while the u-end keeps growing upstream of the on-ramp.

The results (b) and (c) are noisier for NES on-ramp arrivals than constant headway arrivals. This noise permeates to the entire time-space plane, but does not disrupt the basic observations described above. However, the NES arrival introduces large fluctuations throughout the segment that make the key transitions harder to recognize. As discussed in Chapter 5, if these impacts from driver relaxation are not accounted for, they can distort the findings from empirical studies of capacity drop and the FR.
Chapter 5  Unaccounted for impacts of relaxation on empirical studies

Empirical bottleneck studies have to simultaneously deduce the bottleneck capacity, identify the instant that the bottleneck becomes active, and determine where the bottleneck actually forms. Furthermore, the low number of conventional vehicle detectors typically precludes detailed spatial information. The detector stations used in an empirical study could be over a mile apart. As shown in Chapter 4, queuing during the loading period occurred further downstream than conventionally thought and the impacts are diffused over such a large distance that it is very difficult to detect the early queuing.

There are many commonly held assumptions that make it that much more difficult to recognize the faint evidence of queue formation during the loading period. First, the supersaturated loading period data seemingly come from the unqueued regime of a parabolic $q_kFR$ even though in the Chapter 4 simulations the underlying $q_kFR$ was triangular. Second, the commonly used point bottleneck model simply does not apply to the underlying bottleneck mechanism. The point bottleneck model assumes the bottleneck process occurs over a very short distance (e.g., Figure 19 in Appendix A). As shown in Chapter 4, with driver relaxation, the bottleneck process is extended over space. Drivers pass the on-ramp at flows above $R_{Cap}$ and are subject to delays further downstream, but these delays arise because drivers cannot sustain the short headways exhibited immediately after an entrance from the on-ramp. If one attempts to fit the point
bottleneck model to a non-point bottleneck, of course incongruous results should arise in the findings. This chapter discusses the specific impacts of these findings on empirical studies of bottleneck capacity in Section 5.1 and the FR in Section 5.2 if one does not explicitly account for the relaxation process. Finally, this chapter closes with a brief summary in Section 5.3.

### 5.1 Bottleneck capacity

To gain insight into empirical studies of bottleneck capacity, the three columns of Figure 16 respectively show the time series detector data that would be measured 0.1, 0.2, and 0.6 miles downstream of the on-ramp using the data underlying the constant headway ramp arrival scenario in Figure 5 and corresponding NES arrival scenario in Figure 7. Each plot in Figure 16 has one curve for the constant headway ramp arrivals (via Figure 5) and another curve for the NES ramp arrivals (via Figure 7). The top row of Figure 16 shows the rescaled cumulative arrival curve from the individual vehicle arrivals after subtracting a background flow equal to RCap, 2,200 vph (see, e.g., Cassidy and Windover, 1995). Typically one does not know RCap a priori and some other convenient background flow is used. However, in this case we do know RCap and use it as the background flow to highlight the boundary when flow is above or below RCap. So in Figure 16a the resulting curve from this background subtraction technique will be horizontal when q is equal to the background flow. The middle row shows the time series q for the given location using the same time axis as the top row (recall that q is the derivative of the cumulative arrivals). We use a conventional 30 sec sampling period for the moving average in Figure 16, and thus, some aliasing is evident in the middle row.
after 800 sec, where the measured flow fluctuates about RCap. This aliasing arises because RCap falls between resolvable values of flow and thus the samples include a non-integer number of headways, i.e., it reflects the limitations of sampling rather than an actual instability. As can be seen in the top row, the flow is actually at RCap after 800 sec. The bottom row shows the 30 sec space mean speed at the given location, again using the same time axis as the top row.

Figure 16: Time series detector measurements at three locations downstream of the on-ramp (by column). The first row shows the cumulative arrival curve after subtracting a background flow equal to RCap, the second row shows q, and the bottom row shows v, all with a common time axis. Each plot shows one curve for the constant headway arrivals and another for the NES ramp arrivals.

Now consider these measurements in the context of the so-called capacity drop if we did not know RCap a priori. Many researchers have empirically observed the highest

---

14 This aliasing is another confounding factor that is often overlooked in the empirical studies. Care must be taken to control for these sampling issues (e.g., by using the background subtraction technique of Cassidy and Windover, 1995), otherwise, the unbiased measurement error could be as much as 120 vph for a 30 sec sample.
q through a bottleneck just prior to the assumed activation. This high q is commonly taken to be the bottleneck's capacity. Once the bottleneck becomes active in the reported studies, q drops from the assumed capacity by 1% to 18% (Banks, 1991; Cassidy and Bertini, 1999; Hall and Agyemang-Duah, 1991; Hall and Hall, 1990; Persaud and Hurdle, 1991; Zhang and Levinson, 2004, Chung et al, 2007). Most of these studies rely on either on q or cumulative arrival curves to reach this conclusion. Unfortunately, most of these studies also employ the conventional point bottleneck model to determine when the subject bottleneck becomes active. Recall from Chapter 4 that there is no sign of queuing or delay upstream of the on-ramp during the loading period. A conventional point bottleneck model would not indicate that the bottleneck was active until queuing and delays are observed upstream, i.e., sometime after the settling period has begun. By this instant demand has exceeded capacity for some time– at least 300 sec after the bottleneck actually activated in the case of Figure 16. Meanwhile, the supersaturated q downstream of the ramp during the loading period superficially appears to be unqueued due to the fact that the drop in speed is so small and the relationship between q and k are consistent with a parabolic qkFR as discussed in Section 4.2. Even if one constructed a queuing diagram to catch delays between detectors like Cassidy and Bertini (1999), as noted in Section 4.2, the amount of accumulation is so small that it would be hard to detect. As a result, there is no clear indicator in the empirical data that these unsustainably high flows are in fact transient.

With the benefit of knowing RCap, return to the top row of Figure 16; a clear pattern is evident at all three locations regardless of whether constant headway arrivals or
NES arrivals on the ramp. Prior to t=0 there is no ramp flow, so the combined demand is below RCap and the slope is negative. When the ramp begins flowing at 360 vph, the combined demand exceeds RCap (positive slope), but the excess vehicles are being stored further downstream, so even at mile 0.6 we see a supersaturated q in excess of RCap. After the settling period begins, the d-end recedes upstream and many of the vehicles stored downstream of the on-ramp dissipate, consuming some of the RCap that would otherwise be available at the given location. So q drops below RCap (negative slope) during the settling period. Then q stabilizes at RCap (zero slope). Furthermore, the net accumulation after the settling period appears to be very small since the rescaled cumulative arrival curve returns to almost the same value it had when the on-ramp demand first arrived at the given location (consistent with Figure 5d and Figure 7d). The magnitude of the loading period displacement decreases the further downstream one looks, reflecting the fact that vehicles are being stored throughout the segment between the on-ramp and the d-end.

In the very likely scenario where one fails to recognize that the bottleneck activates at t=0 in an empirical study, the high q of the loading period will erroneously be assumed to be capacity, leading to an overestimate of capacity. In reality the q above RCap is simply indicative of the vehicle accumulation between the on-ramp and the d-end due to driver relaxation, but that is very hard to detect in an empirical study. Then when the q drops at the start of the settling period and the active bottleneck is finally detected, of course it would look like there is a drop from capacity. For example, in Figure 16a at mile 0.1 conventional study would consider that the bottleneck becomes
active at about 200 sec, would erroneously measure BCap to be about 2,350 vph, measured right before the flow finally drops to 2,200 vph. On the other hand, if one were somehow able to properly assign the activation time to \( t=0 \), when the supersaturated flow drops around 300 sec, we actually see a drop to capacity, i.e., RCap.

The true BCap cannot exceed RCap even though one should expect to measure sustained \( q \) in excess of RCap at some locations. To avoid the impacts of the loading period one can go beyond the d-end to measure the bottleneck's capacity, but the d-end can extend over 1.8 miles downstream of the on-ramp. Unfortunately it is quite rare for a bottleneck to be sufficiently isolated to be able to go that far downstream; often the impacts of one geometrical or operational feature collide with the next. For example, in studying an on-ramp bottleneck Cassidy and Rudjanakanoknad (2005) found capacity dropped due to lane change maneuvers immediately downstream of the on-ramp.

The fact that \( q \) is at its lowest during the settling period is particularly noteworthy. Several of the cited empirical studies show similar trends, with \( q \) dropping to its lowest value immediately after bottleneck activation is detected and then subsequently recovering to a higher value. In the context of Figure 16, this trend may be indicative of the empirical study location actually being upstream of the d-end for a portion of time. Examples include Persaud et al (1998) [their Figure 1 between 75-87 min] and Cassidy and Bertini (1999) [their Figure 5 between 6:30-6:37]. In fact Cassidy and Bertini also found cyclical surges with a frequency comparable to those in our Figure 9b. Careful inspection of their Figure 2 appears to show accumulation of about 10 vehicles on a segment thought to be downstream of the bottleneck process, several minutes prior to the
reported bottleneck activation time. This accumulation is similar to the accumulation of 6 vehicles that we see during the loading period in Figure 5d, over the same distance downstream of the on-ramp. If so, then the high q observed before breakdown in these studies may actually be supersaturated q from the loading period. However, these similarities may also simply be by chance, since the empirical studies include many other factors not found in our study, e.g., lane change maneuvers as per Cassidy and Rudjanakanoknad (2005). In any event, these ambiguities highlight the need for microscopic empirical data collected at the right locations to tease out the individual contributing factors and the present study underscores the fact that such data collection may have to cover several miles.

5.2 The fundamental relationship

As noted in Section 3.1, the traffic state (q, k, and v) is commonly assumed to fall on some fundamental relationship, FR. The FR is the foundation for much of traffic flow theory, yet debate continues about the shape of the FR over the past several decades.15 Most FR's were derived from empirical data and in this section we consider the impacts of the supersaturated states on the observed FR. Out of convenience the discussion focuses on the qkFR16. Once more using the trajectories underlying Figure 5a with constant headway ramp arrivals, Figure 17 shows the observed flow versus density at the three locations used in Figure 16. The first column uses a 30 sec moving average and the second column uses a 60 sec moving average. Recall that the underlying qkFR is

---

15 See Coifman and Kim (2011) for a review of the literature.
16 The findings translate to the other two bivariate realizations of the FR (q versus v, and v versus k) via the fundamental equation.
triangular with \( v_f = 60 \text{ mph} \), \( RCap = 2,200 \text{ vph} \), and \( w = -12 \text{ mph} \) (shown with dashed lines in the plots), yet the measured \( q \) climbs more than 10% above \( RCap \) due to the supersaturated states. If one strictly used the recorded data, the unqueued regime of the \( qkFR \) appears to trace out a straight line with slope \( v_f \) from the origin to \( RCap \) (following the underlying triangular \( qkFR \)). Then as the supersaturated \( q \) increases above \( RCap \) to the "apparent capacity" (i.e., the maximum supersaturated \( q \) erroneously taken as capacity, as per Section 5.1) the empirical \( qkFR \) bends to the right, with \( v \) dropping to 50 mph. As discussed in Chapter 4, the flow above \( RCap \) is actually measured within the bottleneck process and represents the fact that vehicles are being stored downstream.
Figure 17: q-k relationships at three locations downstream of the on-ramp with constant headway arrivals, by row, when the on-ramp has constant headway arrivals (a)-(c): 30 sec sampling period; (d)-(f): 60 sec sampling period. The underlying triangular qkFR is shown with dashed lines in each plot.

When taking the measurements from Figure 17 in temporal order, the traffic state progresses in a clockwise sequence, starting from near the apex of the underlying
triangular qkFR. As one would expect, the progression is cleaner in the 60 sec data, but the story is the same in the 30 sec data. In either case, the data progresses through the loading period yielding most of the measurements above RCap, then q remains supersaturated but drops below RCap during the settling period. Finally, the sequence returns to RCap at a speed at or near $v_f$ after the vehicles stored downstream discharge.\textsuperscript{17} Without recognizing the fact that the states are supersaturated, the apparent capacity is higher than the real capacity, and the peak q occurs after the bottleneck has become active. This cycle yields several values of q for a given k, likely one of the sources of the noise in empirical qkFR.

This analysis was repeated for NES arrivals on the ramp with very similar results as illustrated in Figure 18. The one key difference is that rather than stabilizing at the apex of the underlying qkFR after the first settling period, at mile 0.1 and 0.2 the traffic state continued to cycle through loading and settling periods with smaller displacements from RCap than the first cycle. These smaller perturbations did not reach mile 0.6 (see, e.g., the last column in Figure 16), and so at that location the traffic state remained at the apex of the underlying qkFR.

\textsuperscript{17} If the location is upstream of the d-end, speeds will be slightly below $v_f$ because traffic is still accelerating at that location, otherwise, the final state will be at the apex of the underlying qkFR.
Figure 18: q-k relationships at three locations downstream of the on-ramp with NES ramp arrivals, by row, when the on-ramp has constant headway arrivals (a)-(c): 30 sec sampling period; (d)-(f): 60 sec sampling period. The underlying triangular qkFR is shown with dashed lines in each plot.

One should see similar cycles below RCap at on-ramps within the queue further upstream. For example, Leclercq et al (2007) studied an on-ramp within a queue from a
downstream bottleneck and examined the impact of driver relaxation from vehicles entering from the ramp on the q and k measurements. In the context of the present work, their plots clearly show supersaturated states immediately downstream of the on-ramp, though they merely referred to these points as being "nonequilibrium". Since their study was strictly within a queue, few of the supersaturated flows were above RCap and they did not consider the implications on an empirically measured qkFR at the bottleneck. In this way, the supersaturation process can impact empirical observations of the entire queued regime, not just those measurements close to capacity. If these supersaturated states go unaccounted for in an empirical study, they will contribute noise throughout the entire queued regime.

The empirical qkFR distorted by the driver relaxation process should be reproducible as long as demands are roughly similar from day to day on the mainline and ramp. In other words, the supersaturated states offers a reproducible mechanism that can pull the empirical qkFR above the underlying qkFR, i.e., shifting away from the origin. This effect can be stable in time because the drivers entering from the on-ramp are constantly being replenished. However, it remains transient in space, after some distance the drivers do obtain their preferred headways, at which point the supersaturated states disappear. The distortion can yield a parabolic-like curve over the entire unqueued regime, making such a location superficially look attractive for empirical study; but because the supersaturated portion of the curve arises from the relaxation process, the resulting qkFR is not representative of most roadway segments.
In the context of empirical FR, Cassidy (1998) developed a technique to extract stationary traffic states by using cumulative vehicle arrivals and smoothed occupancy to identify periods of relatively stable demand. However, the specific requirement for stable conditions would result in very few queued observations close to RCap.

5.3 Summary

This chapter discussed the implications of the relaxation process (via the simulation results from Chapter 4) in the context of empirical studies of capacity drop and the FR. This chapter simulated conventional detector measurements and showed that under conventional assumptions it would be difficult to identify the fact that a bottleneck is active during the loading period. Investigation of the simulated detector measurements reveals that time-series flow downstream of the on-ramp follows a time-series trend typical of many empirical studies that found a capacity drop. In fact, the high flows before the drop come from unsustainable, supersaturated states during the loading period, i.e., they come after the bottleneck first became active. Since most of the empirical studies also employ the conventional point bottleneck model, it is unlikely that they would be able to detect activation until shortly after the settling period, this delay leads to an overestimate of capacity. Instead of \( q \) dropping "from capacity", we see \( q \) drop "to capacity" from super-saturation.

Empirical \( qkFR \) are typically constructed by taking flow and occupancy (or density) immediately downstream of an apparent point bottleneck to ensure the observation of capacity flows. Unfortunately, the point bottleneck assumption distorts the results. When incorporating driver relaxation, the loading period is characterized by
supersaturated traffic states immediately downstream of the on-ramp. We use the time series data to construct empirical qkFRs from the simulation results. Superficially the qkFRs in Figure 18 appear to be representative of a parabolic-like qkFR and exhibit an apparent capacity above RCap; however, the underlying qkFR is actually triangular with maximum sustainable flow of RCap. As discussed in Chapter 4, the supersaturated traffic states come from the loading period, after the bottleneck has become active and thus, distorts the observed qkFR. One would have to take measurements strictly downstream of the d-end to measure the true qkFR. Since the d-end extends more than a mile downstream of the on-ramp, as illustrated in Figure 5a, at most bottlenecks it is likely impossible to go that far before impacts of other traffic or operational features can further impact the traffic state.
Chapter 6  Conclusions and future work

In closing, Section 6.1 summarizes the results and presents the conclusions of this study. Section 6.2 discusses the limitations of the relaxation model and its implementation, and then discusses future research to extend this work.

6.1 Conclusions

This simulation study examined traffic behavior in the vicinity of an on-ramp bottleneck, revisiting commonly held assumptions and uncovering systematic biases that likely have distorted empirical studies of bottleneck formation, capacity drop, and the fundamental relationship. We modify Newell's car following model to include the driver relaxation process. At the macroscopic scale the traffic state for any sample containing one or more of vehicles undertaking the relaxation process will be supersaturated (i.e., resulting in a measured traffic state above the underlying qkFR). In other words the supersaturated traffic states offers a reproducible mechanism that can pull the empirical qkFR above the underlying qkFR, i.e., shifting away from the origin, and in some cases, above RCap.

As an on-ramp bottleneck becomes active, the entering drivers are constantly being replenished and keep the traffic state supersaturated. After the combined demand first exceeds capacity in our simulations, the bottleneck activation progresses through the following steps: (1) Moving bottlenecks occur for a few minutes downstream of the on-
ramp. The bottlenecks are superficially indistinguishable from high flow, non-active conditions, but the supersaturated q is above RCap. (2) A fixed queue forms some distance downstream of the on-ramp and eventually extends up to several miles beyond the ramp. Between the on-ramp and d-end (downstream end of the queue), the supersaturated q remains above RCap (beyond the d-end, q never exceeds RCap). (3) The u-end (upstream end of the queue) grows upstream, eventually reaching the on-ramp several minutes after demand first exceeded capacity. (4) With the ramp drivers now entering at lower speeds, the relaxation distance shrinks, and thus, the d-end recedes upstream. The number of vehicles stored downstream drops, and as they dissipate, they consume some RCap that would otherwise be available at the ramp, i.e., q drops below RCap. (5) Finally the system stabilizes at RCap (or near RCap in the presence of NES ramp arrivals). Steps 1-3 are termed the loading period and step 4 the settling period; both of these periods exhibit supersaturated traffic states downstream of the on-ramp, though during the settling period q is below RCap.

Reinterpreting many empirical studies in the context of our results, during the loading period a conventional point bottleneck model would erroneously indicate that the bottleneck is inactive. In fact during the loading period most of the bottleneck activity actually occurs downstream of the on-ramp, which is inconsistent with a simple point bottleneck model. The bottleneck process occurs over an extended distance, in excess of 1 mile. If one fails to recognize the fact that the bottleneck is already active during the loading period, one would overestimate the bottleneck capacity due to the supersaturated q and the recorded activation time will be too late. Only after the settling period is over
does q return to the actual bottleneck capacity, which is equal to RCap. Instead of q dropping "from capacity", we see q drop "to capacity" from supersaturation. If proven empirically, this finding has important implications for traffic flow theory and traffic control, e.g., the bottleneck process and traffic responsive ramp metering, respectively.

We suspect these confounding effects have largely gone unnoticed due to the ambiguity in defining exactly what constitutes "unqueued" conditions. In fact, measuring q, k, v from our simulation results we see a seemingly parabolic-like qkFR (with the parabolic-like portion coming from the supersaturated states above RCap) more than a mile downstream of the on-ramp due to the driver relaxation while the underlying qkFR is triangular in this case. The simulation reveals that these locations are not strictly downstream of the bottleneck process, and v is only slightly below $v_f$. The distortions in qkFR result from the transient traffic states over the loading period due to driver relaxation. However, as previously argued by Coifman and Kim (2011) any v below $v_f$ may be indicative of a sample that includes queued conditions for a portion of the sample and that appears to be the case in the current study as well: as long as a driver is traveling below $v_f$ they are constrained by downstream conditions. Thus, using a strict $v_f$ criteria for unqueued states would ensure the downstream observation site was past the entire bottleneck process, but it would also put this downstream observation site several miles past the on-ramp in many of our simulations- a distance that is often infeasible due to extraneous features downstream of the on-ramp in empirical studies (e.g., if the next ramp enters or exits before the d-end it will preclude observation of RCap).
The driver relaxation process is a confounding factor far below the resolution of conventional macroscopic data, and empirical studies usually fail to account for it. One thing is clear, however, the bottleneck process appears to occur over a much longer distance than previously thought, with subtle influences arising far downstream of the apparent point bottleneck location. Right now we are faced with the very daunting challenge that there are few data sources with high enough resolution to tease out the individual contributing factors and enable such advances. So the present work is also meant to help focus future data collection in such a way that these necessary data will be collected from the right locations, and ultimately, so that more robust models can eventually developed. None of the existing publicly available, microscopic, empirical traffic data sets span the necessary region (several miles downstream of the apparent bottleneck).¹⁸

6.2 Limitations and Future works

One objective in this dissertation is to present a very simple model that shows beyond a doubt that driver relaxation is an important factor that could very easily have confounded prior empirical studies. The spatio-temporal range and magnitude of the results certainly depend on the uncalibrated model from Chapter 3. This omission is due to a lack of microscopic data for calibration. So likewise, one should not depend upon the precise values reported herein, rather, the general trends, including:

- the loading period superficially appears to be unqueued even though it actually occurs after the bottleneck has activated,

¹⁸ While there are a handful of publicly available empirical microscopic data, e.g., Smith (1985), Kovvali et al (2007), none of them span a recurring bottleneck and thus, these data sets provide little insight into the problem at hand.
• the initial queue formation appears to happen downstream of the on-ramp,
• the d-end can extend over a mile downstream of the on-ramp, and
• the bottleneck process appears to occur over an extended distance that is poorly captured with the point bottleneck model.

There were many other assumptions in this work, discussed below, that should also be accounted for. It is our hope that future research will address these limitations and add greater precision to refine these interpretations. As a first step in this direction, we relaxed one of our assumptions and used NES arrivals on the ramp. We found that the results are noisier than they are for constant headway arrivals. This noise permeates to the entire time-space plane, but does not disrupt the basic relationships described in Section 4.2. We suspect the same would be true if we added NES arrivals to the mainline, an inhomogeneous vehicle fleet, or simulated a merge lane where vehicles were allowed to enter the freeway over a range of distances. The results without the stochastic effects represent a best-case scenario, adding in the stochasticity, the basic findings remain, e.g., the long duration of the loading period where no queues are evident upstream of the on-ramp and q downstream of the on-ramp is supersaturated. The stochasticity introduces large fluctuations throughout the segment that make the key transitions harder to recognize unless you know to look for them. In the present work we are trying to tease out the subtle phenomena at the earliest stages of bottleneck activation- so we skim away many of these distractors.

The one-lane freeway is another such simplification, explicitly excluding the possibility for lane change maneuvers within the bottleneck (Coifman et al, 2003; Laval
and Daganzo, 2006; Duret et al, 2010; Coifman and Kim, 2011), in part so that we can highlight the impacts of driver relaxation at an on-ramp without the confounding effects of secondary lane change maneuvers. On a real freeway one should see several mainline drivers change lanes to avoid the on-ramp flow and thus, carry the driver relaxation process to the inside lanes (see, e.g., Newman, 1963). While the basic process should be similar for lane change maneuvers, the impacts become harder to track because the maneuvers are not constrained to a specific location. Chung et al (2007) showed that ramp metering can reduce the number of lane change maneuvers and thus increase queue discharge flows, i.e., the maximum sustainable throughput depends on driver behavior. In the context of Coifman and Kim (2011) the lane change maneuvers may preclude observing the true capacity altogether (both BCap and RCap). If the freeway segment is operating near RCap, each lane change maneuver sends a "hole" downstream in the exited lane and a brief delay upstream in the entered lane. Combined, these two waves reduce the flow everywhere on the freeway, but they are not a reduction in RCap, instead, they simply represent a brief departure from RCap. Since the disturbances are very small (one headway of delay per maneuver) and the lane change maneuvers are distributed over space, the impacts of these maneuvers are very hard to isolate. Of course lane change maneuvers are also subject to driver relaxation (e.g., Cohen, 2004; Wang and Coifman, 2008; Leclercq et al, 2007; Xuan and Coifman, 2012). So the impacts of driver relaxation that we find in the context of an on-ramp are also likely to translate to lane

---

19 The "hole" propagates downstream in the exited lane because that lane is already near \( v_f \) and the following vehicles cannot close the gap. The brief delay propagates upstream in the entered lane because that lane is already at RCap, so all vehicles upstream of the maneuver must be delayed by one headway for this vehicle to enter the lane.

20 Alternatively, in this context each lane change maneuver represents a transient point bottleneck that lasts only a few seconds.
change maneuvers where drivers enter the new lane at random locations. The above
discussion of lane change maneuvers in the context of Coifman and Kim (2011) assumed
instantaneous driver relaxation. If one used more realistic driver relaxation, the delay and
the associated upstream moving wave would not start until the affected drivers began
relaxation process. Very similar to what we found at the on-ramp, the supersaturated \( q \) in
the entered lane would propagate downstream with the vehicles. As the drivers approach
normal headways via driver relaxation, \( q \) drops, a small delay wave forms (one headway
of delay) in the entered lane at some point downstream of the lane change maneuver
location, and then propagates upstream past the lane change maneuver location.

In our simulation we simplified the merging process for the ramp vehicles. In
practice the merging process should be more complicated. In the case of courtesy
yielding (Wang and Mahmassani, 2005 and Choudhury et al, 2007), for example, a
mainline vehicle slows down and allows the merging vehicle to enter the mainline
comfortably (Daamen et al, 2010). Thus, spacing between the affected vehicles could be
longer than predicted in this study. As a result the affected vehicles might require a
shorter distance and time to reach their preferred speed and spacing, which in turn shrinks
the size of the affected region. However, the adjustment during the merging process
would induce other disturbances that may counterbalance these benefits. In any event,
when the combined demand exceeds \( R_{\text{Cap}} \), there is no way for the affected vehicles to
take comfortable spacing ahead without some relaxation process.

Our simulations also used an abrupt change when the ramp flow switches on and
the combined demand instantaneously jumps above \( R_{\text{Cap}} \). We would expect a slower,
more continuous demand increase at a real bottleneck, which should extend the duration of the loading period. The long-term combined average demand might not even exceed RCap at the onset of queuing, with minor fluctuations occasionally pushing demand above RCap, e.g., Figure 8b and d. Simulation also let us study networks that were not encumbered by confounding downstream features. All of these factors that were simplified in the present work make it that much harder to pinpoint exactly when demand exceeds capacity in empirical studies.

While we believe the overall findings of this work are accurate, the exact form is highly sensitive to several factors, e.g., using Laval and Leclercq (2008) in place of our relaxation model, we found the affected range dropped by 43%. So the present work should not be viewed as a complete model of the very complicated bottleneck process and we feel one should resist the temptation to build detailed models until the details of the process can be measured empirically. The objective of this study is to highlight the impacts of what we believe to be an important factor that has previously gone largely overlooked, as well as motivate future research into the nuances of these issues. One fact is clear, the point bottleneck model is too simplistic. To advance the understanding of bottleneck mechanisms our community needs to devise ways to better handle multiple interacting features rather than assuming a simple point bottleneck.
References


896-913.

Cassidy, M.J., Windover, J.R., 1995. “Methodology for assessing the dynamics of
freeway traffic flow,” Transportation Research Record 1484, pp.73-79.


following models and a proposed fuzzy inference model,” Transportation

following,” Operation Research, vol. 6, no. 2, pp. 165-184.

cooperative lane changing and forced merging behavior,” Transportation

and capacity drop at three freeway bottlenecks,” Transportation Research: Part
B, vol. 41, no 1, pp. 82-95.

microscopic simulation models,” Transportation Research Record 1883, pp.50-
58.

dual loop detectors," Transportation Research: Part A, vol. 36, no. 4, pp. 351-
364.

Coifman, B., Kim, S., (2011). "Extended bottlenecks, the fundamental relationship and
capacity drop," Transportation Research: Part A. vol. 45, no. 9, pp. 980-991.

Coifman, B., Krishnamurthy, S., Wang, X., (2003) "Lane Change Maneuvers Consuming
Freeway Capacity," Proceeding of the Traffic and Granular Flow 2003
Conference, pp. 3-14.


83


Appendix A: Simulation results without the driver relaxation process

Figure 19 shows the results corresponding to Figure 5 except now we omit the driver relaxation process, as would be the case with a point bottleneck model. In Figure 19a there is no sign of queued conditions downstream of the on-ramp because the affected vehicles immediately adjust the shorter spacing upon a merge. As expected with this point bottleneck there is clear distinction in traffic conditions between upstream and downstream of the on-ramp. Figure 19b and c shows the flow and density downstream of the on-ramp corresponds to the state at RCap (compare to Figure 5). One could derive the same state diagram directly using Lighthill Whitham and Richards' macroscopic traffic flow theory (Lighthill and Whitham, 1955; Richards, 1956) with a triangular qkFR.
Figure 19: Repeating the scenario from Figure 5a-c using Newell’s car following model without the driver relaxation process.
Appendix B: Simulation results with 20 mph of wave speed

As discussed 4.4.3, Figure 20-Figure 25 repeat the analysis corresponding to Figure 5-Figure 10 only now with the wave speed, w, set to -20 mph. As the magnitude of w increases, queuing shrinks and generally both the loading and settling periods have a shorter duration and the affected region shrinks due to quicker driver response.
Figure 20: Figure 5 with 20mph wave speed.
Figure 21: Figure 6 with 20mph wave speed.
Figure 22: Figure 7 with 20mph wave speed.
Figure 23: Figure 8 with 20mph wave speed.

Figure 24: Figure 9 with 20mph wave speed.
Figure 25: Figure 10 with 20mph wave speed.
Appendix C: Simulation results using different car following models

As discussed in Chapter 3, Newell’s car following model is a linear form of the more general GM car following model \((m, l)\) are equal to zero in equation 1). For completeness, we repeated the analysis after replacing Newell’s car following model with other non-linear GM car following models. Figure 26 and Figure 27 correspond to Figure 5 after employing a non-linear model calibrated by Gazis et al (1959), and Ozaki (1993) respectively. Note that in Figure 26 the d-end is still far downstream at the end of the simulated period. To verify that traffic state downstream of the on-ramp eventually stabilizes in this case, Figure 28 shows u-end and d-end over an extended period. In both Figure 26 and Figure 27 the d-end and u-end are slightly different from Figure 5 due to new underlying speed-spacing relationship. However, the general trends remain unchanged- initial queue formation downstream of the on-ramp, high speeds and supersaturated flows during the loading period.
Figure 26: Repeating the scenario from Figure 5a-c using the car following model developed by Gazis et al (1959) and including our driver relaxation process from Section 3.3.
Figure 27: Repeating the scenario from Figure 5a-c using the car following model developed by Ozaki (1993) and including our driver relaxation process from Section 3.3.
Figure 28: u-end and d-end associated with Figure 26 over an extended period
Appendix D: Implementation of Laval and Leclercq's relaxation model

In addition to the models discussed in the body of this dissertation, the research also implemented the analysis using another relaxation model developed by Laval and Leclercq (2008). Equation 3 shows Laval and Leclercq's car following rule between vehicle i (leader) and i+1 (follower), which is modified from Newell’s simplified car following model for traffic within a queue.

\[
x_{i+1}(t + \Delta t) = x_i(t) + v_i(t + \Delta t) \times \Delta t - \frac{\Delta N_{i+1}(t + \Delta t)}{k[v_i(t + \Delta t)]}
\]  

[Eq.3]

where,

\(\Delta N_{i+1}(t)\) : Difference in vehicle number between two consecutive vehicle i+1 and i at time t, as defined below.

\(x_i(t), v_i(t)\) : Position and velocity of vehicle i at time t

\(\Delta t\) : Time step

\(k[v_i(t)]\) : Density corresponding to \(v_i(t)\)

Under stationary conditions, \(\Delta N_i(t)\) is always 1 so that a follower vehicle trajectory at \(t + \Delta t\) can be predicted directly from Newell's car following model based on the leader’s position and the follower's preferred spacing (the inverse of k) at time \(t + \Delta t\). After a merge event, however, using our terminology the entering vehicle and following vehicle are supersaturated (or "non-stationary" using Laval and Leclercq's terminology),
taking unsustainably short spacings and under these conditions \( \Delta N_i(t) \) is less than 1 to reflect these truncated spacings adopted by the drivers. Equation 4 (adapted from Laval and Leclercq, 2008) is used to calculate \( \Delta N_i(t) \).

\[
\Delta N_{i+1}(t+\Delta t) = \Delta N_{i+1}(t) \times \left\{ \frac{1}{k[i(t)]} + (v_i(t+\Delta t) - v_i(t) + \varepsilon) \times \Delta t \right\} \times k[i(t+\Delta t)]
\]  
[Eq. 4]

where,

\( \varepsilon \): Difference in speed that the lane changer is willing to maintain with its leader

In Equation 4, the term in the bracket reduces to the inverse of \( k[i(t+\Delta t)] \) when \( \varepsilon \) is zero and in that case, \( \Delta N_{i+1} \) does not change over time. If vehicle \( i+1 \) is a merging vehicle, the initial \( \Delta N_{i+1} \) could be less than 1 when the spacing at time \( t \) is shorter than preferred at a given speed. Thus \( \Delta N_{i+1} \) is less than 1 and it would be a situation where the merging vehicle will tolerate a shorter spacing. Under typical relaxation process, a merge vehicle tends to increase spacing ahead and ultimately reaches preferred speed-spacing relationship downstream as validated in Leclercq et al (2007). Thus \( \varepsilon \) is likely to be positive value, which is consistent with the calibration result from Leclercq et al (2007). The positive \( \varepsilon \) results in increase of \( \Delta N_{i+1} \) as time progresses so that ultimately it will approach to 1, at which point the merge vehicle finishes the relaxation process.

However as Laval and Leclercq described, the single parameter, \( \varepsilon \), in their model is somewhat subjective so that it might be hard to calibrate. Meanwhile, their model triggers the relaxation process right at the end of the merge no matter what the merging vehicle’s speed is. However it would be more reasonable to say that the relaxation process from a merging vehicle initiates after some level of adjustment to the mainline.
traffic speed. For example, if speed at the merge were faster than mainline traffic, the merging vehicle would decelerate first and then start the relaxation process.