UNIVERSITY OF CALIFORNIA
RIVERSIDE

Sequential Hypothesis Testing With Spatially Correlated Presence-Absence Data and the Corridor Problem

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Applied Statistics

by

Elijah Daniel DePalma

December 2011

Dissertation Committee:

Dr. Richard Arnott, Co-Chairperson
Dr. Daniel Jeske, Co-Chairperson
Dr. Matthew Barth
Dr. James Flegal

The Dissertation of Elijah Daniel DePalma is approved:

Committee Co-Chairperson

Committee Co-Chairperson

University of California, Riverside

Acknowledgments

I am grateful to Professor Richard Arnott and Professor Daniel Jeske for their guidance and support, and for the research opportunities they have provided over the last few years.

I am grateful to Professor Matthew Barth for allowing me to participate in his research group, and to his research assistants Kanok Boriboonsomsin, Alex Vu and Guoyuan Wu. I am especially grateful to Professor Alex Skabardonis for granting me permission to obtain the PeMS data used in this dissertation.

The text appearing in Chapter 2 of this dissertation, in part or in full, has been submitted as a working paper for publication to *Journal of Economic Entomology*. The co-author Daniel R. Jeske listed in that working paper directed and supervised the research which forms the basis for this dissertation. The co-authors Jesus R. Lara and Mark S. Hoddle listed in that working paper contributed empirical data, editorial assistance, the graph appearing in Figure 2.1, a portion of the Discussion section, and general research collaboration. This joint work was supported in part by a USDA-NIFA Regional Integrated Pest Management Competitive Grants Program Western Region Grant 2010-34103-21202 to Mark S. Hoddle and Daniel R. Jeske.

The text appearing in Chapter 3 of this dissertation, in part or in full, is a reprint of the material as is appears in *Transportation Research Part B* (2011), 45:743–768. The co-author Richard Arnott listed in that publication directed and supervised the research which forms the basis for this dissertation. We are indebted to Caltrans and the USDOT for research award UCTC FY 09-10, awarded under Caltrans for transfer to USDOT.

The text appearing in Chapter 4 of this dissertation, in part or in full, has

been submitted as a working paper for publication to *Transportation Research Part B*. The co-author Richard Arnott listed in that working paper directed and supervised the research which forms the basis for this dissertation. We are grateful to the US Department of Transportation, grant #DTRT07-G-0009, and the California Department of Transportation, grant #65A0216, through the University of California Transportation Center (UCTC) at UC Berkeley, and the dissertation fellowship award through the University of California Transportation Center (UCTC).

To Professor Richard Arnott and Professor Daniel Jeske...

...for their continuing guidance and innovation.

ABSTRACT OF THE DISSERTATION

Sequential Hypothesis Testing With Spatially Correlated Presence-Absence Data and the Corridor Problem

by

Elijah Daniel DePalma

Doctor of Philosophy, Graduate Program in Applied Statistics
University of California, Riverside, December 2011
Dr. Richard Arnott, Co-Chairperson
Dr. Daniel Jeske, Co-Chairperson

Firstly, we develop a sampling methodology for making pest treatment decisions based on mean pest density. Previous research assumes pest densities are uniformly distributed over space, and advocates using sequential, presence-absence sampling plans for making treatment decisions. Here we develop a spatial sampling plan which accomodates pest densities which vary over space and which exhibit spatial correlation, and we demonstrate the effectiveness of our proposed methodology using parameter values calibrated from empirical data on *Oligonychus perseae*, a mite pest of avocados. To our knowledge, this research is the first to combine sequential hypothesis testing techniques with presence-absence sampling strategies which account for spatially correlated pest densities.

Secondly, we investigate "The Corridor Problem", a model of morning traffic flow along a corridor to a central business district (CBD). We consider travel time cost and schedule delay (time early) cost, and we assume that a fixed number of identical commuters have the same desired work start-time at the CBD and that late arrivals are prohibited. Traffic flow along the corridor is subject to LWR flow congestion with

Greenshields' Relation (i.e., mass conservation for a fluid coupled with a negative linear relationship between velocity and density), and we seek to characterize the no-toll equilibrium or user optimum (UO) solution, in which no commuter can reduce their trip cost, and the social optimum (SO) solution, in which the total population trip cost is minimized.

Allowing for a continuum of entry-points into the corridor we develop a numerical algorithm for constructing a UO solution. Restricting to a single entry-point we provide complete characterizations of the SO and UO, with numerical examples and quasi-analytic solutions. Finally, we develop a stochastic model of incident occurrence on a corridor, calibrated using a recently developed change-point detection algorithm applied to traffic data along a San Diego freeway over the course of a year, coupled with Hierarchical Generalized Linear Model (HGLM) fitting techniques. We use the calibrated incident model in a simulation study to determine the effect of stochastic incidents on the equilibrium solutions to the single-entry point Corridor Problem.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

† In Vino Veritas (In Wine Is Truth)

An Avocado Farmer's Dilemma

"Plucking a leaf from the nearest tree, he looked closely to see any evidence of the tiny mite pest which could devastate his crop. Scanning up and down the leaf vein he managed to count twenty mites until the wind blew dust across his eyes, frustrating his efforts. In desperation he circled the tree, randomly plucking and inspecting leaves to see evidence of even a single mite pest. Of the twelve leaves he gathered, seven had evidence of mite pests. He knew that in a severe pest infestation situation more than three-quarters of the leaves would be infested, yet he also knew that in such a situation not every tree would be infested. If the infestation was too large, then he would have no choice but to treat the entire grove at a cost of many thousands of dollars and damaging environmental impact. However, if the infestation was not severe enough, then leaf drop would never be initiated and there would be no need to treat the orchard. His gaze again drifted to the large orchard of budding avocado trees, as he contemplated whether or not a treatment would be necessary."

Congestion in the Morning Commute

"As traffic slowed to a crawl, she glanced at her clock to reassure herself that she would arrive to work on time. After leaving her home at 8:00am and traveling 5 miles, her clock read 8:20am with another 10 miles to go before reaching downtown. She looked around at the nearby vehicles and wondered how many other commuters faced the same decision each morning: Do I leave home at 8:00am and face an hour commute to arrive at work exactly on time? Or do I leave earlier to beat the traffic and arrive early, but then be forced to wait until 9:00am when the work day starts? As her mind wandered, she heard the dreaded crunch of metal on metal indicating that an accident had occurred, and watched in dismay as traffic came to a complete stop."

The body of this dissertation is composed of four chapters (Ch. 2-5) covering two distinct research problems, the first being the development of a sampling methodology for pest managment (Ch. 2), and the second being an investigation into the traffic dynamics of morning commute patterns into a city center (Ch. 3-5). Of these four chapters, three of them (Ch. 2-4) have been published or have been submitted for publication to an academic journal. Hence, each chapter in the body of this dissertation stands alone as a complete research article with its own abstract, introduction, conclusions, and references, and also possibly notational glossary and appendices. For the purpose of constructing this dissertation, the individual research articles comprising each chapter have been reformatted so that the style and the labeling of figures and tables is consistent throughout the entire document.

The first research problem seeks to develop a sampling strategy to be used by pest managers in making a treatment decision for an orchard. Typically, once the mean pest density in an orchard has crossed a critical threshold the resulting economic damage to the crop warrants the application of a pest treatment. A relationship exists between the mean pest density in an orchard and the proportion of leaves infested with at least one pest, and an understanding of this relationship enables the use of a presence-absence sampling strategy, where a pest manager estimates the mean pest density in an orchard by determining the proportion of leaves in the orchard which are infested with at least one pest. Presence-absence sampling strategies result in huge sampling cost savings, and these savings can be further increased by utilizing sequential sampling strategies in lieu of fixed-size sampling strategies. However, a current limitation in developing sequential, presence-absence sampling strategies has been the necessary assumption that the proportion of infested leaves in an orchard is constant throughout the orchard. From empirical data we find that pest populations vary significantly throughout an orchard and

exhibit significant spatial correlation, invalidating the use of a naive sequential sampling strategy. However, we propose a spatial sampling strategy which diminishes the effect of spatial correlation sufficient to utilize a sequential, presence-absence sampling strategy, and which allows for pest populations to vary throughout the orchard. We demonstrate the effectiveness of our sampling methodology using empirical data from a mite pest of avocado orchards, although the principles behind our proposed sampling methodology are general and may be applied to other pest management or spatial sampling situations.

The second research problem investigates the traffic dynamics and economic equilibrium departure patterns of the morning traffic commute into a city center. Individual commuters attempt to minimize their commuting trip cost by balancing the cost of travel time versus the cost of arriving early (schedule delay). In the absence of any tolling policy, the natural equilibrium state is one in which all commuters suffer the same trip cost, termed the User Optimum (UO). However, in this equilibrium state the aggregate sum of trip costs of all commuters is not minimal, and an alternative equilibrium state is one in which the aggregate sum of trip costs is minimized, termed the Social Optimum (SO). The goal of a governing agency is to implement a tolling policy which increases the trip cost in such a way that the resulting natural equilibrium state is the SO, thereby minimizing the overall social cost of the morning commute.

There are three chapters (Ch. 3-5) which address the second research problem. In all three chapters we use the same macroscopic model of traffic flow, in which traffic flow resembles one-dimensional fluid flow with an imposed equation of state relating velocity to density. In Ch. 3 we determine the UO under the assumption that commuters may enter the roadway at any point, resulting in a continuum of entry points. In Ch. 4 we restrict commuters to enter the roadway at a single point, and building off of earlier work we determine the resulting SO and UO solutions and discuss their economic

properties. In Ch. 5 we develop a model of stochastically occurring incidents (i.e., unusual increases in congestion) on a roadway, calibrating this model by applying an online, change-point detection algorithm to detect incidents in traffic data from a freeway leading into downtown San Diego. We use model fitting techniques from the recent theory of Hierarchical Generalized Linear Models (HGLM), and we apply this model of stochastically occurring incidents to a simulation study to determine how the previously determined SO and UO solutions vary with the introduction of random incidents.

Chapter 2

Sequential Hypothesis Testing With Spatially Correlated Presence-Absence Data

Abstract

A pest management decision to initiate a control treatment depends upon an accurate estimate of mean pest density. Presence-absence sampling plans significantly reduce sampling efforts to make treatment decisions by using the proportion of infested leaves to estimate mean pest density in lieu of counting individual pests. The use of sequential hypothesis testing procedures can significantly reduce the number of samples required to make a treatment decision. Here we construct a mean-proportion relationship for *Oligonychus perseae*, a mite pest of avocados, from empirical data, and develop a sequential presence-absence sampling plan using Bartlett's sequential test procedure. Bartlett's test can accommodate pest population models which contain nuisance parameters which are not of primary interest. However, it requires that population mea-

surements be independent, which may not be realistic due to spatial correlation of pest densities across trees within an orchard. We propose to mitigate the effect of spatial correlation in a sequential sampling procedure by using a tree-selection rule (i.e., maximin) which sequentially selects each newly sampled tree to be maximally spaced from all other previously sampled trees. Our proposed presence-absence sampling methodology applies Bartlett's test to a hypothesis test developed using an empirical mean-proportion relationship coupled with a spatial, statistical model of pest populations, with spatial correlation mitigated via the aforementioned tree-selection rule. We demonstrate the effectiveness of our proposed methodology over a range of parameter estimates appropriate for densities of *O. perseae* that would be observed in avocado orchards in California, USA.

Key Words: Bartlett's Sequential Test, Binomial Sampling, Generalized Linear Mixed Models

2.1 Introduction

Neglecting the spatial structure of pest populations can result in an inaccurate estimation of pest densities. Spatial analyses have been previously used in studies of diverse groups of pests of agricultural crops such as lentils (Schotzko and Okeeffe, 1989), cotton (Gozé et al., 2003), grapes (Ifoulis and Savopoulou-Soultani, 2006) and grape varieties (Ramírez-Dávila and Porcayo-Camargo, 2008). In all these studies, spatial analyses were conducted by first transforming count data so as to resemble continuous, normally-distributed data. Generalized linear mixed models (GLMM), however, are statistical models which are particularly useful for modeling discrete response variables which may be correlated (Breslow and Clayton, 1993), such as spatially corre-

lated count data or presence-absence data. GLMM's have been used across multiple scientific disciplines, including ecological studies of pest populations (Candy, 2000, Bianchi et al., 2008, Takakura, 2009). In this paper, we propose a spatial GLMM for a sequential presence-absence sampling program for *Oligonychus perseae*, Tuttle, Baker, and Abatiello (Acari: Tetranychidae), a pest mite of avocados (Persea americana Miller [Lauraceae]) in California, USA, as an example for developing this modeling approach.

The persea mite, *O. perseae*, is native to Mexico and is an invasive pest in California, Costa Rica, Spain, and Israel. It is a foliar pest of avocados and is most damaging to the popular 'Hass' variety which accounts for 94% percent of the total production acreage in California (CAC, 2009), it is worth ∼$300 million each year, and ∼6,000 growers farm ∼27,000 ha of this cultivar (CAC, 2010). Feeding by high density populations of *O. perseae* can cause extensive defoliation to avocados (Hoddle et al., 2000), and in California this pest is typically controlled with pesticides (Humeres and Morse, 2005). A scientifically based action threshold and economic injury level has not been calculated for *O. perseae* in California. However, work from Israel suggests that the economic injury level lies between 100 and 250 mites per leaf and the recommended action threshold is in the range of 50-100 mites per leaf (Moaz et al., 2011).

Counting *O. perseae* mites with a hand lens in the field is tedious, time consuming, and an inaccurate approach to monitor population densities for making control decisions. An alternative approach is presence-absence or binomial sampling, which estimates pest population density using the proportion of leaves infested with at least one mite versus the proportion of clean leaves with no mites. Presence-absence sampling is fast, simple, and allows large areas to be surveyed quickly to quantify pest damage. Presence-absence sampling programs have been developed for a variety of agricultural pests including other spider mite species, eriophyid mites, aphids, flea beetles, leaf

hoppers, whiteflies, mealybugs and leaf miners (Alatawi et al., 2005, Binns et al., 2000, Galvan et al., 2007, Hall et al., 2007, Hyung et al., 2007, Kabaluk et al., 2006, Martinez-Ferrer et al., 2006, Robson et al., 2006).

Sequential sampling procedures are considered a cost effective approach to assessing pest densities (Mulekar et al., 1993, Young and Young, 1998, Binns et al., 2000). Cost savings accrue in comparison to fixed sample size procedures, because sequential procedures often require a significantly reduced number of sampled observations to reach a treatment decision, which can result in appreciable savings in the cost of sampling. In applications of sequential sampling, Wald's (Wald, 1947) sequential probability ratio test (SPRT) is the most often used approach. Wald's SPRT is useful for sampling programs when it can be assumed that, aside from the primary parameter of interest, there are no additional unknown parameters (i.e., nuisance parameters) in the model. In the case of independent and identically distributed (IID) samples, a modification to Wald's SPRT results in Bartlett's (Bartlett, 1946) SPRT, which can be applied to pest count models containing nuisance parameters (Shah et al., 2009). However, spatial correlation of pest populations violates the independence assumption required for Bartlett's SPRT. In related work on spatially correlated pest count data, Li et al., 2011 proposed a first-stage initial sample used to assess the effective range of spatial correlation, followed by a second-stage sampling procedure in which each sampled observation is outside of the effective range of all previously sampled observations. Sampling outside of the effective range eliminates any spatial correlation so that Bartlett's SPRT may be applied. In this paper, we propose to sequentially sample observations for *O. perseae* so that each sampled observation is maximally spaced from all other previously sampled observations, thereby eliminating spatial correlation. This sampling strategy eliminates the necessity of an initial, first-stage sample as proposed by Li et al., 2011, and we demonstrate its

effectiveness for mitigating spatial correlation sufficiently to allow the application of Bartlett's SPRT for a range of parameter estimates appropriate to *O. perseae* in California avocado orchards. To our knowledge, this paper is the first to combine sequential hypothesis testing techniques with presence-absence sampling strategies which account for spatial correlation of pest densities.

2.2 Materials and Methods

2.2.1 Mean-Proportion Relationship

The essential component of a presence-absence sampling plan is an accurate relationship between the mean pest density, M, and the proportion of leaves infested with at least one pest individual, P. The mean-proportion relationship can be modeled using an empirical equation (Kono and Sugino, 1958, Gerrard and Chaing), which has been used to develop binomial sampling plans for pests as in Hall et al., 2007 and Martinez-Ferrer et al., 2006,

$$\ln\left(-\ln\left(1-P\right)\right) = a + b \cdot \ln\left(M\right) \tag{2.1}$$

The parameters a and b can be fit using linear regression.

To construct a mean-proportion relationship for *O. perseae*, Hass avocado leaves were collected randomly from 9 avocado orchards in Southern California across various years (Table 2.1), and counts of all *O. perseae* stages (except eggs) were performed using stereomicroscopes. 72 mite count data sets (incorporating 30,656 leaves with a density range of 0-342 mites per leaf) were used to fit (2.1), with resulting parameter estimates $a = -1.72762$ and $b = 0.66527$. This relationship is shown in Fig. 2.1

Orchard	County	Year sampled	Trees sampled	Number of leaves	Number of sets
1	Ventura	1997	42	6,469	16
2	Orange	1999	66	5,280	8
2	Orange	2000-01	42	17,220	41
3	San Diego	2009	31	247	1
4	Santa Barbara	2009	30	240	1
5	Santa Barbara	2009	30	240	1
6	Santa Barbara	2009	30	240	1
7	Santa Barbara	2009	30	240	1
8	Ventura	2010	30	240	1
9	Ventura	2010	30	240	1

Table 2.1: Summary information for the avocado orchards in California from which count data were collected to construct a mean-proportion relationship for *Oligonychus perseae*. A total of 72 data sets, each measuring the proportion of infested leaves and the mean leaf mite density, were used to fit the empirical equation, (2.1). The 72 data points and the resulting fitted curve are graphed in Fig. 2.1.

where we plotted the 72 data pairs of mean pest density per leaf and proportion of infested leaves, along with the fitted empirical equation, (2.1).

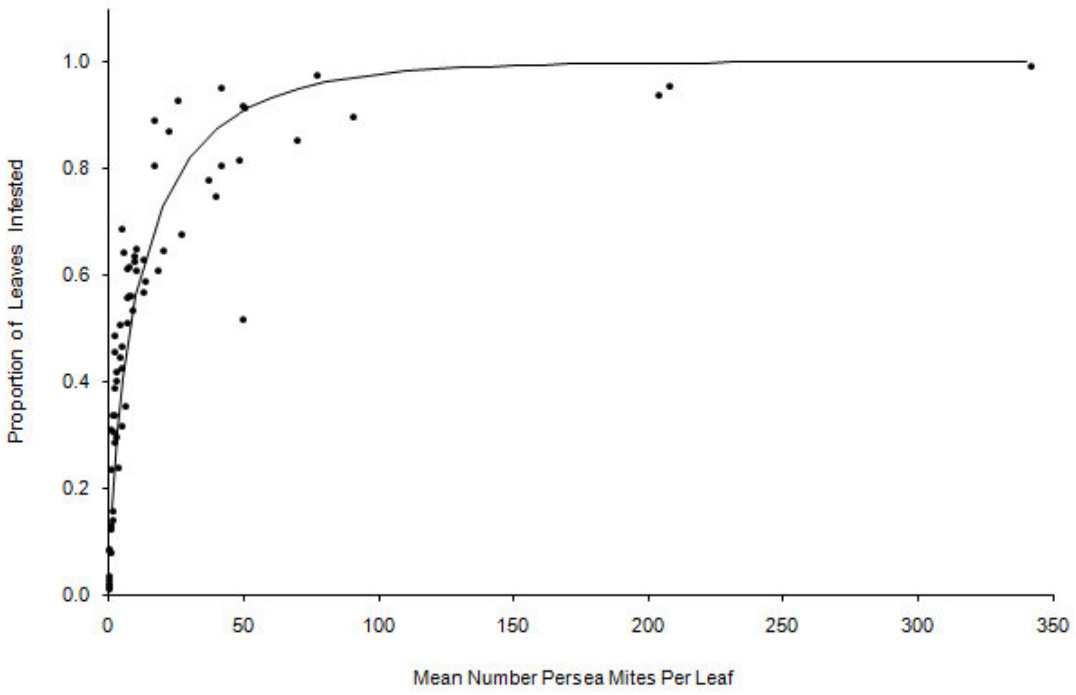

Figure 2.1: Plotted values of data sets for *Oligonychus perseae*, each measuring the proportion of infested leaves and the mean mite density per leaf, and a graph of the fitted empirical equation, (2.1).

2.2.2 Presence-Absence Sampling Hypothesis Test

The mean-proportion relationship allows a pest control adviser to estimate the mean density of mites per leaf without counting individual mites. This is achieved by sampling a number of leaves and determining the proportion of leaves for which at least one mite is present. In our context, we use the mean-proportion relationship to convert an action threshold for mite densities per leaf into an action threshold for proportion of infested leaves.

Moaz et al., 2011 determined an action threshold range for *O. perseae* mite densities to be 50-100 mites per leaf. Using the lower bound of this range we construct a statistical decision problem for intervention treatment by proposing the following hypothesis test for mean mite densities, M:

$$H_0 : 25 \frac{\text{mites}}{\text{leaf}} \quad \text{vs.} \quad H_1 : 75 \frac{\text{mites}}{\text{leaf}}$$

Since the midpoint of this range is 50 mites/leaf, if the null hypothesis, H_0, is rejected in favor of H_1, then mite densities are above 50 mites/leaf and treatment is recommended. Failure to reject H_0 in favor of H_1 implies that mite densities are below 50 mites/leaf and treatment is unnecessary.

Using (2.1) we may explicitly write P as a function of M,

$$P = 1 - \exp\left\{-\exp\left(a + b \ln M\right)\right\}, \tag{2.2}$$

and using the above parameter estimates for a and b we convert the hypothesis test for

M into the following hypothesis test for P:

$$H_0 : P = 0.78 \quad \text{vs.} \quad H_1 : P = 0.95.$$

In a binomial sampling plan, the number of infested leaves (i.e., leaves with at least one mite present) in a randomly selected sample of leaves follows a binomial distribution, with the only unknown model parameter being the proportion of infested leaves, P. The hypothesis test for P may be evaluated using Wald's SPRT, which is the most efficient procedure for evaluating a hypothesis test of simple hypotheses when the underlying model contains only a single, unknown parameter. However, from our field observation and analysis of *O. perseae* count data (Li et al., 2011), mite populations were shown to cluster on individual avocado trees, with neighboring trees having similar population densities. This spatial correlation did not exist once sampled trees were 3-4 trees distant from the last sampled tree (Li et al., 2011). Thus, to more accurately evaluate the above hypothesis test and make a decision regarding treatment, a model must be developed that accounts for the aggregation of mites on individual trees and the spatial correlation of mite densities across trees.

2.2.3 Spatial GLMM

To account for the aggregation of mites on individual trees we constructed a Bernoulli response GLMM in which the proportion of infested leaves varies by tree, as determined by a fixed effect common to all trees and a random effect which varies from tree to tree. To account for the spatial correlation of mite densities among trees we allow the random tree effects in the GLMM to be spatially correlated. Specifically, suppose that we select n trees to be sampled, and that on each tree we randomly sample m leaves.

For $i = 1, \ldots, n$, on the ith tree let p_i be proportion of infested leaves, $0 \leq p_i \leq 1$, and let Y_{ij} be the corresponding Bernoulli(p_i) response for the jth leaf sampled, $j = 1, \ldots, m$, where $Y_{ij} = 1$ if at least one mite is present and $Y_{ij} = 0$ otherwise. Let γ denote the fixed effect common to all trees, let $\mathbf{S} = (S_1, \ldots, S_n)'$ denote the spatially correlated random tree effects for the n trees sampled, and let Y_i equal the sum of the m Bernoulli responses for the ith tree, $Y_i = \sum_{j=1}^{m} Y_{ij}$. Therefore, our proposed spatial GLMM is defined as:

$$\begin{aligned} Y_i \,|\, \mathbf{S} &\sim \text{Binomial}(m, p_i) \\ \text{logit}(p_i) \equiv \log\left(\frac{p_i}{1-p_i}\right) &= \gamma + S_i \\ \mathbf{S} &\sim MVN\,(\mathbf{0},\, \Sigma) \end{aligned} \tag{2.3}$$

where Σ is the $n \times n$ covariance matrix for the random tree effects whose off-diagonal elements determine the correlation structure. We propose allowing for a spatially symmetric correlation structure in which the correlation between the random effects of two trees decreases exponentially with the distance between the trees, known as a spatial exponential correlation structure (Schabenberger and Gotway, 2005). With this correlation structure, the (i, i') element of Σ is $\sigma^2 \exp\left(\frac{-d_{i,i'}}{\rho}\right)$, where $d_{i,i'}$ is the Euclidean distance between the i-th and i'-th trees, ρ is a scale parameter that dictates the strength of the spatial correlation, and σ^2 is a scale parameter that determines the variability of the random tree effect on an individual tree. Under this parameterization it can easily be shown that the effective range of the spatial correlation is 3ρ, and that for tree-separation distances beyond this range the spatial correlation is essentially diminished (Schabenberger and Gotway, 2005).

2.2.4 Spatial GLMM Hypothesis Test

It follows from (2.3) that for each tree the proportion of infested leaves, p_i, is a logit-normal random variable with parameters γ and σ^2. Although the mean of a logit-normal random variable cannot be analytically related to its parameters, a simple analytic relation exists between γ and the median of p_i,

$$\gamma = \log\left(\frac{\text{median}(p_i)}{1-\text{median}(p_i)}\right). \tag{2.4}$$

In the spatial GLMM model the proportion of infested leaves varies from tree to tree, and a pest manager seeking to make a treatment decision for an entire orchard may use median(p_i) as a measure of the proportion of infested leaves over the entire orchard. Thus, using the spatial GLMM the hypothesis test we previously derived for *O. perseae* in terms of P may be converted into a hypothesis test for γ as follows:

$$H_0 : \gamma = \log\left(\frac{0.78}{1-0.78}\right) = 1.27 \quad \text{vs.} \quad H_1 : \gamma = \log\left(\frac{0.95}{1-0.95}\right) = 2.94. \tag{2.5}$$

Hence, the median pest density over an entire orchard is determined by the spatial GLMM parameter γ , whereas σ^2 and ρ are nuisance parameters.

2.2.5 Bartlett's SPRT

In a model without nuisance parameters, Wald's SPRT is the most efficient test of simple hypotheses, requiring the minimum number of expected samples among all hypothesis tests with the same Type-I (falsely reject H_0) and Type-II (falsely fail to reject H_0) error rates. In a model which contains nuisance parameters, Bartlett, 1946 proved that, if the samples are independent and identically distributed (IID), then the Type-I,

II error rates are asymptotically preserved if the nuisance parameters are replaced with their conditional maximum likelihood estimates at each stage of the sequential testing procedure.

In the context of this study, the IID assumption of Bartlett's SPRT is achieved if spatial correlation is not present, and in a subsequent section we propose a tree-selection rule which effectively diminishes any spatial correlation. Hence, throughout this section we presume that our proposed spatial GLMM has been reduced to a GLMM with no spatial correlation ($\rho = 0$) to which Bartlett's SPRT may be applied.

We apply Bartlett's SPRT to the observations $\{Y_1, Y_2, \ldots\}$, where Y_i is the number of mite-infested leaves on the i-th tree among the m leaves sampled, with m determined in the next section. The sequential test for subsequent sampling occasions is based on the log-likelihood ratio,

$$\lambda_n = \log \left\{ \frac{f\left(\mathbf{Y_n};\, \gamma_1, \hat{\sigma}_n^2(\gamma_1)\right)}{f\left(\mathbf{Y_n};\, \gamma_0, \hat{\sigma}_n^2(\gamma_0)\right)} \right\}. \tag{2.6}$$

Here, n denotes the current number of trees sampled in the sequential procedure, $\mathbf{Y_n} = (Y_1, \ldots, Y_n)'$ are the current observed responses, and the likelihoods are obtained from

(2.3) by integrating out the random effects, $\mathbf{S_n} = (S_1, \ldots, S_n)'$, assuming $\rho = 0$,

$$
\begin{aligned}
f\left(\mathbf{Y_n}; \gamma, \sigma^2\right) &= \int \cdots \int_{\mathcal{R}^n} f\left(\mathbf{Y_n}|\mathbf{S_n}\right) \cdot f\left(\mathbf{S_n}\right) \, d\mathbf{S_n} \\
&= \int \cdots \int_{\mathcal{R}^n} \left[\prod_{i=1}^{n} \binom{m}{Y_i} p_i^{Y_i} (1-p_i)^{m-Y_i} \right] \cdot \left[\frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left(-\frac{1}{2} \mathbf{S_n}' \Sigma^{-1} \mathbf{S_n} \right) \right] d\mathbf{S_n} \\
&= \prod_{i=1}^{n} \left\{ \int_{\mathcal{R}^n} \binom{m}{Y_i} \left(\frac{\exp(\gamma + S_i)}{1 + \exp(\gamma + S_i)} \right)^{Y_i} \left(1 - \frac{\exp(\gamma + S_i)}{1 + \exp(\gamma + S_i)} \right)^{m-Y_i} \right. \\
&\qquad \left. \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} S_i^2 \right) \, dS_i \right\}. \qquad (2.7)
\end{aligned}
$$

(2.7) consists of a product of n one-dimensional integrals, each of which is easily numerically evaluated using Gauss-Hermite quadrature. For $\gamma \in \{\gamma_0, \gamma_1\}$, $\hat{\sigma}_n^2(\gamma)$ denotes the conditional MLE of the unknown nuisance parameter σ^2, obtained by setting γ in (2.7) to γ_0 or γ_1, respectively, and then maximizing the right-hand side with respect to σ^2. For the hypothesis test in (2.7) used to make a treatment decision for *O. perseae*, $\gamma_0 = 1.27$ and $\gamma_1 = 2.94$.

The upper and lower stopping boundaries of Bartlett's SPRT are

$$
A = \log\left(\frac{\beta}{1-\alpha} \right) \quad \text{and} \quad B = \log\left(\frac{1-\beta}{\alpha} \right), \qquad (2.8)
$$

respectively, so that Bartlett's SPRT rejects H_0 in favor of H_1 at the first n for which $\lambda_n \geq B$, fails to reject H_0 in favor of H_1 at the first n for which $\lambda_n \leq A$, and continues by sampling another tree if $A < \lambda_n < B$. The resulting Type-I and Type-II error rates asymptotically satisfy $P\left(\text{Reject } H_0 | H_0\right) \leq \alpha$ and $P\left(\text{Fail to reject } H_0 | H_1\right) \leq \beta$, respectively, so that α and β are Type-I, II error rate upper bounds, respectively.

2.2.6 Leaf-Selection Rules and Sampling Cost

To determine the optimal number of leaves to sample per tree, m, and assuming that our tree-selection rule has effectively diminished spatial correlation ($\rho = 0$), we conducted a simulation study to analyze average sample numbers (ASN) of Bartlett's SPRT applied to the hypothesis test in (2.6), over a range of m and σ^2 parameter values appropriate to *O. perseae*, and Type-I,II error rate upper bounds of $\alpha = 0.10$ and $\beta = 0.10$.

Let N denote the number of sampled trees required to reach the stopping rule in Bartlett's SPRT. As the number of leaves sampled per tree, m, increases, the expected number of sampled trees, $E(N)$, decreases, but the expected total number of sampled leaves, $m \cdot E(N)$, increases (see the Results section for details). To determine an optimal value for m, we constructed a simple sampling cost function which includes a sampling cost for each tree and an additional sampling cost for each leaf:

$$\text{Cost} \;=\; (\text{cost per tree}) \;\cdot N + \;(\text{cost per leaf}) \;\cdot m \cdot N. \tag{2.9}$$

For a given value of m the expected cost, $E(\text{Cost})$, depends on $E(N)$, which varies with γ. For each value of m, we evaluate $E(N)$ at the value of γ, say $\gamma_{\max}$, for which $E(N)$ is maximized. We choose m to minimize the expected cost, which up to a constant of proportionality can be written as:

$$E(\text{Cost}) \propto (1 + cm) \cdot E(N)|_{\gamma_{\max}}, \text{ where } c = \frac{\text{cost per leaf}}{\text{cost per tree}}. \tag{2.10}$$

In practice, the costs associated with selecting an additional leaf should be much less than the costs associated with selecting and locating an additional tree, so that the

leaf-to-tree cost ratio, c, should be much less than one. Given a value of $c < 1$, $E(\text{Cost})$ vs. m is plotted and in the resulting graph an optimal value of m is chosen so as to minimize $E(\text{Cost})$.

2.2.7 Sequential Maximin Tree-Selection Rule

To mitigate spatial correlation of mite counts between adjacent trees we propose to sequentially select each tree to be maximally spaced from all other previously selected trees. We base our notion of 'maximally spaced' on a maximin distance criterion, in which each tree is selected so as to maximize the minimum distance it has to all other previously selected trees. A design constructed by this rule has been referred to as a 'coffee-house design for the similar way in which customers select their tables in a coffee-house (Müller, 2007).

In a non-sequential, fixed-size spatial sampling setting, maximin designs possess optimality properties which we now briefly describe. In a fixed-size sampling setting, a maximin design simultaneously selects all points so that the minimum distance between all pairs of selected points is maximized. The index of a fixed-size maximin design is the number of pairs of points separated by this maximal, minimum distance. For any statistical model in which the correlation between two points is a decreasing function of the distance between the two points, a fixed-size maximin design of smallest index is asymptotically related to an optimal design which minimizes the variances of parameter estimates (Johnson et al., 1990). This result enables the construction of an asymptotically optimal fixed-size sampling design based on geometric criteria alone.

In the context of this paper, we adopt the above notion of 'index' to a sequential, maximin tree-selection rule, as follows. At each stage in the sequential procedure, we define a maximin tree to be a tree (not necessarily unique) whose minimum distance

to all previously selected trees is maximal. The index of a maximin tree is defined to be the number of previously selected trees separated by this maximal, minimum distance. Our proposed sequential, maximin tree-selection rule is to select a maximin tree of smallest index.

In Fig. 2.2 we provide a visual illustration of the sequential, maximin tree-selection rule applied to a 20x20 grid of equally spaced trees, demonstrating how the first 13 trees are selected.

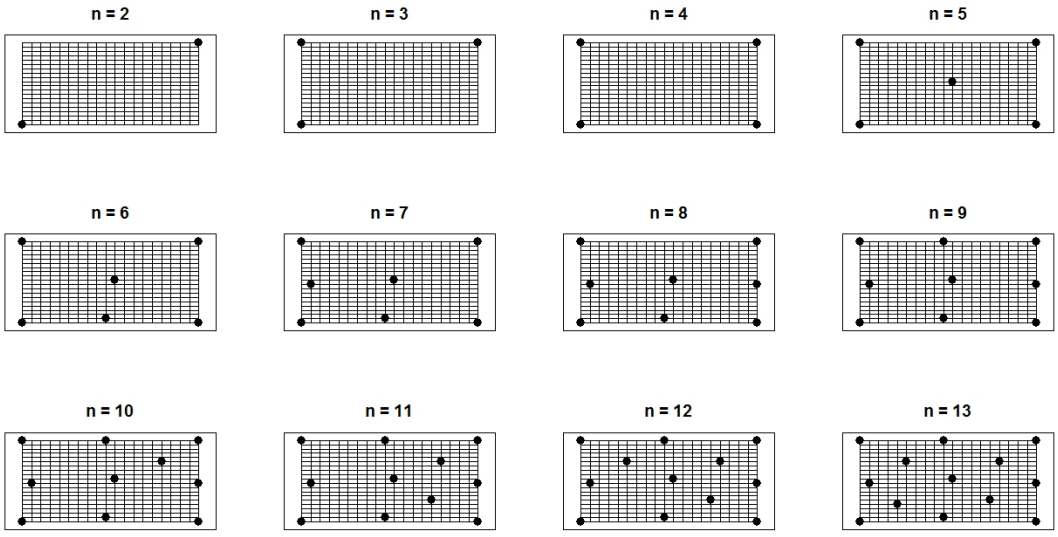

Figure 2.2: Visual illustration of the sequential, maximin tree-selection rule applied to a 20x20 grid of trees, demonstrating how the first 13 trees are selected. Here we arbitrarily chose the first tree selected to be the lower, left-hand corner tree.

2.2.8 Evaluation of Proposed Methodology

Our proposed methodology for developing a presence-absence sampling plan is to use the mean-proportion relationship coupled with the spatial GLMM to construct the treatment decision hypothesis test in (2.5), to which we apply Bartlett's SPRT coupled with the sequential, maximin tree-selection rule and the leaf-selection rule. We validate

the proposed methodology by verifying that the sequential, maximin tree-selection rule successfully diminishes spatial correlation sufficient to preserve the Type-I, II error rates of Bartlett's SPRT applied to the hypothesis test in (2.5), for a range of σ^2 and ρ parameter values appropriate to *O. perseae*.

In a simulation study we simulated presence-absence data from a spatial GLMM for a range of values of the spatial correlation parameter, ρ, and the nuisance parameter, σ^2, appropriate to *O. perseae*. Assuming the optimal leaf selection rule of $m = 6$ leaves per tree (see Results section), we simulated data from a 20x20 grid of 400 equally spaced trees. For each simulation we evaluated the hypothesis test in (2.5) by applying Bartlett's SPRT with Type-I,II error rate upper bounds of $\alpha = 0.10$ and $\beta = 0.10$. However, we truncated Bartlett's SPRT so that the maximum possible number of trees sampled is 10% of the orchard, or 40 trees in this example. If a stopping rule had not been reached after 40 trees had been sampled, then the sequential procedure was halted and a decision made based on whether the sequential hypothesis test statistic, λ_{40}, was closer to B, the stopping rule upper boundary (reject H_0), or closer to A, the stopping rule lower boundary (fail to reject H_0).

We compared the sequential, maximin tree-selection rule to several other tree-selection rules, all of which are illustrated in Fig. 2.3: (A) border selection, where trees were sampled along the orchard borders; (B) diagonal selection, where trees were sampled along a diagonal in the orchard; (C) zigzag selection, where the lower orchard border is sampled, followed by the orchard diagonal, followed by the upper orchard border; (D) grid selection, where trees were sampled on a grid pattern uniformly spaced throughout the orchard; (E) SRS selection, where trees were selected using simple random sampling throughout the orchard. In Fig. 2.3 we indicate the order in which trees were selected by shading from darkest to lightest. Although all 40 trees are designated for each truncated

sequential hypothesis test, in practice the average number of trees sampled to reach a stopping rule typically ranged between 5 and 10 trees.

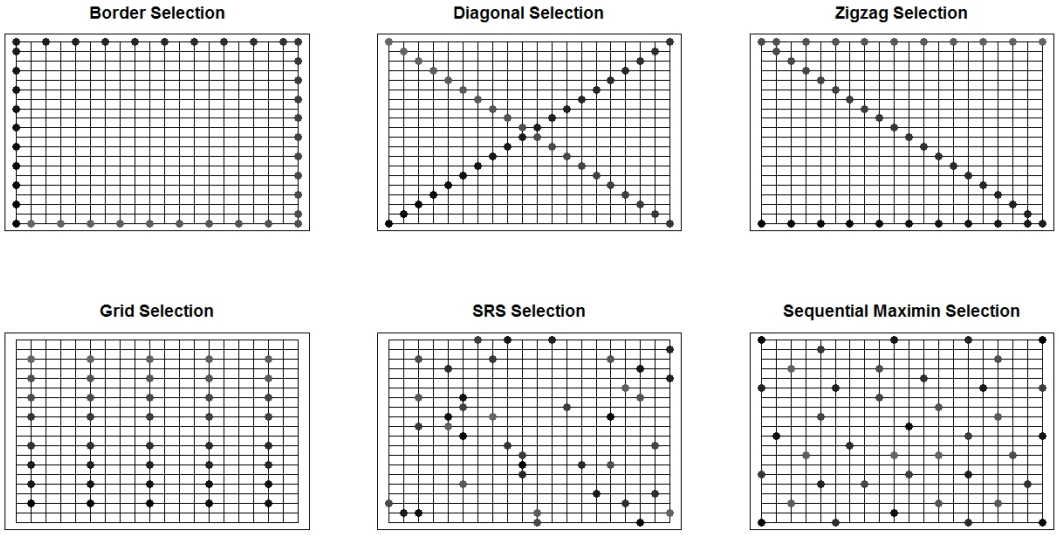

Figure 2.3: The six tree-selection rules for which we evaluated the error rates of Bartlett's SPRT. The shading indicates the order of tree selection from darkest to lightest. We truncated the sequential hypothesis test at an upper bound of 40 trees, all of which are graphed here. However, in all cases the sequential hypothesis test typically terminated after sampling the first 5-10 trees.

We caution the reader to distinguish between the SRS tree-selection rule, which at each sequential step randomly selects a tree from all remaining trees over the entire orchard, and what might be referred to as a random tree-selection rule in which a pest manager walks through a grove haphazardly, randomly selecting trees to sample. Since this latter type of tree-selection rule does not sequentially select trees to be spaced far apart, our results from patterned tree-selection rules suggest that it will not mitigate spatial correlation sufficiently to apply Bartlett's sequential test.

Orchard	n	m	γ	σ^2	ρ
4	30	8	1.53	1.67	1.06
5	60	8	5.48	0.00024	0.073
8	400	4	3.24	1.23	4.37
10	402	4	-0.85	0.87	1.32

Table 2.2: Parameter estimates for four sets of *Oligonychus perseae* presence-absence data fitted to the spatial GLMM model, (2.3). For each data set, n is the number of trees and m is the number of leaves sampled per tree. In each orchard, trees were approximately equally spaced on a grid, and in fitting the spatial GLMM distance is measured in tree-separation units. Note that in the second data set (Orchard 5) mites were present on nearly every sampled leaf.

2.3 Results

2.3.1 Illustrated Examples: Sample Parameter Estimates

Various statistical software packages implement model fitting and parameter estimation for GLMM's, such as SAS Proc Glimmix. To provide realistic parameter estimates for *O. perseae* distributions in avocado orchards we fitted the spatial GLMM model, (2.3), to four presence-absence sets of data, with the fitted parameters provided in Table 2.2. Based on these estimates, in the simulation studies we allowed σ^2 to vary from 0.5 to 2.0, and ρ to vary from 0 to 5.0.

2.3.2 Leaf-Selection Rules: Outcome

Fig. 2.4 shows ASN curves for the expected number of sampled trees, and Fig. 2.5 shows ASN curves for the expected total number of sampled leaves, where each point was obtained using 20,000 simulations. We observe that as more leaves per tree were sampled (i.e., as m increases), one expects to sample fewer trees but more total leaves. The ideal choice for m minimizes the expected cost, $E(\text{Cost})$, which depends upon the leaf-to-tree cost ratio, c. In Fig. 2.6, $E(\text{Cost})$ is plotted for several values of $c < 1$, $c = 0.01, 0.10, 0.25$ and 0.50. We observe that as m increases beyond 6 leaves

per tree the expected cost does not significantly decrease for smaller values of c and σ^2, and increases for larger values of c and σ^2. Thus, for *O. perseae* we conclude that, if spatial correlation has been effectively diminished, then an ideal leaf-selection rule for evaluating the hypothesis test in (2.5) which applies to a range of parameter values and leaf-to-tree sampling cost ratio values is to randomly select $m = 6$ leaves per tree.

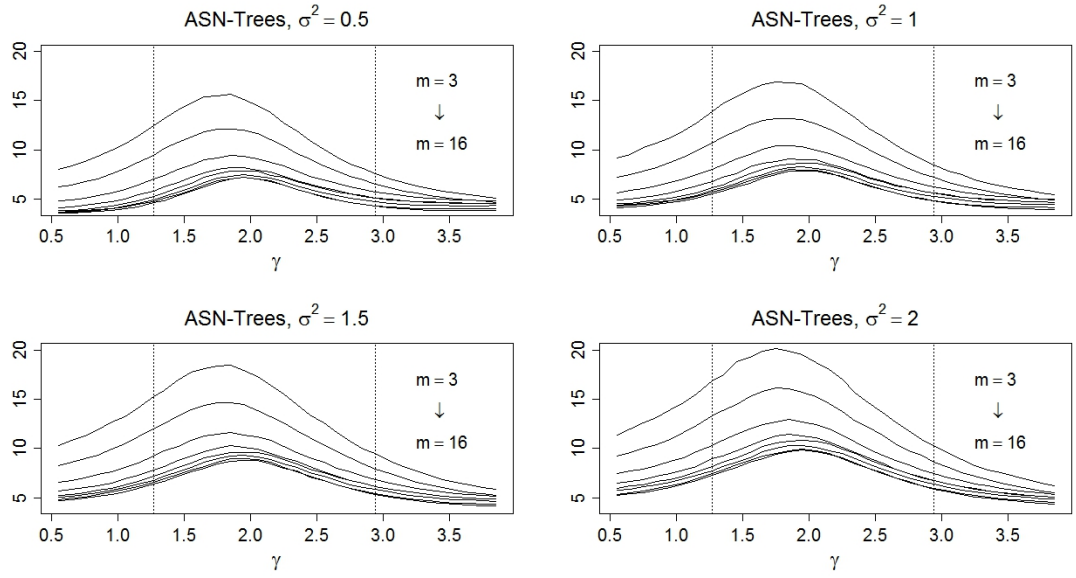

Figure 2.4: ASN curves for the expected number of sampled trees in Bartlett's SPRT. The number of leaves sampled per tree, m, ranges between 3 (upper curve), 4, 6, 8, 10, 12, 14 and 16 (lower curve).

2.3.3 Evaluation of Proposed Methodology: Outcome

Figs. 2.7 and 2.8 display the results of our simulation study, which show how the observed Type-I and Type-II errors vary in the truncated sequential hypothesis test as the strength of spatial correlation increases from 0 to 5.0. Each point was obtained using 20,000 simulations, and the percentage of simulations for which the stopping rule was not reached after sampling 40 trees was negligibly small, never exceeding 1.5%. All

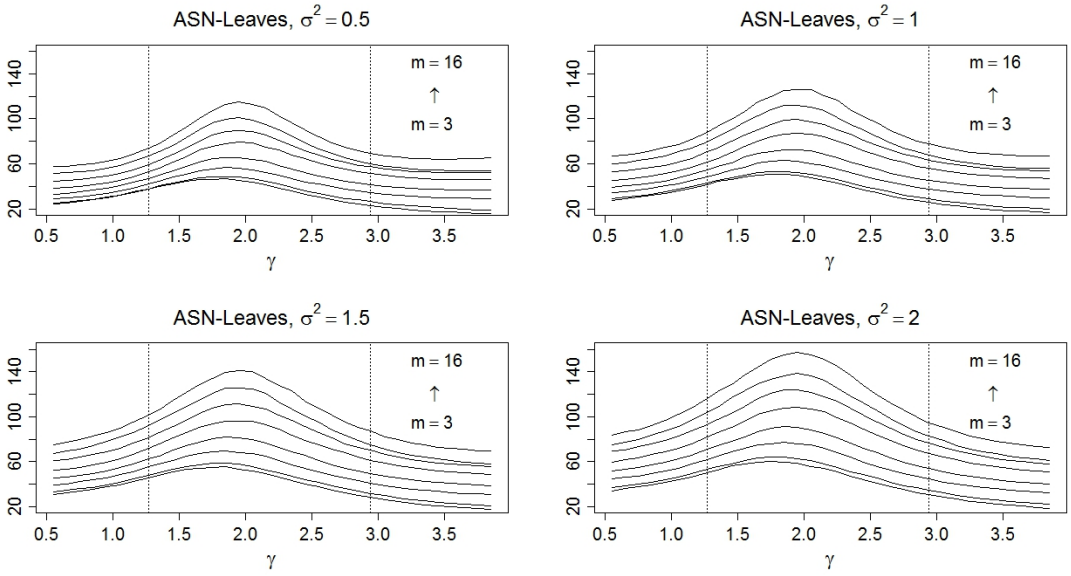

Figure 2.5: ASN curves for the expected number of sampled leaves in Bartlett's SPRT. The number of leaves sampled per tree, m, ranges between 3 (lower curve), 4, 6, 8, 10, 12, 14 and 16 (upper curve).

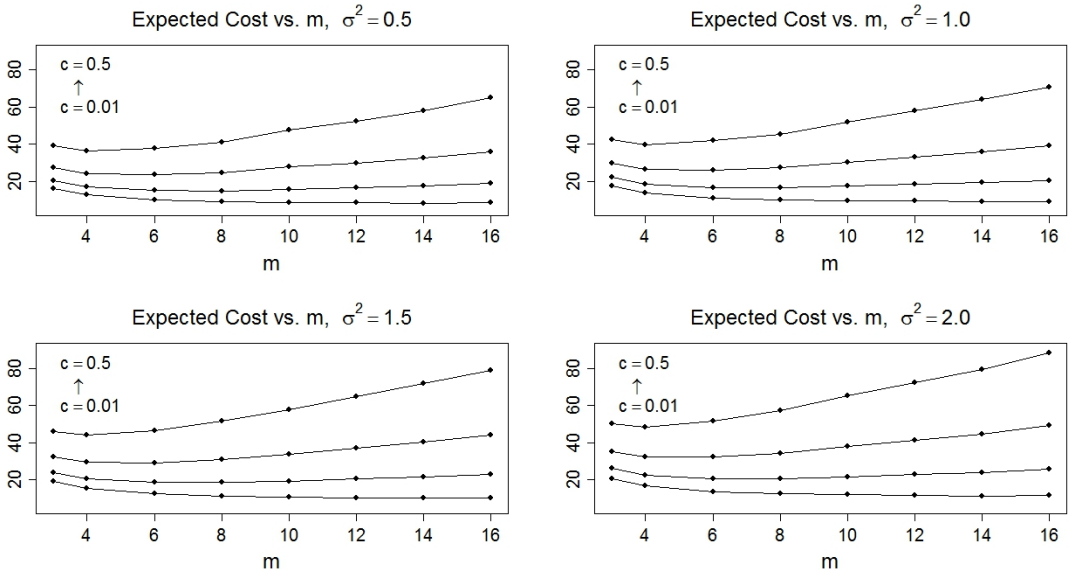

Figure 2.6: Expected sampling cost vs. the number of leaves selected per tree, m, where the leaf-tree sampling cost ratio, $c = \frac{\text{cost per leaf}}{\text{cost per tree}}$, ranges between 0.01 (lower curve), 0.10, 0.25 and 0.50 (upper curve).

of the patterned tree-selection rules show strong inflations of the observed Type-I, II error rates, from which we conclude that patterned tree-selection rules cannot be used in Bartlett's SPRT if spatial correlation is present. Although the SRS tree-selection

rule performs better than the patterned tree-selection rules, the sequential, maximin tree-selection rule outperforms all other tree-selection rules, preserving the 10% Type-II error rate over the range of parameters tested, and preserving the 10% Type-I error rate up to a spatial correlation strength of $\rho = 2.0$.

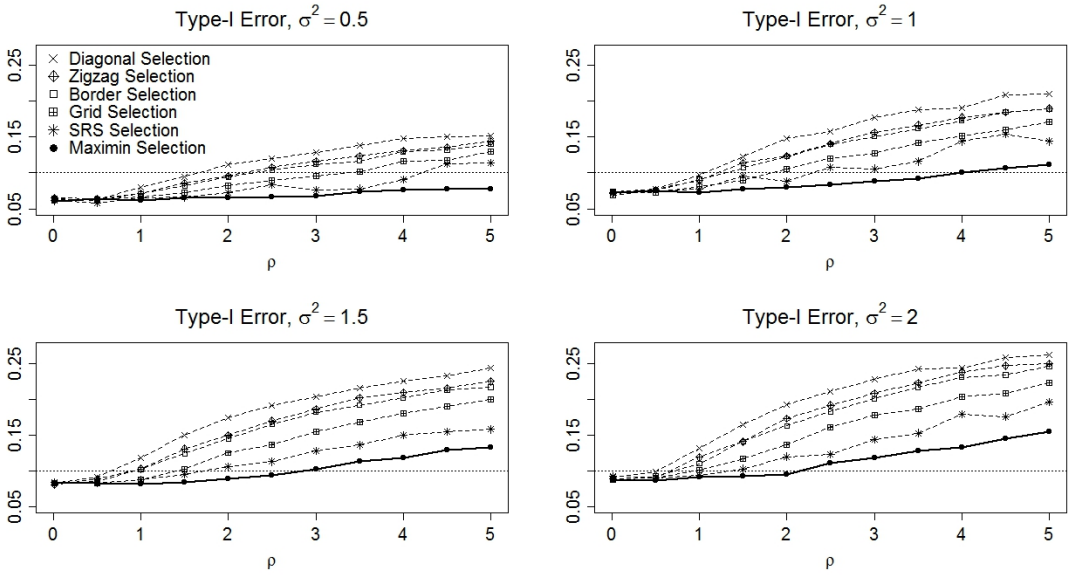

Figure 2.7: Observed Type-I error rates for Bartletts SPRT for data simulated with correlation parameter ρ . The solid curve corresponds to the sequential, maximin tree-selection rule, the dashed curves correspond to the patterned and SRS tree-selection rules, and the horizontal dotted line is the theoretical error rate upper bound of $\alpha = 0.10$.

The estimates for the spatial correlation parameter reported in Li et al. 2011, based on count data, ranged from 0.24 to 1.55, so that a reasonable range of study for ρ was taken to be 0-2.0. Our presence-absence data analyses suggest allowing ρ to increase up to 5.0. More typically, we do not expect to achieve values beyond 2.0, but this extended range was used to introduce robustness into our conclusions. In Figs. 2.7 and 2.8, as ρ ranges from 0-2.0 we see that our proposed methodology consistently preserves the Type-I and Type-II error rates. Even if the spatial correlation is as high as 5.0, the proposed methodology still preserves the Type-II error rates, although the

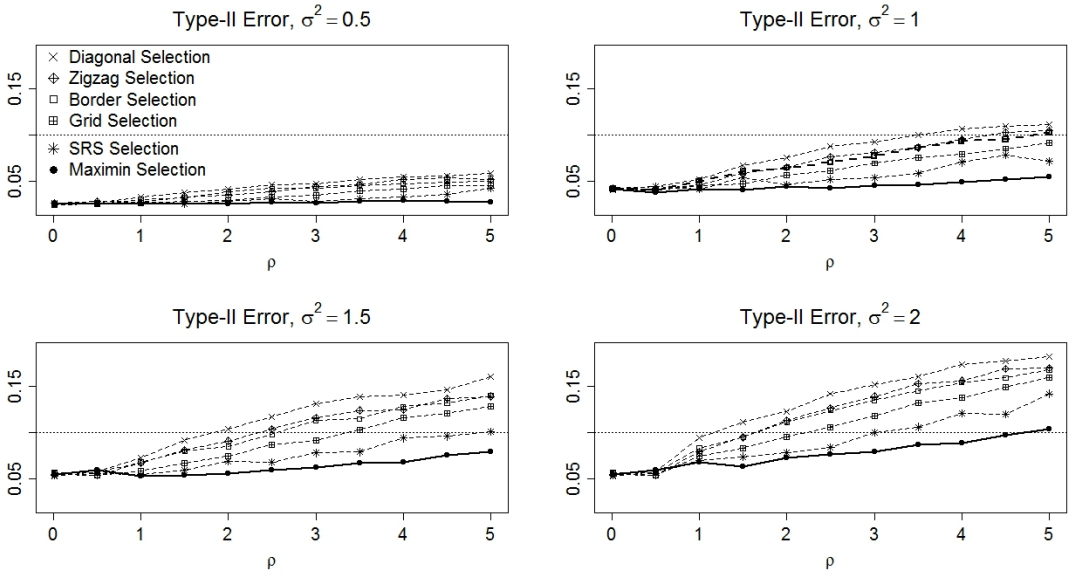

Figure 2.8: Observed Type-II error rates for Bartletts SPRT for data simulated with correlation parameterρ . The solid curve corresponds to the sequential, maximin tree-selection rule, the dashed curves correspond to the patterned and SRS tree-selection rules, and the horizontal dotted line is the theoretical error rate upper bound of $\beta = 0.10$.

Type-I error rates become slightly elevated. In the context of making a treatment decision based on an action threshold, making a Type-II error corresponds to failing to treat an orchard for *O. perseae* when mite densities are greater than 50 mites/leaf and treatment is necessary, and making a Type-I error corresponds to treating an orchard when mite densities are below 50 mites/leaf and treatment is unnecessary. Thus, using our proposed methodology, even under high levels of spatial correlation, a pest manager will not fail to treat a grove needing treatment, but may conservatively treat a grove for which treatment is not required.

This simulation study confirms the effectiveness of the proposed methodology for the range of parameter values appropriate to *O. perseae*. In particular, the methodology proposed here eliminates the need for an initial pilot sample as suggested by Li et al. 2011.

2.4 Discussion

The ultimate purpose of developing a sampling plan is to provide an easy to use tool for pest managers to use to allow them to quickly and accurately reach decisions on whether or not avocado orchards need to be treated for *O. perseae*, an important foliar mite pest of avocados in California, Mexico, Costa Rica, Spain, and Israel. Because a reliable sampling tool does not exist, IPM programs for *O. perseae* in California are relatively non-existent and it is likely that numerous pesticide applications are applied annually for the control of this pest when they are not needed. Analysis of pesticide use trends in California avocados shows a remarkably rapid increase in pesticide applications following the invasion of *O. perseae* in 1990 (Hoddle, 2004), and the adoption of a sampling plan similar to that proposed here may help reverse this trend by reducing the rate of unnecessary applications for this pest.

The work presented here is the first statistical application of spatial analyses coupled with sequential sampling for the development of a sampling plan for pest management. Our proposed presence-absence sampling methodology for *O. perseae* evaluates a sequential hypothesis test of pest population densities which, 1) accounts for aggregation of pest populations on individual trees, and 2) mitigates spatial correlation of pest populations on adjacent trees using a tree-selection rule which sequentially selects trees to be maximally spaced from all other previously selected trees (sequential, maximin tree-selection). Based on a simulation study we determined that the expected sampling cost is essentially minimized with a random selection of $m = 6$ leaves per tree, and based on a separate simulation study of Bartlett's SPRT with 10% Type-I, II error rates, we demonstrated that the sequential, maximin tree-selection rule preserves the error rates in the presence of spatial correlation, with average sample numbers for

the sequential test being 5-10 trees. Although our results demonstrate the effectiveness of our presence-absence sampling methodology for parameter estimates relevant to *O. perseae*, the methodology can easily be applied to other pests, and even other non-pest spatial sampling situations.

The results of the simulations conducted here demonstrate the effectiveness of our spatial presence-absence sampling methodology for parameter estimates relevant to *O. perseae* in California avocado orchards. With further research involving field validation, our sampling model has the potential to be customized as a reliable decision-making tool for pest control advisers and growers to use for control of this mite in commercial avocado orchards. To meet this goal, software would be needed to help a pest manager with tree selection and with evaluating the treatment decision hypothesis test at each sequential step. A component of any new technology is end-user adoption, especially if underlying concepts appear difficult and application potentially complicated. With the widespread ownership and use of smart phones, sampling programs like that developed here could be made available as a downloadable "app." This has several major attractions for users: (1) by following simple sampling instructions on a screen (such as GPS directions to the next tree to sample) and punching in sampling data (yes or no for the presence or absence of *O. perseae* for each sampled leaf), user uncertainty about sampling methodology (both tree and leaf selections) and correct calculations and interpretation of outcomes are potentially minimized. (2) Smart phone apps would return management decisions in real time and can be immediately emailed to a supervisor. Photos and GPS coordinates generated by the smart phone could also be included in reports if extra details are useful for decision-making. (3) All sampling events have the potential to be archived electronically eliminating the need for expensive "triplicate docket books" and storage space for these paper records. (4) Because the popularity

of smart phone apps is increasing, a well-developed app that is attractive in appearance and easy to use may help greatly with the adoption of sampling plans, like that developed here for *O. perseae*, for IPM programs.

In this paper we employed a sampling cost function, (2.9), which includes a fixed cost for each tree sampled. Future work might include a more sophisticated per tree sampling cost which varies during the sequential sampling process to account for both the distance and the land topography between subsequently sampled trees, which may be of interest to a pest manager seeking to minimize their distance traveled and seeking to avoid sampling from trees which are difficult to reach (e.g., trees on steep hillsides). Additionally, the spatial GLMM model of pest populations which we employed assumes that pest individuals are distributed randomly within a tree, and that correlations of pest populations on adjacent trees are spatially symmetric. Future research on sequential sampling with spatial components which extends beyond these model assumptions may address issues pertaining to pest populations that are systematically distributed within trees, and may include anisotropic (i.e., asymmetric) correlation structures of pest populations, allowing for stronger correlation along orchard edges or within orchard rows (see Ifoulis et al. 2006).

References

Alatawi, F. J., G. P. Opit, D. C. Margolies, D. C., and J. R. Nechols. 2005. Within-plant distribution of twospotted spider mites (Acari: Tetranychidae) on impatiens: development of a presence-absence sampling plan. *J. Econ. Entomol.*, 98:1040–1047.

Bartlett, M. S. 1946. The large sample theory of sequential tests. *Proc. Cambridge Philos. Soc.*, 42:239–244.

Bianchi E. J. J. A., P. W. Goedhart, and J. M. Baveco. 2008. Enhanced pest control in cabbage crops near forest in The Netherlands. *Landsc. Ecol.*, 23:595–602.

Breslow, N. E., and D. G. Clayton. 1993. Approximate Inference in Generalized Linear Mixed Models. *J. Amer. Statistical Assoc.*, 88:9–25.

Binns, M. R., J. P. Nyrop, and W. Van Der Werf. 2000. Sampling and monitoring in crop protection: the theoretical basis for developing practical decision guidelines. CABI Publishing, Wallingford, United Kingdom.

CAC (California Avocado Commission). 2009. Acreage inventory summary, 2009 update using remote sensing technology. California Avocado Commission, 3pp., http://www.avocado.org/acreage-inventory-summary/ (last accessed 3 June 2011).

CAC (California Avocado Commission). 2010. 2009-10 Annual Report. California Avocado Commission 29 pp., http://www.californiaavocadogrowers.com/annual-report/ (last accessed 3 June 2011).

Candy, S. G. 2000. The application of generalized linear mixed models to multi-level sampling for insect population monitoring. *Environ. Ecol. Stat.*, 7:217–238.

Galvan, T. L., E. C. Burkness, and W. D. Hutchison. 2007. Enumerative and binomial sequential sampling plans for the multicolored Asian lady beetle (Coleoptera:Coccinellidae) in wine grapes. *J. Econ. Entomol.*, 100:1000–1010.

Gerrard, D.J., and H.C. Chaing. 1970. Density estimation of corn rootworm egg populations based upon frequency of occurrence. *Ecology*, 51:237–245.

Gozé, E., S. Nibouche, and J. P. Deguine. 2003. Spatial and probability distribution of helicoverpa armigera (Hubner) (Lepidoptera : Noctuidae) in cotton: systematic sampling, exact confidence intervals and sequential test. *Environ. Entomol.*, 32:1203–1210.

Hall, D. G., C. C. Childers, and J. E. Eger. 2007. Binomial sampling to estimate rust mite (Acari: Eriophyidae) densities on orange fruit. *J. Econ. Entomol.*, 100:233–240.

Hoddle, M. S. 2004. Invasions of leaf feeding arthropods: why are so many new pests attacking California-grown avocados? *California Avo. Soc. Yrbk.*, 87:65–81.

Hoddle, M. S., L. Robinson, and J. Virzi. 2000. Biological control of Oligonychus perseae on avocado: III. Evaluating the efficacy of varying release rates and release frequency of Neoseiulus californicus (Acari: Phytoseiidae). *Int. J. Acarol.*, 26:203–214.

Humeres, E. C., and J. G. Morse. 2005. Baseline susceptibility of persea mite (Acari: Tetranychidae) to abamectin and milbemectin in avocado groves in southern California. *Exp. Appl. Acarol.*, 36:51–59.

Hyung, D. L., J. J. Park, J-H. Lee, K-I. Shin, and K. Cho. 2007. Evaluation of binomial sequential classification sampling plan for leafmines of Liriomyza trifoli (Diptera: Agromyzidae) in greenhouse tomatoes. *Int. J. Pest Manag.* 53:59–67.

Ifoulis, A. A. and M. Savopoulou-Soultani. 2006. Use of geostatistical analysis to characterize the distribution of Lobesia botrana (Lepidoptera: Tortricidae) larvae in northern Greece. *Environ. Entomol.*, 35:497–506.

Johnson, M. E., L. M. Moore, and D. Ylvisaker. 1990. Minimax and maximin distance designs. *Journal of Statistical Planning and Inference*, 26:131–148.

Kabaluk, J. T., M. R. Binns, and R. S. Vernon. 2006. Operating characteristics of full count and binomial sampling plans for green peach aphid (Hemiptera: Aphididae) in potato. *J. Econ. Entomol.*, 99:987–992.

Kono, T., and T. Sugino. 1958. On the estimation of the density of rice stems infested by the rice stem borer. *Jpn. J. Appl. Entomol. Zool.*, 2:184–188.

Li, J. X., D. R. Jeske, J. L. Lara, and M. S. Hoddle. 2011. Sequential hypothesis testing with spatially correlated count data. *Integration: Mathematical Theory and Applications*, (in press)

Martinez-Ferrer, M. T., J. L Ripolles, and F. Garcia-Mari. 2006. Enumerative and binomial sampling plans for citrus mealybug (Homoptera: Psuedococcidae) in citrus groves. *J. Econ. Entomol.*, 99:993–1001.

Moaz, Y., S. Gal., M. Zilberstein, Y. Izhar, V. Alchanatis, M. Coll., and E. Palevsky. 2011. Determining an economic injury level for the persea mite, Oligonychus perseae, a new pest of avocado in Israel. *Entomol. Exp. Appl.*, (in press).

Mulekar, M. S., L. J. Young, and J. H. Young. 1993. Testing insect population density relative to critical densities with 2-SPRT. *Environ. Entomol.*, 22: 346-351.

Müller, W. 2007. Collecting spatial data: optimum design of experiments for random fields. Springer-Verlag, Berlin, Germany.

Ramírez-Dávila, J. F., and E. Porcayo-Camargo. 2008. Spatial distribution of the nymphs of Jacobiasca lybica (Hemiptera: Cicadellidae) in a vineyard in Andalucia, Spain. *Rev. Colomb. Entomol.*, 34:169–175.

Robson, J. D., M. G. Wright, and R. P. P. Almeida. 2006. Within-plant distribution and binomial sampling of Pentalonia nigronervosa (Hemiptera: Aphididae) on banana. *J. Econ. Entomol.*, 99:2185–2190.

Schabenberger, O., and C. A. Gotway. 2005. Statistical methods for spatial data analysis. Chapman & Hall/CRC Press, Boca Raton, Florida.

Schotzko, D. J., and L. E. Okeeffe. 1989. Geostatistical description of the spatial-distribution of Lygus hesperus (Heteroptera: Miridae) in lentils. *J. Econ. Entomol.*, 82:1277–1288.

Shah, P., D. R. Jeske, and R. Luck. 2009. Sequential hypothesis testing techniques for pest count models with nuisance parameters. *J. Econ. Entomol.*, 102:1970–1976.

Takakura, K. 2009. Reconsiderations on evaluating methodology of repellent effects: validation of indices and statistical analyses. *J. Econ. Entomol.*, 102:1977–1984.

Wald, A. 1947. Sequential analysis. Wiley and Sons, New York.

Young, L. J., and J. H. Young. 1998. Statistical ecology. Kluwer Academic Publishers, Boston.

Chapter 3

The Corridor Problem: Preliminary Results on the No-toll Equilibrium

Abstract

Consider a traffic corridor that connects a continuum of residential locations to a point central business district, and that is subject to flow congestion. The population density function along the corridor is exogenous, and except for location vehicles are identical. All vehicles travel along the corridor from home to work in the morning rush hour, and have the same work start time but may arrive early. The two components of costs are travel time costs and schedule delay (time early) costs. Determining equilibrium and optimum traffic flow patterns for this continuous model, and possible extensions, is termed "The Corridor Problem". Equilibria must satisfy the trip-timing condition, that at each location no vehicle can experience a lower trip price by departing at a different time. This paper investigates the no-toll equilibrium of the basic Corridor Problem.

Key Words: Morning Commute, Congestion, Corridor, Equilibrium

3.1 Introduction

In recent years, considerable work has been done examining the equilibrium dynamics of rush-hour traffic congestion. The central feature is the trip-timing condition, that no vehicle can experience a lower trip price by departing at a different time, where trip price includes the cost of travel time, the cost of traveling at an inconvenient time (termed schedule delay cost), and the toll, if applicable. The theoretical work on the topic has been in the context of Vickrey's model of a deterministic queue behind a single bottleneck (Vickrey, 1969), with some papers treating extensions to very simple networks, with each link containing a bottleneck.

While insightful, the work does not provide much insight into the spatial dynamics of rush-hour traffic congestion. Start by visualizing a departure rate surface over a metropolitan area. What does it look like at a point in time, and how does it change over the rush hour? Similarly, what do the flow, density, and velocity surfaces look like, and how do they evolve?

This paper takes a modest step forward in examining the spatial equilibrium dynamics of rush-hour congestion. It lays out perhaps the simplest possible model with continuous time and space that can address the issue. The metropolitan area is modeled as a single traffic corridor of uniform width joining the suburbs to the central business district (CBD), a point in space; the population entering each point along the corridor over the rush hour is taken as given; except for their locations, vehicles are identical, contributing equally to congestion, having a common work start-time, and having a common trip price function that is linear in travel time and schedule delay; congestion

takes the form of classic flow congestion; and there is no toll. The paper poses the simple question: What pattern(s) of departures satisfy the trip-timing condition? We term the corresponding problem and extensions, including determination of socially optimal allocations, "The Corridor Problem".

Unless some insight has eluded us, answering this question in the context of even so basic a model is surprisingly difficult (but if it were not difficult, it would likely have been solved). We have not yet succeeded in obtaining a complete solution, but because of the problem's difficulty feel justified in reporting what progress we have made.

There are good reasons to believe that the Corridor Problem is important. On the practical side, solving the problem would provide a point of entry to understanding the spatial dynamics of rush-hour traffic congestion, which is surely important in the enlightened design of road and mass transit networks. On the theoretical side, the problem has posed a stumbling block to the development of three lines of theoretical literature on the economics of traffic congestion. During the 1970s several papers were written on the economics of traffic congestion in the context of the monocentric city and related models (Solow and Vickrey, 1971; Solow, 1972; Kanemoto, 1976; and Arnott, 1979), assuming that traffic flow is constant over the day. Their focus was on second-best issues, in particular on how the underpricing of urban auto congestion distorts land use and affects capacity investment rules. Are the insights from that literature substantially modified when account is taken of the ebb and flow of traffic? At around the same time, Beckmann and Puu (e.g., Beckmann and Puu, 1985) started work on two-dimensional, steady-state continuous flow models of traffic congestion. Solving the Corridor Problem might provide insight into how to extend their work to non-stationary traffic flow. Later, Arnott et al., 1994 attempted to generalize the bottleneck model to

a traffic corridor, modeled as a series of bottlenecks with entry points between them. Because of the model's linearity, the solution degenerated into the treatment of multiple cases, the number rising geometrically with the number of bottlenecks. Thus, despite its elegant simplicity in other contexts, the bottleneck model does not appear well suited to examining the spatial dynamics of traffic congestion.

There is some prior work on the equilibrium spatial dynamics of urban traffic congestion. In the context of the monocentric model, Yinger, 1993 assumed that vehicles at the urban boundary are the first to depart and depart together, and are followed by successive cohorts from increasingly more central locations, and solved for the implied spatial dynamics of congestion over the rush hour. Ross and Yinger, 2000 proved that the departure pattern assumed in Yinger, 1993 does not satisfy the trip-timing condition, and that no other simple departure pattern does either. In earlier work, Arnott, 2004 conjectured an equilibrium departure pattern but, since he was unable to prove his conjecture, investigated a discretized variant of the problem, with buses and bus stops, termed the "bus-corridor problem". Congestion takes the form of bus speed varying inversely with the number of passengers. The numerical examples of the bus-corridor problem presented there are consistent with the form of departure set conjectured for the corridor problem proper, but do not prove the conjecture since the discretization alters the problem. Tian et al., 2007 derive the equilibrium properties of a variant of the bus-corridor problem in which congestion takes the form of crowding costs that increase in the number of passengers, provide some solution algorithms, and present numerical examples. The numerical examples of this variant of the bus-corridor problem are also consistent with the form of the departure set conjectured for the Corridor Problem proper in Arnott, 2004, but again do not prove the conjecture because the problem is somewhat different.

Section 2 presents the basic model and states the problem. Section 3 derives some implications of the trip-timing condition. Section 4 states the heuristic reasoning underlying an initial proposed solution. Section 5 undertakes the mathematical analysis of the initial proposed solution, in the process demonstrates that the initial proposed solution is not consistent in one respect with the trip-timing condition, and modifies the proposed solution. Section 6 develops an algorithm to solve numerically for a departure pattern consistent with the modified proposed solution. The algebraic calculations used to derive the results in Section 6 are placed in the Appendix at the end of the paper. Section 7 presents the results of the numerical algorithm and a graph of an observed commuting pattern on a Los Angeles freeway. Section 8 concludes.

3.2 Model Description

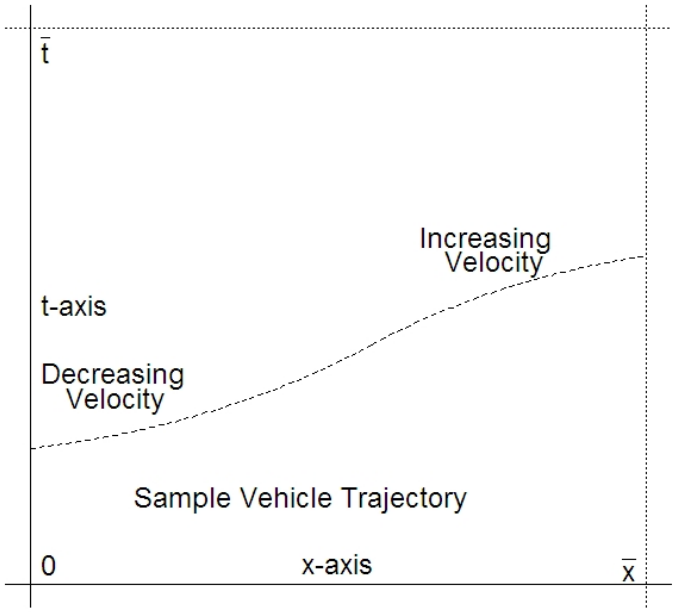

Figure 3.1: Traffic Corridor, Space-time Diagram

Consider a traffic corridor of constant width that connects a continuum of residential locations, "the suburbs," to a point central business district (CBD) that lies at the eastern end of the corridor, as shown in Figure 3.1. Location is indexed by x, the distance from the outer boundary of residential settlement towards the CBD, that is located at $\bar{x}$. $N(x)dx$ denotes the exogeneous number of vehicles departing between x and $x+dx$ over the rush hour. It is assumed that $N(x)$ is strictly positive for $x \in (0, \bar{x})$.

A glossary listing all the notation used in the paper is located at the end of the paper, following the references.

3.2.1 Trip Cost

Each morning all vehicles travel from their departure location to the CBD and have the common desired arrival time, $\bar{t}$. Late arrivals are not permitted, and in the absence of a toll the common travel cost function is

$$C = \alpha\,(\text{travel time}) + \beta\,(\text{time early})\,,$$

where α is the value or cost of travel time, and β is the value or cost of time early. It is assumed that $\alpha > \beta$, which is supported by empirical evidence (Small, 1982). Let $T(x,t)$ denote the travel time of a vehicle that departs from x at time t. Then $t+T(x,t)$ is the vehicle's arrival time, so that $\bar{t}-[t+T(x,t)]$ is its time early, so that the trip cost may be written as

$$C(x,t) = \alpha T(x,t) + \beta\left(\bar{t} - [t + T(x,t)]\right). \tag{3.1}$$

There are no tolls, so that, at each time and location, trip price equals trip cost.

3.2.2 Continuity Equation

Classical flow congestion is assumed, which combines the equation of continuity with an assumed relationship between velocity and density. Recall that we have assumed the road to be of constant width. Accordingly, at location x and time t, let $\rho(x,t)$ denote the density of vehicles per unit length, and let $v(x,t)$ denote velocity. The relationship between velocity and density is written as

$$v(x,t) = V(\rho(x,t)),$$

with $V' < 0$. It is typically assumed that: i) V steadily decreases with ρ, so that flow, $F = \rho V$, is a smooth convex function of ρ; ii) flow is zero with zero density and also with jam density. The equation of continuity is simply a statement of conservation of mass for a fluid, that the change in the number of vehicles on a section of road of infinitesimal length equals the inflow minus the outflow. Letting $n(x,t) \geq 0$ denote the entry rate onto the road, the equation of continuity is

$$\frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x}(\rho V) = n(x,t).$$

3.2.3 Trip-Timing Condition (TT)

There are two equilibrium conditions. The first is that all vehicles commute. If we let $\mathcal{D}$ denote the departure set, i.e., the set of (x,t) points for which departures occur in equilibrium, then the condition that all vehicles commute can be written as

$$\int_{(x,t)\,\in\,\mathcal{D}} n(x,t)\,dt = N(x) \quad \forall\, x \in [0,\bar{x}]. \tag{3.2}$$

The second equilibrium condition is the trip-timing condition (TT), that no vehicle can experience a lower trip price by departing at a different time. Letting $p(x)$ denote the equilibrium trip price at location x, the TT condition can be written as

$$\begin{aligned} &C(x,t) = p(x) \quad \forall\,(x,t) \in \mathcal{D} \qquad \text{(Equality Component of the TT)} \\ &C(x,t) \geq p(x) \quad \forall\,(x,t) \notin \mathcal{D} \qquad \text{(Inequality Component of the TT).} \end{aligned} \tag{3.3}$$

It states that at no location can the trip price be reduced by traveling outside the departure set at that location. A no-toll equilibrium is a departure pattern, $n(x,t) \geq 0$, and a trip price function, $p(x)$, such that the equilibrium conditions are satisfied, with

$T(x,t)$ obtained from the solution to the continuity equation.

3.3 Implications of the Trip-Timing Condition

3.3.1 Relation between Arrival and Departure Times

From (3.1) and (3.3),

$$T(x,t) = \frac{p(x) - \beta(\bar{t} - t)}{\alpha - \beta} \quad \forall\, (x,t) \in \mathcal{D}$$

Hence, over the departure set at each location, travel time increases linearly in the departure time at the rate $\frac{\beta}{\alpha-\beta}$. In particular, if two vehicles leave the same location, x, one at time t, the other at time $t + \Delta t$, and if both (x,t) and $(x, t + \Delta t)$ are in the departure set, then the difference between their arrival times is

$$\Delta a = \frac{\alpha}{\alpha - \beta}\Delta t. \tag{3.4}$$

The same is true for any time and location, (x', t') such that (x', t') and $(x', t' + \Delta t)$ are in the departure set. How is this possible? One way this result can be achieved, and we shall show is the only one, is for the departure function for a later cohort to be identical to that for an earlier cohort, except for the addition of vehicles closer to the CBD, since cars entering at all locations represented in the earlier cohort are then slowed down by the same amount. This result can be illustrated by considering the bus-corridor discretization of the same problem, for which the speed of a bus is related to its number of passengers.

Suppose that the first bus to depart picks up passengers at stops 1 and 2, and that the second bus to depart picks up passengers at stops 1, 2, and 3. The trip-timing

equilibrium condition requires that travel time on the second bus be higher than travel time on the first bus by the same amount for passengers boarding at stop 1 as for those boarding at the stop 2. The travel time increase for those boarding at stop 1 equals the travel time increase between stops 1 and 2, 2 and 3, —. The travel time increase for those boarding at stop 2 equals the travel time increase between stops 2 and 3, —. For these travel time increases to be the same requires that the travel time between stops 1 and 2 be the same for the first and second buses, which requires that at bus stop 1 the same number of passengers board the first and second buses. The argument in the next section formalizes this intuitive argument.

3.3.2 Constant Departure Rate within Interior of Departure Set

It will prove convenient at this point to make the transformation of variables

$$a(x,t) = t + T(x,t)$$

where $a(x,t)$ is the arrival time at the CBD of a vehicle that departs location x at time t. If $\hat{T}(x,a)$ is the travel time of a vehicle that arrives at the CBD at time a, then the inverse transformation is

$$t(x,a) = a - \hat{T}(x,a)$$

which relates departure time to arrival time. The trip-timing condition, expressed in terms of arrival time, is

$$p(x) = \alpha\hat{T}(x,a) + \beta(\bar{t} - a) \quad \forall\,(x,a) \in \mathcal{A}, \tag{3.5}$$

where $\mathcal{A}$, the arrival set, is the set of all (x, a) for which the arrival rate is positive. The advantage of working in terms of arrival time is that $\hat{T}(x, a)$ tracks the cohort of vehicles that arrives at time a. Since a vehicle with arrival time a passes location x at time $a - \hat{T}(x, a)$,

$$\hat{T}(x, a) = \hat{T}(x + dx, a) + \frac{dx}{v(x, a - \hat{T}(x, a))}$$

and so

$$\hat{T}_x(x, a) = -\frac{1}{v(x, a - \hat{T}(x, a))} = -\frac{1}{V(\rho(x, a - \hat{T}(x, a)))}. \tag{3.6}$$

Note that we use a subscript to indicate a partial derivative, where appropriate. Differentiation of (3.5) with respect to a and then x yields

$$\hat{T}_a = \frac{\beta}{\alpha} \quad \forall\ (x, a) \in \text{int}(\mathcal{A})$$
$$\hat{T}_{ax} = 0 \quad \forall\ (x, a) \in \text{int}(\mathcal{A}),$$

while differentiation of (3.6) with respect to a yields

$$\hat{T}_{xa}(x, a) = \frac{V' \rho_t (1 - \hat{T}_a)}{V^2} \tag{3.7}$$

From equality of mixed partial derivatives, it follows that the right-hand side of (3.7) equals zero, and since V', $1 - \hat{T}_a$ and V are all strictly nonzero, it follows that

$$\rho_t(x, a - \hat{T}(x, a)) = 0 \quad \forall\ (x, a) \in \text{int}(\mathcal{A}),$$

or

$$\rho_t(x, t) = 0 \quad \forall\ (x, t) \in \text{int}(\mathcal{D}). \tag{3.8}$$

(3.8) states that traffic density is constant at a particular location over the interior of the departure set at that location. Since $\rho_t = 0$, the continuity equation reduces to

$$\frac{\partial}{\partial x}(\rho V(\rho)) = n(x,t).$$

Differentiating this equation with respect to t yields

$$\frac{\partial n}{\partial t} = \frac{\partial}{\partial t}(\rho V)_x = \frac{\partial}{\partial x}(\rho V)_t = \frac{\partial}{\partial x}0 = 0.$$

Thus, at each location, the departure rate is constant over the interior of the departure set:

$$n(x,t) = n(x) \quad \forall\, (x,t) \in \text{int}(\mathcal{D}). \tag{3.9}$$

3.4 Proposed Departure Set

Consider two vehicle trajectory segments, running from some x' to x'', both of which are in the interior of the departure set, an earlier one and a later one. For each x, traffic density is the same for both trajectories (from (3.8)), as is the departure rate (from (3.9)). Furthermore, at all locations between x' and x'' the travel time of the later trajectory exceeds the travel time of the earlier trajectory by the same amount. This requires that travel time between x'' and the CBD be higher for the later trajectory, that in turn requires that more vehicles enter the road between x'' and the CBD for the later trajectory. One way this can be achieved is for the first departure time at each location to be later for more central locations. Put alternatively, later trajectories pick up vehicles at increasingly central locations.

Figure 3.2 displays a departure set consistent with this reasoning. Time is

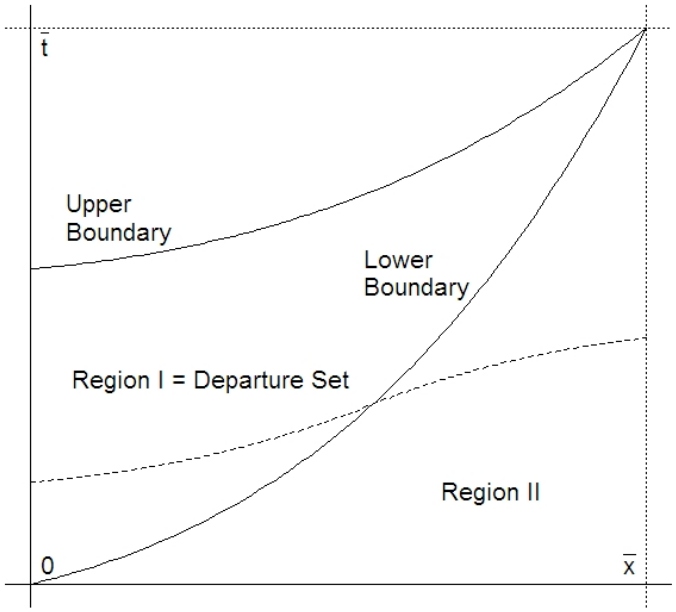

Figure 3.2: Horn-shaped *proposed departure set*. The upper boundary is a vehicle trajectory corresponding to the final cohort of vehicles to arrive at the CBD. The dashed line indicates a sample vehicle trajectory, with decreasing velocity within the departure set (Region I) and increasing velocity outside the departure set (Region II). Note that the slope of the dashed line equals the slope of the upper boundary, up to the point of leaving the departure set.

renormalized so that $t = 0$ corresponds to the start of the morning commute (at $x = 0$.) The departure set is connected. The lower boundary gives the time of the first departure at each location, and the upper boundary the time of the last departure at each location. A sample trajectory is shown as the dashed line. The first trajectory contains vehicles from only the most distant location. Succeeding trajectories contain vehicles from locations successively closer to the CBD, as well as from all more distant locations. The last trajectory, which arrives at the CBD exactly at the desired arrival time, $\bar{t}$, contains vehicles from all locations. We refer to the departure set as Region I, and the region below the departure set as Region II.

Since the pattern of density by location in the interior of the departure set does not change over time, at any location the number of vehicles entering at more distant locations must equal the flow at that location. At more central locations therefore, within the departure set the flow rate must be higher, which is inconsistent with hypercongestion. Thus, along a vehicle trajectory velocity decreases from $x = 0$ to the lower boundary of the departure set, and then speeds up from the lower boundary of the departure set to the CBD as traffic thins out since no vehicles are entering the road. Thus, a vehicle trajectory is convex in the interior of Region I and concave in the interior of Region II.

The above line of reasoning leaves open the properties of the boundary of the departure set. The rest of this section will sketch a more formal derivation of the properties of the departure set.

3.4.1 General Properties of the Departure Set

We shall argue that the departure set has the following properties:

Property 1 The upper boundary of the departure set is a vehicle trajectory.

Property 2 At any location, the departure rate on the upper boundary of the departure set is the same as in the interior of the departure set.

Property 3 At any location, the departure set at that location is a connected set.

Property 4 The departure set is connected and does not contain holes.

Property 1: Upper Boundary of the Departure Set Is a Vehicle Trajectory.

If the upper boundary of the departure set is a vehicle trajectory, the trajectory must arrive at the CBD exactly at $\bar{t}$. Suppose not, and that the trajectory arrives at

the CBD at $t' < \bar{t}$. Then at any location there is a departure time for which a vehicle can depart, experience no traffic congestion, and arrive at the CBD between t' and $\bar{t}$, experience less travel time and arrive less early than all other vehicles departing from that location, which is inconsistent with the trip timing condition.

Suppose that the final cohort of vehicles to arrive at the CBD does not contain vehicles that depart from $x = 0$, and that the latest departure from $x = 0$ in the departure interval is at t'. Then a vehicle departing $x = 0$ slightly after t' can travel at free-flow travel speed until it meets the cohort of vehicles that departs from $x = 0$ at t', hence experiencing a lower trip cost than the vehicle that departs $x = 0$ at t'.

With some modification, the same line of reasoning can be applied to establish that the final cohort must contain vehicles from every location.

Property 2: At any Location, the Departure Rate on the Upper Boundary of the Departure Set Is the Same as in the Interior.

(3.8) indicates that, at each location, density and therefore velocity are constant in the interior of the departure set. Hence, in the interior of the departure set velocity can be written as $v(x)$ and density can be written as $\rho(x)$. If the departure rate were different on the upper boundary, then the velocity as a function of location for the last cohort would not be $v(x)$, which can be shown to imply violation of the trip-timing condition.

Property 3: Over Any Infinitesimally Small Location Interval, $[x, x + dx]$, the Departure Set is a Connected Set

We have proved that, at a particular location in the interior of the departure set, density (3.8) and the entry rate (3.9) must be constant. Now suppose that the

departure set over an infinitesimally small location interval, $[x, x + dx]$, is disconnected. Then a vehicle departing from x outside of the departure set will experience a lower traffic density, and hence a higher velocity, over $[x, x + dx]$, than a vehicle departing within the departure set at the same location. Therefore, to a vehicle that departs x inside the departure set at an earlier time, a vehicle that departs x outside the departure set at a later time will incur less travel time cost up to the point when it either enters the departure set (i.e., joins a cohort of vehicles that depart within the departure set) or reaches the CBD. This is inconsistent with the TT-condition, that requires that vehicles that depart the same location but at a later time incur greater travel time cost.

Property 4: The Departure Set Is Connected and Does Not Contain Holes.

Property 4 easily follows from Properties 1 and 3, and the requirement that the population density be nonzero at all locations up to the edge of the metropolitan area. Since the population density is nonzero at all locations the departure set is nonempty at all locations. From Property 3 the departure set at a given location is a connected set. From Property 1 the upper boundary of the departure set is a vehicle trajectory, which is a connected set. Thus, the departure set is the union of connected sets, each of which has nonempty intersection with a connected set, and is therefore connected.

3.4.2 Proposed Departure Set

In his earlier work on the Corridor Problem, Arnott, 2004 had established Properties 1 and 2, and conjectured Properties 3 and 4. Furthermore, he was able to solve numerically for equilibrium of the (discretized) bus-corridor problem on the assumption that, at each bus stop, the same number of passengers board each bus picking up passengers at that stop, except perhaps the first. Accordingly, Arnott conjectured

that the departure set takes the form shown in Figure 3.2, with the start of the rush hour endogenous. We refer to this as the *proposed departure set*. Region I is the departure set, and Region II is the region in the $(0,0) \rightarrow (\bar{x}, \bar{t})$ rectangle below the departure set. The upper boundary of the departure set is the vehicle trajectory arriving at the CBD at $\bar{t}$, and includes departures from all locations. The departure set is connected; a vehicle trajectory (shown in Figure 3.2 as a dashed line) starts at $x = 0$, in Region I is parallel to the upper boundary of the departure set (since, at each location, velocity is constant over that location's departure time interval), and in Region II accelerates. He conjectured that adjustment of the lower and upper boundaries provides enough freedom for the trip-timing condition to be satisfied for an arbitrary distribution of population along the corridor. For obvious reasons, Arnott termed the proposed departure set, a "horn-shaped" departure set.

As we shall see, Arnott's proposed departure set was incorrect, as there must be a zone bordering the CBD with no departures. Refer to a departure set that modifies the proposed departure set, taking this feature of the solution into account, the *modified departure set*. Like the proposed departure set, the modified departure set does not provide enough freedom for an equilibrium solution to exist with an arbitrary distribution of population along the corridor. Accordingly, we shall address the question: What distributions of population along the corridor are consistent with the modified departure set?

3.5 Mathematical Analysis: Analytic Results

Section 3.5 derives the full set of equations that together determine an equilibrium departure rate distribution for the modified departure set. Sections 3.5.1 through

3.5.3 analyze the continuity equation within each of two regions, the interior of Region I (the departure set), and Region II. Section 3.5.4 lays out the procedure we employ for describing trajectories and points on the lower boundary of the departure set. Section 3.5.5 derives one of three governing equations, and applies it to demonstrate that the proposed departure set is inconsistent with equilibrium. Section 3.5.6 presents the modified departure set, which modifies the proposed departure set to be consistent with the governing equation. Section 3.5.7 states the three governing equations and discusses how they work together, and section 3.5.8 derives the three governing equations.

The central insight on which the analytical results and the numerical analysis build is that once the lower boundary of the departure set, as well as the flow rate at every point along the boundary, are solved for, the rest of the solution follows straightforwardly. Thus, the focus is on solving for the lower boundary of the departure set consistent with the three governing equations.

3.5.1 Continuity Equation: Method of Characteristics

The continuity equations in Regions I and II are first-order, quasi-linear partial differential equations for the density, ρ, as a function of (x, t) within each region. We may solve each equation by the method of characteristics, that converts the PDE into a system of ODE's, whose solution yields characteristic curves in the x-t plane, with ρ determined in each region as a function along these curves (Evans, 2002, Rhee et al., 1986). In the following two sections we apply this method of characteristics to Regions I and II.

3.5.2 Region I

We showed earlier that the TT condition implies that, at each location, the departure rate is constant over the interior of the departure set and along its upper boundary, and shall assume furthermore that the departure rate is the same along the lower boundary as well. The continuity equation in Region I is therefore

$$\frac{\partial \rho}{\partial t} + (\rho V)' \frac{\partial \rho}{\partial x} = n(x)$$

where $'$ indicates a derivative with respect to ρ, i.e., $(\rho V)' \equiv \frac{d}{d\rho}(\rho V)$. The characteristic curves for this PDE satisfy

$$\frac{dt}{1} = \frac{dx}{(\rho V)'} = \frac{d\rho}{n(x)}.$$

In particular, the flow, $F = \rho V$, satisfies

$$(\rho V)' \, d\rho = n(x) \, dx,$$

and, since initially $F = 0$,

$$F(x) = (\rho V)(x) = \int_0^x n(x') \, dx'. \tag{3.10}$$

Thus, within the departure set, the flow at a particular location equals the sum of the departure rates at upstream locations. The flow function can be solved for, given the departure rate function, and *vice versa*.

At any location, the population equals the departure rate there times the width of the departure set there. Knowledge of any two allows for the solution of the third.

Since population is nonzero at all locations, (3.10) implies that flow is a strictly

increasing function of x on the interior of the departure set, reaching at most capacity flow at the entry point closest to the CBD. Hence, we conclude that an equilibrium solution to the Corridor Problem does not permit hypercongestion. Furthermore, this implies a one-to-one relationship on the interior of the departure set between flow, velocity and density. Newell, 1988 proved that a traffic corridor with a single entry point does not admit hypercongestion. The result here is more general since entries are possible at every point along the corridor. Recall that we have assumed a road of constant width; hypercongestion could presumably occur in a more general model, e.g., if road capacity were to shrink as the CBD is approached.

3.5.3 Region II

The continuity equation in Region II is the well-known homogeneous equation

$$\frac{\partial \rho}{\partial t} + (\rho V)' \frac{\partial \rho}{\partial x} = 0.$$

$(\rho V)'$ is the rate at which flow changes with density, and its reciprocal is the rate at which density changes with flow. As discussed in Newell, 1993, the characteristic curves in Region II are straight lines which are iso-density curves with slope $\frac{dt}{dx} = \frac{1}{(\rho V)'}$. These characteristic lines emanate from the lower boundary of the departure set, and completely determine the density field in Region II (see Figure 3.3). Since flow is non-decreasing on the lower boundary of the departure set, the slopes of the characteristic lines are non-decreasing, and therefore the characteristic lines are non-intersecting. This excludes the possibility of shocks occurring in Region II, and implies that density is continuous in Region II. Note that the initial vehicle trajectory, with $V = V_0$, coincides with a characteristic line. Also note that the slope of a characteristic line at a point on

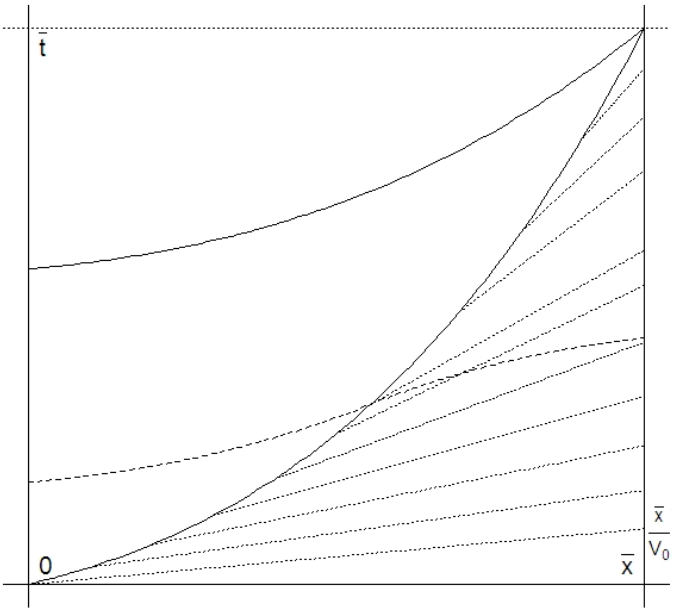

Figure 3.3: Characteristic lines in Region II, which are iso-density curves, along with a sample vehicle trajectory.

the lower boundary is greater than the slope of a trajectory curve at that point, since $\frac{1}{(\rho V)'} = \frac{1}{V+\rho V'} > \frac{1}{V}$.

3.5.4 Parametrization of Lower Boundary Curve

We first normalize time so that the rush hour starts at $t = 0$; $\bar{t}$ is then the length of the rush hour, which is determined as part of the overall solution. We parametrize the lower boundary curve of the departure set as follows (refer to Figure 3.4). A cohort of vehicles departs $x = 0$ at time u, reaches the lower boundary curve of the departure set at location $x = b(u)$, and arrives at the CBD at time $a(u)$. From (3.4), is $a(u) = \frac{\alpha}{\alpha-\beta}u + \frac{\bar{x}}{V_0}$.

The travel time to the lower boundary curve for a vehicle departing $x = 0$ at time u is

$$T_I(u) = \int_0^{b(u)} \frac{1}{v(x')}\, dx'$$

where $v(x)$ is the velocity in the departure set at location x. Thus, the (x, t) coordinates along the lower boundary curve are parametrized as $(b(u), u + T_I(u))$.

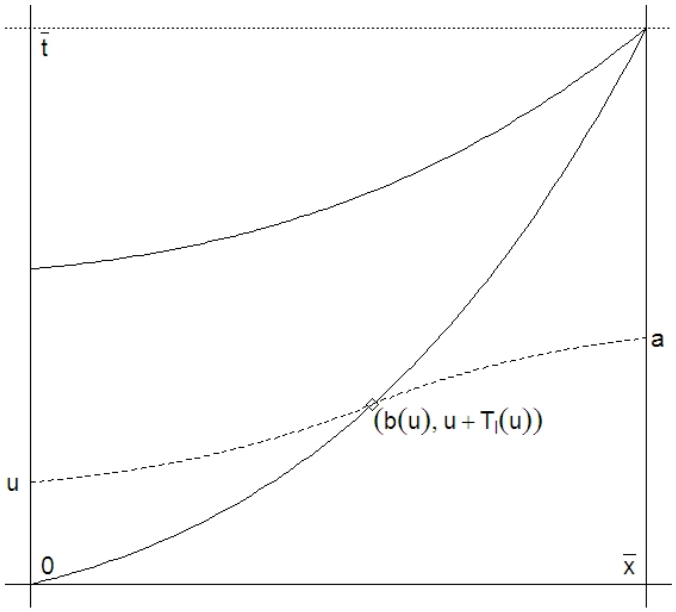

Figure 3.4: Trajectory departing $x = 0$ reaches the lower boundary at location $x = b(u)$. Velocity function within the departure set at location x is $v(x) = v(b(u))$.

3.5.5 Arrival Flow Rate

An important consequence of the TT condition is that it enables us to determine the flow rate at the CBD in terms of the flow rate at the lower boundary of the departure set. Let $F(b(u))$ denote the flow within the departure set at location

$x = b(u)$, so $F(b(u)) = \int_0^{b(u)} n(x')\,dx'$. If we follow the vehicle trajectory that departs $x = 0$ at time u, leaves the departure set at location $b(u)$, and arrives at the CBD at time $a(u) = \frac{\alpha}{\alpha-\beta}u + \frac{\bar{x}}{V_0}$, then we may write the cumulative number of arrivals at the CBD by time a, $A(a)$ as

$$\begin{aligned} A(a) &= \int_0^{u(a)} \int_0^{b(u')} n(x)\,dx\,du' \\ &= \int_0^{u(a)} F(b(u'))\,du'. \end{aligned} \tag{3.11}$$

Since $\frac{du}{da} = \frac{\alpha-\beta}{\alpha}$, we may determine the arrival flow rate at the CBD as

$$\begin{aligned} \text{Flow at } (\bar{x}, a) \;&=\; \frac{dA}{da} \;=\; F(b(u(a)))\frac{du}{da} \\ &=\; \frac{\alpha-\beta}{\alpha}F(b(u(a))) \end{aligned}$$

Hence, flow for a cohort of vehicles strictly decreases from the lower boundary to the CBD by the multiplicative factor $\frac{\alpha-\beta}{\alpha}$.

3.5.6 Modification of Proposed Departure Set

The flow of a cohort of vehicles decreases from the lower boundary of the departure set to the CBD by a factor of $\frac{\alpha-\beta}{\alpha}$. Also, flow must be continuous from the lower boundary of the departure set to the CBD. With the proposed departure set, these two properties cannot be simultaneously satisfied for the last cohort of vehicles. Thus, the proposed departure set is inconsistent with equilibrium. If the departure set is modified so that there is zero population density over an interval before the CBD, with the interval being determined so as to satisfy the first condition, then both conditions can be satisfied. Thus, in what follows we consider a *modified departure set*, which is

still horn-shaped, but has zero population density near the CBD. A modified departure set is shown in Figure 3.5.

Since flow is an increasing function along the lower boundary curve and since hypercongestion does not occur, the maximum flow must occur at the tip of the horn, and this maximum flow must be less than or equal to capacity flow.

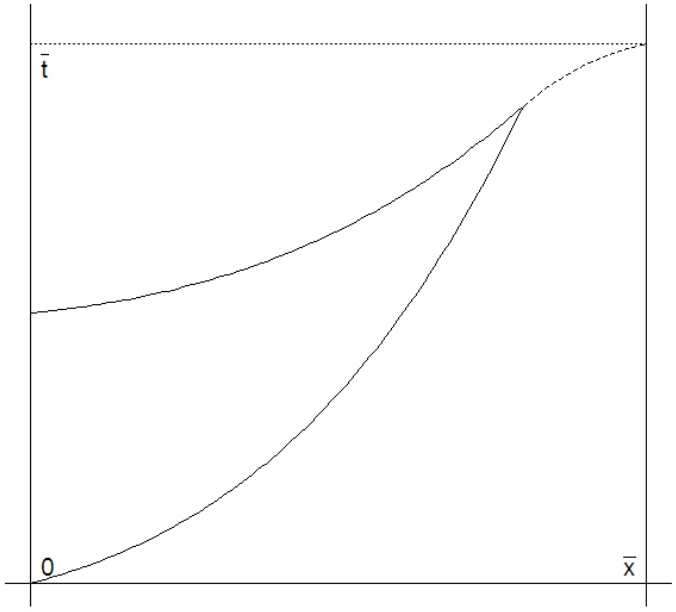

Figure 3.5: The flow for a cohort must decrease from the tip of the departure set to the CBD by the multiplicative factor $\frac{\alpha-\beta}{\alpha}$.

3.5.7 Three Governing Equations: Summary

The TT condition implies that, at each location, the departure rate is constant over the interior and upper boundary of the departure set. We have made an assumption concerning the departure rate along the lower boundary of the departure set such that, at each location, the departure rate is constant over the entire departure set, including

the lower boundary. Using this result, we shall derive three equations, (3.12), (3.13) and (3.14), that must be satisfied by an equilibrium solution to the Corridor Problem with the modified departure set.

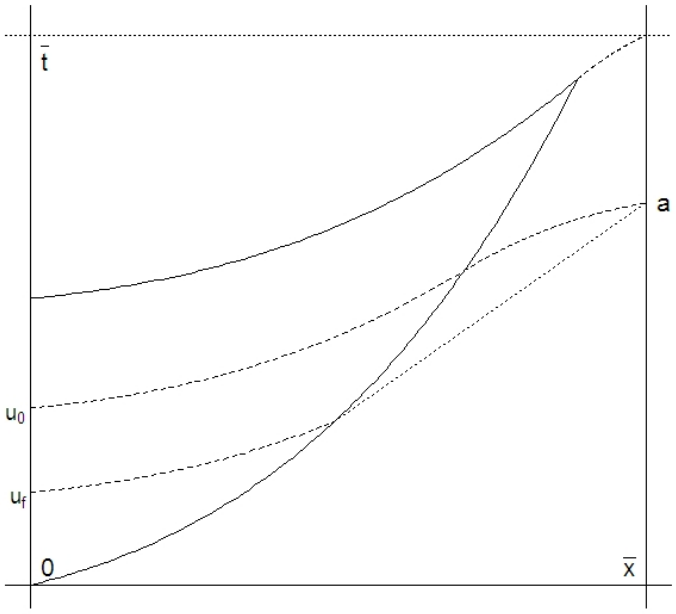

Figure 3.6: Arrival of the vehicle trajectory departing $x = 0$ at time u_0 intersects the characteristic line that originates from $(b(u_f), u_f + T_I(u_f))$, where $u_f < u_0$.

Consider a vehicle trajectory departing $x = 0$ at time u_0 and arriving at the CBD at time $a(u)$ (Figure 3.6). The characteristic curve through this arrival point originates from the lower boundary at location $b(u_f)$, where $u_f < u_0$ (see Figure 3.6). Since the characteristic curve is a straight line of constant flow, we may equate the flow at the lower boundary for the cohort u_f to the flow at the CBD for the cohort u_0. Hence, for each departure time u_0 there will correspond a unique value of u_f. We now state the three governing equations that must be satisfied for each pair of u_0 and u_f values, and we derive these equations in the following sections.

$$u_0 = \frac{\alpha - \beta}{\alpha}\left[\frac{\bar{x} - b(u_f)}{\frac{d}{d\rho}(\rho V)|_{b(u_f)}} + u_f + \int_0^{b(u_f)} \frac{1}{v(x')}dx' - \frac{\bar{x}}{V_0}\right] \tag{3.12}$$

$$F(b(u_f)) = \frac{\alpha - \beta}{\alpha}F(b(u_0)) \tag{3.13}$$

$$\int_{u_f}^{u_0} F(b(u'))\,du' = \left(-\rho + \frac{\rho V}{(\rho V)'}\right)\Bigg|_{b(u_f)} (\bar{x} - b(u_f)). \tag{3.14}$$

Since we have established a one-to-one correspondence within the departure set between density, flow and velocity, we can eliminate all the terms involving density, ρ, and velocity, V, and rewrite all three equations in terms of only flow, F. A solution to these three equations consists of the function $b(u)$ describing the x-coordinates of the lower boundary curve, and the flow function along the lower boundary curve, $F(b(u))$. Since the outer boundary of residential settlement is at $x = 0$, the function $b(u)$ must satisfy $b(0) = 0$, and since initially there is no traffic, the flow function must satisfy $F(0) = 0$. Note that the initial width of the departure set (i.e., the duration of time over which individuals depart at $x = 0$) and the work start-time at the CBD, $\bar{t}$, will be determined as part of the solution and are not given *a priori*.

3.5.8 Derivation of the Three Governing Equations

First Governing Equation

Referring back to Figure 3.6, since the TT condition implies that $a = \frac{\alpha}{\alpha-\beta}u_0 + \frac{\bar{x}}{V_0}$, and since the slope of the characteristic line is $\frac{1}{\frac{d}{d\rho}(\rho V)|_{b(u_f)}}$, by calculating the slope

directly we have

$$\frac{\frac{\alpha}{\alpha-\beta}u_0 + \frac{\bar{x}}{V_0} - [u_f + T_I(u_f)]}{\bar{x} - b(u_f)} = \frac{1}{\frac{d}{d\rho}(\rho V)|_{b(u_f)}}.$$

We may solve for u_0 in terms of u_f to obtain the first governing equation

$$u_0 = \frac{\alpha-\beta}{\alpha}\left[\frac{\bar{x} - b(u_f)}{\frac{d}{d\rho}(\rho V)|_{b(u_f)}} + u_f + \int_0^{b(u_f)} \frac{1}{v(x')}dx' - \frac{\bar{x}}{V_0}\right].$$

Second Governing Equation

Flow on the lower boundary decreases along a trajectory to the CBD by the multiplicative factor $\frac{\alpha-\beta}{\alpha}$, i.e., if $F(b(u))$ is the flow at the lower boundary of the cohort that departed at time u, and $a(u) = \frac{\alpha}{\alpha-\beta}u + \frac{\bar{x}}{V_0}$ is the arrival time at the CBD of this cohort, then the flow at $(\bar{x}, a(u))$ equals $\frac{\alpha-\beta}{\alpha}F(b(u))$. Referring to Figure 3.6, since characteristic lines in Region II are iso-flow lines, this allows us to express the flow within the departure set at location $b(u_f)$ in terms of the flow within the departure set at location $b(u_0)$, yielding the second governing equation

$$F(b(u_f)) = \frac{\alpha-\beta}{\alpha}F(b(u_0)).$$

Since hypercongestion does not occur within the departure set, there is a one-to-one relation between flow, density and velocity. We may therefore also use this second governing equation to determine the density (or velocity) at $b(u_f)$ in terms of the density (or velocity) at $b(u_0)$.

Third Governing Equation

Consider a trajectory that departs $x = 0$ at time u_0, and parameterize the (x, t) coordinates of the portion of this trajectory in Region II as $(\tilde{x}(u), \tilde{t}(u))$, $u_f \leq$

$u \leq u_0$, where $(\tilde{x}(u), \tilde{t}(u))$ is the point on the trajectory in Region II that intersects the characteristic line emanating from the lower boundary of the departure set at the point $(b(u), u + T_I(u))$ (see Figure 3.7).

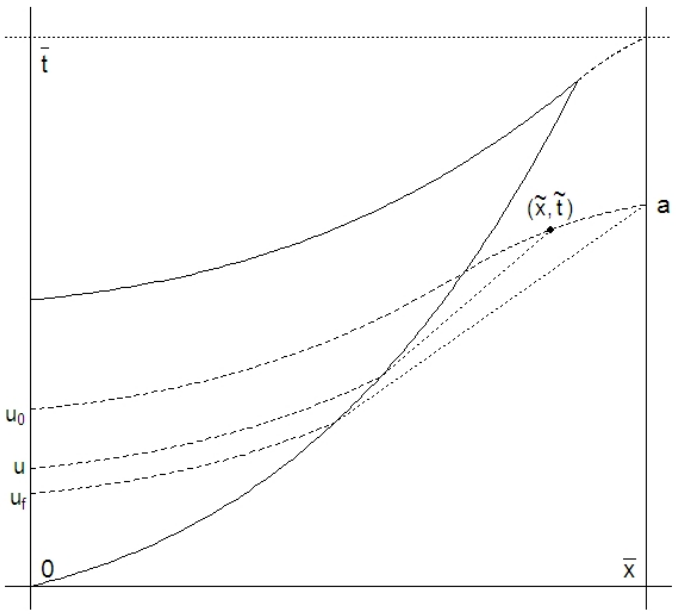

Figure 3.7: Trajectory in Region II parametrized as $(\tilde{x}(u), \tilde{t}(u))$, where $u_f \leq u \leq u_0$.

Hence,

$$(\tilde{x}(u_0), \tilde{t}(u_0)) = (b(u_0), u_0 + T_I(u_0))$$

$$(\tilde{x}(u_f), \tilde{t}(u_f)) = (\bar{x}, \frac{\alpha}{\alpha - \beta} u_0 + \frac{\bar{x}}{V_0}).$$

The cumulative flow, or cumulative number of arrivals, is constant along a trajectory in Region II. By (3.11), for the cohort that departs $x = 0$ at time u the cumulative flow

along the cohort's trajectory in Region II is

$$A(u) = \int_0^u F(b(u'))\, du'.$$

Denote the cumulative flow as a function of the space-time coordinate in Region II as $\hat{A}(x,t)$, to distinguish it from the cumulative flow along a trajectory in Region II, $A(u)$. Following Newell, 1993, if we move along a characteristic line in Region II from (x,t) to $(x+dx, t+dt)$, then the cumulative flow along this line satisfies

$$\frac{d\hat{A}}{dx} = -\rho + \frac{\rho V}{(\rho V)'} \tag{3.15a}$$

$$\frac{d\hat{A}}{dt} = -\rho(\rho V)' + \rho V, \tag{3.15b}$$

where ρ and V are the constant density and constant velocity along the characteristic line. If we integrate (3.15a) along the characteristic line from $(b(u), u + T_I(u))$ to $(\tilde{x}(u), \tilde{t}(u))$, we obtain

$$\hat{A}(\tilde{x}(u), \tilde{t}(u)) - \hat{A}(b(u), u + T_I(u)) = \left(-\rho + \frac{\rho V}{(\rho V)'}\right)\Bigg|_{b(u)} (\tilde{x}(u) - b(u)), \quad u_f \le u \le u_0.$$

Since along the trajectory $(\tilde{x}(u), \tilde{t}(u))$ the cumulative flow is the constant value $A(u_0) = \int_0^{u_0} F(b(u'))\, du'$, and since on the lower boundary at $b(u)$ the cumulative flow is $A(u) = \int_0^u F(b(u'))\, du'$, we may rewrite this expression as

$$\int_u^{u_0} F(b(u'))\, du' = \left(-\rho + \frac{\rho V}{(\rho V)'}\right)\Bigg|_{b(u)} (\tilde{x}(u) - b(u)), \quad u_f \le u \le u_0. \tag{3.16}$$

In particular, when $u = u_f$, we obtain the third governing equation

$$\int_{u_f}^{u_0} F(b(u'))\,du' = \left(-\rho + \frac{\rho V}{(\rho V)'}\right)\Bigg|_{b(u_f)} (\bar{x} - b(u_f)).$$

This equation relates the integral of the flow along the lower boundary from $b(u_f)$ to $b(u_0)$, to the distance from the CBD to the lower boundary at $b(u_f)$, and the density and velocity at $b(u_f)$.

3.6 Numerical Analysis (Greenshields')

Ideally, given an arbitrary population density we would like to construct a solution to the Corridor Problem and, if multiple solutions exist, characterize all possible solutions. However, it is not clear if a solution will exist for an arbitrary population density, e.g., we have already shown that the population must be zero some finite distance before the CBD. In the last section we showed how a solution with the modified departure set can be determined by solving for a lower boundary curve of the departure set and a flow along this lower boundary curve. We also showed how a specific solution uniquely determines a population density. Thus, we seek to characterize all possible solutions consisting of lower boundary curves and flows along these lower boundary curves, from which we could extract all possible population densities that admit solutions to the Corridor Problem with the modified departure set. We have determined that any solution must satisfy the three governing equations. What is not clear, however, is whether or not the three governing equations admit a solution, and, if so, whether or not they admit a unique solution.

In this section we consider a specific velocity-density relation (Greenshields'), and provide a numerically constructive proof that, for a given ratio of parameters, $\frac{\beta}{\alpha}$,

the three governing equations admit a unique solution for a modified departure set and a flow distribution on that departure set that reaches capacity flow (note that $\bar{x}$ and V_0 are scale parameters that will only determine the scaling along the distance and time axes). As discussed, this solution will uniquely determine a population density. Thus, in this section we will conclude that, using Greenshields' relation and given $\frac{\beta}{\alpha}$, there is a unique population density that admits a solution to the Corridor Problem with the modified departure set that reaches capacity flow.

Each one of these solutions admits a continuous family of truncated solutions that do not reach capacity flow. To see this, suppose that we have a departure set solution that reaches capacity flow (that must necessarily occur at the tip of the departure set, since flow is non-decreasing). Consider any vehicle trajectory that departs $x = 0$ before the last departure, i.e., below the upper boundary of the departure set. This vehicle trajectory intersects the lower boundary of the departure set at some midway point, and traverses Region II to reach the CBD. If we now let this vehicle trajectory be the upper boundary of a new, truncated departure set, removing all other trajectories that depart after it, then we obtain a truncated departure set solution that is identical to the original departure set over the regions not truncated. The flow at the tip of this truncated departure set is less than capacity flow. Based on the results of this section, we can further conclude (with Greenshields' relation) that, for a given flow value less than or equal to capacity flow at the tip of the horn and a given $\frac{\beta}{\alpha}$, there is a unique population distribution solving the Corridor Problem with the modified departure set.

We begin by introducing Greenshields' Relation, choosing appropriate scale parameters, and then restating the three governing equations with this Relation and scale parameters implemented. We then give a broad overview of our numerical strategy before presenting the details. All algebraic calculations used to derive the results in this

section have been placed in the Appendix at the end of the paper.

3.6.1 Greenshields' Velocity-Density Relation

Greenshields' linear velocity-density relation is

$$V = V_0(1 - \frac{\rho}{\rho_J})$$
$$\rho = \rho_J(1 - \frac{V}{V_0}),$$

where V_0 is the free-flow velocity and ρ_J is the jam density. We write the flow in terms of velocity as

$$F = \rho V = \frac{\rho_J}{V_0} V(V_0 - V),$$

that achieves its maximum value, capacity flow, at $V = \frac{V_0}{2}$, Since we have shown that hypercongestion does not occur, $\frac{V_0}{2} \leq V \leq V_0$, and we may write velocity in terms of flow as

$$V = \frac{V_0}{2}\left(1 + \sqrt{1 - \frac{4F}{\rho_J V_0}}\right).$$

Also,

$$(\rho V)' \equiv \frac{d}{d\rho}(\rho V) = 2V - V_0.$$

3.6.2 Scale Parameters and Notation

The natural units of distance and time are $\bar{x}$ and $\frac{\bar{x}}{V_0}$, respectively. Thus, we choose our units such that $\bar{x} = 1$ and $\frac{\bar{x}}{V_0} = 1$, so that $V_0 = 1$ and $\frac{1}{2} \leq V \leq 1$. We also

choose the units of population so that the jam density, $\rho_J = 4$, that results in the flow, F varying from 0 to a capacity flow value of 1. The only relevant parameter, then, is $\frac{\beta}{\alpha} < 1$, that is the ratio of the unit time early cost to the unit travel time cost. Finally, let w denote the slope of the flow vs. density curve, $w = 2V - 1$ where $0 \leq w \leq 1$. Newell, 1993 refers to w as the "wave velocity," and although it is not necessary to introduce this additional function, it is useful in simplifying the algebraic manipulations that follow. Using these units and notation we restate the three governing equations in a form that will be useful for our numerical procedure:

$$u_0 = \frac{\alpha - \beta}{\alpha}\left[\frac{1 - b(u_f)}{w(b(u_f))} + u_f + \int_0^{b(u_f)} \frac{2}{1 + \sqrt{1 - F(x')}}\, dx' - 1\right] \tag{3.17a}$$

$$F(b(u_f)) = \frac{\alpha - \beta}{\alpha} F(b(u_0)) \tag{3.17b}$$

$$\int_{u_f}^{u_0} F\left(b(u')\right)\, du' = \frac{(1 - w(b(u_f)))^2}{w(b(u_f))}(1 - b(u_f)). \tag{3.17c}$$

3.6.3 Overview of Numerical Solution

To determine any solutions to the Corridor Problem with the modified departure set under Greenshields' relation, we must simultaneously solve (3.17) subject to the constraints $b(0) = 0$ and $F(0) = 0$. We will first seek a departure set solution that reaches capacity flow. As mentioned earlier, the existence of such a solution will imply a continuous family of truncated solutions that do not reach capacity flow.

(3.17) involve a natural pairing of u values, u_0 and u_f. We utilize this pairing to discretize the problem, using the second governing equation (3.17b) to exactly determine the flow values at each of the discretization points, with flow values ranging from 0 to the capacity flow value of 1. If our discretization is fine enough, then we can linearly approximate both the lower boundary curve and the flow over each discretized subin-

terval, that enables us to "discretize" the first and third governing equations, (3.17a) and (3.17c), i.e., to restate them in a form on each discretized subinterval that does not involve integrals. Our numerical procedure takes the lower boundary curve and flow values at one discretized subinterval, inputs them into the discretized versions of the first and third governing equations for the next discretized subinterval, yielding a pair of linear equations for the lower boundary curve and flow values for the next discretized subinterval. The unique solution to this pair of linear equations yields the lower boundary curve and flow values for the next discretized subinterval. Futhermore, the initial seed values for this numerical procedure are uniquely determined by linearly approximating the lower boundary curve and flow values on the first discretized subinterval. Hence, by making valid linear approximations we numerically construct a solution to the three governing equations, and at each step of our numerical procedure the solution values we obtain are uniquely determined. Thus, this procedure provides a numerically constructive proof that, given Greenshields' relation and a ratio of parameters $\frac{\beta}{\alpha}$, there is a unique solution to the Corridor Problem with the modified departure set that reaches capacity flow.

A complication arises when we attempt to construct the final segment of the lower boundary curve, since the u_f values in this segment do not have a corresponding u_0 value with which they can be paired, and thus the first and third governing equations will no longer be applicable. We alleviate this problem by guessing a value of the lower boundary curve for the subsequent discretized subinterval, using our guess to numerically construct a vehicle trajectory that should theoretically intersect the lower boundary curve exactly at the point that we guessed, and choosing repeated guesses until we find that the vehicle trajectory does intersect our point at some desired level of tolerance. This procedure is re-iterated until the final segment has been constructed.

The following sections provide the details of the numerical procedure, with all algebraic calculations relegated to the appendix.

3.6.4 Iterated Sequence of Discretized Flow Values

Over the departure set, we seek a solution such that the flow increases from 0 to a capacity flow value of 1. We choose an initial flow value, $0 < F_0 \leq 1$, determine the point on the lower boundary curve of the departure set where this flow value is attained, track the trajectory curve through this point until it reaches the CBD, and then backtrack along the characteristic line intersecting this point until reaching the lower boundary curve. This has already been graphically illustrated in Figure 3.6. By (3.17b), the flow at this new iterated value is $F_1 = F_0 \left(\frac{\alpha-\beta}{\alpha}\right)$. Continuing this iterative procedure, as we approach the residential boundary ($x = 0$), the nth iterated flow value is $F_n = F_0 \left(\frac{\alpha-\beta}{\alpha}\right)^n$, that approaches 0.

We graphically illustrate this iterative procedure in Figure 3.8, with an initial flow value of $F_0 = 1$, capacity flow. Note that (3.17b) allows us to develop a sequence of flow iterates, $F_i \equiv F(b(u_i))$, without knowing the corresponding u_i and $b_i \equiv b(u_i)$ values.

As illustrated in Figure 3.8, if we begin with $F_0 = 1$ and iterate this procedure N times, then we partition the lower boundary curve into $N+1$ segments, where the first segment begins with a flow value of 1 and ends with a flow value of $\frac{\alpha-\beta}{\alpha}$. We further subdivide the flow values on this first segment into k subdivisions, equally spaced between the values $\frac{\alpha-\beta}{\alpha}$ and 1. We apply the iterative procedure to each of the k flow values within this segment, generating k flow values within each of the subsequent segments. Hence, if we consider the first N segments constructed, each having k subdivisions, then we obtain a sequence with a total of $N_f = Nk$ flow values. We illustrate this idea in Fig-

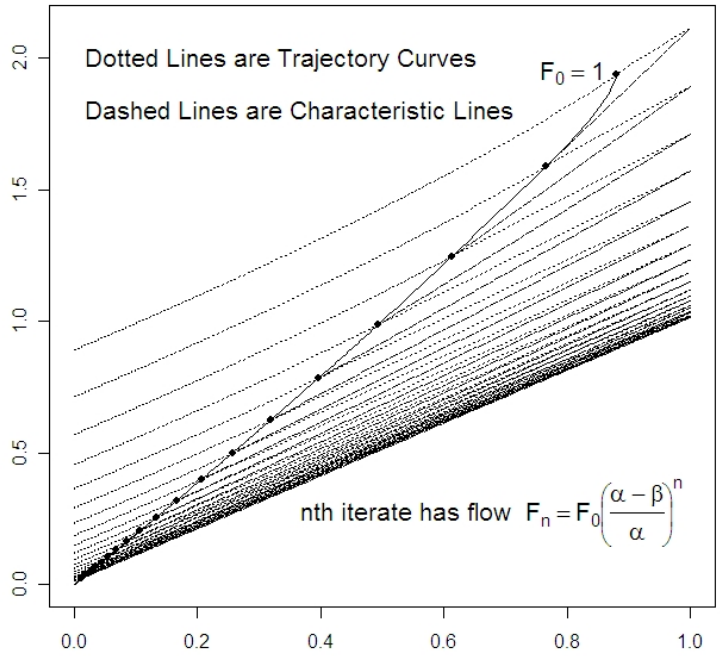

Figure 3.8: Iterated flow values. Iterative procedure tracks the intersection at the CBD of a trajectory curve with a characteristic line.

ure 3.9, where we have reordered the sequence of flow values such that F_1 is the smallest flow value in the sequence, increasing to the capacity flow value, $F_{N_f} = F_{Nk} = 1$. The generation of this sequence of flow iterates derives from (3.17b), and does not require knowledge of the corresponding u_i or $b(u_i)$ values. We state a closed form formula for all flow values in this sequence, given values of N and k:

$$F_{i+(N-j)k} = \left(\frac{\alpha-\beta}{\alpha}\right)^{j-1}\left(\frac{\alpha-\beta}{\alpha} + \frac{i}{k}\left[1 - \frac{\alpha-\beta}{\alpha}\right]\right), \quad \forall\, 1 \le j \le N,\ 1 \le i \le k \tag{3.18}$$

Since $V = \frac{1}{2}\left(1+\sqrt{1-F}\right)$ and $w = \sqrt{1-F}$, we may use this sequence to obtain corresponding sequences of $v_i \equiv v(b(u_i))$ and $w_i \equiv w(b(u_i))$ values.

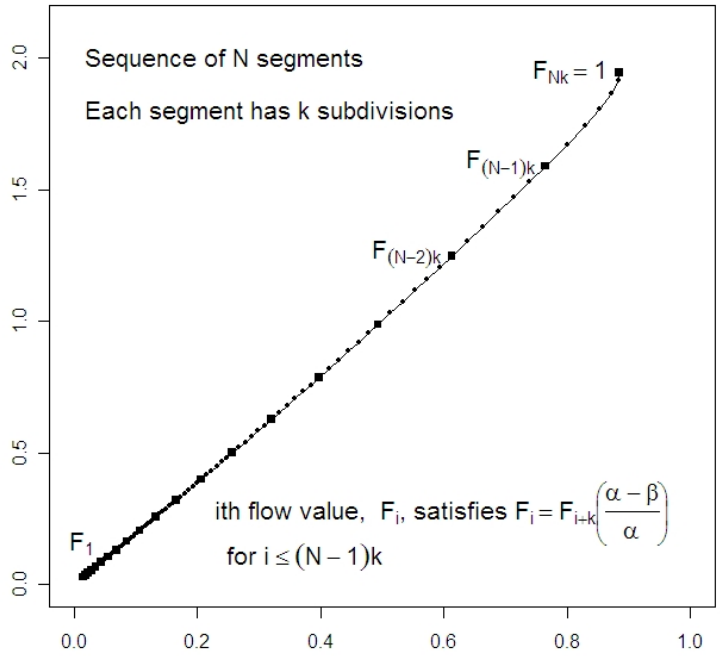

Figure 3.9: Sequence of flow values from F_1 to $F_{N_f} = F_{Nk} = 1$. For all i, $1 \leq i \leq (N-1)k$, the ith flow value satisfies $F_i = F_{i+k}\left(\frac{\alpha-\beta}{\alpha}\right)$.

3.6.5 Discretization of the First and Third Governing Equations

The (x,t) coordinates of the lower boundary curve are parametrized in terms of u as $(b(u), u+T_I(u)) = \left(b(u), u + \int_0^{b(u)} \frac{dx'}{v(x')}\right)$. Corresponding to our sequence of flow values, we denote the sequence of coordinates of the lower boundary curve as (b_i, t_i), $1 \leq i \leq N_f$. Similarly, we denote the iterated sequence of cumulative flow values as $A_i \equiv A(u_i) = \int_0^{u_i} F(b(u'))\, du'$.

If we choose k large enough, then over each subinterval (u_{i-1}, u_i), we may

approximate $F(b(u))$ and $b(u)$ as linear functions of u:

$$F(b(u)) \approx F_{i-1} + \frac{F_i - F_{i-1}}{u_i - u_{i-1}}(u - u_{i-1}), \quad u \in (u_{i-1}, u_i)$$

$$b(u) \approx b_{i-1} + \frac{b_i - b_{i-1}}{u_i - u_{i-1}}(u - u_{i-1}), \quad u \in (u_{i-1}, u_i). \tag{3.19}$$

As shown in the Appendix, we may use these linear approximations to determine expressions for the t_i and A_i values, that may be substituted into (3.17a) and (3.17c) to yield discretized versions of these two equations. The algebraic manipulations have been placed in the Appendix, and we present here only the final results, namely the discretized version of the first governing equation,

$$\begin{aligned} u_{i+k} = &\left(\frac{\alpha - \beta}{\alpha}\right)\left\{\frac{1 - b_i}{w_i} + t_{i-1} + (u_i - u_{i-1})\right. \\ &\left. + \frac{4(b_i - b_{i-1})}{F_i - F_{i-1}}\left[\log\left(\frac{1 + w_i}{1 + w_{i-1}}\right) - (w_i - w_{i-1})\right] - 1\right\}, \quad 1 \le i \le N_f - k, \end{aligned} \tag{3.20}$$

and the discretized version of the third governing equation,

$$u_{i+k} = u_{i+k-1} + \frac{2}{F_{i+k} + F_{i+k-1}}\left[\frac{(1 - w_i)^2}{w_i}(1 - b_i) - A_{i+k-1} + A_i\right], \forall i,\ 1 \le i \le N_f - k. \tag{3.21}$$

Although the three governing equations, (3.12), (3.13) and (3.14), are valid for any monotonic velocity-density relationship, if we attempt to discretize the first and third governing equations then we will generally obtain two non-linear equations in two unknowns (e.g., using Underwood's relation). However, if we use Greenshields' velocity-density relationship then the discretized versions of the first and third governing equations are a pair of *linear* equations in two unknowns, that may be solved exactly.

3.6.6 Iterative Procedure

Using (3.18) we may determine all F_i and w_i values, for $1 \leq i \leq N_f$. Suppose, that for a given value of i, $i < N_f - k$, we know the values $b_1, \ldots, b_{i-1}$, $t_1, \ldots, t_{i-1}$, $u_1, \ldots, u_{i+k-1}$ and $A_1, \ldots, A_{i+k-1}$. Then the discretized versions of the first and third governing equations, (3.21) and (3.20), are a pair of linear equations in the unknown quantities b_i and u_{i+k}, in terms of known quantities. We may solve these equations to obtain the values of b_i and u_{i+k}, and then use these values in (A-3.3) and (A-3.1) to determine A_{i+k} and t_i. This procedure may be iterated until $i = N_f - k$. The algebraic details are provided in the Appendix, where we derive (A-3.4) and (A-3.5), which completely summarize the core of our iterative procedure. Specifically, given the initial seed values b_1, t_1, $u_1, \ldots, u_{1+k}$ and $A_1, \ldots, A_{1+k}$, we iteratively use (A-3.4) and (A-3.5) to determine $b_1, \ldots, b_{N_f-k}$, $t_1, \ldots, t_{N_f-k}$, $u_1, \ldots, u_{N_f}$ and $A_1, \ldots, A_{N_f}$. At the conclusion of this procedure, the only undetermined quantities will be the (b_i, t_i) values in the last segment, from $i = N_f - k + 1$ to $i = N_f$.

3.6.7 Initializing Seed Values

To initiate the above iterative procedure, we must provide values for b_1, t_1, $u_1, \ldots, u_{1+k}$ and $A_1, \ldots, A_{1+k}$. If we choose N large enough, then $F_1, \ldots, F_{1+k}$ will be close to 0, and over the interval $(0, u_{1+k})$ we can approximate $F(b(u))$ as a linear function of u, $F(b(u)) \approx \frac{F_1}{u_1} u$. Using this linear approximation we then apply the discretized versions of the first and third governing equations, that enables us to solve for u_1 and b_1, and, hence, determine all necessary initializing seed values. The algebraic details are provided in the Appendix, with the initial seed values given in (A-3.8).

3.6.8 Final Segment

After implementing the above iterative procedure, we will have determined all F_i, w_i, u_i and A_i values for $i = 1, \ldots, N_f$, and will have determined all b_i, t_i values for $i = 1, \ldots, N_f - k$. The only remaining values to determine are $(b_{N_f-k+1}, t_{N_f-k+1}), \ldots, (b_{N_f}, t_{N_f})$, corresponding to the (x, t) coordinates of the lower boundary curve in the final segment. Furthermore, since we have determined u_{N_f} (the departure time at $x = 0$ of the final cohort of vehicles that arrives at the CBD exactly at time $\bar{t}$), from (3.4) we can calculate $\bar{t}$ as $\bar{t} = \frac{\alpha}{\alpha-\beta} u_{N_f} + 1$.

In constructing the final segment, the first governing equation will no longer be applicable, since the characteristic lines emanating from the final segment will intersect the CBD at a point greater than $\bar{t}$, that does not contain the intersection of any vehicle trajectories. Recall that in deriving the third governing equation, we first determined the change in the cumulative flow along a characteristic line from the lower boundary of the departure set up to a vehicle trajectory (3.16). We simplified this equation by only considering the point where the vehicle trajectory intersects the CBD, i.e., by setting $u = u_f$ so that $\tilde{x}(u) = \bar{x}$ in (3.16), yielding the third governing equation. To determine the final segment of the lower boundary curve we use (3.16) in its more general form.

To construct this final segment of k pieces, given a known value of b_i on one piece we guess a value for b_{i+1} on the next piece. We then further subdivide this single piece into m subdivisions. The characteristic lines emanating from the endpoints of these m subdivisions partition the vehicle trajectory curve through b_{i+1} into m pieces. We assume that, over each of the m partitions, the vehicle trajectory curve can be approximated by a linear function, that will be valid if the subdivisions are chosen fine enough, i.e., if m is chosen large enough. We then use this linear approximation for the

vehicle trajectory curve on each subdivision to explicitly calculate the vehicle trajectory curve. If our initial guess for b_{i+1} was correct, then the vehicle trajectory curve we calculate should intersect the lower boundary curve at b_{i+1}. We try different values of b_{i+1} until obtaining one such that the vehicle trajectory curve we calculate based on the b_{i+1} value intersects the lower boundary curve at b_{i+1} with sufficiently small error. Using this b_{i+1} value we calculate t_{i+1} based on our linear approximation for $b(u)$ on the $i+1$th piece. This procedure is iterated until all b_i, t_i values have been determined.

The algebraic details of this procedure are provided in the Appendix, along with a graph illustrating this procedure (Figure 3.14).

3.7 Numerical Results (Greenshields)

3.7.1 Departure Set Solutions

We implement the numerical procedure for the ratio of parameter values $\frac{\beta}{\alpha} = 0.2, 0.4, 0.6$ and 0.8. A graph of each departure set, along with the vehicle trajectory corresponding to the final cohort of vehicles, is displayed in Figure 3.10. We have plotted all graphs with the same axis, to discern the behaviour of the solution with respect to the ratio of parameter values. As the unit time early cost, β, approaches the unit travel time cost, α, the width and length of the departure set decreases, approaching free-flow condition of zero traffic departing $x = 0$ at time $u = 0$ and arriving at the CBD at time $\bar{t} = 1$. In each of these graphs the flow reaches the capacity flow value of 1 at the tip of the departure set. At this point the slope of the characteristic curves will be infinite, and since the slope of the lower boundary curve must be greater than the slope of the characteristic curves, it too must be infinite. We observe this behaviour in our numerical solutions, as the tip of the departure set becomes vertical.

To understand why the departure set solutions are so sensitive to the value of $\frac{\beta}{\alpha}$, recall from (3.4) that $\Delta a = \frac{\alpha}{\alpha-\beta}\Delta t$. The higher is $\frac{\beta}{\alpha}$, the more rapidly must congestion increase to satisfy the trip-timing condition. Since the maximum level of congestion corresponds to capacity flow, the higher is $\frac{\beta}{\alpha}$ the shorter must be the rush and the smaller must be the population.

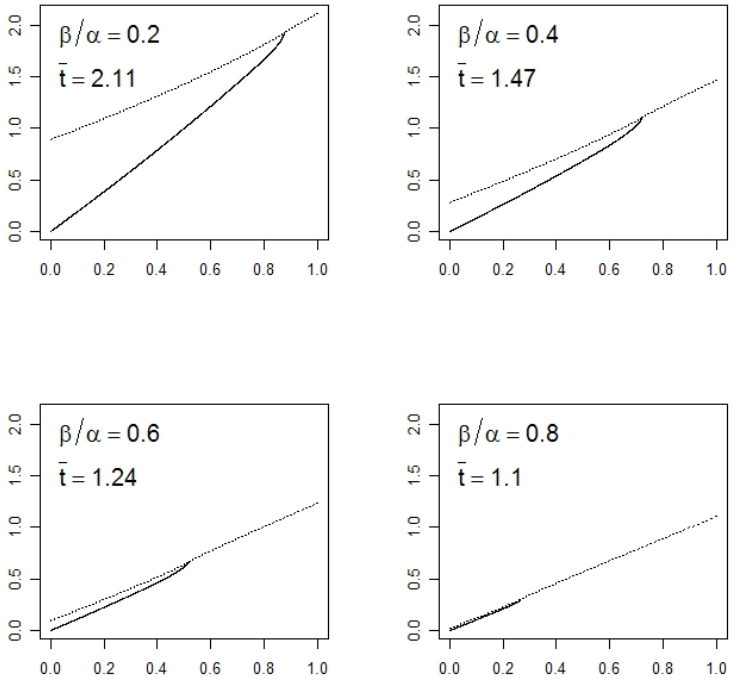

Figure 3.10: Numerically constructed departure set solutions for various ratios of parameter values, $0 < \frac{\beta}{\alpha} < 1$. The lower boundary curve of the departure set is graphed with a solid line. The upper boundary curve of the departure set, that corresponds to the trajectory of the final cohort of vehicles (that arrives at the CBD at $\bar{t}$), is graphed with a dashed line.

The following table lists the numerical values (to two decimal places) of several important features of our numerical solutions.

Numerical Results with $\bar{x} = 1$, $V_0 = 1$ and $\rho_J = 4$				
	$\frac{\beta}{\alpha}$ Values			
	0.2	0.4	0.6	0.8
Width of Departure Set at $x = 0$ (time units)	0.89	0.28	0.10	0.02
Tip of Departure Set (distance, time) units	(0.88, 1.93)	(0.72, 1.10)	(0.52, 0.67)	(0.27, 0.31)
$\bar{t}$ (time units)	2.11	1.47	1.24	1.10
Total Population (population units)	0.43	0.13	0.03	0.005

3.7.2 Corresponding Population Densities

We previously showed how the TT condition implies that, at each location, within the interior of the departure set and on the upper boundary of the departure set the departure rate, $n(x,t)$, is constant (3.9). Hence, once we have numerically determined the flow, we may numerically differentiate (3.10) to determine the constant departure rate at each location. From (3.2), we may determine the population density at each location by multiplying the departure rate at that location by the width of the departure set at that location.

In Figure 3.11 we have graphed the population densities corresponding to the sample departure set solutions in Figure 3.10. The integral of the population density is the total population, which also equals the cumulative flow value along the trajectory in Region II for the final cohort of vehicles. Note that if we considered a truncated departure set solution that did not reach the capacity flow value of 1, then the corresponding population densities would have the same general shape but would include less overall population and would reach zero earlier.

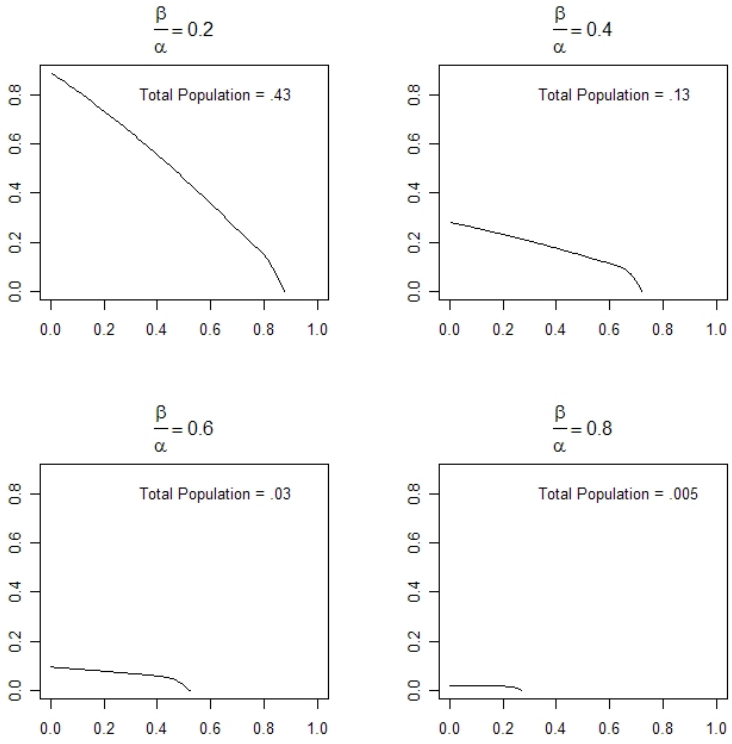

Figure 3.11: Population densities corresponding to the departure set solutions in Figure 3.10. We calculated the departure rate at each location by numerically differentiating the flow values with respect to location. The population density is obtained by multiplying the departure rate by the departure set width at that location.

3.7.3 Interpretation of Results

To gain a more intuitive feeling for our results, we transform our results using more realistic values of $\bar{x}$, V_0 and ρ_J. Suppose the residential settlement is $\bar{x} = 10$ mi long. To satisfy $\bar{x} = 1$ we must choose distance units so that

$$1 \text{ distance unit } = 10 \text{ mi.}$$

Suppose free-flow velocity is $V_0 = 50$ mi/hr. To satisfy $V_0 = 1$ we must choose time units so that

$$50 \text{ mi/hr } = \frac{10 \text{ mi}}{\frac{1}{5} \text{ hr}} = \frac{1 \text{ distance unit}}{\frac{1}{5} \text{ hr}} = 1.$$

Thus,

$$1 \text{ time unit } = \frac{1}{5} \text{ hr } = 12 \text{ min.}$$

Suppose the jam-density for a single traffic lane is $\frac{1 \text{ vehicle}}{16 \text{ ft}}$. If the road has a constant width of four lanes, then the jam density of the road is $\frac{4 \text{ vehicles}}{16 \text{ ft}}$. To satisfy $\rho_J = 4$ we must choose population units so that

$$\frac{4 \text{ vehicles}}{16 \text{ ft}} = 1320 \frac{\text{vehicles}}{\text{mi}} = \frac{13,200 \text{ vehicles}}{10 \text{ mi}} = 4 \times \frac{3300 \text{ vehicles}}{1 \text{ distance unit}} = 4.$$

Thus,

$$1 \text{ population unit} = 3300 \text{ vehicles.}$$

To put this number into perspective, let us suppose that the city is 4 miles wide and that 5% of the city area is used for roads from the suburbs to the CBD. Since the typical lane width is 11 feet, there would be $[(5280)(5)(0.05)] \div [(4)(11)] = 30$ such roads in the city, in which case a population unit for the entire city would be 99,000 vehicles.

For each of the four ratios of parameters $\frac{\beta}{\alpha} = 0.2, 0.4, 0.6, 0.8$, the departure set solutions and corresponding population densities have the same graphs as shown in Figures 3.10 and 3.11, except that we must use the above units of distance, time and population along the axes in those figures. The following table lists the numerical values of the same features presented in the previous table, but using the current system of units.

Numerical Results with $\bar{x} = 10$ mi, $V_0 = 50$ mi/hr and $\rho_J = \frac{4 \text{ vehicles}}{16 \text{ ft}}$				
	$\frac{\beta}{\alpha}$ Values			
	0.2	0.4	0.6	0.8
Width of Departure Set at $x = 0$ (minutes)	10.7	3.4	1.2	0.2
Tip of Departure Set (miles, minutes)	(8.8, 23.2)	(7.2, 13.2)	(5.2, 8.0)	(2.7, 3.7)
$\bar{t}$ (minutes)	25.3	17.6	14.9	13.2
Total Population per Four-Lane Road (vehicles)	1419	429	99	16.5

3.7.4 Comparison with Empirical Data

A model is just a model. There are several good reasons to believe that the spatial dynamics of rush-hour traffic congestion are considerably more complex than our model suggests. First, the equilibrium traffic dynamics we characterized applied for only a family of population distributions along the traffic corridor, and the actual population distribution may be quite different. Second, in US metropolitan areas, an increasingly large proportion of trips do not have a commuting purpose, and our analysis does not apply to non-work trips. Third, in almost all metropolitan areas, the CBD is becoming less dominant as an employment center, and employment locations are becoming increasingly decentralized, dispersed, and polycentric. Fourth, employers respond to increased traffic congestion by offering less concentrated work start times. And fifth, the bottleneck model implies that, in early morning rush-hour travel from a common origin to a common destination, workers with higher $\frac{\alpha}{\beta}$ ratios commute earlier since they are relatively more averse to congestion than an inconvenient arrival time. Since those with higher $\frac{\alpha}{\beta}$ ratios tend to have higher income, the spatial dynamics of traffic flow should depend on the distribution of income along the corridor. Nevertheless, a model is not very useful unless it has predictive content. A similar litany of caveats can be given about the monocentric city model, and yet many of its predictions are supported by

empirical evidence.

To compare the predicted equilibrium departure pattern with an actual commuting pattern, Figure 3.12 presents[1] a surface of smoothed flow rates along interstate 60W leading into downtown Los Angeles from 4 am to 10 am on Tuesday, May 4, 2010. Data are from 23 traffic stations irregularly spaced along a 15-mile segment of freeway. Each traffic station records the number of vehicle counts per 5-minute interval. Linear local regression is used to smooth the traffic flow rates over time and space. According to the theory, the traffic flow surface should have a horn-shaped base, with the wide end of the horn extending along the time axis and with the tip on the far upper corner (at least at 9am), and its height should be increasing from the wide end of the horn to the tip. The actual surface looks quite different.

Figure 3.13 displays the data in a form that provides a less discriminating test. According to the theory, the flow-weighted mean time at a location should increase with proximity to the CBD, and indeed it does. Even though the theory passes this weak test, it is clear that much more research will be needed before we can say that we understand the spatial dynamics of rush-hour traffic congestion reasonably well.

3.8 Concluding Comments

3.8.1 Directions for Future Research

This paper takes a preliminary step in developing a theory of the spatial dynamics of metropolitan traffic congestion. Even the preliminary step is really only half

[1]We are very thankful to Jifei Ban for collecting the raw data and producing the graphs. Data was collected from the Caltrans Performance Measurement System (PeMS, http://pems.dot.ca.gov/), and the local linear regression smoothing and the graphs were produced using the R-packages *locfit* and *lattice*.

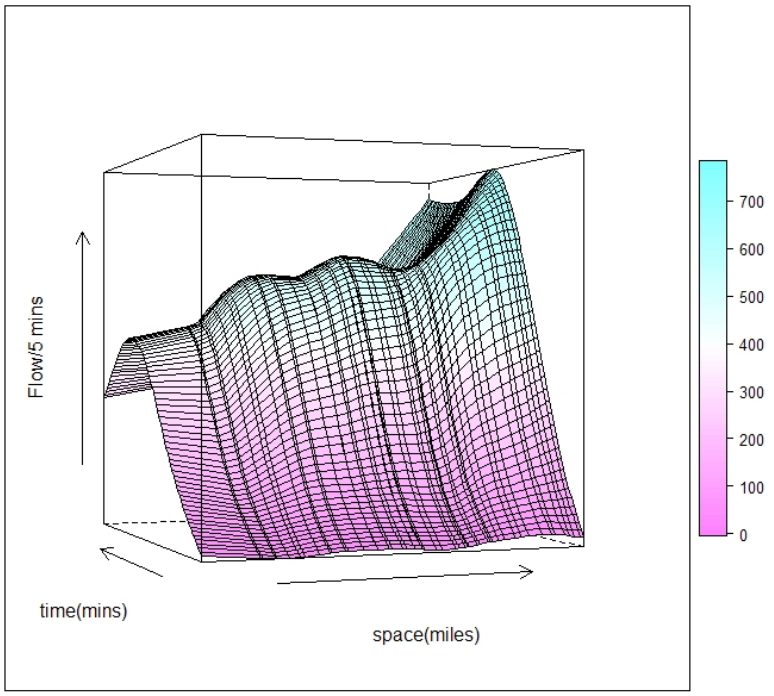

Figure 3.12: Flow rates along freeway 60-westbound extending a distance of 15 miles from downtown Los Angeles, collected from 4am to 10am on a Tuesday morning. Flow rates are obtained from vehicle counts per 5 minute interval, collected at 23 irregularly spaced stations. Linear local regression is used to smooth the raw flow values over time and space.

a step. When we started this research, we hoped to provide a complete solution to the Corridor Problem. Instead, we have succeeded in providing a solution for only a restricted family of population distributions along the traffic corridor. What remains to be done to obtain a complete solution? Frankly, we dont know. Initially we conjectured that there are mass points along the lower boundary of the departure set, but these cannot be accommodated with LWR traffic congestion. Then we conjectured that there is queuing at entry points, but in a decentralized environment queuing occurs only when there is a capacity constraint on entry, which is not a feature of the model. If not mass points or queuing, what? To deal with this puzzle, we have adopted the strategy of solving a sequence of simpler models culminating in the Corridor Problem. Building on

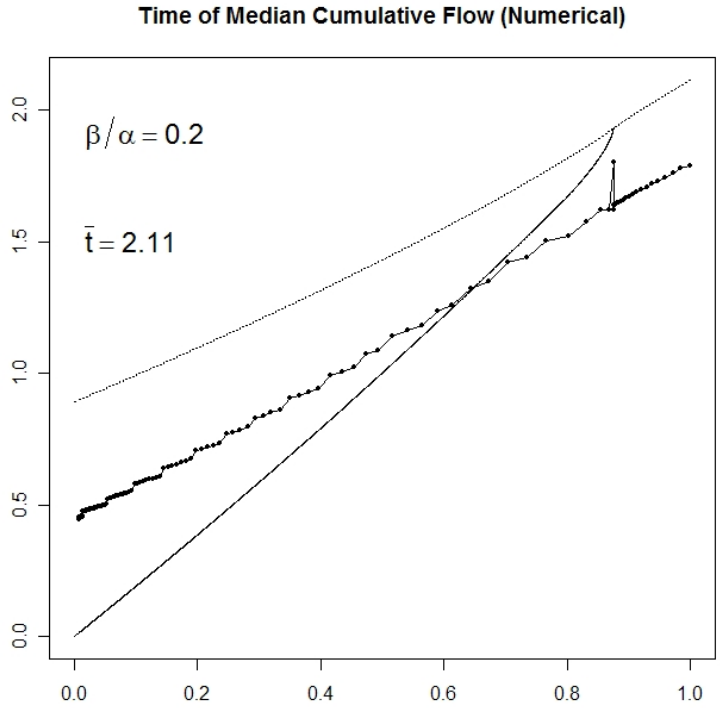

Figure 3.13: Flow-weighted mean times at each location from the raw data collected for Figure 3.12. The flow-weighted mean time steadily increases as we near the CBD, which is in agreement with our analytic solution.

Newell, 1988, we have solved for the equilibrium and optimum for a single-entry corridor (Arnott and DePalma, 2011) for a particular form of the congestion technology The next step is to solve for the equilibrium and optimum for a two-entry corridor (the origin and further along the Corridor), the step after that for an n-entry corridor, and the step after that a continuum-entry corridor as a limit. The Corridor Problem is intriguing partly because it is so difficult. In the real world, the spatial dynamics of traffic congestion are well behaved, being much the same from day to day, strongly suggesting that an equilibrium exists. Why then is it so difficult to characterize?

In searching the literature, we uncovered no empirical papers on the spatial dynamics of rush-hour traffic congestion at the level of a metropolitan area. There is a branch of literature in traffic flow analysis that looks at the dynamics of traffic

congestion along individual freeways, with a view towards understanding how traffic jams form and dissipate, and how they affect flow, but all the papers in this literature take as given the pattern of entries onto the road. There is also an emerging literature (Geroliminis and Daganzo, 2007) that documents the existence of a stable macroscopic diagram for rush-hour traffic congestion in the center of metropolitan areas, but it does not examine the relationship between rush-hour traffic dynamics at different locations. Most large metropolitan areas, at least in Europe and North America, have undertaken at least one large travel diary survey. Using these data, it should not be difficult to document the empirical regularities of rush-hour traffic dynamics.

To achieve analytical tractability, theory employs toy models. These models elucidate basic principles but can take us only so far in understanding empirical phenomena. Bridging the gap are theory-based programming models. The trip-timing condition is a variational inequality. Given the current state of the art, it should be possible to model the equilibrium spatial dynamics of rush-hour traffic congestion as the solution to variational inequality problems, taking as given the actual spatial patterns of residential locations and of employment locations and their work start-times. The major difficulty will be incorporating household activity scheduling and non-commuting travel.

More generally, household and firm location decisions, household travel-time decisions, and firm work-start-time decisions are all interdependent. Addressing these interdependencies requires a full-blown general (economic) equilibrium model, in which prices adjust to clear land, labor, and other markets at all locations and times. Current computable general equilibrium models of metropolitan transportation and land use (Anas and Liu, 2007) are conceptually sophisticated and descriptively rich. They have not been applied to the study of intra-day traffic dynamics, but could be extended to do so.

3.8.2 Conclusion

Determining equilibrium traffic flow over the course of the day for an entire metropolitan area is an important unsolved problem in urban transport economic theory and in transportation science. The problem appears to be an important one. Traditional network equilibrium analysis, as well as the theory of land use, employs models that abstract from intra-day variation in traffic. Since traffic congestion is an inherently nonconvex phenomenon, one wonders how much bias and loss the aggregation implicit in this abstraction entails. This paper investigated a simple model of the spatial dynamics of morning rush-hour congestion in the absence of tolls. There is a single corridor connecting a continuum of residential locations to the central business district. There is an exogenous density of identical commuters along the corridor, all of whom make a morning trip to the CBD and have a common desired arrival time. The constant-width road is subject to classical flow congestion. A vehicle's trip price is linear in travel time and early arrival time (late arrival is not permitted). Equilibrium satisfies the trip-timing condition that no vehicle can lower its trip price by altering its departure time. What is the equilibrium pattern of departures? We termed this problem and related extensions (such as the social optimum, and the equilibrium with heterogeneous commuters and price-sensitive demand), *The Corridor Problem*.

Even though the no-toll equilibrium Corridor Problem assumes away many features of a much more complex reality, it appears very difficult to solve. We have not yet succeeded in obtaining a complete solution to the problem, and this paper presented preliminary results.

Consider two locations x' and x'', with x' further from the CBD than x''. Vehicles from both locations travel in one cohort (a group of cars arriving at the CBD

at the same time) arriving at the CBD at an earlier time a' and another at a later time a''. Since the vehicles departing x' and x'' experience the same decrease in schedule delay when traveling in cohort a'' rather than in cohort a', to satisfy the trip-timing equilibrium condition they must experience the same increase in travel time. Building on this observation applied to all locations and cohorts, we proved that successive cohorts are identical except for the addition of vehicles at more central locations. The first cohort contains only vehicles from the locations most distant from the CBD, the second cohort is identical (in entry rate, flow, and density at the most distant locations) except that it adds vehicles at the next most distant locations, and so on, until the last cohort contains vehicles from all locations.

Based on these results, Arnott had conjectured that there would exist an equilibrium with this qualitative departure pattern for an arbitrary distribution of population along the road. However, in constructing numerical solutions of equilibria with this qualitative departure pattern, DePalma discovered that such an equilibrium exists (and when it is exists, is unique) only for a two-parameter family of population distributions along the road. Thus, either an equilibrium does not exist for other population distributions, or our specification of the problem is incomplete, or our solution method somehow overconstrains the problem. We have been unable to uncover the root of the difficulty, which is why we are unable to provide a complete solution.

Once a full solution to the no-toll equilibrium of the Corridor Problem is obtained, there are many interesting applications and extensions to be investigated. But a full solution to the basic problem must come first. The previous subsection proposed a research strategy with this goal.

References

Anas, A. and Liu, Y. (2007). A regional economy, land use, and transportation model (Relutran): formulation, algorithm design, and testing. *Journal of Regional Science*, 47:415-455

Arnott, R. (1979). Unpriced transport congestion. *Journal of Economic Theory*, 21:294–316.

Arnott, R. (2004). The corridor problem: No-toll equilibrium. *Mimeo.*

Arnott, R., de Palma, A., and Lindsey, R. (1994). Bottlenecks in series. *Notes.*

Arnott, R. and DePalma, E. (2011). Morning commute in a single-entry traffic corridor with no late arrivals. Working paper

Beckmann, M. and Puu, T. (1985). *Spatial Economics: Density, Potential and Flow.* Studies in Regional Science and Urban Economics 14, Amsterdam: North-Holland.

Evans, L. (2002). *Partial Differential Equations.* American Mathematical Society, Providence, Rhode Island.

Geroliminis, N. and Daganzo, C. (2007). Macroscopic modeling of traffic in cities. *86th Annual Meeting of the Transportation Research Board*, Paper #07-0413.

Kanemoto, Y. (1976). Cost-benefit analysis and the second-best use of land for transportation. *Journal of Urban Economics*, 4:483–503.

Newell, G. (1988) Traffic flow for the morning commute. *Transportation Science*, 22(1):47–58.

Newell, G. (1993). A simplified theory of kinematic waves in highway traffic, part 1: General theory. *Transpn. Res.-B*, 27B:281–287.

Rhee, H., Aris, R., and Amundson, N. (1986). *First-Order Partial Differential Equations: Volume 1 Theory and Applications of Single Equations.* Prentice-Hall, Englewood Cliffs, New Jersey.

Ross, S. and Yinger, J. (2000). Timing equilibria in an urban model with congestion. *Journal of Urban Economics*, 47:390–413.

Small, K. (1982). The scheduling of consumer activities: Work trips. *American Economic Review*, 72:467–479.

Solow, R. (1972). Congestion, density, and the use of land in transportation. *Swedish Journal of Economics*, 74:161–173.

Solow, R. and Vickrey, W. (1971). Land use in a long, narrow city. *Journal of Economic Theory*, 3:430–447.

Tian, Q., Huang, H., and Yang, H. (2007). Equilibrium properties of the morning peak-period commuting in a many-to-one transit system. *Transportation Research B: Methodological*, 41:616–631.

Vickrey, W. (1969). Congestion theory and transport investment. *American Economic Review*, 59:251–260.

Yinger, J. (1993). Bumper to bumper: A new approach to congestion in an urban model. *Journal of Urban Economics*, 34:249–275.

Notational Glossary

CBD	Central Business District
TT	Trip-Timing condition
x	distance
t	time
$\bar{x}$	location of CBD
$\bar{t}$	work start time at the CBD
$N(x)$	population density
C	total travel cost
α	unit cost of travel time
β	unit cost of time early arrival
$T(x,t)$	travel time from (x,t) to the CBD
$\rho(x,t)$	density (vehicles/length) at (x,t)
$v(x,t)$	velocity at (x,t)
$V(\rho)$	velocity as a function of density
F	flow, density times velocity
$n(x,t)$	entry rate, or departure rate, at (x,t)
$\mathcal{D}$	set of (x,t) points at which departures occur
$p(x)$	equilibrium trip price of a departure at location x
a	arrival time at the CBD
$\hat{T}(x,a)$	travel time of a departure from location x and arriving at time a
$\mathcal{A}$	set of (x,a) points for which arrival rate is positive
u	departure time from location $x=0$

Region I	(x, t) points within the departure set
Region II	(x, t) points below the departure set
$b(u)$	x-coordinate of the lower boundary
$T_I(u)$	travel time to the lower boundary
$A(u)$	cumulative flow along the lower boundary
$\hat{A}(x, t)$	cumulative flow in Region II
u_0, u_f	pair of departure times from $x = 0$, as in Figure 3.6
$(\tilde{x}, \tilde{t})$	space-time coordinates of a trajectory in Region II
V_0	maximum, free-flow velocity (Greenshields')
ρ_J	jam density at which velocity is zero (Greenshields')
w	"wave-velocity" or slope $\frac{d}{d\rho}(\rho V)$
i, j	dummy indices
N	number of segments of lower boundary
k	number of subdivisions within each segment
$N_f = Nk$	number of points on lower boundary curve
$u_i, b_i, F_i, v_i, w_i, A_i, t_i$	function values along the lower boundary curve
$Q, Q_1, \ldots, Q_4$	quantities calculated in the iterative procedure
m	number of subintervals in final segment
$(F_i)_j, (w_i)_j, (v_i)_j, (b_i)_j, (t_i)_j$	values on jth subinterval, ith subdivision of the final segment of the lower boundary curve

Appendix

Discretization of the First Governing Equation

The (x,t) coordinates of the lower boundary curve are parametrized in terms of u as $(b(u), u+T_I(u)) = \left(b(u), u + \int_0^{b(u)} \frac{dx'}{v(x')}\right)$. Corresponding to our sequence of flow values, the sequence of coordinates of the lower boundary curve are (b_i, t_i), $1 \le i \le N_f$. The time coordinates, t_i, can be expressed as

$$\begin{aligned} t_i &= u_i + \int_0^{b_i} \frac{2}{1+\sqrt{1-F(x')}}\,dx' \\ &= u_{i-1} + \int_0^{b_{i-1}} \frac{2}{1+\sqrt{1-F(x')}}\,dx' + (u_i - u_{i-1}) + \int_{b_{i-1}}^{b_i} \frac{2}{1+\sqrt{1-F(x')}}\,dx' \\ &= t_{i-1} + (u_i - u_{i-1}) + \int_{b_{i-1}}^{b_i} \frac{2}{1+\sqrt{1-F(x')}}\,dx'. \end{aligned}$$

If we choose k large enough, then over each subinterval (u_{i-1}, u_i), we may approximate $F(b(u))$ and $b(u)$ as linear functions of u, as in (3.19):

$$F(b(u)) \approx F_{i-1} + \frac{F_i - F_{i-1}}{u_i - u_{i-1}}(u - u_{i-1}), \quad u \in (u_{i-1}, u_i)$$

$$b(u) \approx b_{i-1} + \frac{b_i - b_{i-1}}{u_i - u_{i-1}}(u - u_{i-1}), \quad u \in (u_{i-1}, u_i).$$

We now use these linear approximations to approximate t_i:

$$\begin{aligned} t_i &= t_{i-1} + (u_i - u_{i-1}) + \int_{u_{i-1}}^{u_i} \frac{2\frac{db}{du}}{1+\sqrt{1-F(b(u))}}\,du \\ &\approx t_{i-1} + (u_i - u_{i-1}) + \int_{u_{i-1}}^{u_i} \frac{2\frac{b_i - b_{i-1}}{u_i - u_{i-1}}}{1+\sqrt{1-F_{i-1} - \frac{F_i - F_{i-1}}{u_i - u_{i-1}}(u - u_{i-1})}}\,du \\ &= t_{i-1} + (u_i - u_{i-1}) + \frac{4(b_i - b_{i-1})}{F_i - F_{i-1}}\left[\log\left(\frac{1+w_i}{1+w_{i-1}}\right) - (w_i - w_{i-1})\right]. \end{aligned} \tag{A-3.1}$$

Substituting this approximation for t_i into the first governing equation, (3.17a), yields a discretized version of the first governing equation, presented earlier in the paper in (3.20):

$$u_{i+k} = \left(\frac{\alpha-\beta}{\alpha}\right)\left\{\frac{1-b_i}{w_i} + t_{i-1} + (u_i - u_{i-1}) \right.$$
$$\left. + \frac{4(b_i - b_{i-1})}{F_i - F_{i-1}}\left[\log\left(\frac{1+w_i}{1+w_{i-1}}\right) - (w_i - w_{i-1})\right] - 1\right\}, \quad 1 \le i \le N_f - k.$$

Note that this equation is a linear equation in the unknown quantities u_{i+k} and b_i. If, instead of using Greenshields' relation, we had used another velocity-density relation, then the discretized version of the first governing equation would have been a nonlinear equation in these two unknown quantities.

Discretization of the Third Governing Equation

Let $A_i \equiv A(u_i) = \int_0^{u_i} F(b(u'))\,du'$. From the third governing equation, (3.17c), for all i, $1 \le i \le N_f - k$,

$$\int_{u_i}^{u_{i+k}} F(b(u'))\,du' = A_{i+k} - A_i = \frac{(1-w_i)^2}{w_i}(1-b_i). \tag{A-3.2}$$

If we choose k large enough, then over each subinterval (u_{i-1}, u_i), we may approximate $F(b(u))$ as a linear function of u, as in (3.19):

$$F(b(u)) \approx F_{i-1} + \frac{F_i - F_{i-1}}{u_i - u_{i-1}}(u - u_{i-1}), \quad u \in (u_{i-1}, u_i).$$

We use this linear approximation to directly approximate A_{i+k} as

$$\begin{aligned} A_{i+k} &= \int_0^{u_{i+k}} F(b(u'))\,du' \\ &= A_{i+k-1} + \int_{u_{i+k-1}}^{u_{i+k}} F(b(u'))\,du' \\ &\approx A_{i+k-1} + \frac{u_{i+k} - u_{i+k-1}}{2}\left(F_{i+k} + F_{i+k-1}\right). \end{aligned} \tag{A-3.3}$$

Substituting this expression for A_{i+k} into (A-3.2) yields

$$A_{i+k-1} + \frac{u_{i+k} - u_{i+k-1}}{2}\left(F_{i+k} + F_{i+k-1}\right) - A_i = \frac{(1-w_i)^2}{w_i}(1-b_i).$$

We rearrange this equation to solve for u_{i+k}, yielding our discretized version of the third governing equation, presented earlier in the paper in (3.21):

$$u_{i+k} = u_{i+k-1} + \frac{2}{F_{i+k} + F_{i+k-1}}\left[\frac{(1-w_i)^2}{w_i}(1-b_i) - A_{i+k-1} + A_i\right],\ \forall i,\ 1 \le i \le N_f - k.$$

As with the discretized version of the first governing equation, this equation is a linear equation in the unknown quantities u_{i+k} and b_i, and if we had used another velocity-density relation, then the discretized version of the third governing equation would have been a nonlinear equation in these two unknown quantities.

Iterative Procedure

We may solve the discretized versions of the first and third governing equations, (3.20) and (3.21), for the unknown quantities b_i and u_{i+k}, and then use these values in (A-3.3) and (A-3.1) to determine A_{i+k} and t_i. This procedure may be iterated until $i = N_f - k$. Specifically, at each step calculate the quantities $Q, Q_1, \ldots, Q_4$ in terms of

the known quantities:

$$Q = \frac{4}{F_i - F_{i-1}} \left[\log\left(\frac{1+w_i}{1+w_{i-1}} \right) - (w_i - w_{i-1}) \right]$$

$$Q_1 = u_{i+k-1} + \frac{2}{F_{i+k} + F_{i+k-1}} \left[\frac{(1-w_i)^2}{w_i} - A_{i+k-1} + A_i \right]$$

$$Q_2 = \frac{2(1-w_i)^2}{w_i(F_{i+k} + F_{i+k-1})} \tag{A-3.4}$$

$$Q_3 = \left(\frac{\alpha - \beta}{\alpha} \right) \left[\frac{1}{w_i} + t_{i-1} + (u_i - u_{i-1}) - Qb_{i-1} - 1 \right]$$

$$Q_4 = \left(\frac{\alpha - \beta}{\alpha} \right) \left[-\frac{1}{w_i} + Q \right]$$

The updated values may be calculated as follows:

$$\begin{aligned} b_i &= \frac{Q_1 - Q_3}{Q_2 + Q_4} \qquad & t_i &= t_{i-1} + (u_i - u_{i-1}) + (b_i - b_{i-1})Q \\ u_{i+k} &= Q_1 - b_i Q_2 \qquad & A_{i+k} &= A_i + \frac{(1-w_i)^2}{w_i}(1-b_i) \end{aligned} \tag{A-3.5}$$

(A-3.4) and (A-3.5) completely summarize the core of our iterative procedure. Specifically, given the initial seed values b_1, t_1, $u_1, \ldots, u_{1+k}$ and $A_1, \ldots, A_{1+k}$, we iteratively use (A-3.4) and (A-3.5) to determine $b_1, \ldots, b_{N_f-k}$, $t_1, \ldots, t_{N_f-k}$, $u_1, \ldots, u_{N_f}$ and $A_1, \ldots, A_{N_f}$. At the conclusion of this procedure, the only undetermined quantities will be the (b_i, t_i) values in the last segment, from $i = N_f - k + 1$ to $i = N_f$.

Initializing Seed Values

To initiate the above iterative procedure, we must provide values for b_1, t_1, $u_1, \ldots, u_{1+k}$ and $A_1, \ldots, A_{1+k}$. If we choose N large enough, then $F_1, \ldots, F_{1+k}$ will be close to 0, and over the interval $(0, u_{1+k})$ we can approximate $F(b(u))$ as a linear function of u. We apply the discretized versions of the first and third governing equations, that enable us to solve for u_1 and b_1, and, hence, determine all necessary initializing seed values. Specifically, suppose that over the interval $(0, u_{1+k})$ we approximate $F(b(u)) \approx \frac{F_1}{u_1} u$. Therefore, for $i = 1, \ldots, k+1$, $u_i = \frac{u_1}{F_1} F_i$. From the discretized version of the first governing equation, (3.20),

$$u_{1+k} = \left(\frac{\alpha - \beta}{\alpha}\right) \left\{ \frac{1 - b_1}{w_1} + u_1 + \frac{4b_1}{F_1} \left[\log\left(\frac{1 + w_1}{2}\right) - (w_1 - 1) \right] - 1 \right\}.$$

If we replace u_{1+k} with $\frac{u_1}{F_1} F_{1+k}$ and recall that $\frac{F_{1+k}}{F_1} = \frac{\alpha}{\alpha - \beta}$, then after some algebraic simplification we obtain

$$u_1 \left[\left(\frac{\alpha}{\alpha - \beta}\right)^2 - 1 \right] = \frac{1 - w_1}{w_1} + \left[\frac{4}{F_1} \left\{ \log\left(\frac{1 + w_1}{2}\right) - w_1 + 1 \right\} - \frac{1}{w_1} \right] b_1. \quad \text{(A-3.6)}$$

Based on our linear approximation for $F(b(u))$ over the interval $(0, u_{1+k})$, we may calculate the cumulative flow on this interval as

$$A(u) = \int_0^u F(b(u'))\, du' = \frac{F_1}{u_1} \frac{u^2}{2} \qquad u \in (0, u_{1+k}).$$

In particular, for $i = 1, \ldots, 1 + k$, $A_i = \frac{u_1}{2F_1} F_i^2$. From the third governing equation, (A-3.2),

$$A_{1+k} - A_1 = \frac{(1 - w_1)^2}{w_1} (1 - b_1).$$

Replacing A_{i+k} with $\frac{u_1}{2F_1}F_{1+k}^2$ and A_1 with $\frac{u_1F_1}{2}$, and recalling that $\frac{F_{1+k}}{F_1} = \frac{\alpha}{\alpha-\beta}$, we obtain

$$\frac{u_1F_1}{2}\left[\left(\frac{\alpha}{\alpha-\beta}\right)^2 - 1\right] = \frac{(1-w_1)^2}{w_1}(1-b_1). \tag{A-3.7}$$

(A-3.6) and (A-3.7) are a pair of linear equations that may be solved to obtain u_1 and b_1. Since $u_i = \frac{u_1}{F_1}F_i$ for $i = 1, \ldots, 1+k$, we may determine $u_2, \ldots, u_{1+k}$. Since our linear approximation implies that $A_i = \frac{F_1}{u_1}\frac{u_i^2}{2}$ for $i = 1, \ldots, 1+k$, we may determine $A_1, \ldots, A_{1+k}$. Finally, using (A-3.1) we may determine t_1. To summarize,

$$b_1 = \frac{(1-w_1)(1-w_1-\frac{F_1}{2})}{2w_1\log\left(\frac{1+w_1}{2}\right) + 1 - w_1^2 - \frac{F_1}{2}}$$

$$u_1 = \frac{2(1-w_1)^2(1-b_1)}{F_1w_1\left[\left(\frac{\alpha}{\alpha-\beta}\right)^2 - 1\right]}$$

$$t_1 = u_1 + \frac{4b_1}{F_1}\left[\log\left(\frac{1+w_1}{2}\right) - w_1 + 1\right] \tag{A-3.8}$$

$$u_i = \frac{u_1}{F_1}F_i, \quad i = 2, \ldots, 1+k$$

$$A_i = \frac{u_i}{2F_1}F_i^2, \quad i = 1, \ldots, 1+k.$$

Final Segment

After implementing the above iterative procedure, we will have determined all F_i, w_i, u_i and A_i values for $i = 1, \ldots, N_f$, and will have determined all b_i, t_i values for $i = 1, \ldots, N_f - k$. The only remaining values to determine are $(b_{N_f-k+1}, t_{N_f-k+1})$, …, (b_{N_f}, t_{N_f}), corresponding to the (x, t) coordinates of the lower boundary curve in the

final segment. Furthermore, since we have determined u_{N_f} (the departure time at $x = 0$ of the final cohort of vehicles that arrives at the CBD exactly at time $\bar{t}$), from (3.4) we can calculate $\bar{t}$ as $\bar{t} = \frac{\alpha}{\alpha-\beta} u_{N_f} + 1$.

Suppose that (b_i, t_i) is known for some $N_f - k \leq i < N_f$, and we wish to determine (b_{i+1}, t_{i+1}). Subdivide the flow from its value F_i at (b_i, t_i) to its value F_{i+1} at (b_{i+1}, t_{i+1}) into m equal values, and let $(F_i)_j$ denote the flow value at the jth subdivision, so $(F_i)_j = F_i + \frac{j}{m}(F_{i+1} - F_i)$, $j = 0, \ldots, m$. Based on these values we may calculate $(w_i)_j = \sqrt{1 - (F_i)_j}$ and $(v_i)_j = \frac{1}{2}\left(1 + \sqrt{1 - (F_i)_j}\right)$, for $j = 0, \ldots, m$. If we use the linear approximations for $F(b(u))$ and $b(u)$, (3.19), and the resulting time coordinate along the lower boundary curve, (A-3.1), then we may approximate the space-time coordinates along the lower boundary curve corresponding to the sequence with $j = 0, \ldots, m$ as $((b_i)_j, (t_i)_j)$, where

$$\begin{aligned} (b_i)_j &= b_i + \frac{j}{m}(b_{i+1} - b_i) \\ (t_i)_j &= t_i + \frac{j}{m}(u_{i+1} - u_i) + \frac{4(b_{i+1} - b_i)}{F_{i+1} - F_i}\left[\log\left(\frac{1 + (w_i)_j}{1 + w_i}\right) - \{(w_i)_j - w_i\}\right]. \end{aligned} \tag{A-3.9}$$

Consider the vehicle trajectory that passes through the lower boundary at (b_{i+1}, t_{i+1}), and denote its space-time coordinates in Region II as $(\tilde{x}, \tilde{t})$. The characteristic line emanating from $((b_i)_j, (t_i)_j)$ intersects this vehicle trajectory at the point $(\tilde{x}_j, \tilde{t}_j)$, $j = 0, \ldots, m$. Beginning with $j = 0$, proceed as follows. Approximate the trajectory curve through the point $(\tilde{x}_j, \tilde{t}_j)$ as a straight line with slope $\frac{dt}{dx} = \frac{1}{(V_i)_j}$, whose equation is given by

$$x - \tilde{x}_j = (V_i)_j(t - \tilde{t}_j).$$

Now consider the characteristic line emanating from $((b_i)_{j+1}, (t_i)_{j+1})$, that has slope

$\frac{dt}{dx} = \frac{1}{(w_i)_{j+1}}$, and whose equation is given by

$$x - (b_i)_{j+1} = (w_i)_{j+1}(t - (t_i)_{j+1}).$$

The (x, t) intersection of these two lines is taken as the $(j+1)$st point along the vehicle trajectory, $(\tilde{x}_{j+1}, \tilde{t}_{j+1})$, i.e.,

$$\tilde{t}_{j+1} = \frac{(w_i)_{j+1}(t_i)_{j+1} - (b_i)_{j+1} + \tilde{x}_j - (V_i)_j \tilde{t}_j}{(w_i)_{j+1} - (V_i)_j}$$

$$\tilde{x}_{j+1} = (w_i)_{j+1}(\tilde{t}_{j+1} - (t_i)_{j+1}) + (b_i)_{j+1}.$$

If m is chosen large enough, and we iterate this procedure from $j = 0, \ldots, m-1$, then the final point on our trajectory, $(\tilde{x}_m, \tilde{t}_m)$, should coincide with the point on the lower boundary curve, (b_{i+1}, t_{i+1}). The above procedure depends upon our initial choice for b_{i+1}, and the magnitude of our error in choosing b_{i+1} can be measured by the difference between $\tilde{x}_m$ and b_{i+1}. Thus, to determine b_{i+1}, we construct a function that calculates this error for various values of b_{i+1}, and choose the value of b_{i+1} that minimizes this error. Once we have determined b_{i+1} we update t_{i+1}, and then iteratively repeat this procedure until obtaining (b_{N_f}, t_{N_f}). A graph illustrating these concepts is provided in Figure 3.14.

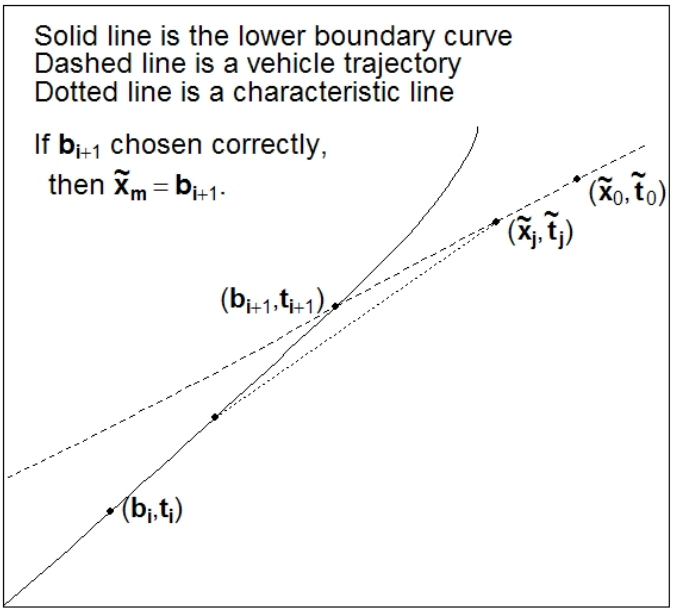

Figure 3.14: For $i = N_f - k + 1$ to $i = N_f - 1$, proceed as follows. Guess a value of b_{i+1}, and subdivide the lower boundary curve from b_i to b_{i+1} into m segments, $(b_i)_0 = b_i, \ldots, (b_i)_j, \ldots, (b_i)_m = b_{i+1}$. Iteratively construct the vehicle trajectory through (b_{i+1}, t_{i+1}) from the point $(\tilde{x}_0, \tilde{t}_0)$ to $(\tilde{x}_m, \tilde{t}_m)$. If b_{i+1} was chosen correctly, then $\tilde{x}_m = b_{i+1}$.

Chapter 4

Morning Commute in a Single-Entry Traffic Corridor with No Late Arrivals

Abstract

This paper analyzes a model of early morning traffic congestion, that is a special case of the model considered in Newell, 1988. A fixed number of identical vehicles travel along a single-lane road of constant width from a common origin to a common destination, with LWR flow congestion and Greenshields' Relation. Vehicles have a common work start time, late arrivals are not permitted, and trip cost is linear in travel time and time early. The paper explores traffic dynamics for the Social Optimum, in which total trip cost is minimized, and for the User Optimum, in which no vehicle's trip cost can be reduced by altering its departure time. Closed-form solutions for the Social Optimum and quasi-analytic solutions for the User Optimum are presented, along with numerical examples, and it is shown that this model includes the bottleneck model (with

no late arrivals) as a limit case where the length of the road shrinks to zero.

Key Words: Corridor, Morning Commute, Social Optimum, Congestion

4.1 Introduction

In recent years considerable theoretical work has been done on the dynamics of rush-hour traffic congestion. Most of this work has applied the basic bottleneck model (Vickrey, 1969, as simplified in Arnott and Lindsey, 1990), in which congestion takes the form of a queue behind a single bottleneck of fixed flow capacity. A strength of the bottleneck model is its simplicity, which permits many extensions. A weakness is that the technology of traffic congestion may not be well described by queues behind bottlenecks. This paper replaces bottleneck congestion with LWR (Lighthill and Whitham, 1955 and Richards, 1956) flow congestion, which combines the equation of continuity (conservation of mass for a fluid) with an assumed technological relationship between local velocity and local traffic density, and covers bottleneck congestion as a limiting case.

Newell, 1988 (herein referred to as Newell) considered a model of the morning commute in which a fixed number of identical commuters must travel along a road of constant width subject to LWR flow congestion, from a common origin to a common destination, and in which trip costs are a linear function of travel time and schedule delay. He allowed for a general distribution of desired arrival times and a general technological relationship between local velocity and local density, and precluded late arrivals by assumption. He obtained qualitative properties of both the social optimum (SO) in which total trip cost is minimized, and the user optimum (UO) in which no commuter can reduce their trip cost by altering their departure time. While a tour de force, his paper has been overlooked by the literature, likely because of the density of its discussion

and analysis.

Our paper provides a detailed analysis of a special case of Newell's model, assuming that commuters have a common desired arrival time and that local velocity is a negative linear function of local density (Greenshields' Relation). It complements Newell's paper in three respects. First, it is more accessible, laying down arguments in greater detail and providing numerical solutions. Second, by restricting the analysis to Greenshields' Relation, the paper obtains a closed-form solution to the SO problem, and a quasi-analytic solution to the UO problem, which provide additional insight. Third, unlike Newell, it discusses economic properties of the solutions. Furthermore, the Single-Entry Corridor Problem model with no late arrivals includes the bottleneck model (with no late arrivals) as a limit case where the length of the road shrinks to zero.

We came to the topic of this paper indirectly. Interested in the spatial-temporal dynamics of rush-hour traffic congestion in a metropolitan area, Arnott posed the Corridor Problem, which is identical to the model considered in this paper except that traffic enters at all locations along the road. In the process of studying the Corridor Problem (Arnott and DePalma, 2011 gives some preliminary results), Arnott and DePalma came to appreciate its difficulty, and decided to investigate a sequence of simpler problems, the first having cars enter at only one location. Thus, the problem addressed in this paper might be termed the *Single-Entry Corridor Problem.*

Section 2 presents the model and notation, introduces the components of LWR traffic flow theory used in the paper, and illustrates how they are applied by solving for the traffic dynamics along a uniform point-entry, point-exit road, in response to an increase in the entry rate from zero to a constant level for a fixed time period. Sections 3 and 4 contain the main results of the paper, presenting the SO and UO solutions respectively. Section 5 investigates the model's economic properties, and contrasts them

with those of the bottleneck model. Section 6 discusses directions for future research and concludes.

4.2 Model Description

Consider a road of constant width that connects a single entry point to a point central business district (CBD) that lies at the eastern end of the corridor. Since the primary results of this paper are derived from Newell, we adopt the notation and terminology used there, with the following exception: We use the word *departure* to indicate a vehicle's departure from the origin and consequent entry into the corridor, and the word *arrival* to indicate a vehicle's arrival at the CBD and consequent exit from the corridor, whereas Newell uses these words in the opposite manner[1].

Location is indexed by x, with the entry point located at $x = 0$ and the CBD at $x = l$. The corridor consists of a road from $x = 0$ to $x = l$, and possibly a queue at $x = 0$ as well. At time t, let $A(t)$ denote the cumulative inflow into the corridor (including those that may be in a queue), $A_R(t)$ the cumulative inflow into the road, and $Q(t)$ the cumulative outflow at $x = l$. Let $a(t) = \frac{dA}{dt}$ denote the inflow rate into the corridor, $a_R(t) = \frac{dA_R}{dt}$ the inflow rate into the road, and $q(t) = \frac{dQ}{dt}$ the outflow rate at $x = l$.

Let N denote population, t_f the time of the last vehicle departure into the corridor, t_R the time of the last vehicle departure into the road, and $\bar{t}$ the time of the last vehicle arrival at the CBD. If a queue is present at time t_f, then $t_R > t_f$. We will normalize time such that the first departure occurs at time $t = 0$; consequently, t_f, t_R and $\bar{t}$ are endogenous times that provide alternative measures of the duration of the

[1] As indicated in Lago and Daganzo, 2007, in queueing theory it is customary to use the terms *arrivals* to and *departures* from a server, whereas in the economics literature the terms are used in reverse, referring to *departures* from an origin and *arrivals* at a destination.

morning rush hour and are not specified *a priori*.

4.2.1 Trip Cost

All vehicles have a common desired arrival time, the work start-time, which coincides[2] with the time of the last vehicle arrival, $\bar{t}$.

Let $\tau(t)$ denote the travel time of the vehicle departing at time t (which includes time spent in a queue), α_1 the per unit cost of travel time, α_2 the per unit cost of schedule delay (i.e., time early arrival), and $C(t)$ the trip cost:

$$\begin{aligned} \text{Trip Cost} \;&=\; \text{Travel Time Cost} \;+\; \text{Schedule Delay Cost} \\ C(t) &= \alpha_1 \tau(t) + \alpha_2 \left(\bar{t} - [t + \tau(t)]\right), \quad 0 \le t \le t_f. \end{aligned} \tag{4.1}$$

Denote the sum of all trip costs as the *total trip cost*, TTC:

$$\begin{aligned} \text{Total Trip Cost} \;&=\; \text{Total Travel Time Cost} \;+\; \text{Total Schedule Delay Cost} \\ TTC &= \alpha_1(\text{Total Travel Time}) + \alpha_2(\text{Total Schedule Delay}) \\ &= \alpha_1 \int_0^{t_f} \tau(t) a(t)\, dt \;+\; \alpha_2 \int_0^{t_f} \left(\bar{t} - [t + \tau(t)]\right) a(t)\, dt. \end{aligned} \tag{4.2}$$

Throughout this paper we assume that $\alpha_1 > \alpha_2$, which is supported by empirical evidence in Small, 1982, and which is a necessary condition for constructing a UO solution, as is explained in a more general setting in Section 4 of Newell.

[2]The final departure at time t_f arrives at the CBD at time $\bar{t}$. Since late arrivals are not permitted, $\bar{t}$ cannot occur later than the work start-time. If $\bar{t}$ occurs earlier than the work start-time, then a constant schedule delay cost will be added to each vehicle's trip cost, which can be eliminated without changing the traffic dynamics by uniformly translating the entire system in time so that $\bar{t}$ coincides with the work start-time.

4.2.2 Traffic Dynamics

Following Newell, let $k(x,t)$, $v(x,t)$ and $q(x,t)$[3] denote traffic density, velocity and flow, respectively. As does most of the literature, we assume that the velocity-density relationship has the following features:

- v achieves a maximum value of v_0 (free-flow velocity) when $k = 0$.
- v is a non-increasing function of k and equals zero when $k = k_j$ (jam density).
- q achieves a maximum value of q_m (capacity flow) at a unique density value, $k = k_m$.
- q is an increasing function of k for $0 \leq k \leq k_m$ (ordinary flow).
- q is a decreasing function of k for $k_m \leq k \leq k_j$ (congested flow[4]).

Throughout the paper all analytic results are derived using Greenshields' Relation, i.e., the linear velocity-density relationship, $\frac{v}{v_0} = 1 - \frac{k}{k_j}$. Under Greenshields' Relation it can easily be shown that capacity flow is $q_m = \frac{1}{4} v_0 k_j$ (achieved at density value $k_m = \frac{1}{2} k_j$), and that in the ordinary flow régime velocity is related to flow via:

$$\frac{v}{v_0} = \frac{1}{2}\left[1 + \sqrt{1 - \frac{q}{q_m}}\right]. \tag{4.3}$$

4.2.2.1 Continuity Equation, Characteristics and Shocks

The continuity equation, its solution using the method of characteristics, and the formation of shock and rarefaction waves, are explained in detail in books such as

[3] As in Newell, the symbol q has different interpretations depending on context: i) $q(x,t)$ denotes the flow rate at the spacetime point (x,t); ii) $q(t)$ denotes the outflow rate at $x = l$; iii) q denotes the flow rate used as a dependent or independent variable with other quantities, such as v and k.

[4] The terminology "ordinary flow" and "congested flow" is standard in the engineering literature, whereas the usual terminology in the economics literature is "congested flow" and "hypercongested flow". Throughout this paper we use the former terminology.

LeVeque, 1992. Here we briefly sketch the concepts that are relavant to our paper.

The LWR traffic flow model is formulated in terms of a hyperbolic partial differential equation known as the continuity equation, which is a statement of conservation of mass of a fluid:

$$\frac{\partial k(x,t)}{\partial t} + \frac{\partial q(x,t)}{\partial x} = 0.$$

Inserting a functional relationship between flow and density, $q(k)$, into the continuity equation yields

$$\frac{\partial k(x,t)}{\partial t} + q'(k)\frac{\partial k(x,t)}{\partial x} = 0. \tag{4.4}$$

The method of characteristics for solving (4.4) is to reduce the PDE to a pair of ODE's: i) $\frac{dx}{dt} = q'\left(k(x,t)\right)$; and ii) $\frac{dk(x,t)}{dt} = 0$. The solutions to this pair of ODE's are called *characteristic curves*, or *characteristics*, and are straight lines in the spacetime plane along which k is constant. In t-x space, a characteristic line with density k has slope $\frac{\Delta x}{\Delta t} = q'(k)$, which is termed its wave velocity. As in Newell, it is more convenient to work in x-t space and to deal with the reciprocal of the wave velocity normalized by free-flow velocity, $w = \frac{v_0}{q'(k)}$. Greenshields' Relation yields $w = \frac{1}{1-\frac{k}{k_m}}$, and in the ordinary flow régime, $k \leq k_m$, we can write w in terms of q as

$$w = \frac{1}{\sqrt{1-\frac{q}{q_m}}}. \tag{4.5}$$

When characteristic lines intersect, the density becomes discontinuous, resulting in what is called a *shock wave*, which propagates through spacetime along a curve called the *shock wave path*, i.e., a path of intersecting characteristic lines. Although the continuity equation, (4.4), is not satisfied along a shock wave path, a weak, integral-form of the continuity equation is satisfied. However, to uniquely determine a solution to the weak

form of the continuity equation an additional condition is required, called the *entropy condition*.[5] This condition requires that, for fixed t, as x increases from left to right across a shock wave path the density across the shock wave path must discontinuously increase, i.e., $k_l < k_r$ where k_l is the density to the left of the shock wave path and k_r is the density to the right of the shock wave path. As shown in LeVeque, 1992, the speed of the shock wave path must satisfy

$$\left(\frac{dx}{dt}\right)_{\text{Shock}} = \frac{q_r - q_l}{k_r - k_l}, \tag{4.6}$$

where flows q_l and q_r are defined similarly to k_l and k_r.

4.2.2.2 Trajectory of Last Vehicle Departure is a Shock Wave Path

If the last vehicle to depart does not travel at free-flow velocity, then its trajectory coincides with a shock wave path. To see this, consider a point on the trajectory of the last vehicle to depart. For fixed t, to the left of this point flow and density are both zero, whereas to the right flow and density are both greater than zero, $q_r > 0$ and $k_r > 0$, so that the point lies on a shock wave path. Since, from (4.6), the speed of the shock wave at this point is $\frac{q_r}{k_r}$, which coincides with the velocity of the vehicle at this point, the vehicle's trajectory is a shock wave path.

4.2.2.3 Corridor and Road Inflow Rates, and Queue Development

Since the inflow rate into the road cannot exceed capacity flow, a queue develops if and only if the inflow rate into the corridor, $a(t)$, is greater than capacity flow. If a queue is present, then the inflow rate into the road, $a_R(t)$, is capacity flow. Furthermore,

[5] As discussed in Section 4.4.2 of Daganzo, 1997, the entropy condition is equivalent to requiring that each interior spacetime point be connected to a point on the boundary through a truncated characteristic that does not intersect any other characteristics or shocks.

since entries do not occur at any other point along the road, the traffic dynamics along the road will be in the ordinary flow régime for all time.

4.2.2.4 Method of Characteristics for the Single-Entry Corridor Problem

The following discussion is relevant to Figures 4.1, 4.7 and 4.11. Given a road inflow rate, $a_R(t)$ for $0 \leq t \leq t_R$, to determine the traffic dynamics throughout spacetime we draw a half-line characteristic emanating from each point $(0, t)$ on the t-axis with slope $\frac{\Delta t}{\Delta x} = \frac{1}{v_0\sqrt{1-\frac{a_R(t)}{q_m}}}$. Along this line density, flow and velocity are all constant, with flow being the constant value $a_R(t)$, and density and velocity being derived from the flow value via Greenshields' Relation. Provided that this line does not intersect any other characteristic line, the density along this line is constant from $x = 0$ to $x = l$.

In spacetime regions of zero density the characteristic lines have slope $\frac{1}{v_0}$ and coincide with vehicle trajectory curves corresponding to free-flow travel. Since the corridor is initially empty, the first vehicle departing at $t = 0$ travels at free-flow velocity, arriving at the CBD at time $t = \frac{l}{v_0}$.

If the inflow rate discontinuously increases at a point in time, then the slope of the characteristic lines emanating just below and just above that point on the t-axis also discontinuously increases. A fan of characteristic lines emanates from this point, referred to as a *rarefaction wave.*

After drawing all characteristics and determining all shock paths, density is determined for all spacetime points (x, t), $0 \leq x \leq l$ and $-\infty < t < \infty$. Using Greenshields' Relation we can determine the velocity at each spacetime point, from which we can obtain vehicle trajectory curves and trip costs. Although this solution method works in principle, it is computationally difficult to calculate shock paths; however, some recent work has been done in designing a robust algorithm for use with Greenshields' Relation,

Wong and Wong, 2002. A much simpler method is to use the cell-transmission model, (Daganzo, 1995), which uses a finite-difference-equation approximation to the continuity equation to determine the density at all spacetime points. The cell-transmission model does not permit the exact solution of shock wave paths, but given *any* initial inflow rate and *any* flow-density relationship, it permits numerical determination of the outflow rate, trajectory curves and trip costs. Throughout the paper we have repeatedly used the cell-transmission model to numerically verify the theoretically derived results.

4.2.2.5 Cumulative Inflow and Outflow Curves

The following discussion is relevant to Figures 4.2, 4.4, 4.5, 4.6 and 4.10. Section 2 of Newell determines relations which, in the absence of shocks, must be satisfied by cumulative inflow and outflow curves, and we restate those relations here. Let $(t, A(t))$ and $(t', Q(t'))$ be points on the cumulative inflow and outflow curves, respectively, such that the flow rates at both points are equal, i.e., $a(t) = q(t')$. These two points are related via equations (2.7) and (2.14) from Newell:

$$t = t' - \frac{l}{v_0} w(q) \tag{4.7a}$$

$$A(t) = Q(t') - \frac{l}{v_0} q \left[w(q) - \frac{v_0}{v(q)} \right], \tag{4.7b}$$

where $w(q)$ is given in (4.5). In a graph of cumulative inflow and outflow curves, we use a dashed line to connect these related points of equal flow rate, and therefore these dashed lines have slope

$$\frac{Q(t') - A(t)}{t' - t} = q \left[1 - \frac{v_0}{w(q) v(q)} \right]. \tag{4.8}$$

4.2.3 Scaled Units

Throughout the paper we utilize the following system of scaled units:

- Choose length units such that $l = 1$.
- Choose time units such that $\frac{l}{v_0} = 1$, so effectively $v_0 = 1$.
- Choose population units such that $\frac{q_m l}{v_0} = 1$, so effectively $q_m = 1$. Under Greenshields' Relation, this choice of population units is equivalent to $k_j = 4$.
- Choose cost units such that $\alpha_1 \frac{q_m l}{v_0} \frac{l}{v_0} = 1$, so effectively $\alpha_1 = 1$.

Given an equation in scaled units, we can recover the unscaled equation by dividing each term by the appropriate scaling factor. For example, in the SO we determine the time of the last vehicle arrival at the CBD in scaled units:

$$\bar{t} = 1 + \frac{1}{2}N + \sqrt{\frac{1}{\alpha_2}N + \left(\frac{1}{2}N\right)^2}.$$

Since $\bar{t}$ has units of time, N has units of population and α_2 has units of cost per population per time, to recover the unscaled equation we replace $\bar{t}$ with $\frac{\bar{t}}{\frac{v_0}{l}}$, N with $\frac{N}{\frac{q_m l}{v_0}}$, and α_2 with $\alpha_2 \frac{\frac{q_m l}{v_0}\frac{l}{v_0}}{\alpha_1 \frac{q_m l}{v_0}\frac{l}{v_0}} = \frac{\alpha_2}{\alpha_1}$. Inserting these replacements into the previous equation yields

$$\bar{t} = \frac{l}{v_0} + \frac{1}{2}\frac{N}{q_m} + \sqrt{\frac{\alpha_1}{\alpha_2}\frac{l}{v_0}\frac{N}{q_m} + \left(\frac{1}{2}\frac{N}{q_m}\right)^2}.$$

Unless otherwise specified, all of the results which follow are in scaled units.

4.2.4 Constant Inflow Rate

In this section we analyze the traffic dynamics resulting from a constant inflow rate into the corridor, $a(t) = q_c$ from $t = 0$ to $t = t_f$. We analytically derive vehicle

trajectories and their associated trip costs, the outflow rate, and total trip cost.

If $q_c > 1$, a queue develops at the entry point and the road inflow rate is capacity flow, 1. In this case, for $t > t_f$ the corridor inflow ceases, the queue dissipates, and the road inflow rate is capacity inflow until time $t_R > t_f$, at which point the entire population has entered the road. The resulting traffic dynamics are identical to a situation in which an identical size population enters the corridor at capacity inflow from $t = 0$ to $t = t_R$. The only differences between the two situations are the individual vehicle trip costs and total trip cost, since in the former situation vehicles in a queue incur an additional travel time cost. In the following derivations we initially presume that the constant inflow rate satisfies $q_c \leq 1$, and later remark on how total trip cost is affected if $q_c > 1$.

4.2.4.1 Characteristic Curves

Vehicles depart at a constant inflow rate, $q_c \leq 1$, during the time interval $0 \leq t \leq t_f$. Since a population N departs, $t_f = \frac{N}{q_c}$. For $t < 0$ and $t > t_f$ the characteristic lines emanating from the t-axis have slope 1, representing free-flow traffic, and for $0 < t < t_f$ the characteristic lines have slope $w_c = \frac{1}{\sqrt{1-q_c}} > 1$. Thus, at $t = 0$ we obtain a rarefaction wave, and at $t = t_f$ we obtain a shock wave whose path coincides with the trajectory of the last vehicle departure. The following analysis is split into two separate cases:

Case 1 All vehicle trajectories intersect the rarefaction wave.

Case 2 Not all vehicle trajectories intersect the rarefaction wave.

Vehicles departing at time $t > 0$ travel with constant velocity

$$v_c = \frac{1 + \sqrt{1 - q_c}}{2} = \frac{w_c + 1}{2w_c},$$

until reaching either the upper boundary of the rarefaction wave (Cases 1 and 2) or the CBD (Case 2 only). Since the upper boundary line of the rarefaction wave has slope w_c, the time point at which the upper boundary line reaches the CBD is w_c. Let t_c denote the departure time of a vehicle that would reach the CBD at time w_c if it traveled at constant velocity v_c, so that

$$t_c = w_c - \frac{1}{v_c} = \frac{w_c(w_c - 1)}{w_c + 1}.$$

Thus, a vehicle departing earlier than t_c travels at constant velocity v_c until reaching the upper boundary of the rarefaction wave (Cases 1 and 2), whereas a vehicle departing later than t_c travels at constant velocity v_c until reaching the CBD (Case 2 only). Case 1 occurs if $t_f \leq t_c$, or, equivalently, $\frac{N}{q_c} \leq \frac{w_c(w_c-1)}{w_c+1}$. Since $w_c = \frac{1}{\sqrt{1-q_c}}$, this condition reduces to

$$q_c \geq 1 - \frac{1}{\left(1 + \frac{N}{2} + \sqrt{N + (\frac{N}{2})^2}\right)^2}. \quad \text{(Case 1)}$$

Thus, Case 1 occurs if the inflow rate q_c is sufficiently large relative to the population. Figure 4.1 plots characteristic lines and the trajectory of the final vehicle departure for Cases 1 and 2.

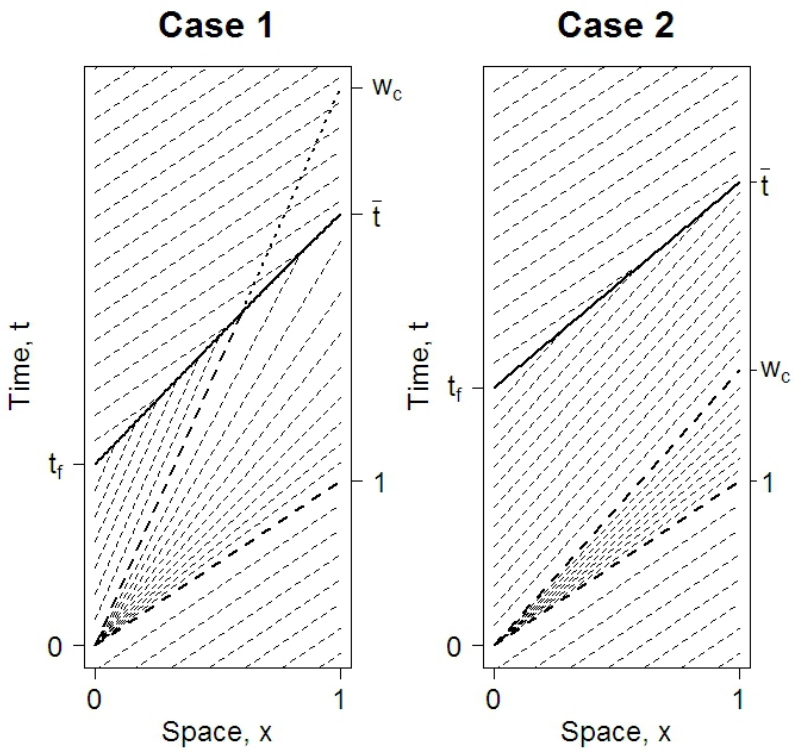

Figure 4.1: Dashed lines are characteristics corresponding to a constant inflow rate from $t = 0$ to $t = t_f$, and the solid line is the trajectory of the final vehicle departure. The bold, dashed lines are characteristics corresponding to the boundaries of the rarefaction wave arising from the initial discontinuous increase in the inflow rate. Case 1 occurs if the inflow rate is sufficiently large relative to the population. In Case 1 the upper boundary of the rarefaction wave does not reach the CBD since it intersects the shock wave corresponding to the trajectory of the final vehicle trajectory, although we indicate its intended path using a dotted line.

4.2.4.2 Vehicle Trajectories

Denote the departure time of a vehicle as t_d, $0 \leq t_d \leq t_f$, and its arrival time as t_a, $1 \leq t_a \leq \bar{t}$. The first vehicle departure occurs at $t_d = 0$, with corresponding arrival at $t_a = 1$, and its trajectory is a straight line coinciding with the lower boundary of the rarefaction wave. If $t_d > 0$, then the vehicle initially travels at constant velocity $v_c = \frac{w_c+1}{2w_c}$, so that the trajectory is a straight line whose spacetime coordinates, (x, t), satisfy

$$x(t) = \frac{w_c + 1}{2w_c}(t - t_d).$$

If $t_d \geq t_c$ (Case 2 only), then the vehicle's straight-line trajectory arrives at the CBD at time

$$t_a = t_d + \frac{2w_c}{w_c + 1}.$$

If $t_d < t_c$ (Cases 1 and 2), then the vehicle's straight-line trajectory intersects the upper boundary of the rarefaction wave at the spacetime point

$$(x_0, t_0) = \left(t_d \frac{w_c + 1}{w_c(w_c - 1)}, \, t_d \frac{w_c + 1}{w_c - 1} \right).$$

Upon entering the interior of the rarefaction wave the vehicle's speed at a spacetime point depends upon the flow value at that spacetime point through Greenshields' Relation, and the flow value at that spacetime point is determined by the characteristic line upon which it lies. If we denote the spacetime point of the vehicle trajectory as (x, t), then since the characteristic line along which it lies is generated from a fan of characteristic lines emanating from the origin, the flow along the characteristic line, q, satisfies $\frac{1}{\sqrt{1-q}} = \frac{t}{x}$. From Greenshields' Relation (4.3) the vehicle's speed satisfies $v = \frac{1}{2}\left[1 + \sqrt{1-q}\right]$, and combining these two equations yields a differential equation for the vehicle's trajectory curve in the interior of the rarefaction wave:

$$\frac{dx}{dt} = \frac{1}{2}\left[1 + \frac{x}{t}\right]. \tag{4.9}$$

This equation can be solved by making the substitution $y = \frac{x}{t}$, resulting in the general solution $x(t) = t + A\sqrt{t}$, where A is an arbitrary constant determined from the initial condition that the trajectory curve originates from the spacetime point (x_0, t_0). After some algebraic simplifications we obtain the trajectory within the interior of the

rarefaction wave as

$$x(t) = t - \sqrt{t_d q_c}\sqrt{t},$$

which arrives at the CBD at time

$$t_a = 1 + \frac{t_d q_c}{2} + \sqrt{t_d q_c + \left[\frac{t_d q_c}{2}\right]^2}.$$

In Table 4.1 we summarize, providing expressions for the vehicle trajectories and their arrival times in terms of their departure times, t_d. The last departure occurs at time $t_f = \frac{N}{q_c}$, and arrives at time $\bar{t}$. Replacing t_d with t_f in the arrival time expressions determines $\bar{t}$ for Cases 1 and 2, and these results are also provided in Table 4.1.

4.2.4.3 Outflow Rate

Figure 4.1 provides a graphical method to determine the outflow rate at the CBD, $q(t)$, by determining the constant flow value on each characteristic line. The characteristic lines corresponding to free-flow travel have zero flow value, so that $q(t) = 0$ for $t < 1$ or $t > \bar{t}$. The characteristic lines that are generated from the rarefaction wave at the origin have flow value q and slope $w = \frac{1}{\sqrt{1-q}} = \frac{t}{1}$, so that $q(t) = 1 - \frac{1}{t^2}$ for $1 < t < \bar{t}$ (Case 1) and $1 < t < w_c$ (Case 2). We conclude that the outflow rate in this time period depends only upon the existence of a discontinuity in the inflow rate at the origin, and does not depend upon the magnitude of the discontinuity, q_c. In Case 2, for $w_c < t < \bar{t}$ the outflow rate is q_c. Integrating $q(t)$ yields the cumulative outflow curve, $Q(t)$. These results are summarized in Table 4.1.

In Figure 4.2 we plot the cumulative inflow and outflow curves for Cases 1 and 2. The dashed lines are constant-flow lines (whose slope is determined from (4.8)), and show how the inflow generates the outflow. In Case 1 the cumulative outflow curve is

generated by the initial discontinuity in the inflow rate, and the subsequent inflow does not affect the outflow rate. In Case 2 the first portion of the cumulative outflow curve from time $t = 1$ to $t = w_c$ is generated by the initial discontinuity in the inflow rate, and the second portion of the cumulative outflow curve from $t = w_c$ to $t = \bar{t}$ is generated by the inflow from $t = 0$ up to some point in time, which we denote as t'. The subsequent inflow after $t = t'$ does not affect the outflow rate.

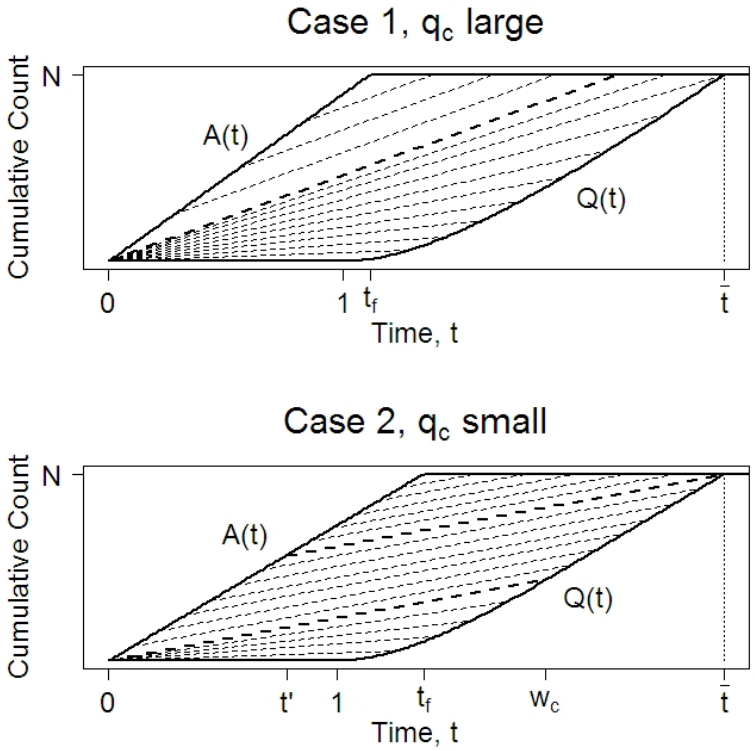

Figure 4.2: Sample cumulative outflow curves corresponding to a constant inflow rate for Cases 1 and 2. The dashed lines are constant flow lines, and show how the inflow generates the outflow. Note that in Case 1 the outflow is completely determined by a portion of the rarefaction wave arising from the initial discontinuity in the inflow rate, and does not depend upon the magnitude of the inflow rate, q_c.

4.2.4.4 Trip Costs

For a departure at time t_d, the travel time cost is $t_a - t_d$, schedule delay cost is $\alpha_2 (\bar{t} - t_a)$, and trip cost is $C = t_a - t_d + \alpha_2(\bar{t} - t_a)$. In Figure 4.3 we graph these

costs as functions of departure time for capacity inflow and $\alpha_2 = 0.5$.

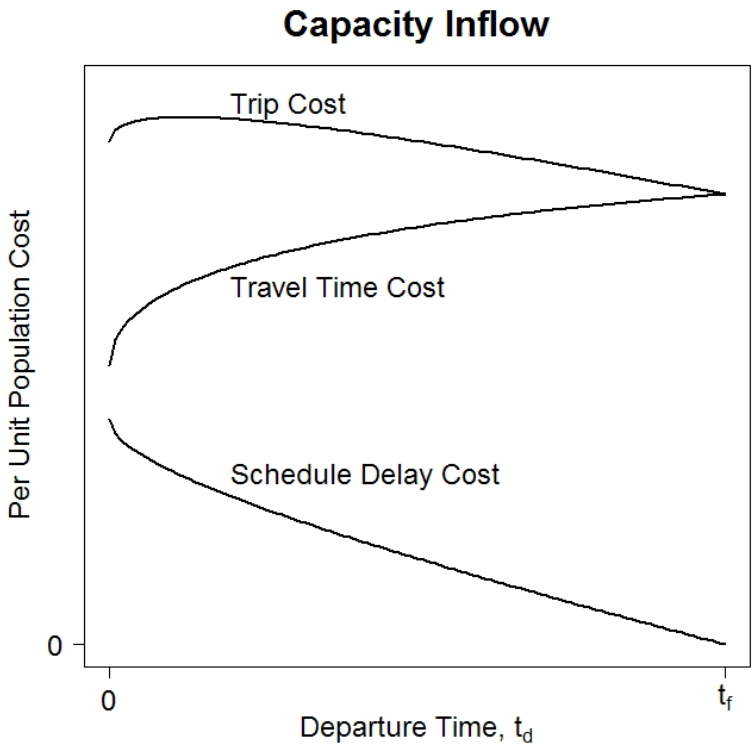

Figure 4.3: Schedule delay cost, travel time cost and trip cost as functions of departure time for capacity inflow rate, with $N = 1$ and $\alpha_2 = 0.5$ (Case 1 applies).

In Figure 4.2 the total schedule delay is the area under the cumulative outflow curve from $t = 1$ to $t = \bar{t}$,

$$\text{Total Schedule Delay} = \int_1^{\bar{t}} Q(t)\, dt, \tag{4.10}$$

the total travel time is the area between the cumulative inflow and outflow curves,

$$\text{Total Travel Time} = \int_0^{t_f} A(t)\, dt + N\,(\bar{t} - t_f) - \int_1^{\bar{t}} Q(t)\, dt, \tag{4.11}$$

and the total trip cost is

$$TTC = \text{Total Travel Time} + \alpha_2 \left(\text{Total Schedule Delay}\right). \tag{4.12}$$

Since $A(t) = q_c t$ for $0 < t < t_f$, using the expression for $Q(t)$ given in Table 4.1, we analytically determine the total schedule delay, total travel time and total trip cost, which are also provided in Table 4.1.

4.2.4.5 Queue Development

Suppose that a population of size N enters the corridor at an inflow rate greater than capacity flow, $q_c > 1$. Since the inflow into the road cannot exceed capacity flow, a queue develops at the entry point. The road inflow rate remains constant at capacity inflow until the queue has dissipated. Thus, the total population enters the corridor by time $t_f = \frac{N}{q_c}$, but does not enter the road until the later time, $t_R = \frac{N}{1} = N$. We can write the cumulative inflow curve into the corridor as

$$A(t) = q_c t,\ 0 \leq t \leq t_f,$$

and the cumulative inflow curve into the road as

$$A_R(t) = t,\ 0 \leq t \leq t_R.$$

The resulting traffic dynamics along the road are identical to a situation in which the total population enters the corridor at capacity inflow from $t = 0$ to $t = N$, so that the entire flow propagation region is covered by a rarefaction wave (i.e., Case 1 of Figure

	Case 1: $t_f \leq t_c \Leftrightarrow q_c$ large	Case 2: $t_f > t_c \Leftrightarrow q_c$ small
Inflow Rate, q_c	$q_c \geq 1 - \frac{1}{\left(1+\frac{N}{2}+\sqrt{N+\left(\frac{N}{2}\right)^2}\right)^2}$	$0 < q_c < 1 - \frac{1}{\left(1+\frac{N}{2}+\sqrt{N+\left(\frac{N}{2}\right)^2}\right)^2}$
Arrival Time if $0 \leq t_d \leq t_c$	$t_a = 1 + \frac{t_d q_c}{2} + \sqrt{t_d q_c + \left[\frac{t_d q_c}{2}\right]^2}$	
Vehicle Trajectory if $0 \leq t_d \leq t_c$	$x(t) = \begin{cases} \left(\frac{w_c+1}{2w_c}\right)(t-t_d), & t_d \leq t \leq t_d\left(\frac{w_c+1}{w_c-1}\right) \\ t - \sqrt{t_d q_c}\sqrt{t}, & t_d\left(\frac{w_c+1}{w_c-1}\right) < t \leq t_a \end{cases}$	
Arrival Time if $t_d > t_c$	-	$t_a = t_d + \frac{2w_c}{1+w_c}$
Vehicle Trajectory if $t_d > t_c$	-	$x(t) = \left(\frac{w_c+1}{2w_c}\right)(t-t_d), \quad t_d \leq t \leq t_a$
Time of Last Vehicle Arrival	$\bar{t} = 1 + \frac{N}{2} + \sqrt{N + \left(\frac{N}{2}\right)^2}$	$\bar{t} = \frac{N}{q_c} + \frac{2}{1+\sqrt{1-q_c}}$
Outflow Rate	$q(t) = \begin{cases} 0, & t \leq 1 \text{ or } t \geq \bar{t} \\ 1 - \frac{1}{t^2}, & 1 < t < \bar{t} \end{cases}$	$q(t) = \begin{cases} 0, & t \leq 1 \text{ or } t \geq \bar{t} \\ 1 - \frac{1}{t^2}, & 1 < t \leq w_c \\ q_c, & w_c < t \leq \bar{t} \end{cases}$
Cumulative Outflow Rate	$Q(t) = \begin{cases} 0, & t \leq 1 \\ t + \frac{1}{t} - 2, & 1 < t \leq \bar{t} \\ N, & t > \bar{t} \end{cases}$	$Q(t) = \begin{cases} 0, & t \leq 1 \\ t + \frac{1}{t} - 2, & 1 < t \leq w_c \\ (t-w_c)q_c + \frac{(w_c-1)^2}{w_c}, & w_c < t \leq \bar{t} \\ N, & t > \bar{t} \end{cases}$
Total Schedule Delay	$TSD = \frac{1}{2}\left(\bar{t}^2 - 1\right) - 2\left(\bar{t} - 1\right) + \log \bar{t}$	$TSD = \frac{1}{2}\left(w_c^2 - 1\right)\left(\frac{\bar{t}}{w_c}\right)^2 - 2\left(w_c - 1\right)\left(\frac{\bar{t}}{w_c}\right) + \log w_c$
Total Travel Time	$TTT = N\bar{t} - \frac{N^2}{2q_c} - TSD$	$TTT = N\bar{t} - \frac{N^2}{2q_c} - TSD$
Total Trip Cost	$TTC = N\bar{t} - \frac{N^2}{2q_c} - (1-\alpha_2)TSD$	$TTC = N\bar{t} - \frac{N^2}{2q_c} - (1-\alpha_2)TSD$

Table 4.1: Table of results for a constant inflow rate, q_c (scaled units).

4.1 with $w_c = \infty$, so that the upper boundary of the rarefaction wave is a vertical line). However, queueing time adds to travel time. In Figure 4.4 we plot the cumulative corridor and road inflow curves, along with the cumulative outflow curve. The traffic dynamics along the road are determined by the inflow rate into the road, indicated by the lightly dashed lines. For a departure at time t, the horizontal distance between the cumulative corridor and road inflow curves is $(q_c - 1)t$, which is the queueing time. The total queueing time is the area between the two cumulative inflow curves, $\frac{1}{2}N^2\left(1 - \frac{1}{q_c}\right)$. Thus, for an inflow rate of $q_c > 1$, the total trip cost is greater than that with capacity flow by $\frac{1}{2}N^2\left(1 - \frac{1}{q_c}\right)$.

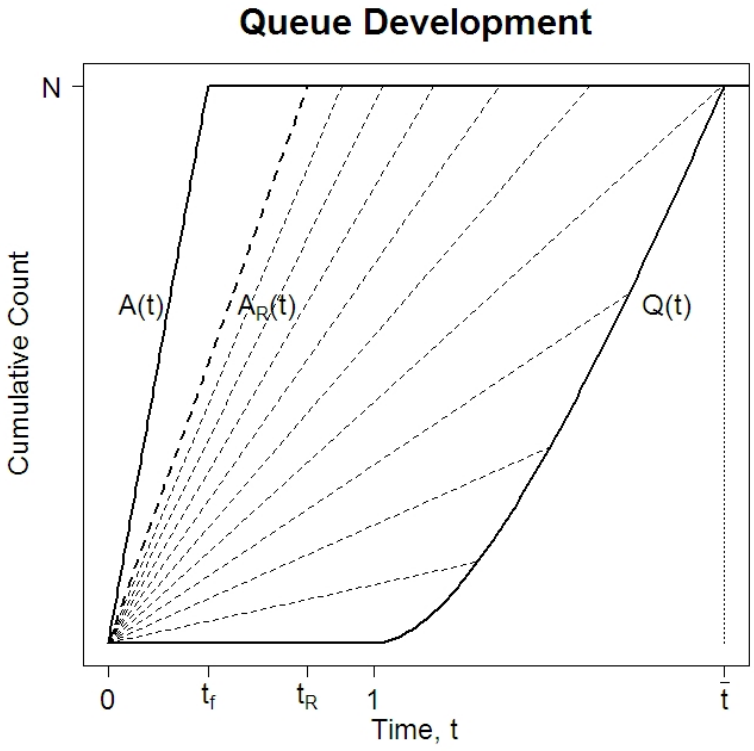

Figure 4.4: If the corridor inflow rate is greater than capacity inflow, then a queue develops and the road inflow rate equals capacity inflow for the duration of the queue. The traffic dynamics along the road are identical to the entire population entering at capacity inflow, but an additional total travel time is incurred equalling the area between the curves $A(t)$ and $A_R(t)$.

4.3 Social Optimum

Given population, N, and unit schedule delay cost, α_2, the social optimum (SO) solution for the Single-Entry Corridor Problem with no late arrivals is the inflow rate function, $a(t)$, that minimizes total trip cost. Newell provides a method for constructing the solution for an arbitrary arrival demand curve. He initially constructs a cumulative outflow curve of maximal growth and a corresponding cumulative inflow curve which is a SO solution if and only if $\alpha_2 \geq 1$, and he then modifies this solution to accomodate $\alpha_2 < 1$. In this section we apply his results to the Single-Entry Corridor Problem in which the demand curve consists of zero arrivals before the work-start time, $\bar{t}$, and the total population at the work-start time. As noted earlier, we normalize time so that the first departure occurs at $t = 0$; consequently, the length of the corridor rush hour, $\bar{t}$, is endogeneous.

4.3.1 Outflow Curve of Maximal Growth

Section 3 of Newell first determines the inflow rate function that minimizes total schedule delay, showing that the area under the cumulative outflow curve is minimized if $Q(t)$ is a cumulative outflow curve of maximal growth. To construct this curve, Newell first rewrites the flow-density relationship in terms of w instead of k to obtain flow as a function of w, $q^*(w)$, and subsequently determines the *nondimensional cumulative outflow curve of maximal growth*, denoted as $Q^*(w) = \int_1^w q^*(z)\, dz$. Greenshields' Relation, (4.5), allows Newell to determine

$$q^*(w) = 1 - \frac{1}{w^2},$$

from which Newell obtains

$$Q^*(w) = \int_1^w q^*(z)\, dz = w + \frac{1}{w} - 2. \tag{4.13}$$

Section 2 of Newell outlines a geometric procedure for constructing the cumulative outflow curve of maximal growth, $Q(t')$, using $Q^*(w)$. In this procedure, a segment of the Q^* curve must be drawn so that it is tangent to the (scaled) arrival-demand curve at some time t'_0, and intersects the arrival-demand curve at some later time. Equation (2.12) from Newell determines the cumulative outflow curve, restated here using our scaled units:

$$Q(t') \;=\; Q(t'_0) + Q^*\left(w_0 + t' - t'_0\right) - Q^*(w_0). \tag{4.14}$$

In the Single-Entry Corridor Problem, the arrival-demand curve is a horizontal line with value zero for $t < \bar{t}$, and a horizontal line with value N for $t \geq \bar{t}$. Thus, the Q^* curve must be drawn with slope zero at $(t'_0, 0)$, and must intersect the (scaled) arrival-demand curve at $(\bar{t}, N)$. The first condition implies that $q(t'_0) = 0$, so $w_0 = 1$. Since the first departure travels at free-flow velocity, the first arrival occurs at time $t'_0 = 1$. Inserting these values and (4.13) into (4.14) yields the cumulative outflow curve of maximal growth,

$$Q(t') = t' + \frac{1}{t'} - 2, \quad 1 \leq t' \leq \bar{t}, \tag{4.15}$$

where $\bar{t}$ is determined such that $Q(\bar{t}) = N$. (4.15) is identical to the cumulative outflow rate for Case 1 in Table 4.1, which is generated as a result of a sufficiently large discontinuous increase in the inflow rate. Thus, the cumulative outflow curve of maximal growth, (4.15), does not uniquely generate an inflow rate. However, as indicated at the end of Section 2 in Newell, since the outflow rate at $\bar{t}$ discontinuously drops to

zero, a cumulative inflow curve can be uniquely generated as a backwards fan of waves originating from the point of discontinuous decrease in outflow rate. Specifically, let $(t, A(t))$ be a point on the cumulative inflow curve generated from a backwards fan of waves emanating from the discontinuity in the outflow rate at $(\bar{t}, N)$ on the cumulative outflow curve. From (4.7a) we obtain:

$$t = \bar{t} - \frac{1}{\sqrt{1 - a(t)}} \qquad \Rightarrow \qquad a(t) = 1 - \frac{1}{(\bar{t} - t)^2}. \tag{4.16}$$

Since $A(0) = 0$, we can integrate $a(t)$ to determine $A(t)$,

$$A(t) = N - \left(\bar{t} - t + \frac{1}{\bar{t} - t} - 2\right) \text{ for } 0 \leq t \leq t_f, \tag{4.17}$$

and $A(t_f) = N$ implies $t_f = \bar{t} - 1$. In Figure 4.5 we graph the cumulative outflow curve of maximal growth, along with the uniquely generated cumulative inflow curve. Comparing Figure 4.5 and Figure 4.2, Case 1 enables the following observations:

- The outflow curves in both figures are identical, each being generated by the discontinuous inflow rate at $t = 0$. Figure 4.2, Case 1 occurs if the inflow rate is sufficiently large, $q_c \geq 1 - \frac{1}{\left(1+N+\sqrt{N+\left(\frac{N}{2}\right)^2}\right)^2}$, which is precisely the slope at $t = 0$ of the cumulative inflow curve in Figure 4.5. As indicated in the last paragraph of Section 2 in Newell, any cumulative inflow curve lying above the cumulative inflow curve in Figure 4.5 generates the same cumulative outflow curve, being the cumulative outflow curve of maximal growth, but with a deceleration shock wave forming at some x, $0 \leq x \leq 1$. For the cumulative inflow curve in Figure 4.2, Case 1, a shock forms at $(x, t) = (0, t_f)$, and the shock wave path corresponds to the trajectory of the final vehicle departure. For the cumulative inflow curve

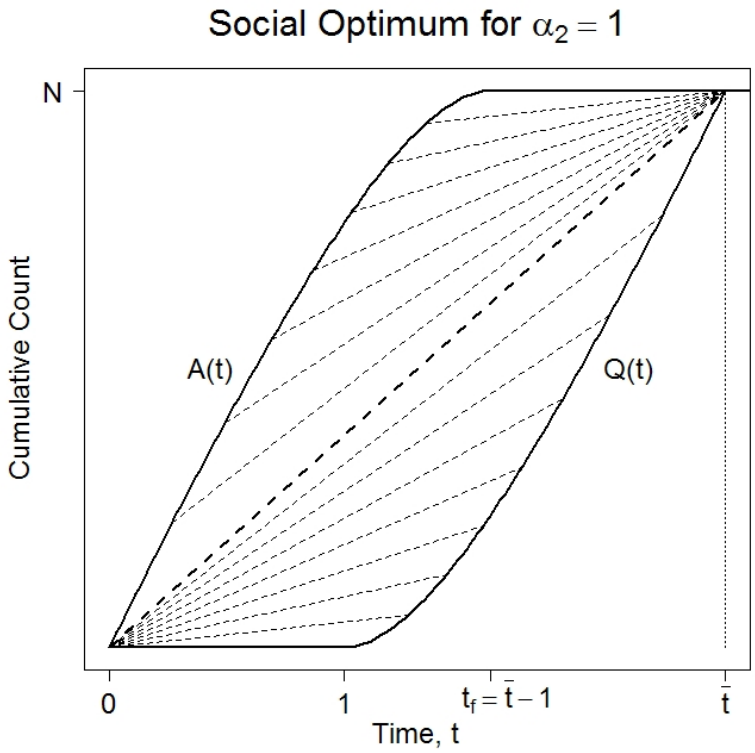

Figure 4.5: $Q(t)$ is the cumulative outflow curve of maximal growth. $A(t)$ is a cumulative inflow curve generated from the discontinuity in the outflow rate at $\bar{t}$. Any other cumulative inflow curve which lies above $A(t)$ generates the same $Q(t)$; however, the $A(t)$ shown is the only cumulative inflow curve for which no shock waves are present in the system, and thus the traffic dynamics of the system are symmetric in the sense that we can equivalently view the time reversal of the outflow rate as generating the inflow rate. Furthermore, the $A(t)$ shown is the SO solution in the case when $\alpha_2 = 1$, i.e., the $A(t)$ shown minimizes the sum of the area between $A(t)$ and $Q(t)$ plus the area under $Q(t)$, from $t = 0$ to $t = \bar{t}$.

in Figure 4.5, the shock forms at spacetime point $(x, t) = (1, \bar{t})$ (indicated in Figure 4.5 by the converging dashed lines), the final departure travels at free-flow velocity, and a shock does not occur on its trajectory.

- If shocks are not present for all t and $0 < x < 1$, then (4.4) will be strictly satisfied. As indicated in Section 2 of Newell, since (4.4) is invariant to changing (x, t) to $(1 - x, \bar{t} - t)$, if shocks are not present then there is a one-to-one correspondence between the inflow rate, $a(t)$, and the time reversal of the outflow rate, $q(\bar{t} - t)$.

Intuitively, in a system with no shocks we can equivalently consider the time reversal of the outflow rate as generating the inflow rate.

- If the unit travel time cost equals the unit schedule delay cost ($\alpha_2 = 1$ in scaled units), then the SO solution minimizes the sum of the area under the cumulative outflow curve and the area between the cumulative inflow and outflow curves. Section 3 of Newell proves that this area is minimized if the cumulative inflow curve is given as in Figure 4.5. Thus, this cumulative inflow curve is also the SO solution for $\alpha_2 = 1$.

4.3.2 Inflow and Outflow Rates

If $\alpha_2 < 1$, then minimizing the total trip cost is equivalent to minimizing the sum of the area between the cumulative inflow and outflow curves, plus α_2 multiplied by the area under the cumulative outflow curve. Newell reduces this minimization problem to a calculus of variation problem for the optimal cumulative outflow curve, which he solves in his Appendix, resulting in (3.2) of his paper which we restate in our scaled units:

$$\begin{aligned} Q(t') &= Q(t_0) + \frac{1}{\alpha_2}\left[Q^*\left(w_0 + \alpha_2\left(t' - t'_0\right)\right) - Q^*(w_0)\right] \\ &= \frac{1}{\alpha_2} Q^*\left(1 + \alpha_2\left(t' - 1\right)\right), \end{aligned}$$

Here Q^* is the same nondimensional cumulative outflow curve of maximal growth from the previous section. Using the functional form in (4.13) determined from Greenshields' Relation, we determine the optimal cumulative outflow curve, $Q(t)$, take its derivative to determine the corresponding outflow rate, $q(t)$, and solve the relation $Q(\bar{t}) = N$ to determine the length of the rush-hour in the corridor, $\bar{t}$. These results are provided in

Table 4.2.

Inserting these outflow relationships into (4.7) determines the corresponding cumulative inflow curve, where $v(q)$ and $w(q)$ are determined via Greenshields' Relation, (4.3) and (4.5), respectively. Combining (4.5) and $q(t)$ from Table 4.2 yields

$$w\left(q\right) = \frac{1}{\sqrt{1 - q(t')}} = 1 + \alpha_2(t' - 1), \quad 1 \leq t' \leq \bar{t}. \tag{4.18}$$

Inserting this relation into (4.7a) yields

$$t = (1 - \alpha_2)\left(t' - 1\right), \quad 1 \leq t' \leq \bar{t} \Leftrightarrow 0 \leq t \leq (1 - \alpha_2)\left(\bar{t} - 1\right), \tag{4.19}$$

from which we can rewrite w in terms of t:

$$w\left(q(t)\right) = 1 + \frac{\alpha_2 t}{1 - \alpha_2}, \quad 0 \leq t \leq (1 - \alpha_2)\left(\bar{t} - 1\right). \tag{4.20}$$

To determine the cumulative inflow curve up to time $t = (1 - \alpha_2)\left(\bar{t} - 1\right)$, insert (4.18) into $Q(t)$ from Table 4.2 to write $Q(t')$ in terms of w, use this expression in (4.7b) to write $A(t)$ in terms of w, and finally use (4.20) to determine $A(t)$:

$$\begin{aligned} A(t) &= \frac{1}{\alpha_2}\left[w + \frac{1}{w} - 2\right] - \left(1 - \frac{1}{w^2}\right)\left[w - \frac{2}{1 + \frac{1}{w}}\right] \\ &= \left(\frac{1}{\alpha_2} - 1\right)\left[w + \frac{1}{w} - 2\right] \\ &= \frac{t^2}{t + \frac{1}{\alpha_2} - 1}, \quad 0 \leq t \leq (1 - \alpha_2)(\bar{t} - 1). \end{aligned}$$

The remaining portion of the cumulative inflow curve from $t = (1-\alpha_2)(\bar{t}-1)$ to $t = t_f$ is generated from the discontinuity in the outflow rate at the point $(\bar{t}, N)$ on the cumulative

outflow curve, yielding the same expressions as in (4.16) and (4.17), and yielding the same conclusion that $t_f = \bar{t} - 1$, since $A(t_f) = N$. We summarize in Table 4.2, providing the cumulative inflow curve, $A(t)$, along with its derivative, $a(t)$.

In Figure 4.6 we graph the cumulative inflow and outflow curves for the social optimum for $\alpha_2 < 1$. Consider the limiting case as $\alpha_2 \to 0$, so that the total trip cost is determined entirely by total travel time. From the results in Table 4.2, as $\alpha_2 \to 0$, $\bar{t} \to \infty$ and $a(t) \to 0$. Thus, as $\alpha_2 \to 0$ the inflow is spread over an infinitely long departure interval, so that all vehicles travel at free-flow velocity, thereby minimizing total travel time.

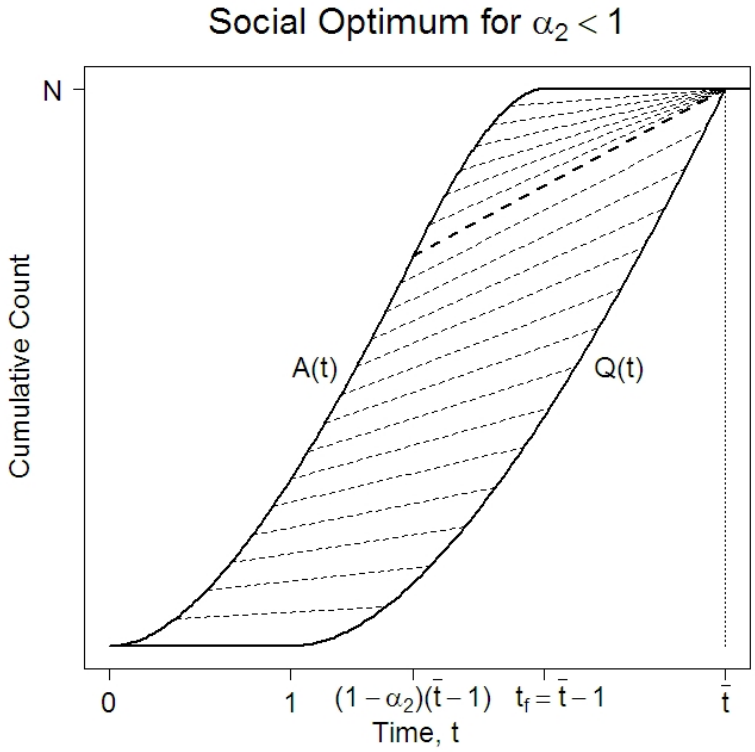

Figure 4.6: Social optimum cumulative inflow and outflow curves for $\alpha_2 < 1$. As α_2 decreases, the area between the inflow and outflow curves (total travel time) decreases, approaching the limiting case in which all vehicles travel at free-flow velocity over an infinitely long rush hour.

4.3.3 Vehicle Trajectories

Each point, (x, t), along a vehicle trajectory intersects a characteristic curve with slope w. Under Greenshields' Relation, (4.3) and (4.5) determine a differential equation for the trajectory path,

$$\frac{dx}{dt} = \frac{1}{2}\left[1 + \frac{1}{w}\right]. \tag{4.21}$$

In Figure 4.7 we graph the characteristic curves in a spacetime diagram for the social optimum, which naturally divide the traffic dynamics into two regions: an upper region consisting of a backwards fan of characteristics, and a lower region. In the upper region

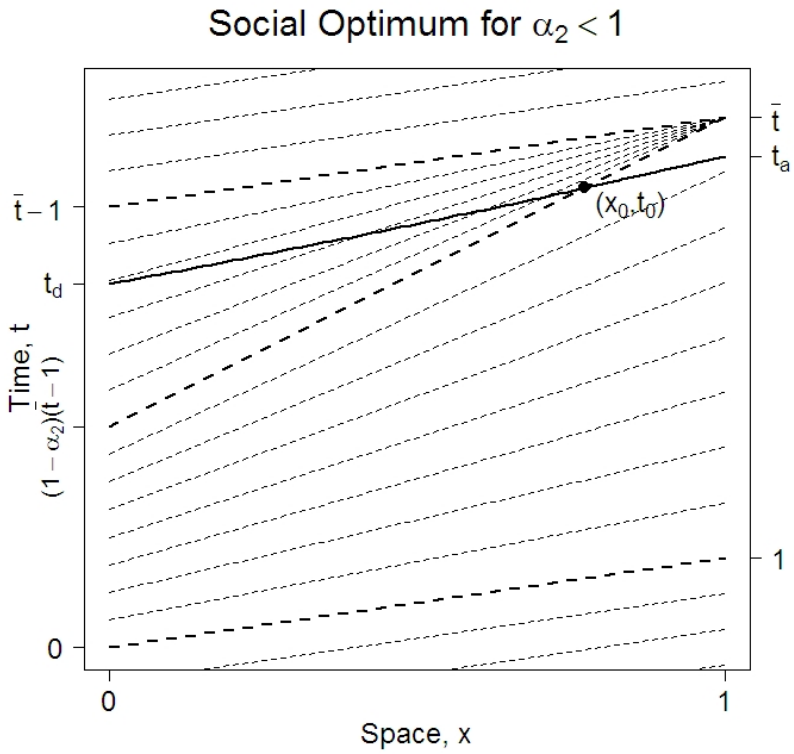

Figure 4.7: Social optimum spacetime diagram. Dashed lines are characteristic lines, which subdivide the traffic dynamics into the two regions indicated with bold dashed lines: an upper region consisting of a backwards fan of characteristics originating from the discontinuity in the flow rate at $\bar{t}$, and a lower region. We have graphed in bold the trajectory curve of a vehicle that departs at $t_d > (1-\alpha_2)(\bar{t}-1)$, traverses the upper region, enters the lower region at point (x_0, t_0), and arrives at time t_a.

the slope of a characteristic line through a point (x, t) satisfies

$$w = \frac{\bar{t} - t}{1 - x},$$

and from (4.21) a trajectory curve in this region satisfies the differential equation

$$\frac{dx}{dt} = \frac{1}{2}\left[1 + \frac{1-x}{\bar{t}-t}\right].$$

The general solution to this equation is

$$1 - x = \bar{t} - t + A\sqrt{\bar{t} - t}, \tag{4.22}$$

where the arbitrary constant A is determined from the starting point of the trajectory.

In the lower region, by (4.20), a characteristic line that originates from $(0, t_0)$ has slope $w = 1 + \frac{\alpha_2 t_0}{1-\alpha_2}$. If the characteristic line passes through the spacetime point (x, t), then

$$\frac{t - t_0}{x} = 1 + \frac{\alpha_2 t_0}{1 - \alpha_2},$$

from which we can solve for t_0, and thus express the slope of the characteristic in terms of x and t as

$$w = 1 + \frac{t - x}{x + \frac{1}{\alpha_2} - 1}.$$

Inserting this expression into (4.21) yields the differential equation that is satisfied by trajectory curves in the lower region,

$$\frac{dx}{dt} = \frac{1}{2}\left[1 + \frac{x + \frac{1}{\alpha_2} - 1}{t + \frac{1}{\alpha_2} - 1}\right],$$

whose general solution is

$$x = t + A\sqrt{t + \frac{1}{\alpha_2} - 1}, \tag{4.23}$$

where the arbitrary constant A is determined by the trajectory starting point.

For a departure at time $t_d \leq (1-\alpha_2)(\bar{t}-1)$, the trajectory is specified in (4.23), where the arbitrary constant is chosen so that the trajectory originates from $(0, t_d)$. For a departure at time $t_d > (1-\alpha_2)(\bar{t}-1)$, the trajectory is specified in (4.22) until it enters the lower region at some point (x_0, t_0) (see Figure 4.7), at which point the trajectory

is specified in (4.23). We summarize in Table 4.2, providing explicit expressions for the vehicle trajectories and their arrival times in terms of their departure times, t_d.

4.3.4 Trip Costs

For a departure at time t_d and arrival at time t_a, travel time is $t_a - t_d$, schedule delay is $\bar{t} - t_a$, and trip cost is $C = t_a - t_d + \alpha_2(\bar{t} - t_a)$. Clearly, schedule delay is a decreasing function of departure time. However, travel time is not an increasing function of departure time, since both the first and last departures travel at free flow velocity. In Figure 4.8 we graph schedule delay cost, travel time cost and trip cost as functions

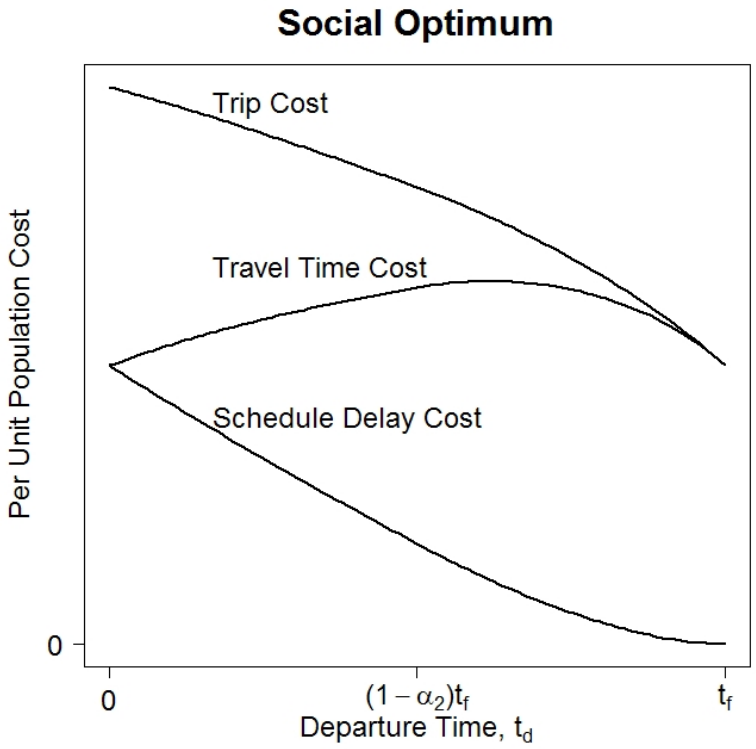

Figure 4.8: Travel time cost, schedule delay cost and trip cost as functions of departure time for the social optimum, with $N = 1$ and $\alpha_2 = 0.5$.

of departure time. For any choice of N and α_2, the graph of these costs has the same qualitative features.

Using the expressions for $A(t)$, $Q(t)$ and $\bar{t}$ from Table 4.2, the area under the cumulative outflow curve and between the cumulative inflow and outflow curves can be calculated, and hence the total schedule delay, total travel time and total trip cost can be determined, from (4.10), (4.11) and (4.12), respectively. The results are provided in Table 4.2.

4.3.5 Reduction to Bottleneck Model

The bottleneck model (Arnott and Lindsey, 1990) is the special case of the Single Entry Corridor Problem with $l = 0$, so that travel time costs are incurred only from time spent in a queue. Arnott and Lindsey, 1990 provides a heuristic derivation of the social optimum, concluding that the departure rate, with no late arrivals, is capacity flow and that the total travel time cost is $\frac{\alpha_2}{2}\frac{N^2}{q_m}$. We now show how our results for the social optimum with $l = 0$ reduce to these results.

In unscaled units, the expressions for $\bar{t}$, t_f, $q(t)$ and TTC from Table 4.2

become:

$$\bar{t} = \frac{l}{v_0} + \frac{N}{2q_m} + \sqrt{\frac{\alpha_1}{\alpha_2}\frac{l}{v_0}\frac{N}{q_m} + \left(\frac{N}{2q_m}\right)^2}$$

$$t_f = \bar{t} - \frac{l}{v_0}$$

$$q(t) = \begin{cases} q_m \left(1 - \frac{1}{\left[1+\frac{\alpha_2}{\alpha_1}(\frac{v_0}{l}t-1)\right]^2}\right) & \frac{l}{v_0} \le t \le \bar{t} \\ 0 & \text{otherwise} \end{cases}$$

$$TTC = \alpha_1 q_m \left[\frac{N}{q_m}\left(\frac{l}{v_0} + \frac{\alpha_2}{\alpha_1}t_f\right) - \frac{1}{2}\frac{\alpha_2}{\alpha_1}t_f^2 + \frac{l}{v_0}t_f - \left(\frac{l}{v_0}\right)^2 \frac{\alpha_1}{\alpha_2}\log\left(1 + \frac{\alpha_2}{\alpha_1}\frac{v_0}{l}t_f\right)\right].$$

Taking the limit as $l \to 0$ yields $\bar{t} = \frac{N}{q_m}$ and $q(t) = q_m$ for $0 \le t \le \bar{t}$, so that the departure rate is capacity flow, as in the bottleneck model. As $l \to 0$, $t_f = \frac{N}{q_m}$ and, using the fact that $\lim_{x\to 0} x^2 \log\left(1 + \frac{1}{x}\right) = 0$, total trip cost reduces to $TTC = \frac{\alpha_2}{2}\frac{N^2}{q_m}$, as in the bottleneck model.

4.4 User Optimum

Given population, N, and unit schedule delay cost, α_2, the user optimum, or no-toll equilibrium, is an inflow rate, $a(t)$, such that no individual has an incentive to change their departure time. This condition can be achieved by imposing the *trip-timing (TT) condition* that no vehicle can experience a lower trip cost by departing at a different time. A simple consequence of the TT-condition is that trip cost must be identical for all departure times, t_d.

Time of Last Vehicle Arrival	$\bar{t} = 1 + \frac{N}{2} + \sqrt{\frac{1}{\alpha_2}N + \left(\frac{N}{2}\right)^2}$
Time of Last Vehicle Departure	$t_f = \bar{t} - 1$
Inflow Rate	$a(t) = \begin{cases} \frac{t\left[t+2\left(\frac{1}{\alpha_2}-1\right)\right]}{\left[t+\frac{1}{\alpha_2}-1\right]^2}, & 0 < t < (1-\alpha_2)(\bar{t}-1) \\ 1 - \frac{1}{(\bar{t}-t)^2}, & (1-\alpha_2)(\bar{t}-1) \le t < t_f \\ 0, & \text{otherwise} \end{cases}$
Cumulative Inflow Rate	$A(t) = \begin{cases} 0, & t \le 0 \\ \frac{t^2}{t+\frac{1}{\alpha_2}-1}, & 0 < t \le (1-\alpha_2)(\bar{t}-1) \\ N - \left(\bar{t} - t + \frac{1}{\bar{t}-t} - 2\right), & (1-\alpha_2)(\bar{t}-1) < t \le t_f \\ N, & t > t_f \end{cases}$
Outflow Rate	$q(t) = \begin{cases} 1 - \frac{1}{[1+\alpha_2(t-1)]^2}, & 1 \le t \le \bar{t} \\ 0, & \text{otherwise} \end{cases}$
Cumulative Outflow Rate	$Q(t) = \begin{cases} 0, & t \le 1 \\ \frac{1}{\alpha_2}\left[1 + \alpha_2(t-1) + \frac{1}{1+\alpha_2(t-1)} - 2\right], & 1 < t \le \bar{t} \\ N, & t > \bar{t} \end{cases}$
Arrival Time if $t_d \le (1-\alpha_2)(\bar{t}-1)$	$t_a = 1 + \frac{1}{2}\left(\frac{t_d^2}{t_d+\frac{1}{\alpha_2}-1}\right) + \sqrt{\frac{1}{\alpha_2}\left(\frac{t_d^2}{t_d+\frac{1}{\alpha_2}-1}\right) + \frac{1}{4}\left(\frac{t_d^2}{t_d+\frac{1}{\alpha_2}-1}\right)^2}$
Vehicle Trajectory if $t_d \le (1-\alpha_2)(\bar{t}-1)$	$x(t) = t - t_d\sqrt{\frac{t+\frac{1}{\alpha_2}-1}{t_d+\frac{1}{\alpha_2}-1}}$
Spacetime Point where Trajectory Enters Region $(1-\alpha_2)(\bar{t}-1) < t_d$	$(x_0, t_0) = \left(1 - \frac{\alpha_2(\bar{t}-1)+1}{(\alpha_2(\bar{t}-1))^2}\frac{(\bar{t}-1-t_d)^2}{\bar{t}-t_d}, \bar{t} - \left(1 + \frac{1}{\alpha_2(\bar{t}-1)}\right)^2\frac{(\bar{t}-1-t_d)^2}{\bar{t}-t_d}\right)$
Arrival Time if $(1-\alpha_2)(\bar{t}-1) < t_d$	$t_a = 1 + \frac{1}{2}\left[\frac{(t_0-x_0)^2}{t_0+\frac{1}{\alpha_2}-1}\right] + \sqrt{\frac{1}{\alpha_2}\left[\frac{(t_0-x_0)^2}{t_0+\frac{1}{\alpha_2}-1}\right] + \frac{1}{4}\left[\frac{(t_0-x_0)^2}{t_0+\frac{1}{\alpha_2}-1}\right]^2}$
Vehicle Trajectory if $(1-\alpha_2)(\bar{t}-1) < t_d$	$x = \begin{cases} 1 - (\bar{t}-t) + (\bar{t}-1-t_d)\sqrt{\frac{\bar{t}-t}{\bar{t}-t_d}}, & t_d \le t \le t_0 \\ t - (t_0 - x_0)\sqrt{\frac{t+\frac{1}{\alpha_2}-1}{t_0+\frac{1}{\alpha_2}-1}}, & t_0 < t \le t_a \end{cases}$
Total Schedule Delay	$TSD = \frac{1}{2}t_f^2 - \frac{1}{\alpha_2}t_f + \frac{1}{\alpha_2^2}\log(1+\alpha_2 t_f)$
Total Travel Time	$TTT = N(1+\alpha_2 t_f) - \alpha_2 t_f^2 + 2t_f - \frac{2}{\alpha_2}\log(1+\alpha_2 t_f)$
Total Trip Cost	$TTC = N(1+\alpha_2 t_f) - \frac{1}{2}\alpha_2 t_f^2 + t_f - \frac{1}{\alpha_2}\log(1+\alpha_2 t_f)$

Table 4.2: Table of results for the social optimum (scaled units).

Let t_d, t'_d denote the departure times of two vehicles that arrive at times t_a, t'_a, respectively. Since their trip costs must be equal,

$$t_a - t_d + \alpha_2 (\bar{t} - t_a) = t'_a - t'_d + \alpha_2 \left(\bar{t} - t'_a\right).$$

Assume $t_d < t'_d$, so $t_a < t'_a$, and let $\Delta t_d = t'_d - t_d$ and $\Delta t_a = t'_a - t_a$. The above equation can be rewritten as

$$\Delta t_a = \frac{1}{1 - \alpha_2} \Delta t_d. \tag{4.24}$$

Since Δt_a, $\Delta t_d > 0$, from (4.24) we conclude that the TT-condition can be satisfied only if $\alpha_2 < 1$. Section 4 of Newell argues that, for an arbitary arrival-demand curve, this is a necessary condition for constructing a user optimum. For the specific arrival-demand function for the Single-Entry Corridor Problem with no late arrivals, the previous argument yields the same result.

Since the first departure at time $t = 0$ arrives at time $t = 1$, (4.24) implies that a departure at time $t = t_d$ satisfies

$$t_a = 1 + \frac{1}{1 - \alpha_2} t_d. \tag{4.25}$$

Since cumulative flow is constant along a trajectory,

$$\int_0^{t_d} a(t)\, dt \;=\; \int_1^{t_a} q(t)\, dt,$$

and differentiating with respect to t_d yields

$$a(t_d) = \frac{1}{1 - \alpha_2} q(t_a). \tag{4.26}$$

Thus, along a vehicle trajectory, from the corridor entry-point to the CBD the flow rate decreases by the multiplicative factor $\frac{1}{1-\alpha_2}$.

4.4.1 Qualitative Features

In the following lemmas we show that a consequence of (4.26) is that $a(0) = 0$ and that $a(t)$ is a continuously increasing function from time $t = 0$ up to the final departure at time $t = t_f$, at which time $a(t)$ discontinuously decreases to zero.

Lemma 1 *The corridor inflow and outflow rates discontinuously decrease to zero at times t_f and $\bar{t}$, respectively.*

Proof. Since schedule delay is a decreasing function of departure time, and since trip cost is constant, travel time must be an increasing function of departure time. As a consequence, the final departure does not travel at free-flow velocity; thus, its trajectory coincides with a shock wave path (see Section 4.2.2.2). ■

Lemma 2 *The corridor inflow rate, $a(t)$, must be an increasing function of departure time from $t = 0$ to the final departure at time $t = t_f$.*

Proof. We use proof by contradiction. Since the first departure occurs at $t = 0$, $a(t) = 0$ for $t < 0$. Denote the first time for which $a(t)$ is nonincreasing by t_{d_1}, and let t_{a_1} denote the arrival time of a vehicle that departs at time t_{d_1}. The outflow rate at t_{a_1} is determined by characteristic lines that originate from $x = 0$ at times earlier than t_{d_1}; thus, the outflow rate must be increasing at time t_{a_1}. However, by (4.26), since the inflow rate is nonincreasing at t_{d_1}, the outflow rate is also nonincreasing at t_{a_1}, which is a contradiction. ■

Lemma 2 implies that, except for the trajectory of the last departure, shocks do not occur in the UO solution.

Lemma 3 *The corridor inflow rate, $a(t)$, does not discontinuously increase.*

Proof. By Lemma 2, since $a(t)$ is increasing, the outflow rate, $q(t)$, is also increasing. If $a(t)$ discontinuously increased, then $q(t)$ would be obtained from a rarefaction fan of characteristic waves, so $q(t)$ would increase continuously. Thus, $q(t)$ increases continuously, whether or not $a(t)$ has a discontinuous increase. From (4.26), the continuity of $q(t)$ implies the continuity of $a(t)$. ■

In particular, since $a(t) = 0$ for $t < 0$, Lemma 3 implies that $a(0) = 0$.

4.4.2 Inflow and Outflow Curves

For the following discussion, refer to Figure 4.9. A departure at time t_d arrives at the CBD at time t_a. Using (4.7) we can draw a line of constant flow from $(t_a, Q(t_a))$ to a point $(t_0, A(t_0))$, where $a(t_0) = q(t_a)$. (4.7a) yields t_0 in terms of t_a as

$$t_0 = t_a - \frac{1}{\sqrt{1 - q(t_a)}},$$

which can be rewritten in terms of t_d using (4.25) and (4.26)

$$t_0 = \frac{1}{1-\alpha_2} t_d - \left(\frac{1}{\sqrt{1-(1-\alpha_2)a(t_d)}} - 1 \right). \tag{4.27}$$

Since $a(t_0) = q(t_a)$, (4.26) can be rewritten as $a(t_d) = \frac{1}{1-\alpha_2} a(t_0)$, and substituting t_0 with (4.27) yields an implicit expression for $a(t_d)$ in terms of t_d:

$$a(t_d) = \frac{1}{1-\alpha_2} a \left(\frac{1}{1-\alpha_2} t_d - \left[\frac{1}{\sqrt{1-(1-\alpha_2)a(t_d)}} - 1 \right] \right). \tag{4.28}$$

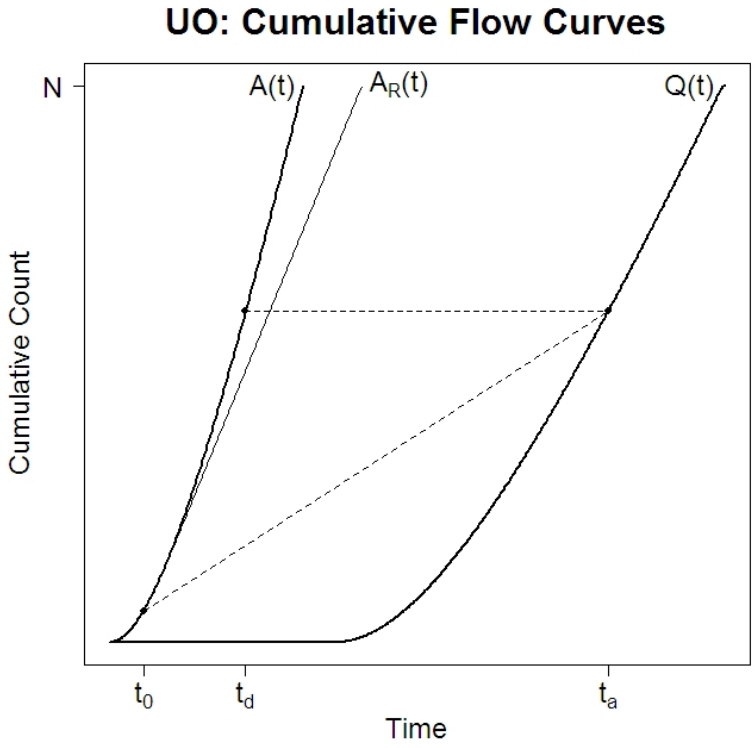

Figure 4.9: Cumulative corridor inflow, road inflow, and corridor outflow curves, $A(t)$, $A_R(t)$, and $Q(t)$, respectively. The slope of $A_R(t)$ cannot exceed capacity flow. The horizontal distance between $A(t)$ and $A_R(t)$ is time spent in a queue. A departure at time t_d arrives at the CBD at time t_a, indicated by the horizontal dashed line. The diagonal dashed line is a line of constant flow, so the slope of $A(t)$ at t_0 equals the slope of $Q(t)$ at t_a.

(4.28) is satisfied by the corridor inflow rate, $a(t)$, for all departure times, t_d. If $a(t)$ exceeds capacity flow, then a queue develops since the road inflow rate cannot exceed capacity flow.

We use an iterative procedure[6] to solve (4.28) for $a(t_d)$, $0 < t_d < t_f$. To simplify the notation, we replace t_d with t, denote $\frac{1}{1-\alpha_2}$ by σ, and denote $\frac{1}{\sqrt{1-a(t)}} - 1$ by $g\left(a(t)\right)$, so (4.28) becomes

$$a(t) = \sigma a\left(\sigma t - g\left(\frac{1}{\sigma}a(t)\right)\right). \tag{4.29}$$

[6]The following iterative procedure is based on work in Danqing Hu's Ph.D. dissertation, *Three Essays on Urban and Transport Economics*, Department of Economics, University of Wisconsin-Madison.

In Table 4.3 we repeatedly iterate (4.29) n times, yielding the following n-th iterated expression for $a(t)$:

$$a(t) = \sigma^n a\left(\sigma^n\left[t - \sum_{j=1}^{n} \frac{1}{\sigma^j} g\left(\frac{1}{\sigma^j} a(t)\right)\right]\right). \tag{4.30}$$

From the lemmas of the previous section, $a(t)$ is a continuously increasing function on $0 \leq t < t_f$; thus, its inverse is well-defined on this interval, so (4.30) may be rewritten as

$$(1-\alpha_2)^n a^{-1}\left[(1-\alpha_2)^n a(t)\right] = t - \sum_{j=1}^{n} (1-\alpha_2)^j \left[\frac{1}{\sqrt{1-(1-\alpha_2)^j a(t)}} - 1\right].$$

Since $a(0) = 0$, taking the limit as $n \to \infty$, the left-hand side becomes zero and we obtain an expression for the inverse of the inflow rate, $t = t(a)$,

$$t(a) = \sum_{j=1}^{\infty} (1-\alpha_2)^j \left[\frac{1}{\sqrt{1-(1-\alpha_2)^j a}} - 1\right], \tag{4.31}$$

which is a convergent series for $0 \leq a < \frac{1}{1-\alpha_2}$. The inflow rate, $a(t)$, increases up to some final value, a_f, at time t_f, when the entire population has departed, i.e., $\int_0^{t_f} a(t)\, dt = N$, which is equivalent to

$$\int_0^{a_f} t(a)\, da + N = t_f a_f.$$

Integrating (4.31), replacing t_f with $t(a_f)$, and algebraically solving for N yields

$$N = \sum_{j=1}^{\infty} \frac{2\left[1 - \sqrt{1-(1-\alpha_2)^j a_f}\right] - (1-\alpha_2)^j a_f}{\sqrt{1-(1-\alpha_2)^j a_f}}, \quad 0 \leq a_f < \frac{1}{1-\alpha_2}. \tag{4.32}$$

Given a population size, N, (4.32) must be numerically solved to determine a_f, which

$$
\begin{aligned}
a(t) &= \sigma a\left(\sigma t - g\left(\tfrac{1}{\sigma}a(t)\right)\right) \\
&= \sigma a(u_1), \text{ where } u_1 = \sigma t - g\left(\tfrac{1}{\sigma}a(t)\right) \\
\\
a(t) &= \sigma^2 a\left(\sigma u_1 - g\left(\tfrac{1}{\sigma}a(u_1)\right)\right) \\
&= \sigma^2 a\left(\sigma^2 t - \sigma g\left(\tfrac{1}{\sigma}a(t)\right) - g\left(\tfrac{1}{\sigma^2}a(t)\right)\right) \\
&= \sigma^2 a(u_2), \text{ where } u_2 = \sigma^2 t - \sigma g\left(\tfrac{1}{\sigma}a(t)\right) - g\left(\tfrac{1}{\sigma^2}a(t)\right) \\
\\
a(t) &= \sigma^3 a\left(\sigma u_2 - g\left(\tfrac{1}{\sigma}a(u_2)\right)\right) \\
&= \sigma^3 a\left(\sigma^3 t - \sigma^2 g\left(\tfrac{1}{\sigma}a(t)\right) - \sigma g\left(\tfrac{1}{\sigma^2}a(t)\right) - g\left(\tfrac{1}{\sigma^3}a(t)\right)\right) \\
&= \sigma^3 a(u_3), \text{ where } u_3 = \sigma^3 t - \sigma^2 g\left(\tfrac{1}{\sigma}a(t)\right) - \sigma g\left(\tfrac{1}{\sigma^2}a(t)\right) - g\left(\tfrac{1}{\sigma^3}a(t)\right) \\
&\vdots \\
a(t) &= \sigma^n a\left(\sigma^n\left[t - \sum_{j=1}^{n} \tfrac{1}{\sigma^j} g\left(\tfrac{1}{\sigma^j}a(t)\right)\right]\right) \\
&= \sigma^n a(u_n), \text{ where } u_n = \sigma^n\left[t - \sum_{j=1}^{n} \tfrac{1}{\sigma^j} g\left(\tfrac{1}{\sigma^j}a(t)\right)\right]
\end{aligned}
$$

Table 4.3: Repeated iteration of (4.29), which yields the n-th iterated expression for $a(t)$.

is inserted into (4.31) to determine t_f. If $a_f > 1$, then a queue develops at some time, denoted as t_Q. For a given value of α_2, if N is sufficiently large then a queue develops. Let us temporarily denote the critical value of N such that a queue develops as N_c, i.e., if $N > N_c$ then a queue develops. From (4.32),

$$N_c = \sum_{j=1}^{\infty} \frac{2\left[1 - \sqrt{1-(1-\alpha_2)^j}\right] - (1-\alpha_2)^j}{\sqrt{1-(1-\alpha_2)^j}}. \tag{4.33}$$

From (4.33) it can easily be shown that N_c is a decreasing function of α_2 which satisfies $\lim_{\alpha_2 \to 1} N_c = 0$, and $\lim_{\alpha_2 \to 0} N_c = \infty$.

Using the above procedure one may numerically construct the inflow rate for the UO solution, $a(t)$. Numerically integrating $a(t)$ yields the cumulative inflow rate, $A(t)$, from which we obtain the cumulative outflow rate $Q(t)$ via $Q(t_a) = A(t_d)$, where $t_a = 1 + \frac{1}{1-\alpha_2} t_d$. In Figure 4.10 we graph the numerically obtained cumulative corridor inflow, road inflow and outflow curves for given values $N = 0.8$ and $\alpha_2 = 0.5$. We chose N large enough (relative to α_2) so that a queue develops. The dashed lines are lines of constant flow, and show how the population inflow determines the outflow. In particular, note that the outflow is completely determined by the inflow that occurs before the queue develops, i.e., before capacity inflow is reached. The trajectory of the final departure is a shock wave path of decreasing shock strength.

4.4.3 Vehicle Trajectories

Given values for the population size, N, and unit schedule delay cost, α_2, we use the procedure outlined in the previous section to numerically construct the inflow rate. The inflow rate allows us to determine the slope of the characteristic lines in spacetime emanating from the t-axis, from which we numerically determine vehicle

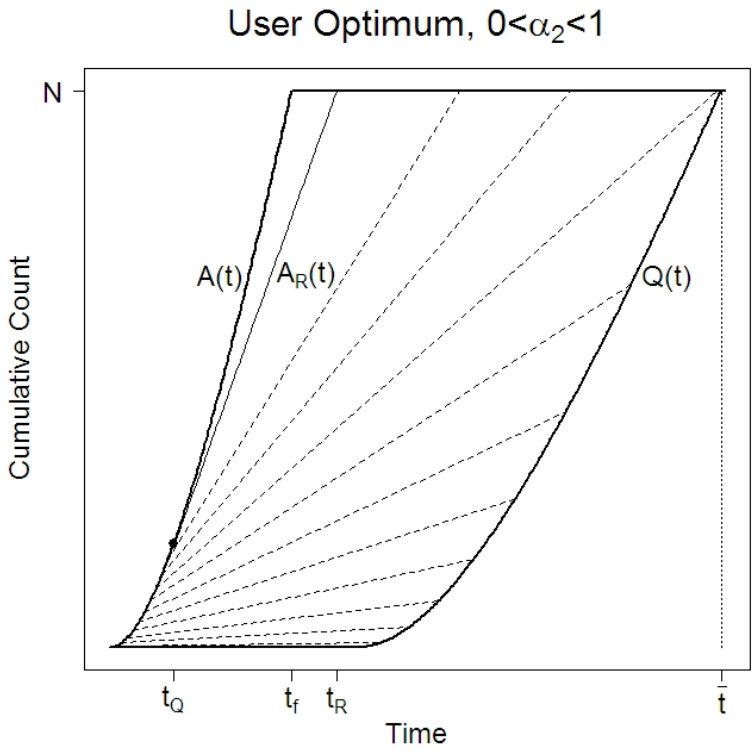

Figure 4.10: Cumulative corridor inflow, road inflow, and corridor outflow curves, $A(t)$, $A_R(t)$, and $Q(t)$, respectively, with $N = 0.8$ and $\alpha_2 = 0.5$. A queue develops at time t_Q, indicated by the solid dot, when the corridor inflow rate reaches capacity flow. As indicated by the dashed lines of constant flow, the outflow is completely determined by the inflow that occurs before the queue develops.

trajectory curves. Since we do not have an analytic expression for the inflow rate, $a(t)$, we are not able to analytically construct the differential equation that is satisfied by vehicle trajectory curves.

Recall that, if N is sufficiently large relative to α_2, then the corridor inflow rate reaches capacity flow at time t_Q, and continues to increase until the final departure at time t_f. The road inflow rate remains at capacity flow from time t_Q until time $t_R > t_f$, where t_R is the time the final vehicle departure enters the road. All departures after t_Q enter a queue before entering the road, and queueing time increases with departure time. In Figure 4.11 we have graphed characteristic lines and vehicle trajectory curves

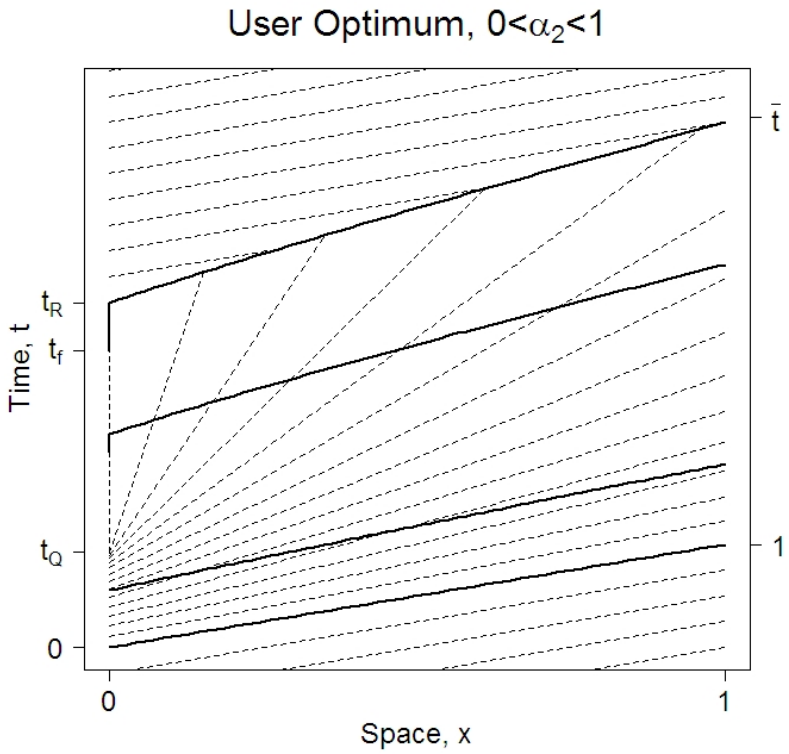

Figure 4.11: Dashed lines are characteristics whose slope depends upon the road inflow rate at the entry point at $x = 0$. For $t_Q < t \leq t_R$ a queue is present and the road inflow rate is capacity flow, so that characteristics are vertical lines. We have plotted four vehicle trajectory curves in bold: The first vehicle departure that travels at free-flow velocity; a departure that occurs before t_Q and does not experience a queue; a departure that occurs after t_Q and experiences a queue; and the final vehicle departure at time t_f that experiences a queue until entering the road at time t_R.

for a user optimum solution in which a queue develops.

4.4.4 Trip Costs

From (4.25) the travel time for a departure at time t_d is a linear function of t_d,

$$t_a - t_d = 1 + \frac{\alpha_2}{1 - \alpha_2} t_d.$$

Since trip cost is constant, schedule delay is also a linear function of t_d, summarized in (4.34), and graphed in Figure 4.12:

$$\begin{aligned} C = & \text{ Travel Time } + \alpha_2(\text{Schedule Delay}) \\ & = 1 + \frac{\alpha_2}{1-\alpha_2} t_d + \alpha_2 \left(\bar{t} - 1 - \frac{1}{1-\alpha_2} t_d \right) \\ & = 1 + \alpha_2(\bar{t} - 1) \end{aligned} \tag{4.34}$$

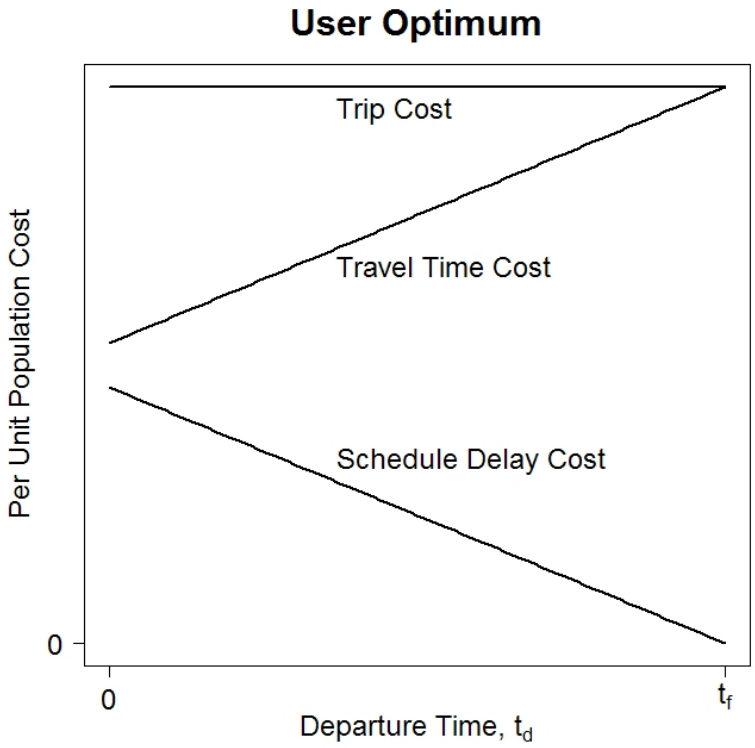

Figure 4.12: Travel time cost, schedule delay cost and trip cost as functions of departure time for the user optimum, with $N = 1$ and $\alpha_2 = 0.5$. The trip-timing condition requires that trip cost is constant and travel time cost is a linear function of departure time, so that schedule delay is also a linear function of departure time.

Since we have not obtained closed-form expressions for the cumulative inflow and outflow curves, we cannot obtain a closed-form expression for the total travel time or schedule delay. However, since the trip cost is constant for all vehicles, we can

analytically determine the total trip cost, in terms of either the final departure time, t_f, or the final arrival time, $\bar{t}$:

$$\begin{aligned} TTC &= N\left[1+\frac{\alpha_2}{1-\alpha_2}t_f\right] \\ &= N\left[1+\alpha_2\left(\bar{t}-1\right)\right]. \end{aligned} \tag{4.35}$$

4.4.5 Reduction to Bottleneck Model

Arnott and Lindsey, 1990 showed that the user optimum for the bottleneck model with no late arrivals entails a constant inflow rate of $\frac{1}{1-\frac{\alpha_2}{\alpha_1}}q_m$, with a resulting total trip cost of $\alpha_2\frac{N^2}{q_m}$. We now show that the UO solution for the Single-Entry Corridor Problem with no late arrivals reduces to the bottleneck model as $l \to 0$.

Since road inflow cannot exceed capacity flow, the outflow, $q(t)$, cannot exceed capacity flow. Rewriting (4.26) in unscaled units, we conclude that corridor inflow cannot exceed $\frac{1}{1-\frac{\alpha_2}{\alpha_1}}q_m$. In (4.31) we provided an expression for the inverse of the inflow rate, which in unscaled units becomes

$$t = \frac{l}{v_0}\sum_{j=1}^{\infty}(1-\frac{\alpha_2}{\alpha_1})^j\left[\frac{1}{\sqrt{1-(1-\frac{\alpha_2}{\alpha_1})^j\frac{a}{q_m}}}-1\right].$$

The infinite sum in this equation converges if $a < \frac{1}{1-\frac{\alpha_2}{\alpha_1}}q_m$. In the limit as $l \to 0$, to ensure that the right-hand side of this equation does not equal zero, the infinite sum must diverge. Since (4.26) implies $a \leq \frac{1}{1-\frac{\alpha_2}{\alpha_1}}q_m$, we conclude that as $l \to 0$, $a(t) = \frac{1}{1-\frac{\alpha_2}{\alpha_1}}q_m$, reproducing the UO solution for the bottleneck model. Since, in the limit as $l \to 0$ the inflow rate is constant, the final departure time is

$$t_f = \frac{N}{\frac{1}{1-\frac{\alpha_2}{\alpha_1}}q_m}.$$

Rewriting the expression for the total trip cost, (4.35), in unscaled units yields

$$TTC = N\left[\alpha_1 \frac{l}{v_0} + \frac{\alpha_2}{1-\frac{\alpha_2}{\alpha_1}} t_f\right],$$

and taking the limit as $l \to 0$ we obtain $TTC = \alpha_2 \frac{N^2}{q_m}$, as in the bottleneck model.

4.4.6 Comparison with a No-Propagation Model

No-propagation models[7] assume a simplified version of LWR flow in which a vehicle travels at constant velocity, depending on the flow or density at some location along the road. In Henderson, 1977, a vehicle's velocity is determined by the flow rate when it enters the road; in Chu, 1992, a vehicle's velocity is determined by the flow rate when it exits the road; and Mun, 1999 applies Henderson's flow congestion specification upstream of a bottleneck.

Recently, Tian et al. (2010) developed a no-propagation model for the morning commute in which a vehicle's constant velocity depends on the density at the entry point, and determined a corresponding UO solution under Greenshields' Relation, assuming no late arrivals. They presented a numerical example where they compared trip cost as a function of total population for their no-propagation model against the standard bottleneck model. In this section we compare their results with the UO results from the single-entry corridor problem results derived assuming LWR flow congestion.

For consistency, we continue to use the notation used throughout this paper, even when presenting the results from Tian et al., 2010. The following parameter values are used in the numerical example from Section 5 of Tian et al., 2010: $l = 50$ km, $v_0 = 150$ $\frac{\text{km}}{\text{hr}}$, $k_j = 200$ $\frac{\text{veh}}{\text{km}}$ (which implies that $q_m = \frac{1}{4} v_0 k_j = 7500$ $\frac{\text{veh}}{\text{hr}}$), $\alpha_1 = 30$ $\frac{\$}{\text{hr}\cdot\text{veh}}$

[7] A vehicle affects the travel speed only of those vehicles traveling in the same cohort. Thus, the congestion caused by a vehicle does not propagate either forwards or backwards.

and $\alpha_2 = 15 \ \frac{\$}{\text{hr}\cdot\text{veh}}$.

Under LWR flow congestion, if the population is sufficiently large then a queue develops in the UO solution. The critical population value such that a queue develops is N_c, and from (4.33) in unscaled units:

$$N_c = \frac{q_m l}{v_0} \sum_{j=1}^{\infty} \frac{2\left[1 - \sqrt{1 - (1 - \frac{\alpha_2}{\alpha_1})^j}\right] - (1 - \frac{\alpha_2}{\alpha_1})^j}{\sqrt{1 - (1 - \frac{\alpha_2}{\alpha_1})^j}} \approx 370 \text{ veh.}$$

To construct a graph of the (per unit population) trip cost function under LWR flow congestion, $C_{\text{LWR}}(N)$, proceed as follows. For each value of N, numerically solve (4.32) in unscaled units to determine a_f, and then insert this value into (4.31) in unscaled units to determine t_f. From (4.25) in unscaled units determine $\bar{t}$ as

$$\bar{t} = \frac{l}{v_0} + \frac{1}{1 - \frac{\alpha_2}{\alpha_1}} t_f,$$

and use this value to determine the trip cost from (4.34) in unscaled units,

$$C_{\text{LWR}}(N) = \alpha_1 \frac{l}{v_0} + \alpha_2 \left(\bar{t} - \frac{l}{v_0}\right).$$

A graph of $C_{\text{LWR}}(N)$ is provided in Figure 4.13, labelled as "LWR flow".

Under the no-propagation model developed by Tian et al., 2010, a queue develops in the corresponding UO solution if the population exceeds the critical value N_{G} (see (15) in Tian et al., 2010),

$$N_{\text{G}} \approx \frac{k_j l}{5} \cdot \frac{\alpha_1 - \alpha_2}{\alpha_2} = 2000 \text{ veh.}$$

The (per unit population) trip cost function under their model is (see (21), (22) and

(29) in Tian et al., 2010)

$$C_{\mathrm{G}}(N) \approx \begin{cases} \frac{\alpha_1 l}{v_0}\left[1+\left(\frac{5v_0}{4l}\cdot\frac{\alpha_2}{\alpha_1-\alpha_2}\cdot\frac{N}{q_m}\right)^{\frac{2}{3}}\right], & N \le N_{\mathrm{G}} \\ \frac{\alpha_1 l}{v_0}\left[1+\left(\frac{5v_0}{4l}\cdot\frac{\alpha_2}{\alpha_1-\alpha_2}\cdot\frac{N_{\mathrm{G}}}{q_m}\right)^{\frac{2}{3}}\right]+\alpha_2\frac{N-N_{\mathrm{G}}}{q_m}, & N > N_{\mathrm{G}} \end{cases}.$$

A graph of $C_{\mathrm{G}}(N)$ is provided in Figure 4.13, labelled as "No-propagation".

Under the standard bottleneck model a queue develops for all population values, and the (per unit population) trip cost function is (see (30) in Tian et al., 2010)

$$C_{\mathrm{B}}(N) = \frac{\alpha_1 l}{v_0} + \alpha_2\frac{N}{q_m}.$$

A graph of $C_{\mathrm{B}}(N)$ is provided in Figure 4.13, labelled as "Bottleneck".

In Figure 4.13 we graph the (per unit population) trip cost function for the above three models, analogous to Fig. 5 from Tian et al., 2010. At any population value, the UO trip costs obtained by the three models have the following relationship of their relative magnitudes:

$$C_{\mathrm{B}}(N) < C_{\mathrm{LWR}}(N) < C_{\mathrm{G}}(N).$$

Tian et al., 2010 concluded that the standard bottleneck model underestimates the user optimum trip cost, attributed to neglecting the effect of flow-dependent congestion. We further conclude that the no-propagation model overestimates the user optimum trip cost. Under LWR flow congestion in the single-entry corridor model, as vehicles enter the roadway their speed gradually increases (with the exception of the first vehicle which travels at free-flow velocity). However, in the no-propagation model vehicles do not change their speed after entering the roadway. Thus, the overestimation of user optimum trip cost in the no-propagation model may be attributed to the increase in

travel time due to vehicles not accelerating.

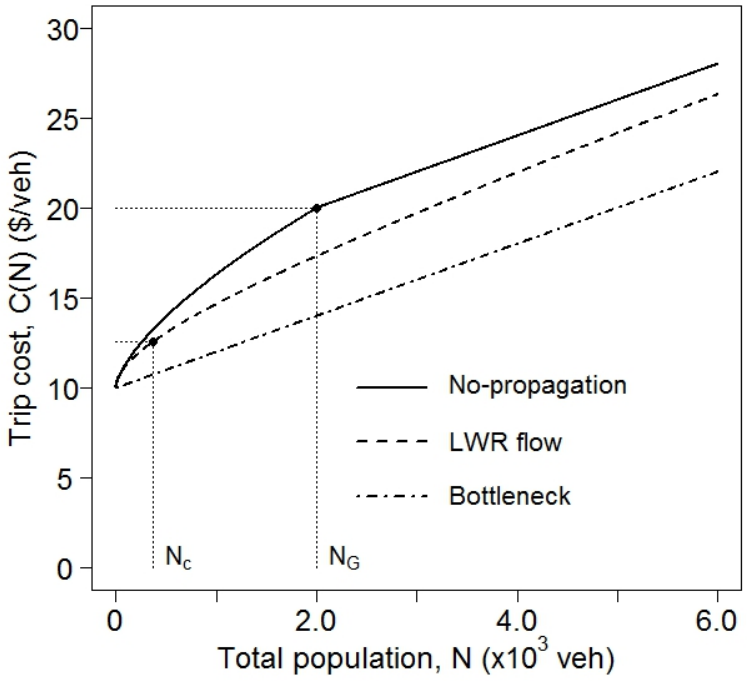

Figure 4.13: Trip cost (per unit population) as a function of total population for the user optimum solution under three different models: A no-propagation model developed by Tian et al., 2010, the single-entry corridor problem model that uses LWR flow congestion, and the standard bottleneck model. In the first two models a queue does not develop until the total population reaches a critical value, N_{G} and N_c, respectively.

4.5 Economic Properties

This section provides a brief discussion of some economic properties of the SO and UO of the single-entry corridor problem. In particular, it will examine: i) the properties of the cost functions for the SO and UO (the supply side of the transportation market); ii) the properties of the time-varying toll that decentralizes the social optimum; and iii) how these properties differ from the corresponding properties of the bottleneck model.

4.5.1 The SO Cost Function

The total trip cost function relates total trip cost to the number of users. The total variable trip cost equals total trip cost minus the cost that would be incurred were there no congestion, which is the cost of users' free-flow travel time. There are two components of total variable trip cost, total schedule delay cost and total variable travel time cost. We employ standard notation with respect to costs. As prefixes to symbols, T denotes total, A average, M marginal, F fixed, and V variable, and as a suffix C denotes cost. The core of a symbol indicates the nature of the quantity or cost. Thus, for example, TTC denotes total trip cost, while $MVTC$ denotes marginal variable trip cost.

To simplify the algebra, we retain the normalization from the earlier sections of the paper. Length units are chosen such that the length of the road equals 1.0, time units such that the free-flow travel time on the road equals 1.0, population units such that $N = 1.0$ corresponds to the road's capacity flow per unit time, and cost units such that the value of travel time equals 1.0.

Table 4.4 records the relevant algebraic results for the social optimum. Results are presented in normalized form, obtained straightforwardly from previous results. Row 1 records the algebra for schedule delay cost, row 2 for variable travel time cost (travel time cost in excess of free flow travel time cost), and row 3 for variable trip cost (the sum of schedule delay and variable travel time cost). Column 1 gives the expression for total cost (e.g., total schedule delay cost), column 2 for marginal cost, column 3 for private cost, and column 4 for externality cost, which equals the difference between marginal cost and private cost. Note that the marginal costs are independent of departure time

	Total Cost
Schedule Delay	$\frac{1}{2}\alpha_2 t_f^2 - t_f + \frac{1}{\alpha_2}\log(1+\alpha_2 t_f)$
Variable Travel Time	$N\alpha_2 t_f - \alpha_2 t_f^2 + 2t_f - \frac{2}{\alpha_2}\log(1+\alpha_2 t_f)$
Variable Trip	$N\alpha_2 t_f - \frac{1}{2}\alpha_2 t_f^2 + t_f - \frac{1}{\alpha_2}\log(1+\alpha_2 t_f)$

	Marginal Cost	Private Cost	Externality Cost
Schedule Delay	$\frac{N(1+\alpha_2 t_f)}{2t_f - N}$	$\alpha_2(\bar{t} - t_a)$	$\frac{N(1+\alpha_2 t_f)}{2t_f - N} - \alpha_2(\bar{t} - t_a)$
Variable Travel Time	$\frac{N}{2t_f - N}$	$t_a - t_d - 1$	$\frac{N}{2t_f - N} - (t_a - t_d - 1)$
Variable Trip	$\alpha_2 t_f$	$\alpha_2(\bar{t} - t_a) + t_a - t_d - 1$	$t_d - (1-\alpha_2)(t_a - 1)$

Table 4.4: Normalized Cost Formulae: Social Optimum, $\bar{t} = t_f + 1$ and $t_f = \frac{N}{2} + \sqrt{\frac{N}{\alpha_2} + \left(\frac{N}{2}\right)^2}$.

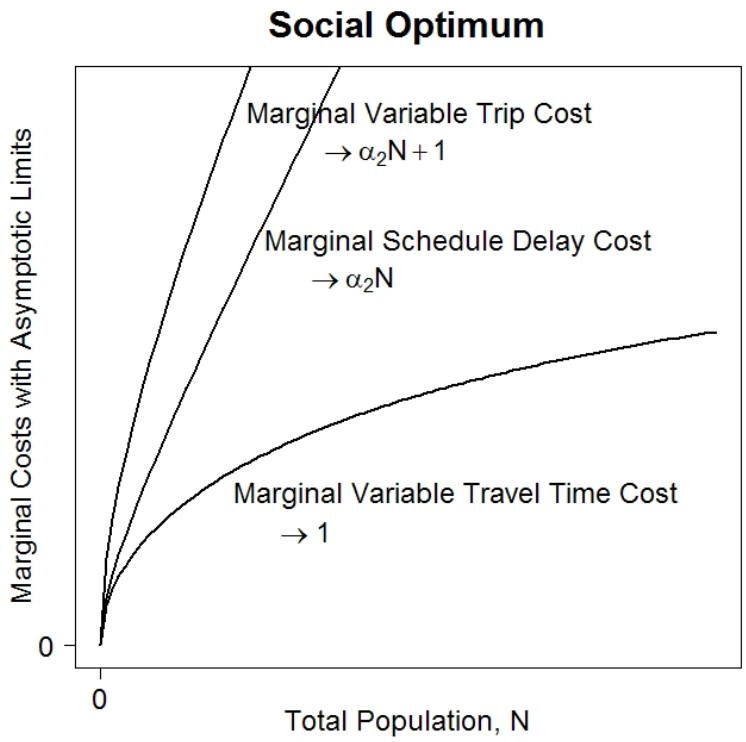

Figure 4.14: Marginal costs as functions of population for the social optimum, with $\alpha_2 = 0.5$.

but that the private costs and the externality costs depend on departure time.

Figure 4.14 plots marginal variable travel time cost, marginal schedule delay cost, and marginal variable trip cost for the social optimum, as functions of population, and records the asymptotic value of each marginal cost as N approaches infinity. Consider first marginal variable travel time cost. In the limit as N approaches zero, vehicles travel at free-flow travel speed, and both private and marginal variable travel time costs are zero. As N increases, mean travel time increases, until in the limit marginal variable travel time cost equals 1.0, which with Greenshields' Relation is the variable travel time cost when traveling the entire length of the road at capacity flow. Consider next marginal schedule delay cost. In the limit as N approaches zero, marginal schedule delay cost is of order $N^{\frac{1}{2}}$ and so approaches zero. In the limit as N approaches infinity,

marginal schedule delay cost equals $\alpha_2 t_f - 1 = \alpha_2 N$.

Since the social optimum entails the minimization of total variable trip costs, the Envelope Theorem can be applied to compute marginal variable trip cost.[8] In particular, it can be computed as the variable trip cost of a vehicle added just before the first vehicle to depart, holding fixed the departure pattern of all other vehicles. This vehicle travels at free-flow travel speed, does not affect the traffic flow of later vehicles, and incurs a schedule delay of t_f. Thus, marginal variable trip cost equals $\alpha_2 t_f$. From the expression for t_f given in Table 4.4, it follows that the elasticity of marginal variable trip cost with respect to N, $E_{MVTC:N}$, increases monotonically in N, rising from 0.5 when N is zero to 1.0 when N is infinite.

An important implication of this result is that at the social optimum the average externality cost imposed by a vehicle on other vehicles is strictly less than the average variable trip cost. Put alternatively, on average, the externality cost a vehicle imposes on other vehicles is less than the increase in cost it experiences due to congestion. In contrast, in the bottleneck model, on average, the externality cost a vehicle imposes on other vehicles equals the increase in cost it experiences due to congestion, while in empirical applications of the conventional static model the externality cost that a vehicle imposes on other vehicles is several times the increase in cost it experiences due to congestion.

4.5.2 The First-Best, Time-Varying Toll

Economists use the term first best to refer to a situation where the only constraints the policy maker/social planner faces are technological and resource constraints,

[8] Note that one cannot apply the Envelope Theorem to derive the marginal variable travel time cost and marginal schedule delay cost. The addition of a vehicle alters the time pattern of inflows and outflows of inframarginal drivers, and these adjustments do not net out.

as is the case in this paper. The first-best, time-varying toll decentralizes the social optimum; that is, imposition of this toll results in a user optimum allocation that coincides with the social optimum allocation. This occurs when each vehicle faces the social costs of its actions, and is achieved by imposing a toll at each point in time equal to the externality cost imposed by the vehicle, evaluated at the social optimum. The externality cost in turn equals the difference between the marginal variable trip cost and the private variable trip cost (the variable travel time cost incurred by a vehicle, $t_a - t_d - 1$, plus the schedule delay cost it incurs, $\alpha_2(\bar{t} - t_a)$). Thus, as a function of departure time, the first-best time-varying toll equals

$$\tau(t_d) = \alpha_2 t_f - \left[\alpha_2 \left(\bar{t} - t_a(t_d)\right) + \left(t_a(t_d) - t_d - 1\right)\right]. \tag{4.36}$$

The expression for t_f is given in Table 4.4, and the form of the function $t_a(t_d)$, which relates arrival time to departure time, is given in Table 4.2.

Figure 4.15 plots marginal variable trip cost, private variable trip cost, and the decentralizing toll, as functions of departure time with $\alpha_2 = 0.5$ and $N = 1$. The toll equals zero for the first vehicle to depart since that vehicle imposes no externality cost on other vehicles, experiencing the entire marginal variable trip cost as its private schedule delay cost. The toll increases at an increasing rate over the morning rush hour. The last vehicle experiences no schedule delay cost and no variable travel time cost, so that its externality cost equals the entire marginal variable trip cost.

Results on first-best optimal capacity with inelastic demand can be obtained straightforwardly. First-best optimal capacity minimizes the total cost associated with transporting the N vehicles from the origin to the destination, including the cost of

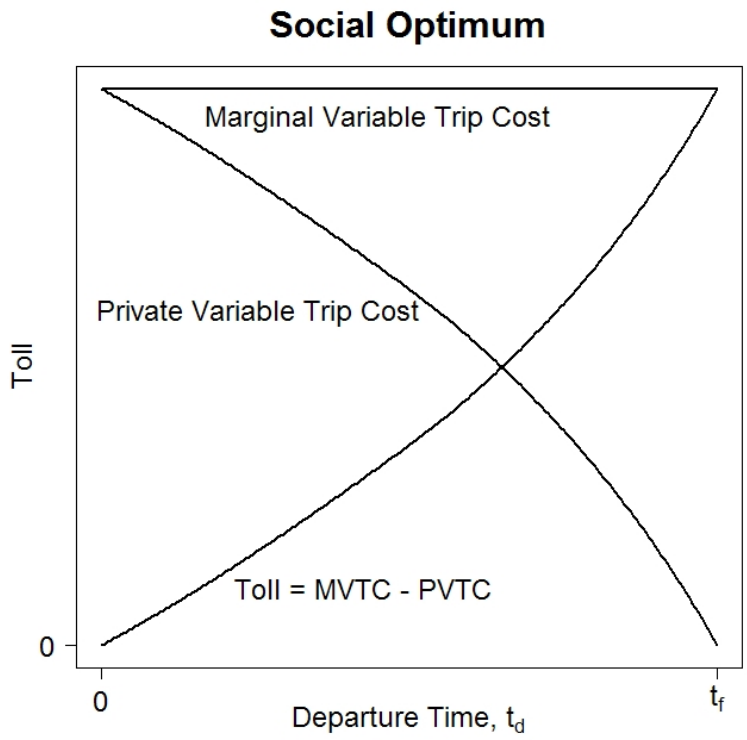

Figure 4.15: Marginal and private variable trip costs, and the toll, as functions of departure time for the social optimum, with $N = 1$ and $\alpha_2 = 0.5$.

constructing the road. The self-financing results carry through. Since the congestion technology has the property that velocity depends on density per unit area, the average trip cost function is homogeneous of degree zero in the volume/capacity ratio, where here volume is population. From familiar statements of the Self-Financing Theorem, it follows that the revenue from the optimal toll covers the cost of constructing optimal capacity if there are constant costs to capacity expansion.

4.5.3 User Optimum/No-Toll Equilibrium

Since the user optimum does not admit a neat analytical solution, we proceed using heuristic argument. We continue to apply the normalizations employed in the previous sections, so that solutions differ according to only two parameters, N and α_2.

The first person to depart experiences zero variable travel time. If it were otherwise, a person departing infinitesimally earlier would experience a lower trip cost, which is inconsistent with equilibrium. Consistency with the trip-timing equilibrium condition requires as well that travel time, as a function of departure time, increase as the rate $\frac{\alpha_2}{1-\alpha_2}$. These two results imply that the departure rate function (with $t = 0$ denoting the time of the first departure) is independent of N. Extra vehicles are accommodated through a lengthening of the rush hour, with the entry rate of these vehicles being determined by the trip-timing condition . Thus, a vehicle imposes a schedule delay externality on those vehicles that depart earlier and a travel time externality on those vehicles that depart later. The common trip price, in excess of free-flow travel time costs, $p(N)$, equals the schedule delay cost of the first person to depart and also the variable travel time cost of the last person to depart:

$$p(N) = \alpha_2 \left(\bar{t}(N) - 1\right) = \bar{t}(N) - t_f(N) - 1. \tag{4.37}$$

From the first of these equations

$$p'(N) = \alpha_2 \bar{t}'(N). \tag{4.38}$$

Since adding a vehicle changes traffic conditions only at the end of the rush hour,

$$\bar{t}'(N) = \frac{1}{q\left(\bar{t}(N)\right)}; \tag{4.39}$$

in words, adding a vehicle increases the length of the rush hour by the reciprocal of the

arrival rate. Thus,

$$p'(N) = \frac{\alpha_2}{q\left(\bar{t}(N)\right)} \tag{4.40}$$

and[9]

$$\begin{aligned} p''(N) &= -\frac{\alpha_2 q'\left(\bar{t}(N)\right)\bar{t}'(N)}{q\left(\bar{t}(N)\right)^2} \\ &= -\frac{\alpha_2 q'\left(\bar{t}(N)\right)}{q\left(\bar{t}(N)\right)^3} \end{aligned} \tag{4.41}$$

Since the arrival rate is increasing over the rush hour, the trip price function is concave.

Total variable trip costs are $Np(N)$. Thus, marginal variable trip cost as a function of N is

$$\begin{aligned} MVTC(N) &= p(N) + Np'(N) \\ &= \bar{t}(N) - t_f(N) - 1 + \frac{\alpha_2 N}{q\left(\bar{t}(N)\right)}. \end{aligned} \tag{4.42}$$

Since trip price and marginal variable trip cost are independent of departure time, the congestion externality cost too is independent of departure time, and equals $\frac{\alpha_2 N}{q(\bar{t}(N))}$. Consider adding the marginal vehicle at the end of the rush hour. Because this additional vehicle has no effect on traffic conditions for inframarginal vehicles, it does not affect their travel time cost, but instead increases each vehicle's schedule delay by $\frac{1}{q(\bar{t}(N))}$ at a social cost of $\frac{\alpha_2 N}{q(\bar{t}(N))}$. For the same reason, marginal schedule delay cost is $\frac{\alpha_2 N}{q(\bar{t}(N))}$ and marginal variable travel time cost is $p(N)$. Since marginal variable travel time cost is $p(N)$ and marginal schedule delay cost in $Np'(N)$, and since $p(N)$ is concave, for each

[9]For some pairs of N and α_2, satisfaction of the trip-timing condition requires the development of a queue. One may distinguish two different trip cost régimes, according to whether the last person to depart does not face a queue (régime 1 - low congestion) or faces a queue (régime 2 - high congestion). In α_2-N space, the boundary between the two régimes is negatively sloped, with high congestion occurring above the boundary and low congestion below it (since an increase in α_2 causes more weight to be placed on schedule delay, which compresses the rush hour).

level of population marginal variable travel time cost exceeds marginal schedule delay cost. And since marginal variable travel time cost equals trip price, and since marginal schedule delay cost equals the congestion externality cost, at each level of population trip price exceeds the congestion externality cost.

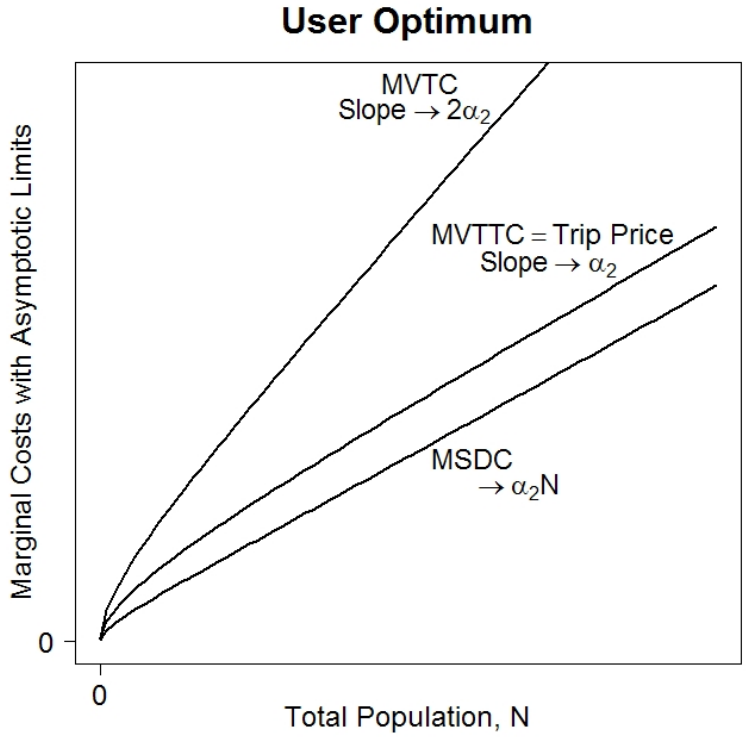

Figure 4.16: Marginal costs as functions of population for the user optimum, with $\alpha_2 = 0.5$.

Figure 4.16 is analogous to Figure 4.14 for the social optimum, except that it adds the trip price function. The marginal variable trip cost function is the vertical sum of the marginal schedule delay cost and marginal variable travel time cost functions, and the trip price function coincides with the marginal variable travel time cost function. All four functions equal zero at zero population. In the limit as N approaches infinity, the arrival rate approaches 1. The marginal schedule delay function approaches $\alpha_2 N$, the slope of the price function (= travel time cost function) approaches α_2, and the

slope of the marginal variable trip cost function approaches $2\alpha_2$.

4.5.4 Comparison of the Social Optimum and User Optimum

Figure 4.17 displays cumulative departures and arrivals, for both the social optimum and the no-toll equilibrium, with $N = 0.8$ and $\alpha_2 = 0.5$. Although $N = 8$ is consistent with the level of rush-hour congestion found in large metropolitan areas that are moderately congested by international standards[10], the qualitative features of the graph can be better viewed with smaller values of N. $\alpha_2 = 0.5$ is consistent with the empirical evidence. The curves for the social optimum are shown with solid lines, those for the no-toll equilibrium with dashed lines. The shapes of the two cumulative arrival curves are similar, though the rush hour is somewhat longer in the social optimum than in the no-toll equilibrium, resulting in higher total schedule delay in the social optimum than in the no-toll equilibrium. The cumulative departure schedules, however, have different shapes. In the social optimum, the schedule has a sigmoid shape, while in the no-toll equilibrium it is convex. Total travel times are considerably lower in the social optimum than in the no-toll equilibrium, partly because travel on the road is more congested in the no-toll equilibrium and partly because queuing occurs in the no-toll equilibrium but not in the social optimum.

Figure 4.18 superimposes Figures 4.14 and 4.16. The solid lines are for the social optimum and the dashed lines for the user optimum. Five points bear note.

1. The marginal schedule delay cost function for the social optimum lies above that

[10]Because the paper is already long, we have avoided providing numerical examples. However, it will be useful to illustrate how the model can be calibrated. Consider a metropolitan area in a poorer country, where infrastructure is low relative to traffic volume over the rush hour. $\frac{N}{q_m}$ is how long the rush hour would be if traffic were at capacity flow for the entire rush hour, and provides an exogeneous measure of infrastructure capacity relative to population. Suppose this equals 2.0 hours. Suppose too that $\frac{l}{v_0} = 0.25$ hours (the average trip would take 15 minutes at free-flow speed). Then, in scaled units the population is $N_{\text{Scaled}} = \frac{N}{q_m \frac{L}{v_0}} = \frac{2.0}{0.25} = 8$.

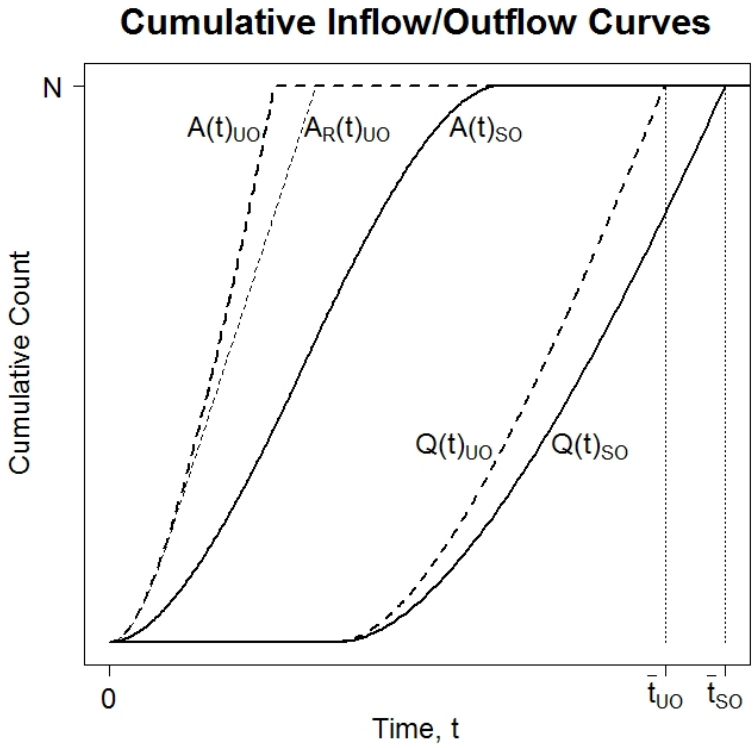

Figure 4.17: Cumulative inflow and outflow curves for the social optimum (solid lines) and user optimum (dashed lines), with $N = 0.8$ and $\alpha_2 = 0.5$. In the user optimum a queue develops when inflow into the corridor exceeds capacity, so that $A(t)_{UO}$ is the cumulative corridor inflow curve, $A_R(t)_{UO}$ is the cumulative road inflow curve, and the horizontal distance between these curves is the queueing time.

for the user optimum. This occurs because, at each level of population, the arrival rate at $\bar{t}$ is lower in the social optimum than in the user optimum.

2. The marginal variable travel time cost function for the social optimum lies below that for the user optimum. For small N, this arises because, at each level of population, travel is less congested in the social optimum than in the user optimum. Furthermore, for large N, there is no queuing in the social optimum, but there is in the user optimum.

3. The marginal variable trip cost function for the social optimum lies below that for the user optimum.

4. At a given level of population, the deadweight loss in the user optimum compared to the social optimum, due to the underpricing of congestion in the user optimum, equals the area between the two marginal variable trip cost functions up to that level of population.

5. In the decentralized social optimum, the trip price equals $MVTC_{SO}$, while in the no-toll equilibrium it equals $MVTTC_{UO}$. Since the $MVTC_{SO}$ curve lies above the $MVTTC_{UO}$ curve for all levels of N, the toll revenue collected from the optimal time-varying toll exceeds the efficiency gain from applying it[11].

4.5.5 Comparison of the Single-Entry Corridor Model to the Bottleneck Model

The basic bottleneck model, with identical individuals, a common desired arrival time, and no fixed component of trip cost, is starkly simple. All congestion takes the form of queuing behind a bottleneck of fixed flow capacity, and the arrival rate equals bottleneck capacity over the arrival interval in both the social optimum and user optimum. In the social optimum, variable travel time costs are zero, and marginal schedule delay cost (and hence marginal variable trip cost) is linear in population. In the user optimum, total variable travel time cost equals total schedule delay cost. The marginal schedule delay cost curve is the same as in the social optimum and coincides with the marginal variable travel time cost curve, so that the marginal variable trip cost curve has double the slope of both. In a decentralized environment, imposition of the optimal

[11] The revenue collected from the optimal time-varying toll can be calculated from the relationship for the decentralized social optimum that toll revenue, R, plus total variable trip costs equals trip price times population. Thus, for the level of population N', $R = MVTC_{SO}(N')\,N' - \int_0^{N'} MVTC_{SO}(N)\,dN$. The efficiency gain from applying the optimal time-varying toll, G, equals the total variable trip cost in the no-toll equilibrium minus total variable trip cost in the social optimum: $G = MVTTC_{UO}(N')\,N' - \int_0^{N'} MVTC_{SO}(N)\,dN$. Thus, $R - G = [MVTC_{SO}(N') - MVTTC_{UO}(N')]\,N'$.

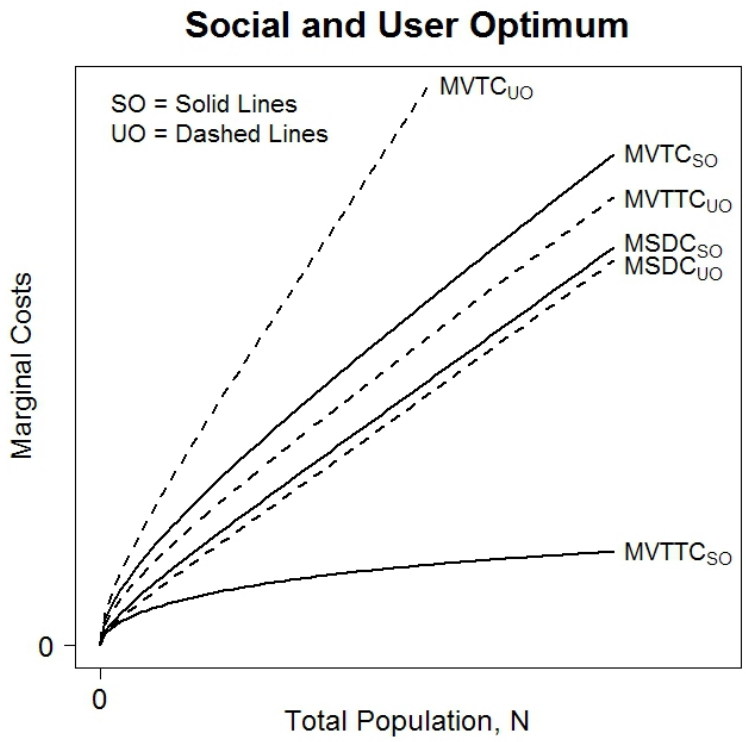

Figure 4.18: Marginal costs for the social optimum (solid lines) superimposed with the marginal costs for the user optimum (dashed lines), as functions of population, with $\alpha_2 = 0.5$.

time-varying toll has no effect on the length of the rush hour or on total schedule delay cost but completely eliminates variable travel time cost, so that the deadweight loss in the user optimum due to unpriced congestion equals total variable travel time cost in the user optimum. Since the optimal time-varying toll simply replaces the user optimum's variable travel time costs - queuing costs - the revenue raised from the optimal time-varying total equals the user optimum's total variable travel time costs and also the deadweight loss in the user optimum due to unpriced congestion. Thus, imposition of the optimal time-varying toll would benefit users as long as part of the revenue it generates is used to their benefit.

The single-entry corridor model takes a step towards realism by treating LWR flow congestion. Like the bottleneck, the road in the single-entry corridor has a fixed

flow capacity; but unlike the bottleneck, the road becomes congested at flow rates below capacity. In a decentralized environment, imposition of the optimal time-varying toll reduces congestion but does not eliminate it and causes the rush hour to lengthen. In the social optimum, marginal schedule delay cost, marginal variable travel time cost, and marginal variable trip cost, are all concave functions of population. Since flow is in the ordinary flow régime and there is no queuing in the social optimum, marginal variable travel time cost has an upper bound, but marginal schedule delay cost and marginal variable trip cost do not. In the user optimum too, marginal schedule delay cost, marginal variable travel time cost, and marginal variable trip cost, are all concave functions of population. Since there is queuing in the user optimum for sufficiently large N (although flow is in the ordinary flow régime), marginal variable travel time cost has no upper bound. The trip price function coincides with the marginal variable travel time cost function, the congestion externality cost function coincides with the marginal schedule delay cost function, and the marginal variable travel time cost function lies above the marginal schedule delay cost function. As in the bottleneck model, the time-varying toll rises monotonically from zero at the beginning of the departure interval to marginal variable trip cost at the end, but is convex in departure time, whereas it is linear in departure time in the bottleneck model. Imposition of the optimal time-varying toll alters the departure function, as in the bottleneck model, but unlike the bottleneck model also alters the arrival function.

The single-entry corridor model provides a more realistic treatment of congestion than the bottleneck model, but this comes at the cost of increased complexity and/or the need to resort to numerical solution. Which model is preferable depends on context. The bottleneck model has the advantage that its simplicity admits numerous analytical extensions, but this same simplicity leads to some unrealistic properties that

can be misleading in policy analysis. Assuming bottleneck congestion facilitates the computation of dynamic network equilibrium but treating each link as being subject to LWR flow congestion should lead to more accurate results, at least for networks of freeways (in contrast to networks of city streets where queuing at intersections account for a large fraction of total delay).

Like the basic bottleneck model, the single-entry corridor model treats demand as being inelastic. And as with the basic bottleneck model, the extension to treat price-sensitive demand would be straightforward. Like the basic bottleneck model, the single-entry corridor model treats the desired arrival time distribution as exogenous, whereas it should be endogenous and derived from employers' profit-maximizing decisions concerning employee start time. In assuming that no vehicle can lower its trip cost by altering its departure time, both the basic bottleneck model and the single-entry corridor model assume that trip-timing decisions are based on perfect information.

Despite their differences, the single-entry corridor model is far more similar to the bottleneck model than either is to the static model of congestion. In the single-entry corridor and bottleneck models with (as assumed) inelastic demand, tolling is effective through altering the timing of departures over the rush hour. In the static model of congestion, in contrast, tolling is effective through altering the number of vehicles. Empirically, since demand for commuting trips is highly inelastic, it appears that the welfare gains from altering the timing of departures are quantitatively more important than those from reducing demand. Furthermore, through ignoring the trip-timing margin of adjustment, application of the standard model has likely resulted in overstating the benefits from reducing overall traffic volume.

4.6 Concluding Remarks

4.6.1 Directions for Future Research

The basic bottleneck model has been extended in numerous ways. Some of these extensions would be straightforward to undertake for the current model, such as allowing for price-sensitive demand, determining optimal capacity, treating users who differ in unit travel time and schedule delay costs, and considering routes in parallel. Others would not be straightforward. The first is to treat late, as well as early, arrival. The difficulty here derives from the analytical complexity of dealing with the discontinuity in the departure rate at the boundary between early and late arrivals. The second is to treat a road of non-uniform width. The third is to treat merges. Merges occur along a traffic corridor with more than one entry point, and also on networks in which two or more directional links meet at a common node. Once the non-uniform road and merge problems have been solved, the stage will be set to tackle the difficult problems of determining the SO and UO with LWR flow congestion on a general corridor and a general network. The fourth is to treat a capacity constraint at the exit point, which may occur when the CBD is congested.

An undesirable property of the current model is that travel is not in the congested flow régime in the UO. But congested travel is ubiquitous. We strongly suspect that congested travel will emerge in the UO but not in the SO once the model is extended to treat either non-uniform road width or merges.

The model assumes that commuters know perfectly how congestion evolves over the rush hour. All congestion is recurrent, and all commuters recognize that it is recurrent. But in fact much congestion is non-recurrent due to variable weather conditions, traffic accidents, and road construction, about which commuters are imperfectly

informed. Some work has been done on non-recurrent congestion in the context of the bottleneck model. DePalma is currently working on extending this paper's model to treat randomly occurring incidents along the road. How they will affect the SO and UO will depend on how well informed commuters are of their occurrence. However well informed they are, in both the SO and UO, incidents will lead to congested travel. Finally, we note that since congested travel does not occur in either the SO and UO solutions to our model, our results apply with a piecewise quadratic flow-density relationship (Lu et al., 2008).

4.6.2 Conclusion

Because of its simplicity, the bottleneck model has been widely employed in theoretical analyses of rush-hour traffic congestion. But this simplicity is attained at the cost of providing an unrealistic treatment of the congestion technology. One wonders which of the bottleneck model's properties are robust, and which derive from the simplicity of the assumed congestion technology. Newell took a step towards remedying this deficiency by replacing the bottleneck with a single-entry corridor of uniform width that is subject to LWR flow congestion, for which the bottleneck is a limiting case. His paper has been rather overlooked by the literature, probably because of its density. This paper considered a special case of Newell's model in which local velocity is a negative linear function of local density, and all commuters have a common desired arrival time at the central business district. These simplifying assumptions permitted a complete closed-form solution for the social optimum and a quasi-analytical solution for the user optimum. Providing detailed derivations and exploring the model's economic properties added insight into rush-hour traffic dynamics in the with this form of congestion, and into how the dynamics differ from those of the bottleneck model.

References

Arnott, R. and de Palma, A. and Lindsey, R. (1990). Economics of a bottleneck. *Journal of Urban Economics*, 27(1), 111–130.

Arnott, R. and DePalma, E. (2011). The corridor problem: preliminary results on the no-toll equilibrium. *Transportation Research Part B*, 45, 743–768.

Chu, X. (1992). Endogeneous trip scheduling: a comparison of the Vickrey approach and the Henderson approach. *Journal of Urban Economics*, 37, 324–343.

Daganzo, C. (1995). A finite difference approximation of the kinematic wave model of traffic flow. *Transportation Research Part B*, 29(4), 261–276.

Daganzo, C. (1997). Fundamentals of transportation and traffic operations. Elsevier Science, New York, NY

Henderson, J.V. (1977). Economic theory and the cities. Academic Press, New York

Hu, D. (2011). Three essays on urban and transport economics. P.h.D. dissertation, Department of Economics, University of Wisconsin-Madison. Available through UMI (ProQuest).

Lago, A. and Daganzo, C. (2007). Spillovers, merging traffic and the morning commute. *Transportation Research Part B*, 41(6), 670–683.

LeVeque, R.J. (1992). Numerical methods for conservation laws. Birkhäuser Verlag, Basel, Switzerland

Lighthill, M.H. and Whitham, G.B. (1955). On kinematic waves II: a theory of traffic flow on long crowded roads. *Proc. Roy. Soc. (Lond.)*, 229(1178), 317–345.

Lu, Y., Wong, S.C., Zhang, M., Shu, C.-W. and Chen, W. (2008). Explicit construction of entry solutions for the Lighthill-Whitham-Richards traffic flow model with a piecewise quadratic flow-density relationship. *Transportation Research Part B*, 42, 355–372.

Mun, S. (1999). Peak-load pricing of a bottleneck with traffic jam. *Journal of Urban Economics*, 46, 323–349.

Newell, G.F. (1988). Traffic flow for the morning commute. *Transportation Science*, 22(1), 47–58.

Richards, P.I. (1956). Shock waves on the highways. *Operations Research*, 4(1), 42–51.

Small, K. (1982). The scheduling of consumer activities: Work trips. *American Economic Review*, 72(3), 467–479.

Tian, Q., Yang, H. and Huang, H. (2010). Novel travel cost functions based on morning peak commuting equilibrium. *Operations Research Letters*, 38, 195–200.

Vickrey, W. (1969). Congestion theory and transport investment. *The American Economic Review*, 59(2), 251–260.

Wong, S.C. and Wong, G.C.K. (2002). An analytical shock-fitting algorithm for LWR kinematic wave model embedded with linear speed-density relationship. *Transportation Research Part B*, 36(8), 683–706.

Notational Glossary

SO	Social Optimum
UO	User Optimum
CBD	Central Business District
x, t	Space and time coordinates
l	Spatial location of the CBD
$a(t)$, $A(t)$	Corridor inflow and cumulative inflow rates
$a_R(t)$, $A_R(t)$	Road inflow and cumulative inflow rates
$q(t)$, $Q(t)$	Corridor outflow and cumulative outflow rates (at $x = l$)
N	Size of population
t_f, t_R	Corridor and road final departure times
$\bar{t}$	Time of final arrival at the CBD
$\tau(t)$	Travel time of a departure at time t
α_1, α_2	Per unit population costs of travel time and schedule delay
$C(t)$	(Per unit population) Trip cost for a departure at time t
TTC	Total Trip Cost (aggregate sum)
$k(x,t)$, $v(x,t)$, $q(x,t)$	Density, velocity and flow rate at spacetime point (x,t)
v_0	Free-flow velocity
k_j	Jam density
q_m	Capacity flow
k_m	Density at which capacity flow is reached
w	Reciprocal of wave velocity normalized by v_0, $w = \frac{v_0}{q'(k)}$
k_l, k_r	Densities to the left and right of a shock wave path
q_l, q_r	Flow rates to the left and right of a shock wave path

q_c	Constant inflow rate
w_c	Characteristic slope for a flow rate of q_c
v_c	Velocity for a flow rate of q_c
t_c	Departure time for arrival at time w_c
t_d, t_a	Departure and arrival times, respectively
(x_0, t_0)	Intersection point of trajectory and wave boundary
A	Arbitrary constant of integration
TSD	Total Schedule Delay (aggregate sum)
TTT	Total Travel Time (aggregate sum)
$q^*(w)$	Nondimensional outflow rate of maximal growth
$Q^*(w)$	Nondimensional cumulative outflow of maximal growth
t_0'	Time of the first arrival in the SO
w_0	Characteristic slope at the CBD at time t_0'
TT	Trip-Timing Condition
σ	Shorthand notation for $\frac{1}{1-\alpha_2}$ in the UO
$g\left(a(t)\right)$	Shorthand notation for $\frac{1}{\sqrt{1-a(t)}} - 1$ in the UO
a_f	Inflow rate at time t_f
t_Q	Time at which a queue develops in the UO
N_c	Critical population value such that a queue develops
T	Total (Section 5)
A	Average (Section 5)
M	Marginal (Section 5)
F	Fixed (Section 5)
V	Variable (Section 5)

C	Cost (Section 5)
E	Elasticity (Section 5)
$\tau(t_d)$	Toll at departure time t_d (Section 5)
$p(N)$	Trip price in excess of free-flow travel time cost (Section 5)

Chapter 5

Stochastic Incidents in the Corridor Problem

Abstract

This paper introduces incidents (i.e., unusual increases in congestion) into the Single-Entry Corridor Problem, a model of morning traffic congestion analyzed in DePalma and Arnott, working paper. We begin by modifying an online, change-point detection algorithm (Lambert and Liu, 2006) to detect incidents in traffic flow data, and apply the resulting incident detection algorithm to weekday morning traffic data along a section of Interstate 5-North from the U.S.-Mexico border to downtown San Diego during 2010. We propose to model the probability of incident occurrence as a logistic model with random effects, and we calibrate our proposed model to the incidents detected in the traffic data set using fitting techniques from the theory of Hierarchical Generalized Linear Models (Lee et al., 2006). We adapt the Cell Transmission Model (Daganzo, 1995) to accomodate incidents, and implement our fitted incident rate model into a simulation study to investigate how stochastic incidents distort equilibrium solu-

tions to the Single-Entry Corridor Problem.

5.1 Introduction

In recent years considerable theoretical work has been done on the dynamics of rush-hour traffic congestion. Most of this work has applied the basic bottleneck model (Arnott and Lindsey, 1990), in which congestion takes the form of a queue behind a single bottleneck of fixed flow capacity. Recent research on the the Corridor Problem (Arnott and DePalma, 2011, DePalma and Arnott, working paper) has replaced bottleneck congestion with LWR (Lighthill and Whitham, 1955 and Richards, 1956) flow congestion, which combines the equation of continuity (conservation of mass for a fluid) with an assumed technological relationship between local velocity and local traffic density, and covers bottleneck congestion as a limiting case. Newell, 1988 considered a model of the morning commute in which a fixed number of identical commuters must travel along a road of constant width subject to LWR flow congestion, from a common origin to a common destination, and in which trip costs are a linear function of travel time and schedule delay. He allowed for a general distribution of desired arrival times and a general technological relationship between local velocity and local density, and precluded late arrivals by assumption. He obtained qualitative properties of both the social optimum (SO) in which total trip cost is minimized, and the user optimum (UO) in which no commuter can reduce their trip cost by altering their departure time. DePalma and Arnott, working paper, provide a detailed analysis of a special case of Newell's model, assuming that commuters have a common desired arrival time and that local velocity is a negative linear function of local density (Greenshields' Relation), which they termed the "Single-Entry Corridor Problem". They determined explicit, analytic

expressions for the UO and SO in the case when late arrivals are not permitted.

Although useful, the research cited above assumes deterministic models of traffic flow and does not allow for randomly occurring traffic accidents or other randomly occurring roadway obstacles which create unusual increases in traffic congestion levels. The random occurrence of unusual traffic congestion leads to unreliable travel time assessments, and recent research suggests that commuters place high value on travel time reliability (Small et al., 2005). To better understand how the random occurrence of congestion affects commuting patterns, in this paper we seek to incorporate stochastic "incidents" (i.e., unusual traffic congestion) into the Single-Entry Corridor Problem.

The usual definition of an "incident" is any accident, vehicle breakdown, roadway obstacle, etc. which leads to an unusal increase in traffic congestion. In this paper, however, we observe traffic flow data at discrete locations along a roadway and do not attempt to correlate unusual increases in congestion at different locations to a single event. Furthermore, the traffic flow data we observe does not provide any indication as to the source or cause of increased congestion levels. Therefore, for the remainder of this paper **we define an "incident" to be any unusual increase in traffic congestion at a particular roadway location, and NOT the event which causes the increase in congestion.** Under this definition, a specific accident or roadway obstruction may result in various "incidents" being observed at different roadway locations.

In this paper we propose a stochastic incident rate model which models the probability of the occurrence of an incident at a particular location based upon the existing traffic conditions at that location and also downstream traffic locations. Using real traffic data we apply an automatic incident detection algorithm to detect the occurrence of incidents, and then use the detected incidents to calibrate the proposed incident rate model. We also propose and fit simple models for the duration of an incident and the

capacity reduction caused by an incident. Lastly, we implement these fitted models into a simulation study to investigate how the introduction of stochastic incidents alters the UO and SO equilibrium solutions to the Single-Entry Corridor Problem.

In Section 2 we discuss the traffic data set we collected to use in fitting the incident rate model. In Section 3 we provide an overview of automatic incident detection algorithms, and outline the online, change-point detection algorithm we modified and subsequently implemented to detect incidents in our data set. In Section 4 we propose and fit an incident rate model, fitted using techniques from the theory of Hierarchical Generalized Linear Models (HGLM). The theory of HGLM's is relatively new and potentially controversial (Meng, 2009), and in Section 4 we also provide an overview of this theory and its accompanying fitting algorithms. In Section 5 we discuss how to modify the Cell Transmission Model (Daganzo, 1995) to numerically obtain traffic flow dynamics in the Single-Entry Corridor Problem model with the occurrence of incidents, and we present the results of our simulation study which investigates how the UO and SO solutions are altered with the introduction of stochastic incidents. Section 6 discusses the limitations of our simulation study and concludes.

5.2 Data Collection

5.2.1 PeMS Overview

The Freeway Performance Measurement System (PeMS) is a software and database tool designed to store and analyze road traffic conditions, and is a joint effort by Caltrans, the University of California, Berkeley, and the Partnership for Advanced Technology on highways. The PeMS database logs data from California freeway detectors, incident-related data from the California Highway Patrol (CHP), and weather

data. PeMS collects raw detector data in real-time and stores, processes and reports this data.

A detector is a vehicle magnetic-sensor imbedded in the pavement and positioned within a travel lane such that vehicles traveling over the sensor are detected. A majority of the loop detectors in the State of California are single loop detectors, which record the following:

- **Flow:** The number of vehicles that cross the detector during a given time period.
- **Occupancy:** The percentage of time that the detector is occupied (in the "on" state) during a given time period.

Speed cannot be measured directly from single loop detector data and has to be estimated using the "G-factor", which is a combination of the average length of the vehicles in the traffic stream and the tuning of the loop detector itself.

Raw detector data in PeMS is handled by special algorithms that detect and impute missing data or data errors. PeMS receives raw, 30-second data from each loop detector, which is aggregated to 5-minute samples per lane. The 5-minute data is then imputed to remove any detected data errors (Urban Crossroads, Inc., 2006).

Caltrans maintains a PeMS database at http://pems.dot.ca.gov, which stores real-time data, tools for data analysis, and a record of CHP incident reports. A separate PeMS database for academic research at http://pemscs.eecs.berkeley.edu is maintained by Berkeley Transportation Systems, Inc. The PeMS academic research database stores real-time detector data which may be accessed through an Oracle SQL database.

5.2.2 Freeway Selection and the Observance of Congestion in Loop-Detector Data

In this work, the goal of gathering PeMS data is to calibrate an incident rate model to be used in a simulation study of the effects of stochastic incidents in the Corridor Problem. With that goal in mind, we set out to locate a freeway whose morning commute pattern characteristics most closely matched the assumptions of the Corridor Problem, namely, all commuters traveling towards a city center on a road of uniform width with the same desired work start-time. Furthermore, we looked for freeways with traffic patterns which were relatively uniform across days, so that the occurrence of an incident (perceived as a spike in traffic congestion) would be clearly discerned. Small highways with low flow rates had very few (if any) clearly discernible congestion spikes, so that we ended up restricting our search to larger freeways. Most of the large urban areas are polycentric with multiple business and commerce districts, and the accompanying freeway diverges and merges result in large fluctuations in congestion which can not be clearly discerned as the result of an incident.

After much searching, we finally selected a 15-mile stretch of Interstate 5 Northbound (I5-N) freeway from the U.S.Mexico border to downtown San Diego. Along this stretch of freeway there are no major freeway diverges, and an inspection of morning commute patterns indicated that commuters generally travel to the downtown San Diego area or to areas just north of downtown San Diego. Along this stretch of freeway there are 23 loop detector stations which collect PeMS data, whose characteristics are summarized in Table 5.1 and whose locations are graphed in Figure 5.1.

We gathered loop detector data for all weekday mornings in the year 2010, for a total of 260 weekdays. For each weekday we gathered loop detector data from 4:00am

Detector	Name	Mainline VDS #	# Lanes	Absolute Postmile
1	N/O CMNO DE LA PLAZA	1114091	6	.057
2	5 NB @ S. Ysidro	1118333	4	.146
3	N/O VIA DE SAN YSIDR	1114709	4	1.412
4	5 NB @ Dairy Mart	1118348	4	2.068
5	N/O DAIRY MART RD	1114720	4	2.56
6	5 NB N/O 905	1118352	4	3.09
7	5 NB @ Coronado Ave	1118743	4	4.004
8	5 NB @ Palm Ave	1118379	4	4.714
9	S/O MAIN ST	1114734	5	5.23
10	5 NB @ J St	1118421	5	7.201
11	5 NB N/O H St	1118735	4	7.82
12	5 NB @ E St	1118450	5	8.524
13	.73 M S/O CIVIC CTR	1114190	5	9.897
14	5 NB Civic Center	1118458	4	10.795
15	5 NB N/O 8th St	1118479	4	11.179
16	.27 M N/O DIVISION	1114205	6	11.847
17	5 NB S/O 15 Radar	1118827	4	12.114
18	.26 M N/O SR-15	1114211	4	12.747
19	NB 5 @ National Ave.	1117762	4	13.022
20	.23 M N/O 28TH ST	1114219	6	13.371
21	NB 5 @ Ceasar Chavez	1117782	5	14.416
22	N/O COMMERCIAL ST	1114100	5	14.676
23	NB 5 @ 94	1117796	4	14.856

Table 5.1: Table of characteristics of the 23 detectors along the 15-mile stretch of I5-N from the U.S.-Mexico border to downtown San Diego. The "Absolute Postmile" marker is the distance in freeway miles from the U.S.-Mexico border to the detector location.

to 10:00am, for a total of 72 entries per loop detector per day (loop detector data is aggregated in 5-minute intervals). Thus, the initial size of our loop detector data set contained $260 \cdot 72 \cdot 23 = 430,560$ entries, with each entry containing the observed flow value, average occupancy rate, and other descriptive characteristics.

At a single loop detector location, we define an incident to be any unusual increase in the congestion level as compared to the normal congestion levels at that location observed in the recent past. Unusual increases in congestion levels are indicated by unusual decreases in flow values and unusual increases in occupancy rates.

When an incident occurs and we observe an unusual increase in the congestion

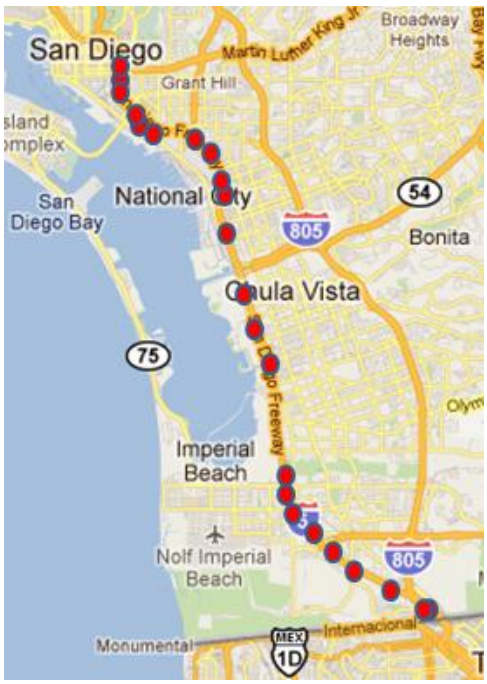

Figure 5.1: Visual graph indicating the locations of the 23 detectors along the 15-mile stretch of I5-N from the U.S.-Mexico border to downtown San Diego.

level at a particular location, we anticipate observing unusual increases in the congestion levels at more upstream locations as time progresses. From a preliminary inspection of the detector data we found that occupancy rates more clearly indicate the occurrence of an incident than flow values. To see this, consider the example provided in Figures 5.2 and 5.3, where we provide a space-time graph of flow values and occupancy rates on two adjacent days. The first day, April 27, 2010, displays normal traffic patterns, whereas the second day, April 28, 2010, displays congested traffic patterns indicating the occurrence of incidents. The flow values graphed in Figure 5.2 significantly drop across the two days, but the drop is not as easily observed as the corresponding increase in occupancy rates graphed in Figure 5.3.

To understand why occupancy rates are more sensitive to congestion than flow values, in Figure 5.4 we provide a scatterplot of the observed occupancy rates and flow

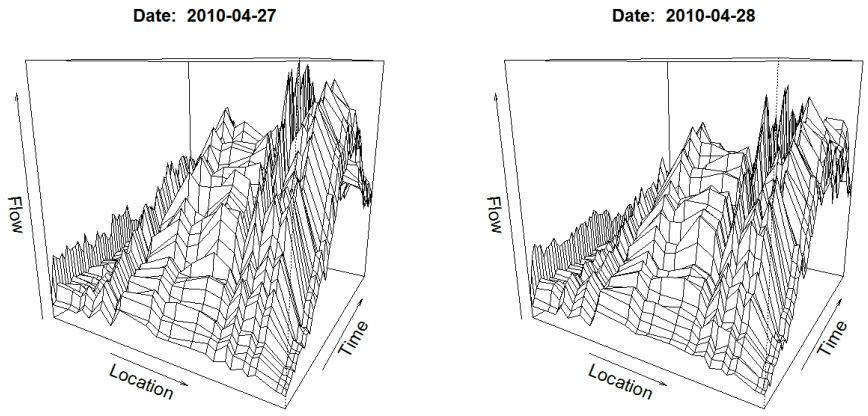

Figure 5.2: Space-time graph of flow values across two adjadcent days. The first day displays usual traffic patterns, whereas the second day displays decreased flow values corresponding to increased congestion levels. The drop in flow values across the two days is not as easily discerned as the corresponding increase in occupancy rates observed in Figure 5.3.

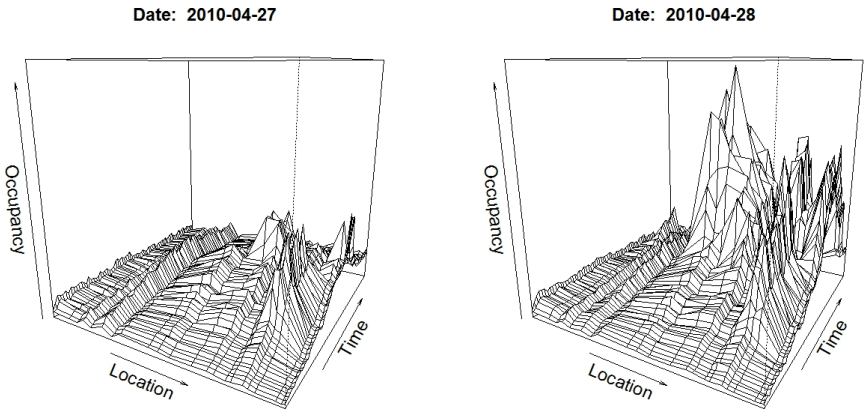

Figure 5.3: Space-time graph of occupancy rates across two adjacent days. The first day displays usual traffic patterns, whereas the second day displays increased occupancy rates corresponding to increased congestion levels. The increase in occupancy rates across the two days is more easily discerned than the corresponding decrease in flow values observed in Figure 5.2.

values for a single detector for the entire 2010 data set. Flow-occupancy (or flow-density) graphs are well studied in transportation research, and can generally be decomposed into two regimes, a free-flow regime and a congested regime. In the free-flow regime flow

values increase linearly with occupancy, which we observe in Figure 5.4. In the congested regime flow generally decreases with occupancy with a wide variance, as we also observe. In Figure 5.4 we observe that many points in the congested regime have large occupancy rates relative to the occupancy rates in the free-flow regime, whereas the flow values in the congested regime are often not much smaller than the maximal flow values observed in the free-flow regime. Therefore, in the following section, to detect incidents we seek to detect unusual increases in occupancy rates instead of unusual decreases in flow values.

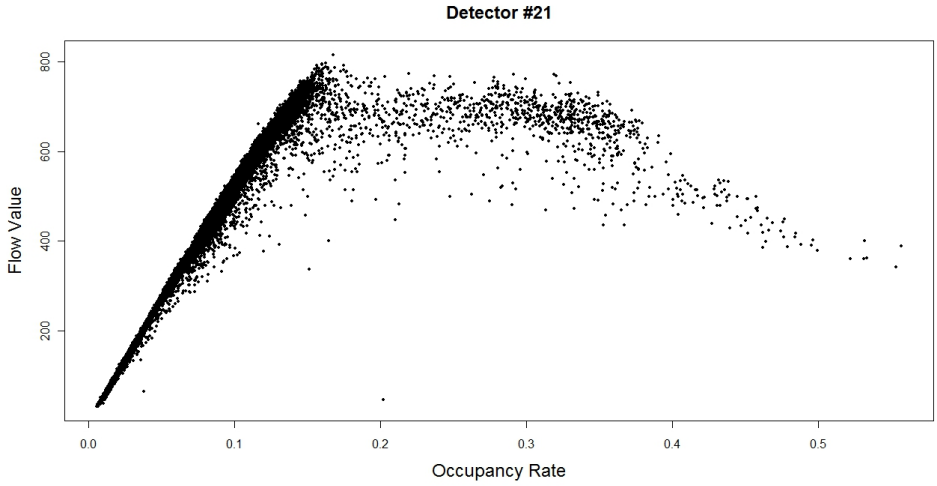

Figure 5.4: Flow-occupancy diagram for all observed data points at a single loop-detector, detector #21. In the free-flow regime flow increases linearly with occupancy, whereas in the congested regime flow decreases with occupancy with wide variance. Relative to the free-flow regime, in the congested regime occupancy rates increase significantly, whereas many flow values exhibit only a small decrease.

5.3 Automatic Incident Detection

Numerous research studies have been performed on incident detection algorithms, with a recent literature review in Parkany, 2005. Most incident detection algorithms can be grouped into four categories (Byrdia et al., 2005):

- comparative algorithms,

- statistical algorithms,
- time-series algorithms, and
- modeling algorithms.

Comparative algorithms are the simplest of all algorithms, and generally compare speed, volume and occupancy measurements from a single detector station against thresholds that define when incident conditions are likely. These comparisons may compare measurements from the same detector across time, or may compare measurements across space from adjacent detectors. The well-known California Algorithms (Payne and Tignor, 1978) fall into this class of algorithms.

Statistical algorithms use statistical techniques to determine whether detector data differ significantly from historical data (PB Farradyne, Inc., 2010). Algorithms in this class use historical, non-incident traffic data to develop mean and expected confidence interval ranges for traffic conditions, and real-time detector data is then analyzed to determine if they fall within these ranges.

Time-series algorithms compare short-term predictions of traffic conditions to real-time traffic conditions. These algorithms generally track what is happening at a detector over time, looking for abnormal changes in the real-time traffic data. The ARIMA Algorithm (Ahmed and Cook, 1982) and the Exponential Smoothing Algorithm (Cook and Cleveland, 1974) fall into this class of algorithms.

Modeling algorithms use traffic flow theories to model expected traffic conditions based on the current detector data. These algorithms model traffic conditions under incident conditions and then examine the detector data to see if predicted values are similar. Modeling algorithms include the well-known McMaster Algorithm (Persuad et al., 1990), and also include modeling approaches using fuzzy logic and neu-

ral networks (Chang and Wang, 1994, Stephanedes and Liu, 1995).

The performance of incident detection algorithms is commonly measured using:

- **Detection Rate:** The number of incident alarms that the algorithm issues relative to the total number of capacity-reducing incidents occurring in the system.
- **False Alarm Rate:** The number of incident alarms that the algorithm incorrectly issues relative to the total number of incident alarms that the algorithm issues.
- **Detection Time:** The average elapsed time from the occurrence of an incident in the system and the detection of the incident by the detection algorithm.

  (Carvell et al., 1997)

Generally, incident detection algorithms seek to have a low detection time while balancing a high detection rate against a high false alarm rate.

We are very fortunate to have received guidance from Dr. Karl Petty, the President of Berkeley Transportation Systems, Inc., who completed a dissertation in Electrical Engineering on freeway incident detection (Petty, 1997). Dr. Petty has visited traffic management centers throughout the country, and found that automatic incident detection algorithms are not reliably used due to their high false alarm rate. In all instances traffic managers rely upon video monitors, CHP feeds and roving tow trucks to manage incidents.

5.3.1 Adaptive Thresholding Algorithm

Our initial attempts to apply incident detection algorithms to the PeMS data were not successful at detecting visually obvious spikes in the occupancy rates. There are several challenges in detecting sudden spikes in our PeMS data: we do not have a large database of historical data which can be reliably used, there are cyclical and

acyclical patterns in the data which require the algorithm to be updated as data is collected, and our data occur in 5-minute aggregated intervals, resulting in relatively small data sets to be used for incident detection.

Recent work (Montes De Oca et al., 2010) has adapted classic, cumulative sum (CUSUM) change-point detection methods to nonstationary time sequences of network data streams which exhibit acyclical patterns varying over time. Based on a recommendation from Yingzhuo Fu, a fellow colleague and graduate student at UC Riverside, we pursued a recently developed adaptive thresholding (AT) algorithm for online, change-point detection (Lambert and Liu, 2006). With several modifications and additional controls, the implementation of the AT algorithm yielded the best results for detecting occupancy rate spikes in our PeMS data, and in the remainder of this section we outline the AT algorithm and the additional controls and modifications we supplemented.

5.3.1.1 Outline of Adaptive Thresholding Algorithm

The AT algorithm outlined below was originally applied to count data for the number of network users on a server, and provided an online, change-point detection algorithm which would signal an alarm if the number of users demonstrated an unusual decrease, indicating the failure of a server (Lambert and Liu, 2006). For our application we adapt the AT algorithm to detect unusual increases in the occupancy rate at a single loop-detector. The AT algorithm signals a data point as an "alarm point" if the occupancy rate is unusually high, and we consider an "incident" to be a sequence of time-consecutive alarm points. Thus, the AT algorithm enables us to identify strings of alarm points (defined to be an incident), and after applying the AT algorithm to our data set we will construct an incident rate model to estimate the probability of the occurrence of an incident. We also estimate the duration of an incident and the average

capacity reduction due to an incident by investigating these strings of alarm points, as described in the following sections.

At time t, let x_t denote the incoming value of the response, which for our application is the occupancy rate at a single detector. The AT algorithm constructs a reference distribution for x_t, whose first two moments are assumed to vary sufficiently smoothly to be interpolated from values on a coarse time grid. The grid values capture the cyclical patterns in the data, and for our application correspond to the six hourly values per day (4:00am to 10:00am). The grid values are updated by exponentially weighted moving averaging (EWMA), and the combination of interpolation from grid values and EWMA tracks both cyclical time patterns and long-term trends.

Each incoming response value is first standardized using the reference distribution, and then a severity metric, S_t, is defined as an EWMA of the standardized responses. One very extreme response can push S_t beyond a threshold, or a sequence of many less extreme but still unusual responses can push S_t beyond a threshold. The AT algorithm consists of four basic steps that are applied to each incoming response x_t at time t:

- Interpolate the stored grid values to obtain the reference distribution in effect at time t.
- Standardize x_t by first computing its p-value under its reference distribution, p_t, and then computing the normal standardized score $Z_t = \Phi^{-1}(p_t)$, where Φ is the cumulative distribution function of the standard normal distribution.
- Threshold the updated severity metric, $S_t = (1-w)S_{t-1} + wZ_t$, against a constant threshold.
- Update the stored grid values with the response x_t, or with a random draw from

the reference distribution if x_t is missing, or with a random draw from the tail of the reference distribution if x_t is an outlier.

In the first step, we assume a reference distribution that is characterized by its mean and variance (for our application we chose a normal distribution). For each hour, h, denote the stored mean and variance grid values as U_h, V_h, respectively. At each time t falling within hour h, the mean and variance of the reference disribution, $\mu_{h,t}$ and $\sigma^2_{h,t}$, are obtained using quadratic interpolation from the stored grid values (U_{h-1}, V_{h-1}), (U_h, V_h) and (U_{h+1}, V_{h+1}).

In the second step, since we choose the reference distribution to be a normal distribution, each incoming response, x_t, is standardized to a standard normal distribution,

$$Z_t = \frac{x_t - \mu_{h,t}}{\sigma_{h,t}}.$$

In the third step, we update the severity metric, S_t, using an EWMA with weight w. If Z_t is approximately distributed as $N(0,1)$, then S_t is approximately distributed as $N(0, \sigma^2_w)$, where $\sigma^2_w = \frac{2}{2-w}$. An alarm is signaled whenever $S_t > L\sigma_w$, and the problem is to choose L and w to balance the detection rate against the false alarm rate.

In the fourth step, if an alarm is signaled or if x_t is an outlier from its reference distribution with p-value greater than 0.9999 or less than 0.0001, then we define X_t to be a random draw from the upper or lower 0.01 tail of the current reference distribution, respectively. Otherwise, we define X_t to be x_t. The mean and variance of the reference

distribution are updated using an EWMA with weight parameter, w_c, as

$$\mu_{h,t}^{\text{new}} = (1 - w_c)\mu_{h,t} + w_c X_t$$

$$\left(\sigma_{h,t}^2\right)^{\text{new}} = (1 - w_c)\sigma_{h,t}^2 + w_c \left(X_t - \mu_{h,t}\right)\left(X_t - \mu_{h,t}^{\text{new}}\right),$$

so that the stored grid values are updated as

$$U_h^{\text{new}} = \frac{MU_h + \mu_{h,t}^{\text{new}}}{M}$$

$$V_h^{\text{new}} = \frac{MV_h + \left(\sigma_{h,t}^2\right)^{\text{new}}}{M}.$$

M is the number of time-steps in each hour, which for our application is $M = 12$ (since each time-step corresponds to a 5-minute interval).

To initiate the algorithm we must provide hourly stored grid values, which we obtain from the sample means and sample variances of the first day of the data set.

The heart of the AT algorithm is based on online estimates of means and variances that capture cyclical patterns which may change over time, and the choice of reference distribution is general and may accomodate count data (e.g., negative binomial distribution).

5.3.1.2 Additional Controls and Threshold Values

We spent a great deal of time adjusting the threshold and EWMA weights to achieve the best success in detecting spikes in occupancy rates. The optimal use of the algorithm was facilitated by adding in several control features which were not discussed in Lambert and Liu, 2006, described below.

At each time-step, the mean and variance of the reference distribution are

obtained using quadratic interpolation from the stored grid values. As a result, it is possible to obtain negative values of the variance even if the stored variance grid values are all positive. To alleviate this problem, if at any time-step we obtain a negative value for the updated variance value, then we ignore this updated value and use the previous variance value without updating.

For each detector, over the entire 2010 data set, we defined a minimum cutoff value for the occupancy rate to be the lower bound of the highest 0.5% occupancy rate values for that detector. When a first alarm is signaled, indicating the occurrence of an incident, we track the occupancy rates for the duration of the incident, i.e., from the first alarm point until the last consecutive alarm point. Throughout the duration of the incident, we check to see if the occupancy rate exceeds the minimum cutoff value. If the occupancy rate never exceeds the minimum cutoff value during the course of the incident, then we remove the alarm signals for all of the points occuring during that incident, i.e., we do not consider any incident to have occurred. Note that this extra control defeats the purpose of an online incident detection algorithm, since we set the minimum cutoff value using the entire data set for that detector over the year. However, in this paper we are not interested the best online detection algorithm, but rather the best detection algorithm to be used for our data set.

Finally, when an incident occurs we track the flow values for the duration of the incident, i.e., from the first alarm point until the last consecutive alarm point, and we calculate the average flow value over the duration of the incident. If the average flow value during the duration of the incident is not less than the flow value at the start of the incident, then we remove the alarm signal for all of the points occuring during that incident, i.e., we do not consider any incident to have occurred.

Using these additional controls we found that the threshold values and EWMA

weights which yield the best results in detecting occupancy rate spikes are

$$w_c = 0.75 \quad \text{and} \quad (w, L) = (0.75,\, 4.0). \tag{5.1}$$

Throughout the remainder of this paper, the term "incident detection algorithm" refers to the AT algorithm with the above additional controls implemented and the above choice of threshold values and EWMA weights.

5.3.2 Example: Comparison with CHP Incident Reports

We applied the incident detection algorithm to each of the 23 detectors to determine the occurrence of spikes in occupancy rates, i.e., to determine the occurrence of incidents. In Figure 5.5 we show a sample graph of occupancy rates over a single day at a single detector location, with the dark colored points indicating alarm points signaled by the incident detection algorithm. The two consecutive strings of alarm points correspond to two separate incidents indicated at that detector by the incident detection algorithm, and we added in the open-circle points to indicate the points just before and just after the incident, which we use to measure the duration of the incident. The horizontal line in this graph corresponds to the lower bound of the upper 0.5% occupancy rate values for that detector, and for an incident to be detected the consecutive string of alarm points must cross that line during the course of the incident. It appears that the first incident in this example is a false alarm since its shape is similar to nearby occupancy rate values which are not indicated as incidents, whereas the second incident in this example demonstrates an unusual increase in occupancy rates.

On the suggestion of Professor Matthew Barth, we compared the incidents signaled by the incident detection algorithm to the CHP feed of incident reports avail-

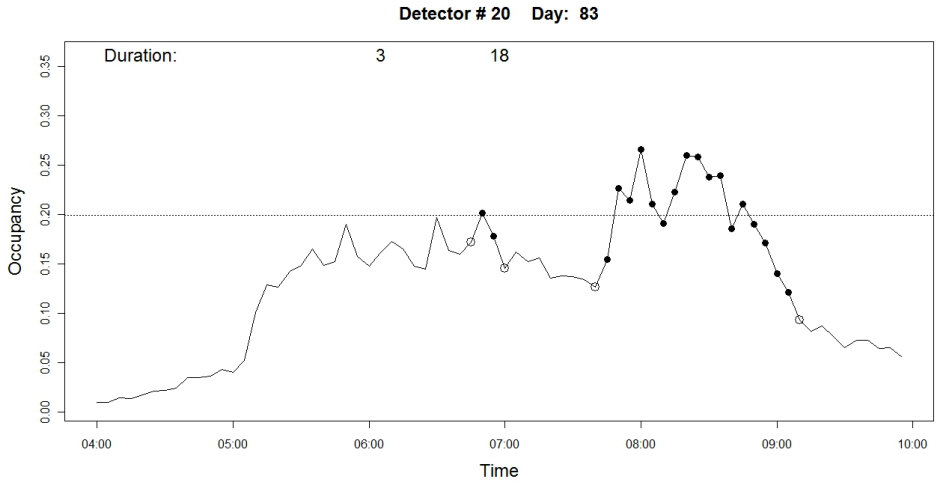

Figure 5.5: Occupancy rates for detector #20 on day 83. On this day the incident detection algorithm signaled two incidents, the first of duration two time-steps (i.e., 10 minutes), and the second of duration 18 time-steps (i.e., 90 minutes). In this graph the black points indicate signaled alarm points, and the open circles indicate the beginning and end of the incident. The horizontal line is the minimum cutoff occupancy rate value, corresponding to the lower bound of the upper 0.5% occupancy rates over the entire year (for this detector).

able through http://pems.dot.ca.gov. In Table 5.2 we list all incidents signaled during January 2010. Using the tools available on http://pems.dot.ca.gov we compiled a listing of all incidents reported through the CHP feed during January 2010, presented in Table 5.3, where we exclude any incidents of duration 0 minutes.

Our incident detection algorithm detected the major traffic collision occurring on January 28, 2010, indicated in Table 5.2 as a series of incidents which move upstream as time progressed. However, the incident detection algorithm did not detect the smaller traffic incidents recorded by the CHP feed. We believe that this is due to the large time resolution of the traffic data, which is aggregated to 5-minute intervals. Furthermore, our incident detection algorithm indicated several incidents which were not recorded in the CHP incident feed. The cause of these spikes in congestion is unknown.

Date	Detector	Start Time	Duration (min)
2010-01-18	12	6:30am	35
2010-01-21	22	7:35am	65
2010-01-26	15	7:35am	15
2010-01-28**	8	7:20am	10
2010-01-28**	11	7:25am	50
2010-01-28**	12	7:15am	70
2010-01-28**	13	7:35am	25
2010-01-28**	14	6:30am	40
2010-01-28**	17	6:20am	50

Table 5.2: Table of incidents detected by the incident detection algorithm during January, 2010. The **'ed incidents detected on 2010-01-28 appear to have come from a single event occurring before 6:20am and downstream of detector 17, as the pattern of incidents moves upstream as time progresses. This single event corresponds to an incident occurring in the CHP feed of incident records appearing in Table 5.3.

Date	Upstream Detector	Time Start	Duration (min)	Description
1/4/2010	13	6:57am	14	Traffic Collision - No Details
1/6/2010	6	7:53am	1	Traffic Hazard
1/6/2010	19	8:38am	2	Defective Traffic Signal
1/8/2010	15	8:15am	1	Traffic Collision - No Injuries
1/11/2010	8	7:05am	11	Traffic Collision - No Details
1/14/2010	20	5:01am	10	Traffic Hazard - Vehicle
1/14/2010	19	6:13am	53	Traffic Collision - No Details
1/15/2010	19	6:18am	48	Hit and Run - No Injuries
1/15/2010	16	9:06am	8	Traffic Hazard
1/18/2010	19	7:15am	80	Traffic Collision - No Details
1/19/2010	8	7:42am	17	Traffic Hazard
1/21/2010	1	6:01am	4	Traffic Hazard
1/22/2010	20	7:18am	17	Roadway Flooding
1/25/2010	9	7:24am	9	Traffic Hazard
1/25/2010	18	8:31am	4	Traffic Hazard
1/25/2010	13	9:01am	5	Traffic Hazard
1/28/2010	13	5:43am	20	Pedestrian on a Highway
1/28/2010**	19	6:03am	37	Traffic Collision - Ambulance

Table 5.3: Table of incidents for January 2010 recorded by the CHP feed, where we exclude any incidents with duration 0 minutes. The **'ed incident on 1/28/2010 corresponds to incidents detected by the incident detection algorithm in Table 5.2.

5.4 Incident Rate Model

Recurrent traffic congestion is the result of excess traffic demand relative to the existing infrastructure. Non-recurrent congestion is the result of irregular occurrences

(i.e., incidents), such as accidents, vehicular breakdowns, obstacles in the roadway, etc. Previous research indicates that non-recurrent congestion constitutes 50-75% of total congestion on U.S. urban freeways, and, furthermore, that only 11% of capacity-reducing incidents are the result of traffic accidents (Giuliano, 1989). In this work we do not seek to distinguish the cause of an incident, but are interested only in the resulting congestion caused by an incident.

Much research has been done on estimating vehicular crash-frequencies, with a recent literature review in Lord and Mannering, 2010. This body of research seeks to understand the factors which affect the likelihood of a crash, and generally seeks to model the number of crashes occurring in a geographical region over a specified time period.

In contrast, however, is an area of research seeking to develop relationships between real-time traffic conditions and the occurrence of accidents, as is done in Oh et al., 2005, Golob et al., 2004 and Golob et al., 2008. In this research area it is well documented that the most significant factor in the occurrence of a freeway traffic accident is the speed variation among different freeway lanes, and that the incident rate per vehicle miles traveled increases with occupancy rate, with the highest proportion of crashes occurring in the most heavily congested flow states. Since the Corridor Problem model contains a single freeway corridor of uniform width which does not account for multiple lanes, we do not use the occupancy rate or flow count variations among different lanes, but we instead use the aggregated flow values and occupancy counts across all lanes. Furthermore, in this paper we do not seek to model accident rates, but rather we seek to model incident rates, without distinguishing between an accident or any other event which leads to increased traffic congestion.

5.4.1 Adjusted Data Set

The first weekday in the initial data set of 260 days is Monday, January 1, 2010, and since the occupancy rates for this day were extremely low (this date is a major holiday with little commuting traffic), we eliminated this date from the data set. The application of the incident detection algorithm requires a single day to calibrate the reference distribution, as discussed in the previous section. Thus, after calibration we applied the incident detection algorithm to the reminaing 258 days in the 2010 data set, resulting in a total of 259 incidents across the 23 detectors, each of varying duration and severity. We caution the reader to recall that we define an incident to be an unusual increase in the occupancy rates at a single detector, so that under our definition the occurrence of an accident or roadway obstacle may lead to several "incidents" occurring at different detectors.

Due to the size and nature of the data set (i.e., binary responses indicating the occurrence of an incident), we restricted the data set to be more homogeneous to aid in the model selection and fitting procedures. In Figures 5.6 and 5.7 we provide histogram plots of the number of incident counts based on the detector and the weekday, respectively. From Figure 5.6 we see that the occurrence of incidents is much lower on detectors #1-9 (which we attribute to the relatively low traffic volumes at these detectors), and from Figure 5.7 we see that the occurrence of incidents is much lower on Fridays. Since we found that a more homogeneous data set greatly assisted with the model fitting procedure, we adjusted our initial data set by removing all data points for detectors #1-9 and for Fridays.

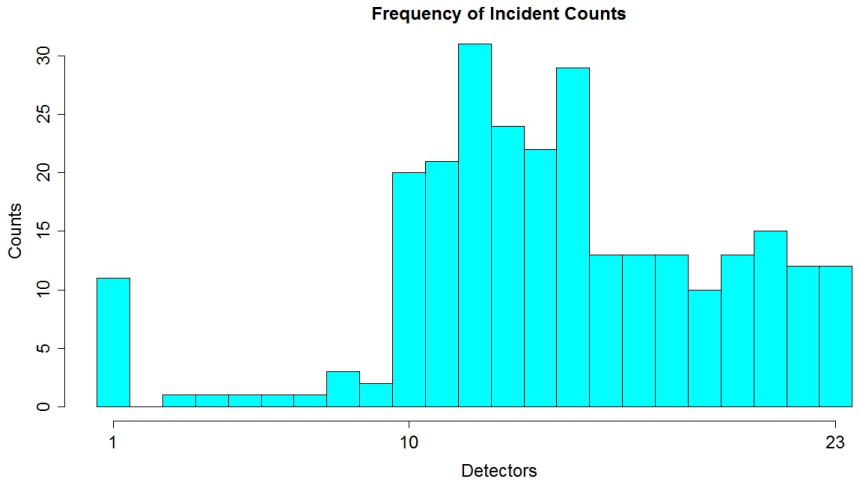

Figure 5.6: Frequency of incident counts per detector. Since the occurrence of incidents is much lower for detectors #1-9, we adjusted the initial data set by removing the data points for these detectors, keeping only data points for detectors #10-23.

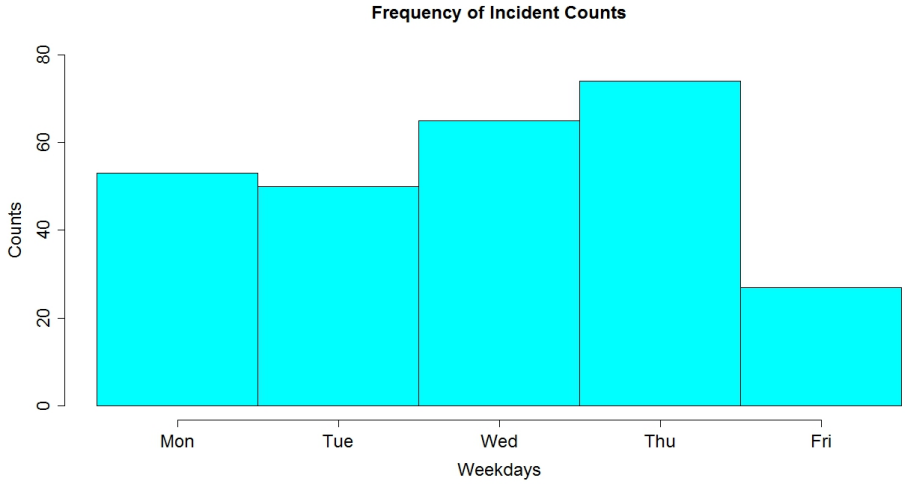

Figure 5.7: Frequency of incident counts per weekday. Since the occurrence of incidents is much lower on Fridays, we adjusted the initial data set by removing the data points for Fridays, keeping only data points for Mondays-Thursdays.

5.4.1.1 Incorporating Spatial Correlation

Since our incident detection algorithm is independently applied to the data for each detector, and since the occurrence of an incident at a single detector will generally lead to incidents at upstream detectors as time increases, we expect to find a correlation in the occurrences of incidents which exhibits directed spatial and directed temporal

behavior. We have explored many different options for incorporating this type of correlation into an incident rate model with varying degrees of success, and have finally chosen to incorporate this correlation among detectors as follows: For each data point we include a "Downstream-Incident Indicator Variable" which indicates whether an incident has occurred or has been in progress at the nearest downstream detector within the last 30 minutes. This indicator variable is treated as a fixed effect in the incident rate model described in the following section. Our decision to look 30 minutes into the past at the nearest downstream detector for the presence of an incident was based on an inspection of the data set, where we observed that spatial correlation between incidents at neighboring detectors did not appear to extend beyond 30 minutes.

We cannot assign downstream-incident indicator variable values for detector #23 (since we do not have data for any further downstream detector), and thus we removed all data points for detector #23 in our adjusted data set. Also, on each day at each detector we recorded 72 data points corresponding to the 5-minute aggregated time intervals from 4:00am to 10:00am. To provide values for the downstream-incident indicator variable we require 6 previous time values (corresponding to 30 minutes in the past), and thus we removed the first 6 data points for each detector on each day in the adjusted data set.

5.4.1.2 Summary of Adjusted Data Set

The adjusted data set contains a data point for each Monday-Thursday weekday in 2010, for detectors #10-22, and for 66 time values from 4:30am-10:00am. Each data point contains six fields: (1) the detector, (2) the day and time, (3) the occupancy rate, (4) the flow value, (5) a 0-1 Incident Indicator indicating whether or not an incident was in progress, (6) and a 0-1 Downstream Incident Indicator indicating whether

or not an incident had been in progress at the nearest downstream indicator anytime within the last 30 minutes.

Since the capacity of the roadway varies among the detector locations, we normalized the occupancy rates and flow values at each detector, dividing by the maximum observed occupancy rates and flow values, respectively, observed at that detector over the entire, 2010 initial data set.

In the incident rate model we develop we seek to model the probability of the occurrence of an incident, which in terms of our adjusted data set is the probability that the 0-1 Incident Indicator switches from a 0 value (no unusual congestion) to a 1 value (unusually high congestion). Over the course of the incident we obtain other descriptive incident characteristics, such as the duration of the incident and the severity of the incident (i.e., the average capacity reduction). However, once an incident is in progress at a detector we no longer utilize those data points until the incident has ceased and the congestion levels have returned to normal. Thus, in our adjusted data set, for each incident we only keep the first data point corresponding to the initial occurrence of the incident, and we remove all other incident data points from the adjusted data set.

The adjusted data set contains a total of 174,183 data points, with 217 of these data points indicating the initial occurrence of an incident. Thus, in the incident rate model we describe below the response is a binary, 0-1 Incident Indicator response, for which only $\frac{217}{174,183} \times 100\% \approx 0.12\%$ of the responses indicate the occurrence of an incident.

5.4.2 Model Selection

An obvious model choice for a binary response variable is a logistic model, which links the expected value of the binary response to a linear combination of fixed

effect predictors through a logit function. To capture the random variations inherent at each detector along the roadway, we also seek to include a random effect for each detector location. Since we have already incorporated spatial and temporal correlation among detectors using the downstream-incident indicator variable, we assume that the random detector effects are independent and identically distributed (iid).

A logistic model with random effects is in the class of Generalized Linear Mixed Models (GLMM), and fitting procedures for GLMM's are implemented in software such as SAS Proc GLIMMIX. However, due to the size and the sparse nature of our binary response data set, the iterative estimation algorithm in SAS Proc GLIMMIX (pseudo-likelihood estimation) failed to converge to yield model parameter estimates, although it did produce meaningful estimates with small subsets of our data set. In the following sections we will describe an alternative fitting procedure which successfully produced meaningful parameter estimates for the entire, adjusted data set. Before proceeding, however, we discuss the significant fixed effects which whose significance we determined using small subsets of the data set.

The following list describes the different fixed effects we considered in constructing a logistic model with random effects to model incident rate:

- Normalized flow value, $Flow_{\text{norm}}$
- Square of normalized flow value, $Flow^2_{\text{norm}}$
- Normalized occupancy rate, Occ_{norm}
- Square of normalized occupancy rate, Occ^2_{norm}
- Interaction term between normalized flow value and normalized occupancy rate, $Flow_{\text{norm}} * Occ_{\text{norm}}$

- Change in normalized flow value over time at the detector location, $\Delta_t Flow_{\text{norm}}$
- Change in normalized occupancy rate over time at the detector location, $\Delta_t Occ_{\text{norm}}$
- Change in normalized flow value from the upstream detector, $\Delta_x Flow_{\text{norm}}$
- Change in normalized occupancy rate from the upstream detector, $\Delta_x Occ_{\text{norm}}$
- Hourly time effect
- Month effect
- Downstream-incident indicator variable, $\mathbf{1}_{\text{Downstream}}$

Using SAS Proc GLIMMIX we fit a variety of small subsets of the adjusted data set to a logistic model with random, iid detector effects to determine the significance of the above fixed effects. Sample SAS code used to fit data subsets is provided in the appendix. We consistently observed that the most strongly significant effect is $\mathbf{1}_{\text{Downstream}}$. We observed a correlation between normalized flow rates and both hourly times and month, which led to the hourly time and month effects being insignificant in a model which included $Flow_{\text{norm}}$. We also observed that normalized occupancy rates were not significant in a model which included $Flow_{\text{norm}}$. Also, the changes in normalized flow values and occupancy rates over both time and space were not significant in any of the models we fitted.

Based on various data subsets, we concluded that the only significant fixed effect terms in a logistic model were $Flow_{\text{norm}}$, $Flow^2_{\text{norm}}$ and $\mathbf{1}_{\text{Downstream}}$.

5.4.3 Model Statement

In this section we provide a precise description of the proposed incident rate model, which models the occurrence of an incident as a logistic model with fixed effects

being the normalized flow values, the square of normalized flow values and a downstream-incident indicator variable, and with iid random effects corresponding to the individual detectors.

Incident Rate Model: Logistic Model with Random Effects (5.2)

i^{th} detector, $i = 1, \ldots, I = 13$ for detectors #10-22

j^{th} timeslot, $j = 1, \ldots, J = 66$ for the 5-min timeslots from 4:30am-10:00am

k^{th} day, $k = 1, \ldots, K = 206$ for all Mondays-Thursdays in 2010

$$Y_{ijk} = \begin{cases} 1, & \text{if an incident is initiated at detector } i,\ \text{timeslot } j,\ \text{day } k \\ 0, & \text{otherwise} \end{cases}$$

$(Flow_{\text{norm}})_{ijk} =$ Normalized flow value at detector i, timeslot j, day k

$\left(Flow^2_{\text{norm}}\right)_{ijk} =$ Square of normalized flow value at detector i, timeslot j, day k

$$(\mathbf{1}_{\text{Downstream}})_{ijk} = \begin{cases} 1, & \text{if an incident was in progress immediately downstream of} \\ & \text{detector } i \text{ within 30 minutes preceeding timeslot } j \text{ on day } k \\ 0, & \text{otherwise} \end{cases}$$

$$Y_{ijk} \,|\, r_i \sim \text{Bernoulli}\,(\mu_{ijk})$$

$$\text{logit}\,(\mu_{ijk}) = \beta_0 + (Flow_{\text{norm}})_{ijk}\,\beta_1 + \left(Flow^2_{\text{norm}}\right)_{ijk}\,\beta_2 + (\mathbf{1}_{\text{Downstream}})_{ijk}\,\beta_3 + r_i$$

$$r_i \sim N\left(0, \sigma_r^2\right)$$

In Model (5.2), r_i is the random effect for detector i, and μ_{ijk} is the expected mean of the response variable conditional on the random effects, $Y_{ijk} \,|\, r_i$, at detector i, timeslot

j and day k.

5.4.4 Model Fitting

To fit Model (5.2) we need to obtain estimates for the fixed effect parameters, $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_3$, and an estimate of the covariance parameter for the random effects, $\hat{\sigma}_r^2$, as well as standard errors of the above estimates.

As previously indicated, the fitting algorithm in SAS Proc GLIMMIX, which employs pseudo-likelihood estimation for the parameter estimates, failed to converge when applied to our entire, adjusted data set. However, using a fitting algorithm developed in the theory of Hierarchical Generalized Linear Models (HGLM), we were able to fit Model (5.2) to the adjusted data set and obtain meaningful parameter estimates. Over the next few sections we provide an overview of the theory of HGLM's and the associated fitting algorithm, describe how we applied the fitting algorithm to Model (5.2), and finally present the fitted parameter estimates.

5.4.4.1 HGLM: Overview

Hierarchical Generalized Linear Models (HGLM) are extensions of generalized linear mixed models (GLMM) in that the distribution of the random effects may be non-normal, and that the dispersion parameters may be modeled as linear functions of fixed effect predictors. The theory of HGLM's was first presented in the late 1990's and early 2000's (Lee and Nelder, 1996, Lee and Nelder, 2001a and Lee and Nelder, 2001b). There has been a profusion of literature addressing HGLM's over the last 10 years, although we belive that the best presentations can be found in Lee and Nelder, 2005 and Lee et al., 2006. There have been attempts to implement the theory of HGLM's into R software, such as in the R-packages *hglm* and *HGLMMM*. However, none of these

packages can accomodate the full generality of an HGLM, e.g., none of the packages fits an HGLM with correlated random effects.

In the theory of HGLM's a key distinction from the classical theory of GLMM's is the use of the h-loglihood (which is just the joint log-likelihood of the responses and the random effects). Lee et al., 2006 argue that the h-loglihood should be used for inference of the random effects, and that, conditional on the responses, the random effects are *estimated* and not *predicted*. Accompanying the theory of HGLM's is a unique fitting procedure, the core of which is a joint mean-dispersion generalized linear model (GLM) fitting procedure (Nelder and Lee, 1991), which in a simple way allows an extension of a GLMM to non-normal random effects, and also provides a simple means for fitting GLMM's with independent random effects.

5.4.4.2 HGLM: Canonical Scale and H-Likelihood

In this discussion of h-likelihoods, let y denote the data, let β denote the fixed effect parameters and let v denote the random effects. The marginal log-likelihood of the data is

$$l(\beta) = \log f_\beta(y),$$

and the extended log-likelihood is defined as the log of the joint distribution of the data and the random effects,

$$l_e(\beta, v) = \log f_\beta(y, v) = \log f_\beta(y) + \log f_\beta(v|y).$$

The h-likelihood is a special type of extended likelihood in which scale of the random effects is "canonical", which is defined as satisfying

$$\frac{f_{\beta_1}\left(\hat{v}_{\beta_1}|y\right)}{f_{\beta_2}\left(\hat{v}_{\beta_2}|y\right)} = 1$$

for all β_1 and β_2, where $\hat{v}$ is the maximum extended-likelihood estimate of v. The following difficulties with obtaining a canonical scale may arise:

- Obtaining a canonical scale for the random effects may require a nonlinear transformation of the random effects, e.g., $v = \log(u)$;
- A canonical scale for the random effects may not exist ;
- Verifying that a random effect scale is canonical may be nontrivial.

Classically, the fixed effect parameters, β, are estimated as maximum likelihood estimates (MLE) of $l(\beta)$, and v is "predicted" so as to minimize the mean-squared prediction error, $E||\hat{v} - v||^2$. If the h-likelihood, $h(\beta, v) = l_e(\beta, v)$, exists (i.e., if a canonical scale for the random effects exists), then Lee et al., 2006 have proven the following:

- The MLE of β from the marginal log-likelihood coincides with the estimate of β obtained from the joint-maximizer of the h-likelihood, $\left(\hat{\beta}, \hat{v}\right)$. In the normal linear mixed model case, the estimate of v from this joint-maximizer coincides with the classical minimizer of mean-squared error. Henderson, 1975 noted that the classical estimates can be obtained by maximizing the joint density function; however, he used joint maximization only as an algebraic device and did not recognize the theoretical implications in terms of extended likelihood inference.

- Partition the observed Fisher Information matrix of the h-likelihood as $I_h^{-1}\left(\hat{\beta},\hat{v}\right) = \begin{pmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{pmatrix}$. The observed information matrix for $\hat{\beta}$ from the marginal log-likelihood equals I_{11}. Furthermore, I_{22} yields an estimate for $\text{var}\left(\hat{v}-v\right)$ which accounts for the inflation of variance caused by estimating β.

Determining if the scale for the random effects is canonical may not be possible. However, Lee and Nelder, 2001a define the weak-canonical scale, which is a scale for the random effects such that the random effects are additively combined with the fixed effects, i.e., $X\beta + Zv$, and they claim that the weak-canonical scale is a canonical scale for a broad class of GLM models.

5.4.4.3 HGLM: Adjusted Profile Log-Likelihood

If a canonical scale for the random effects is not known (or does not exist), then Lee and Nelder, 2005 claim that the random effects and their inference should be estimated from an extended likelihood, but that the fixed effects and their inference must be estimated from the marginal log-likelihood. In this case, they promote using a Laplace approximation for the marginal log-likelihood,

$$l(\beta) \approx p_v\left(l_e\right) \equiv \left[l_e - \frac{1}{2}\log\det\left\{\frac{1}{2\pi}D\left(l_e;v\right)\right\}\right]_{\hat{v}_\beta} \tag{5.3}$$
$$\text{where } \hat{v}_\beta \text{ satisfies } \left.\frac{\partial l_e}{\partial v}\right|_{\hat{v}_\beta} = 0 \text{ and } D\left(l_e;v\right) = -\frac{\partial^2 l_e}{\partial v^2}.$$

$p_v\left(l_e\right)$ is referred to as an adjusted-profiled-log-likelihood (APL) function, in which the extended loglihood (i.e., log-likelihood) has had the random effects, v, profiled out with an adjustment term from a Laplace approximation. The notation $p_*\left(\cdot\right)$ will generally refer to an APL function with an adjustment term using the Laplace approximation.

5.4.4.4 HGLM: Inference of Fixed and Random Effects

Lee et al., 2006 present a summary of inference procedures using extended likelihoods and general likelihood inference for random effects:

- Classical likelihood is for inference about fixed parameters. For general models with latent random variables, the h-likelihood is the fundamental likelihood from which the marginal likelihood may be derived.

- Likelihood inference is justified by large-sample optimality. The likelihood principle asserts that the likelihood contains all information about the fixed effect parameters, but the process of estimation by maximizing the likelihood arises out of convenience and not from the likelihood principle. Similarly, the extended-likelihood principle asserts that the extended-likelihood contains all information about both the fixed effect parameters and random effects, but it does not tell us how to use the likelihood to make estimates. Classically we use non-likelihood methods to estimate random effects (i.e., minimize mean-squared prediction error), but these methods are not easily extendable to non-normal random effects.

- Joint maximization of the extended likelihood is only meaningful if the scale of the random effects is canonical, which heuristically means that the random effects density is information-free for the fixed effect parameters. In this case the extended likelihood is called the h-likelihood, and joint inference for fixed and random effects from the h-likelihood is similar to inference from ordinary likelihood.

- If we do not know the canonical scale, then we still use the h-likelihood for inference of the random effects, but must use the marginal likelihood (or APL approximation) for the estimation of the fixed effects.

- Inference for all nuisance parameters (covariance parameters and correlation parameters) should be determined from the adjusted-profiled-h-loglihood (APHL) function. In the normal linear mixed model case, this function coincides with the profile marginal loglihood function with a REML adjustment.

5.4.4.5 HGLM: Iterative Weighted Least Squares Fitting Algorithm

Implementing the joint mean-dispersion GLM fitting procedure used to fit HGLM's requires a thorough understanding of the iterative weighted least-squares (IWLS) fitting procedure for GLM's. Theoretical discussions of IWLS are simple to read and provide analytic expressions for each step of the procedure; however, these expressions involve inverses of potentially large matrices which cannot be analytically calculated. Furthermore, there are slight differences among authors and software regarding the specific components of GLM's, such as deviances, link functions, etc. To dispel any confusion, in this section we present the details of an IWLS fitting procedure for a Bernoulli response random variable with a logit-link function:

- Let β denote the vector of coefficients of the fixed effects and construct a model matrix, X, so that $\eta = X\beta$ is the linear predictor for the inverse logit of the expected mean of the response variable, μ.

- Define the GLM functions for the Bernoulli distribution:

  Variance function, $V(\mu) = \mu(1-\mu)$

  Link function, $\eta = \log\left(\frac{\mu}{1-\mu}\right)$

  Inverse of the link function, $\mu = \frac{\exp\eta}{1+\exp\eta}$, and its derivative, $\frac{d\mu}{d\eta} = \mu(1-\mu)$

$$\text{Scaled deviance, } d(y) = 2 \times \begin{cases} \log\left(\frac{1}{1-\mu}\right), & y = 0 \\ \log\left(\frac{1}{\mu}\right), & y = 1 \end{cases}.$$

- Initialize the vector of expected means as $\mu = \frac{y+0.5}{2}$, and calculate η.
- Construct the sum of scaled deviances, $D = \sum d(y)$.
- Iterate the following IWLS procedure until the sum of scaled deviances converges: (1) Calculate $\frac{d\mu}{d\eta}$. (2) Calculate $z = \eta + \frac{y-\mu}{\frac{d\mu}{d\eta}}$. (3) Calculate $W = \frac{\left(\frac{d\mu}{d\eta}\right)^2}{V(\mu)}$. (4) Using QR-decomposition (i.e., R-function .dqrls), solve the weighted least-squares equation $X'WX\beta = X'Wz$ to estimate β. (5) Update $\eta = X\beta$ and μ. (6) Calculate the sum of scaled deviances, D.
- After convergence of the IWLS procedure, use the last QR-decomposition to retrieve the diagonal entries of $(X'\beta X)^{-1}$, whose square roots yield estimated standard errors the the estimates of β.

5.4.4.6 HGLM: Augmented GLM

A portion of the HGLM fitting procedure constructs an augmented GLM consisting of a hybrid mixture of two different GLM families (as described below), and then fits the augmented GLM using essentially the same IWLS procedure as described in the previous section. Before discussing augmented GLM's, we first discuss augmented linear models, which are simpler.

In a normal linear mixed model (LMM), the mixed model equations which provide the joint estimation of the fixed and random effects, β and v, can be summarized

into a single, augmented linear model as follows:

$$\begin{pmatrix} y \\ \\ \psi_M \end{pmatrix} = \begin{pmatrix} X & Z \\ \\ 0 & I \end{pmatrix} \begin{pmatrix} \beta \\ \\ v \end{pmatrix} + e^*, \quad e^* \sim N \left(0, \Sigma_a = \begin{pmatrix} \Sigma & 0 \\ \\ 0 & D \end{pmatrix} \right)$$

where Σ and D are the covariance matrices of the error terms and random effect terms in the LMM, respectively, and the quasi-data, ψ_M, satisfies $\psi_M = Ev = 0$. In the above model, if we let y_a denote the vector of augmented responses, T the augmented model matrix and δ the augmented vector of fixed effects, then the above model and corresponding effects estimates may be written as

$$y_a = T\delta + e^*$$

$$\hat{\delta} = \left(T' \Sigma_a^{-1} T \right)^{-1} T' \Sigma_a^{-1} y_a.$$

Thus, determining the fixed and random effects parameter estimates in a linear mixed model is equivalent to determining the fixed effects parameter estimates in an augmented linear model.

The idea of an augmented linear model does not add anything new to the analysis of normal linear mixed models, but is a very useful device in the extension to generalized linear mixed models with non-normal responses, as follows: Allow the possibility for a transformation of the random effects as $v = g_M(u)$, and define the quasi-data $\psi_M = E(u)$. Assuming that the conditional responses and the transformed

random effects both come from an overdispersed GLM family, we may write:

$$E(y|u) = \mu, \quad \text{var}(y|u) = \phi V(\mu)$$

$$E(\psi_M) = u, \quad \text{var}(\psi_M) = \lambda V_M(u).$$

Construct an augmented an linear predictor, an augmented model matrix and an augmented vector of fixed effects parameters as:

$$\eta_{M_a} \equiv \begin{pmatrix} \eta \\ \\ \eta_M \end{pmatrix} = \begin{pmatrix} X\beta + Zv \\ \\ v \end{pmatrix}$$

$$= \begin{pmatrix} X & Z \\ \\ 0 & I \end{pmatrix} \begin{pmatrix} \beta \\ \\ v \end{pmatrix} \equiv T_M \omega.$$

The above model is a augmented GLM model consisting of two GLM families (one for the conditional response and one for the random effects), and may be fitted using an augmented IWLS algorithm by constructing an augmented variable and augmented weight matrix:

$$z_{M_a} = \begin{pmatrix} z \\ \\ z_M \end{pmatrix}, \text{ where } z = \eta + \frac{y - \mu}{\frac{d\mu}{d\eta}} \text{ and } z_M = v + \frac{\psi_M - u}{\frac{du}{dv}}$$

$$W_{M_a} = \begin{pmatrix} \frac{\left(\frac{d\mu}{d\eta}\right)^2}{V(\mu)} & 0 \\ \\ 0 & \frac{\left(\frac{du}{dv}\right)^2}{V(u)} \end{pmatrix}.$$

The scaled deviance is the sum of the scaled deviances for the two GLM families:

$$d = 2\int_{\mu}^{y} \frac{y - s}{V(s)} ds \quad \text{and} \quad d_M = 2\int_{u}^{\psi_M} \frac{\psi_M - s}{V_M(s)} ds.$$

We apply the above IWLS algorithm to the augmented GLM model, where at each iterative step we solve the weighted least-squares equation to solve for ω,

$$T_M' W_{M_a} T_M \omega = T_M' W_{M_a} z_{M_a},$$

with convergence of the algorithm based on the convergence of the sum of the scaled deviances. Although GLM-fitting software in SAS and R cannot fit an augmented GLM, the fitting algorithm is essentially the same as fitting a GLM.

If the scale for the random effects is canonical, which is generally the case for the weak-canonical scale, then the augmented GLM fit yields the joint estimates of fixed and random effects from the h-likelihood. Thus, using this method, if there are no unknown covariance parameters and the random effects are independent, then fitting a GLMM is just as easy as fitting a GLM. Furthermore, we can quite easily extend to non-normal random effects since the above augmented GLM fitting procedure accomodates any GLM family for the random effects.

5.4.4.7 HGLM: Joint Mean-Dispersion GLM Fitting Procedure

In a GLM family with mean μ, the variance can be written as $\phi V(\mu)$, where ϕ is referred to as a "dispersion parameter" (Lee et al., 2006 also refer to it as a prior weight in a GLM). For a normal GLM, $N\left(\mu, \sigma^2\right)$, we have $\phi = \sigma^2$ and $V(\mu) = 1$, and for a Bernoulli GLM, Bernoulli(μ), we have $\phi = 1$ and $V(\mu) = \mu(1-\mu)$.

In an HGLM we have two dispersion parameters, the dispersion parameter for the response distribution (conditional on the random effects), and the dispersion parameter for the random effects distribution. In the model we fit in this paper, Model (5.2), which is a Bernoulli-Normal HGLM (the terminology of Lee et al., 2006 is to refer to

an HGLM by its response distribution followed by the random effects distribution), the only dispersion parameter is the variance of the normal distribution for the random effects.

Nelder and Lee, 1991 developed a joint mean-dispersion GLM fitting algorithm which is used to estimate both the fixed effect coefficients and the dispersion parameter in a GLM, and this algorithm is the core algorithm used to fit HGLM's, as it can be applied to the augmented GLM described in the previous section. We now provide a brief description of this joint mean-dispersion GLM fitting algorithm.

In a normal linear mixed model, the scaled deviances are distributed as $d \sim \phi\chi^2_{(1)}$, which is a gamma distribution with mean ϕ and variance $2\phi^2$. Motivated by this fact, the joint mean-dispersion GLM fitting algorithm iteratively alternates between fitting the mean GLM model for the fixed effect coefficients using the current estimated dispersion parameters as prior weights, and fitting a gamma-GLM for the dispersion parameter using the current estimated deviances as data. This joint mean-dispersion GLM fitting algorithm also allows the dispersion parameter to be modeled as a linear function of fixed effects through a link function (usuually a log-link for a gamma-GLM), which Nelder and Lee, 1991 refer to as "structured dispersion." In Figure 5.8 we provide a schematic diagram (taken from Nelder and Lee, 1991) which demonstrates the relationship between the two GLM's for the mean and the dispersion parameter which are iteratively fitted. In Figure 5.8 the fixed effect predictors for the dispersion parameters are denoted as γ, which are usually a subset of the fixed effect predictors for the mean GLM model, β. A default, however, is to not use any predictors and to fit an intercept-only gamma-GLM model for the dispersion parameters, with log link function $h(\phi) = \log(\phi)$. Also, q are the leverages from the mean GLM model, which theoretically

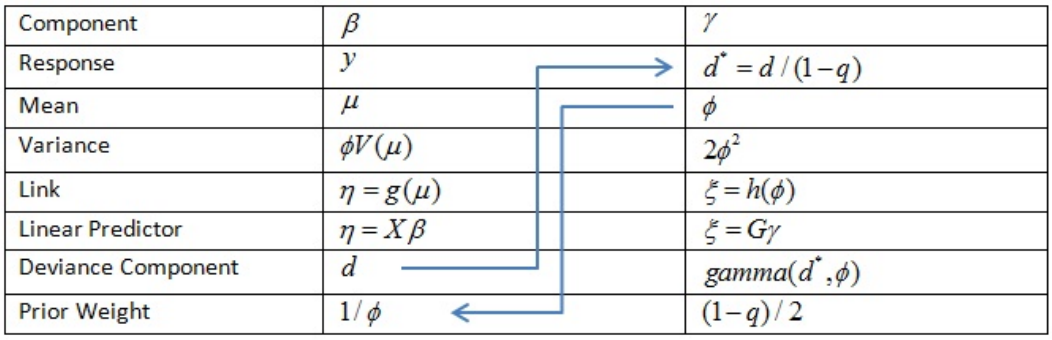

Figure 5.8: Graphical description of the joint mean-dispersion GLM fitting procedure, which iteratively fits a GLM for the mean model and a gamma-GLM for the dispersion parameter. At each stage, the current dispersion parameter estimates are used as prior weights in the mean GLM model, and the current scaled deviances (after dividing by $1-q$ where q are leverages), are used as data in the dispersion paramater gamma-GLM model.

are the diagonal elements of $X\left(X'WX\right)^{-1}X'W$.

5.4.4.8 HGLM: Fitting HGLM's with Independent Random Effects

The HGLM fitting procedure iteratively cycles through three different steps:

- Given values of the dispersion parameters for the response distribution and the random effects distribution, construct an augmented-GLM model for the fixed-effects and random-effects and obtain updated effect estimates, along with deviances and leverages.

- Update dispersion parameter estimates for the response distribution using the deviances and leverages from the first step, by fitting a gamma-GLM as in the joint mean-dispersion fitting procedure.

- Update dispersion parameter estimates for the random effects distribution using the deviances and leverages from the first step, by fitting a gamma-GLM as in the joint mean-dispersion fitting procedure.

For the Bernoulli-Normal HGLM model fitted in this paper the conditional response distribution is Bernoulli, which does not have any dispersion parameter estimates, so that we may eliminate the second step above and iterate between only the first and third steps.

5.4.4.9 HGLM: Fitting HGLM's with Correlated Random Effects

In the case of independent random effects, a great advantage to HGLM's over GLMM's is the easy extension to non-normal random effects and the extension to structured dispersion parameters for both the conditional response distribution and the random effects distribution. Additionally, the fitting procedure for HGLM's repeatedly uses only the IWLS fitting algorithm, and is relatively simple compared to the pseudo-likelihood fitting algorithm for GLMM's.

In the case of correlated random effects, which will not be used in this paper, the fitting procedure for HGLM's becomes much more difficult both theoretically and computationally. Here we briefly touch on the difficulties.

Denote the variance matrix of the random effects distribution as $\text{var}(v) = \lambda\Lambda$, where λ is a dispersion parameter and $\Lambda(\rho)$ is a matrix which captures the correlation structure, with correlation parameters ρ. In principle, Λ may be decomposed as $\Lambda = LL'$, and the random effects, v, may be transformed into uncorrelated random effects, r, as $v = Lr$, so that $\text{var}(r) = \lambda I$ where I is an identity matrix. The linear predictor for the

HGLM becomes

$$\begin{aligned}\eta &= X\beta + Zv \\ &= X\beta + ZLr \\ &= X\beta + Z^*r,\end{aligned}$$

which is the linear predictor for an HGLM with independent random effects, r. However, for an arbitrary random effects distribution for v it may nontrivial to obtain the distribution for the transformed random effects, r, although a simple cases arises if v are normally distributed, in which case r are normally distributed as well.

If the distribution of the transformed random effects can be determined, then Lee et al., 2006 propose the following iterative algorithm for fitting an HGLM with correlated random effects:

- Given estimates for the correlation parameters, fit the transformed HGLM with independent random effects using the HGLM fitting procedure described in the previous section.
- Given estimates for the fixed effects, random effects and dispersion parameters from the previous step, estimate the correlation parameters by maximizing the APHL (which in the normal LMM case is equal to the profiled marginal log-likelihood function with a REML adjustment).

The second step can be a computationally very difficult step, and in my (limited) experience has been the primary hindrance to fitting an HGLM with a large number of random effects and data points.

5.4.4.10 HGLM: Discussion of Computational Aspects

The pseudo-likelihood method (PL) of fitting GLMM's is an iterative procedure which, at each step, uses a linear approximation for the inverse-link function in terms of the fixed and random effects to construct a linear mixed model (LMM) for a pseudo-response, with a normality assumption for the pseudo-response. The fixed and random effect estimates of the LMM are used in the subsequent iterative step. However, if the number of fixed effects, the number of random effects, or the number of responses is large, then implementation of PL is not possible directly. In my experience, the most difficult computational hurdle is minimizing the profiled pseudo-loglihood function to obtain estimates for the nuisance parameters. However, the 'Sweep' method (Wolfinger et al., 1994) is a computational tool developed for the PL method which enables its implementation. Such a tool does not yet exist for the HGLM fitting procedure, as explained below.

In the fitting procedure for an HGLM with correlated random effects, there are two primary steps. The core of the first step is repeated use of the IWLS fitting procedure for augmented GLM's, which may be adapted for potentially large, sparse matrices. The second step, however, is to maximize the adjusted-profiled-loglihood function (APHL) to determine estimates for the correlation parameters, and this step is computationally difficult. A method (such as a 'Sweep' method) must be developed to make this second step of the HGLM fitting procedure tractable for high numbers of parameters or data points.

5.4.4.11 Model Fitting: Details

We fit Model (5.2) to the adjusted data set using the HGLM fitting procedure described in the previous sections (for an HGLM with independent random effects). For Model (5.2) we have a vector of fixed effects, $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)'$, and a vector of random effects, $\mathbf{r} = (r_1, \ldots, r_{13})'$, and as in Section 5.4.4.6 we construct an augmented GLM model with augmented vector of fixed effects, $\omega = (\beta', \mathbf{r}')'$. Applying the joint mean-dispersion GLM fitting algorithm from Section 5.4.4.7 to the augmented GLM allows us to determine parameter estimates for ω, along with estimated standard errors as the square root of the diagonal entries of $\left(T'\Sigma_a^{-1}T\right)^{-1}$ (see Section 5.4.4.5). Thus, using the HGLM fitting procedure, which for our model uses only the IWLS fitting algorithm, we are able to provide estimates for the fixed effects and random effects, $\hat{\beta}$ and $\hat{\mathbf{r}}$, estimated standard errors of their estimates, $\hat{\sigma}_{\hat{\beta}}$ and $\hat{\sigma}_{\hat{\mathbf{r}}}$, and an estimate of the dispersion parameter for the normal, random effects distribution, $\hat{\sigma}_r^2$. However, to obtain a standard error estimate for $\hat{\sigma}_r^2$ requires more work, which we now describe.

If l is the marginal loglihood, then the variance of the estimate for σ_r^2 is given by the inverse of the Fisher Information, so that a standard error estimate for the estimate of σ_r^2 is

$$\hat{\sigma}_{\hat{\sigma}_r^2} = \left[-\frac{\partial^2}{\partial \left(\sigma_r^2\right)^2} l \right]^{-\frac{1}{2}} \Bigg|_{\hat{\sigma}_r^2}.$$

To approximate the marginal loglihood Lee et al., 2006 propose to use the APHL with both the fixed effects and random effects profiled-adjusted out, i.e., $l \approx p_{\beta,\mathbf{r}}(h)$, and we now describe in detail how to obtain the APHL and use it to obtain standard error estimates for $\hat{\sigma}_r^2$ for the incident rate model in this paper.

Model Matrices

The adjusted data set is sorted firstly by detector, secondly by day and lastly

by time.

The data matrix, $X_{174183\times 4}$, contains a column of ones (intercept term), a column of normalized flow values, $Flow_{\text{norm}}$, a column of squared normalized flow values, $Flow^2_{\text{norm}}$, and a column of downstream-incident indicator variable values, $\mathbf{1}_{\text{Downstream}}$.

The model matrix for the random effects, $Z_{174183\times 13}$, is of the form

$$Z = \begin{pmatrix} 1 & 0 & & & 0 \\ \vdots & \vdots & & & \\ 1 & 0 & & & \\ 0 & 1 & & & \vdots \\ & \vdots & & & \\ & 1 & \cdots & & \\ \vdots & 0 & & & \\ & & & & 0 \\ & \vdots & & & 1 \\ & & & & \vdots \\ 0 & 0 & & & 1 \end{pmatrix}_{174183\times 13,}$$

where the number of ones appearing in the ith column equals to the number of data points in the adjusted data set for the ith detector.

H-Loglihood

The extended log-likelihood, which we denote as h for h-likelihood even though

we have not proven that the scale of the random effects is canonical, is given by

$$\begin{aligned} h &= \log f\left(\mathbf{Y} \mid \mathbf{r}\right) \cdot f\left(\mathbf{r}\right) \\ &= \log \prod_{i,j,k} \mu_{i,j,k}^{Y_{i,j,k}} \left(1-\mu_{i,j,k}\right)^{1-Y_{i,j,k}} + \log \frac{1}{\sqrt{|2\pi\sigma_r^2\Lambda|}} e^{-\frac{1}{2}\mathbf{r}'\left(\sigma_r^2\Lambda\right)^{-1}\mathbf{r}}, \end{aligned}$$

where the correlation matrix is the identity matrix, $\Lambda = I_{13\times 13}$, since the random effects are iid. This expression for h may be simplified as

$$h = \sum \left[Y * (X\beta + Z\mathbf{r}) - \log\left(1 + e^{X\beta + Z\mathbf{r}}\right)\right] - \frac{13}{2}\log(2\pi) - \frac{13}{2}\log\sigma_r^2 - \frac{1}{2\sigma_r^2}\mathbf{r}'\mathbf{r},$$

where $*$ denotes a component by component product of vectors.

Adjusted Profile Log Likelihood

To calculate the APHL we use (5.3),

$$p_{\beta,\mathbf{r}}(h) = h - \frac{1}{2}\log\det\frac{1}{2\pi}D\Big|_{\hat{\beta},\hat{\mathbf{r}}},$$

where D is the negative of the matrix of second partial derivatives of h with respect to the components of β and $\mathbf{r}$, and where for simplicity we let $\hat{\beta}$ and $\hat{\mathbf{r}}$ be the estimates provided from the HGLM fitting algorithm.

Estimated Standard Error

After some simplification, the negative of the second derivative of $p_{\beta,\mathbf{r}}(h)$ with respect to σ_r^2 reduces to

$$-\frac{\partial^2}{\partial\left(\sigma_r^2\right)^2} p_{\beta,\mathbf{r}}(h) = \frac{\mathbf{r}'\mathbf{r}}{\left(\sigma_r^2\right)^3} + \frac{1}{2}\mathrm{Trace}\left\{ D^{-1}\frac{\partial^2 D}{\partial\left(\sigma_r^2\right)^2} - \left[D^{-1}\frac{\partial D}{\partial\sigma_r^2}\right]\right\}. \tag{5.4}$$

$D_{17\times 17}$ can be decomposed into a block structure,

$$D = \begin{pmatrix} -\frac{\partial^2 h}{\partial \beta^2}_{4\times 4} & -\frac{\partial^2 h}{\partial \beta \partial \mathbf{r}}_{4\times 13} \\ -\frac{\partial^2 h}{\partial \mathbf{r} \partial \beta}_{13\times 4} & -\frac{\partial^2 h}{\partial \mathbf{r}^2}_{13\times 13} \end{pmatrix},$$

where the individual components of each block are

$$\begin{aligned} \left(-\frac{\partial^2 h}{\partial \beta^2}\right)_{ll'} &= \sum_{i=1}^{174183} \mu_i(1-\mu_i) X_{il} X_{il'} \\ \left(-\frac{\partial^2 h}{\partial \mathbf{r}^2}\right)_{mm'} &= \sum_{i=1}^{174183} \mu_i(1-\mu_i) Z_{im} Z_{im'} + \frac{1}{\sigma_r^2}\delta_{mm'} \\ \left(-\frac{\partial^2 h}{\partial \beta \partial \mathbf{r}}\right)_{lm} &= \sum_{i=1}^{174183} \mu_i(1-\mu_i) X_{il} Z_{im}. \end{aligned}$$

In these expressions $\mu_{i,j,k}$ has been relabelled with a single index, i, and $\delta_{mm'}$ is the usual delta indicator function which equals one if $m = m'$ and zero otherwise. Using these expressions it is easy to take derivatives of D with respect to σ_r^2, for which we obtain

$$\frac{\partial D}{\partial (\sigma_r^2)} = -\frac{1}{(\sigma_r^2)^2} \begin{pmatrix} 0 & 0 \\ 0 & I_{13\times 13} \end{pmatrix} \text{ and } \frac{\partial^2 D}{\partial (\sigma_r^2)^2} = \frac{2}{(\sigma_r^2)^3} \begin{pmatrix} 0 & 0 \\ 0 & I_{13\times 13} \end{pmatrix}.$$

Inserting the above expressions into (5.4) allows us to determine the estimated standard error for $\hat{\sigma}_r^2$ as

$$\hat{\sigma}_{\hat{\sigma}_r^2} = \left[-\frac{\partial^2}{\partial (\sigma_r^2)^2} p_{\beta,\mathbf{r}} \right]^{-\frac{1}{2}} \Bigg|_{\hat{\sigma}_r^2}.$$

5.4.4.12 Fitted Model Results

In the Appendix we provide the R-code used to fit the incident rate model to the adjusted data set, and estimate all standard errors. Furthermore, we note that for

the small subsets of the adjusted data set which could be fit by SAS Proc Glimmix, the results of the HGLM fitting algorithm nearly always exactly matched the results of SAS Proc Glimmix, with the only occasional variation being in the estimated standard error for the estimate of the dispersion parameter, $\hat{\sigma}_r^2$.

In Table 5.4 we display the fitted, parameter estimates for Model (5.2), which required 14 outer iterations to achieve four-digit accuracy in all parameter estimates. The required running time was under two minutes.

Parameter	Estimate	Standard Error
β_0	-16.2498	2.36
β_1	20.1345	6.62
β_2	-9.9452	4.58
β_3	4.5870	0.15
σ_r^2	0.2294	0.066
r_1	-0.1674	0.289
r_2	0.03859	0.261
r_3	-0.1961	0.270
r_4	0.1562	0.249
r_5	0.3531	0.230
r_6	-0.4265	0.327
r_7	-0.8378	0.361
r_8	-0.4843	0.312
r_9	0.08462	0.252
r_{10}	0.4895	0.220
r_{11}	0.5535	0.217
r_{12}	0.3495	0.233
r_{13}	0.0870	0.278

Table 5.4: Parameter estimates and their estimated standard errors for the incident rate model, Model (5.2), fit to the adjusted data set.

Using the fitted parameter estimates, and letting F denote the normalized flow value, at a single detector the probability of the occurrence of an incident, μ, is a logit-normal random variable which satisfies

$$\text{logit}(\mu) = \hat{\beta}_0 + F\hat{\beta}_1 + F^2\hat{\beta}_2 + \mathbf{1}_{Downstream}\hat{\beta}_3 + r, \text{ where } r \sim N\left(0, \hat{\sigma}_r^2\right).$$

Although there is no analytic relation for the mean of a logit-normal random variable, a simple analytic relation exists for the median of a logit-normal random variable,

$$\text{median}\,(\mu) = \frac{\exp\left(\hat{\beta}_0 + F\hat{\beta}_1 + F^2\hat{\beta}_2 + \mathbf{1}_{Downstream}\hat{\beta}_3\right)}{1 + \exp\left(\hat{\beta}_0 + F\hat{\beta}_1 + F^2\hat{\beta}_2 + \mathbf{1}_{Downstream}\hat{\beta}_3\right)},$$

which provides an analytic description for how the probability of the occurrence of an incident at a single detector varies with flow. In Figure 5.9 we graph this relationship for the median of the probability of incident occurrence with respect to normalized flow, for both values of the downstream incident indicator variable.

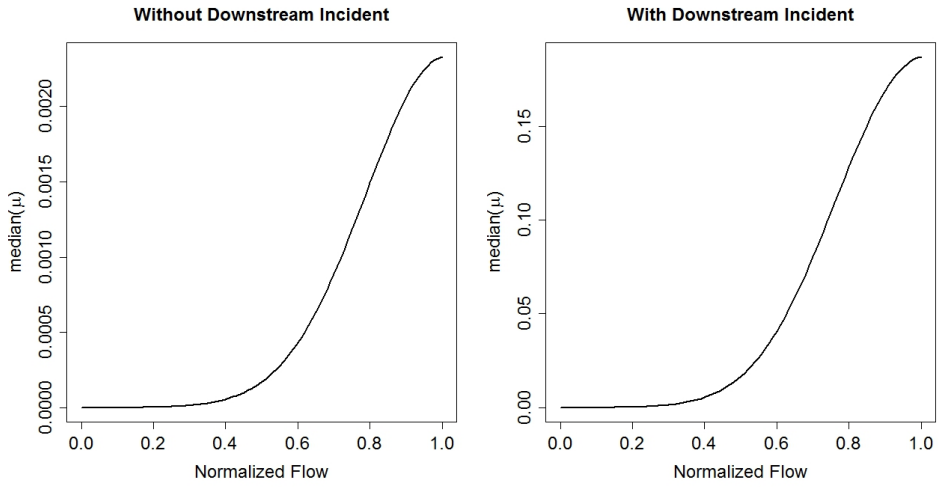

Figure 5.9: Graph of the median of the probability of incident occurrence at a single detector, μ, with respect to normalized flow values, F, using the fitted parameter estimates from Table 5.4. We provide graphs for both values of the downstream incident indicator variable, which indicates whether or not an incident has been in progress at the immediate downstream detector within the previous 30 minutes.

5.4.5 Incident Duration and Capacity Reduction

As discussed in Giuliano, 1989, the impact of incidents on freeway operations depends upon the incident rate, incident type, incident duration and capacity reduction caused by the incident. For this paper we are not interested in incident type, but are

only interested in the congestion increase caused by the incident. In order to incorporate stochastically occurring incidents into the Corridor Problem, we must have a model for incident rate, incident duration, and capacity reduction. The incident rate model developed in the previous section satisfies the first of these three needs.

A review of models of incident duration is presented in Wang et al., 2005, where they generally conclude that incident duration is best modelled as a random variable following a log-normal distribution. However, the models they review do not consider how incident duration depends upon the traffic conditions at the start of the incident. Similarly, Smith et al., 2003 discusses the impact of incidents on capacity reduction, concluding that capacity reduction is best modelled as a random variable following a beta distribution. They also do not consider how capacity reduction depends upon the traffic conditions at the start of the incident.

In the following two sections we investigate the duration and capacity reduction caused by the incidents detected in our 2010 PeMS data, and develop simple linear regression models for incident duration and capacity reduction based upon the normalized occupancy rate at the start of the incident.

5.4.5.1 Model for Incident Duration

As will be shown in Equation 5.5 in Section 5.5.1.1, the system of time-units used in the Corridor Problem satisfies 1.5 time-units = 0.5 hr. For the 217 incidents detected in the adjusted PeMS data set, we convert the duration of those incidents from 5-min intervals into time-units, so that the model we develop for incident duration may be directly applied to the Corridor Problem.

In Figure 5.10, for each of the 217 incidents detected we plot the (scaled) duration of the incident vs. the normalized occupancy rate at the start of the incident.

We observe that incident duration generally decreases with occupancy rate, which we interpret as follows: Incidents occurring in more heavily congested traffic are of shorter duration than incidents occurring in less congested traffic. We did not attempt to explain this phenomenon, but instead fit a simple linear regression model for incident duration in terms of the normalized occupancy rate at the start of the incident. The fitted regression line is also graphed in Figure 5.10.

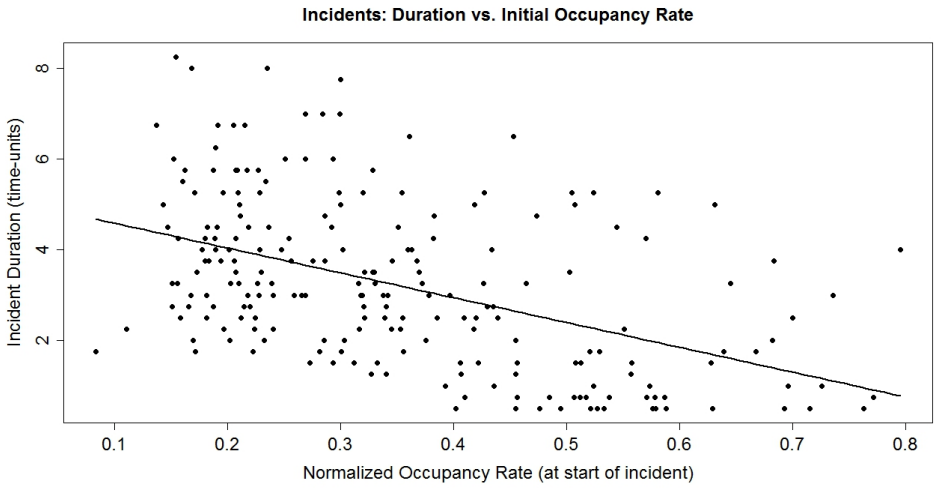

Figure 5.10: For each of the 217 incidents detected, we plot the incident duration (scaled to the Corridor Problem time-units) vs. the normalized occupancy rate at the start of the incident, along with a fitted linear regression line.

5.4.5.2 Model for Incident Capacity Reduction

For each incident we calculate the average normalized flow value over the duration of the incident, and we assume that the reduced capacity of the roadway due to the incident is equal to the average flow value during the incident. Recall that in the incident detection algorithm we inserted a control that did not permit the detection of an incident unless the average flow value over the duration of the incident was less than the flow value at the start of the incident.

In Figure 5.11, for each of the 217 incidents detected we plot the reduced

capacity due to the incident vs. the normalized occupancy rate at the start of the incident. We observe that the reduced capacity generally decreases with occupancy rate, which we interpret as follows: Incidents occurring in more heavily congested traffic are less severe than incidents occurring in less congested traffic. We did not attempt to explain this phenomenon, but only fit a simple linear regression model for incident capacity reduction in terms of the normalized occupancy rate at the start of the incident. The fitted regression line is also graphed in Figure 5.11.

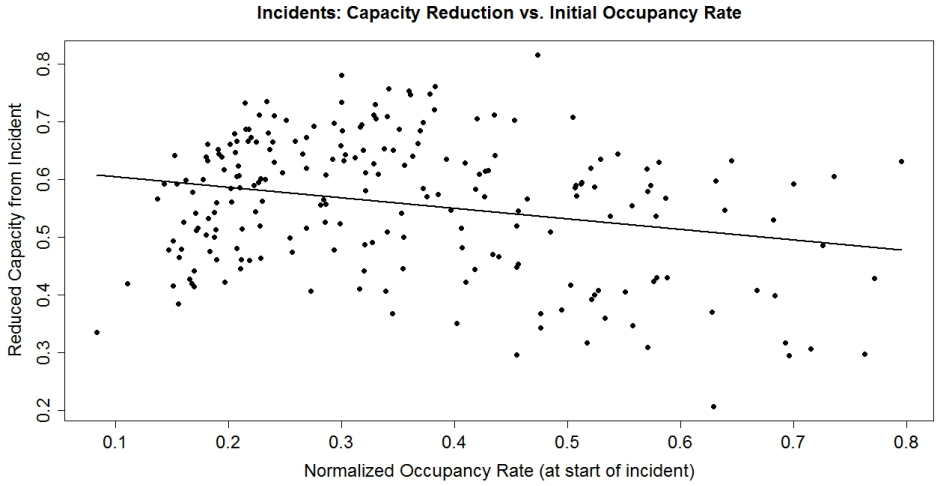

Figure 5.11: For each of the 217 incidents detected, we plot the reduced capacity due to the incident vs. the normalized occupancy rate at the start of the incident, along with a fitted regression line. Here we assume that the reduced capacity is equal to the average, normalized flow value over the duration of the incident.

5.5 Simulation Study: Stochastic Incidents in the Corridor Problem

In this section we use the incident rate model developed in the previous section to introduce randomly occurring incidents into the Corridor Problem. Simulation studies to understand the impact of incidents on traffic flow operations have been undertaken

in related research, such as in Sinha, 2006. In earlier work on the Single-Entry Corridor Problem (DePalma and Arnott, working paper), we obtained analytic expressions for two equilibrium departure patterns (corresponding to the Social Optimum and User Optimum), and in this section we explore how these equilibrium departure patterns are distorted with the introduction of stochastic incidents. We do not anticipate being able to obtain analytic results, but instead conduct a simulation study to determine how the traffic flow dynamics are altered with the occurrence of incidents.

We begin by reviewing the equilibrium solutions to the Single-Entry Corridor Problem from DePalma and Arnott, working paper, followed by a discussion of how to implement the incident rate model into a simulation study of traffic dynamics using the Cell-Transmission model, followed by a presentation and discussion of our simulation results.

5.5.1 Single-Entry Corridor Problem

In the Single-Entry Corridor Problem, a fixed number of identical commuters travel along a single-lane corridor of constant width from a common origin to a common destination, the central business district (CBD), with the same desired work-start time at the CBD. Traffic flow dynamics along the corridor are modeled using LWR flow congestion with Greenshields' Relation (i.e., mass conservation for a fluid coupled with a negative linear relation between traffic speed and traffic density), and trip cost is a linear combination of travel time cost plus schedule delay (time early or time late) cost. We seek to characterize two equilibrium solutions, the user optimum (UO) solution, in which no commuter can reduce their trip cost, and the social optimum (SO) solution, in which the total population trip cost is minimized.

Assuming that late arrivals are not permitted (which is equivalent to an infinite

cost for late arrivals), and building off of work by Newell, 1988, we determined analytic expressions for the UO and SO corridor inflow rates. In Figures 5.12 and 5.13 we reproduce the graphs from DePalma and Arnott, working paper, which show the cumulative inflow and cumulative outflow curves for the UO and SO solutions, respectively (note that the corridor inflow rate is the time derivative of the cumulative inflow curve). The

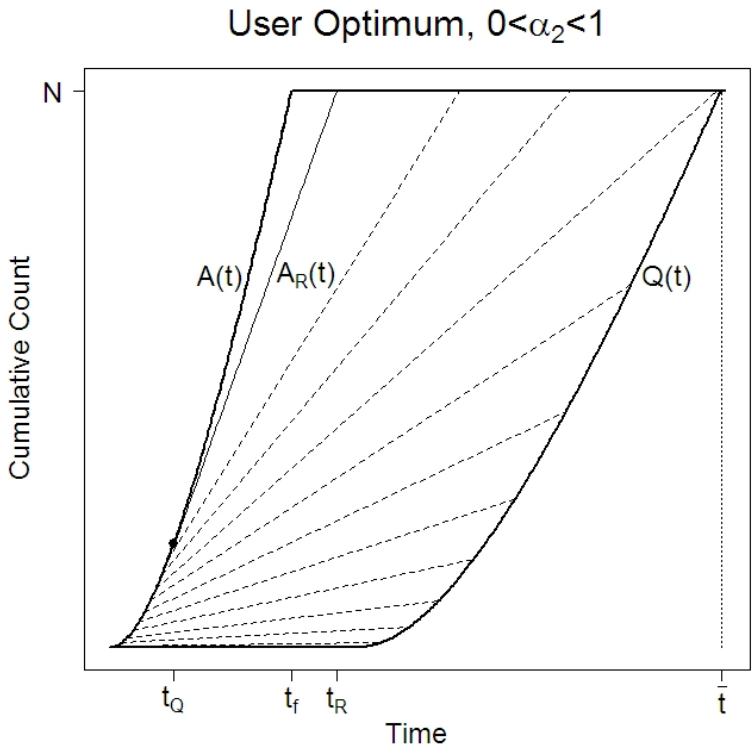

Figure 5.12: Cumulative inflow curves, $A(t)$ and $A_R(t)$, and cumulative outflow curve, $Q(t)$, for the UO solution (taken from DePalma and Arnott, working paper), where α_2 is the ratio of the unit schedule delay cost to unit travel time cost. $A(t)$ and $A_R(t)$ are the cumulative corridor and roadway inflow curves, respectively, so that the horizontal distance between the two curves is the queuing time for a commuter before entering the roadway. t_Q is the time at which a queue develops, t_f and t_R are the times of the final vehicle departures into the corridor and roadway, respectively, and $\bar{t}$ is the time of the final vehicle arrival at the CBD.

UO and SO solutions are completely determined by their corridor inflow rates. In the UO the inflow rate steadily increases, and since traffic flow in the roadway cannot exceed

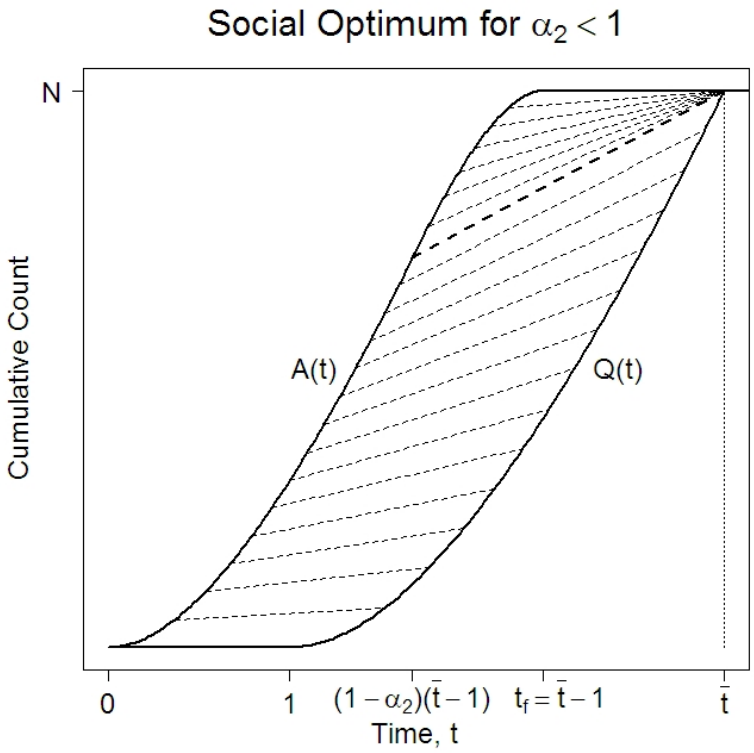

Figure 5.13: Cumulative inflow curve, $A(t)$, and cumulative outflow curve, $Q(t)$, for the SO solution (taken from DePalma and Arnott, working paper), where α_2 is the ratio of the unit schedule delay cost to unit travel time cost. t_f is the time of the final vehicle departure into the corridor and $\bar{t}$ the time of the final vehicle arrival at the CBD. In the SO the first and last vehicle departures travel at free-flow velocity, i.e., they encounter no congestion.

capacity, after a finite time a queue develops at the entrance to the roadway, with the length of the queue increasing over time. In the SO the length of the rush-hour, $\bar{t}$, is longer, so that traffic is more spread out, and the first and last vehicle departures travel at free-flow velocity, i.e., they encounter no congestion.

5.5.1.1 Units

In the Single-Entry Corridor Problem all units are scaled as follows:

- Choose length units such that the length of the corridor, $l = 1$.
- Choose time units such that corridor length to free-flow velocity, $\frac{l}{v_0} = 1$.

- Under the above units, choose population units such that capacity flow, $q_m = 1$.

  Under Greenshields' Relation, this choice of population units is equivalent to jam density, $k_J = 4$ (see DePalma and Arnott, working paper).

- Under the above units, choose cost units such that unit travel time cost, $\alpha_1 = 1$.

Under this choice of units, the only free parameters are N, the total population, and α_2, the unit schedule delay cost ($\alpha_2 < 1$ since DePalma and Arnott, working paper, take schedule delay cost to be less than travel time cost).

We now discuss how to adapt the incident rate model, calibrated from the PeMS data, to the Single-Entry Corridor Problem. We first previewed the daily occupancy rate graphs for detectors #10-23, and observed that the rush-hour traffic occurs roughly between 6:20am and 8:40am, for a total rush-hour duration of $\bar{t} = 2\frac{1}{3}$ hr. Capacity flow rates for a detector were approximately $q_m = 10,000$ vehicles/hr, and the total population travelling during the rush-hour at a single detector was around $N = 16,000$ vehicles. We may obtain these corresponding values in the UO solution (with $\alpha_2 = \frac{1}{2}$) if we set $N = 5$ population-units, which results in a rush-hour duration of $\bar{t} \approx 7$ time-units (this calculation comes from the theory developed in DePalma and Arnott, working paper, which we do not reproduce here). To see this, note that

$$\bar{t} = 7 \text{ time-units } = 2\frac{1}{3} \text{ hr} \quad \Rightarrow \quad 1 \text{ time-unit } = \frac{1}{3} \text{ hr },$$

so that,

$$q_m = 1\frac{\text{population-unit}}{\text{time-unit}} = 10,000\frac{\text{vehicles}}{\text{hr}} = 3,333\frac{1}{3}\frac{\text{vehicles}}{\text{time-unit}}$$
$$\Rightarrow \quad 1 \text{ population-unit } = 3,333\frac{1}{3} \text{ vehicles},$$

and, thus,

$$N = 16,000 \text{ vehicles} \quad \Rightarrow \quad N \approx 5 \text{ population-units} .$$

Based on these units, in the simulation study which follows, we assume that $\alpha_2 = \frac{1}{2}$ and $N = 5$, and we repeatedly make use of the time-unit conversion,

$$1.5 \text{ time-unit } = \frac{1}{2} \text{ hr.} \tag{5.5}$$

5.5.2 Cell-Transmission Model

The Cell-Transmission Model (CTM) is a finite-difference approximation to the kinematic wave theory of traffic flow (Daganzo, 1995), which is the model of traffic flow assumed in the Single-Entry Corridor Problem. The CTM provides an optimally spaced discretization structure of space-time into "cells", and sequentially updates the traffic density within each cell as time progresses. DePalma and Arnott, working paper, previously adapted the CTM to the Single-Entry Corridor Problem with great success, as the CTM numerically determines the traffic dynamics for an arbitrary corridor inflow rate, including the determination of queues at the roadway entrance.

The CTM requires the specification of a relationship between flow, q, and density, k, which for Greenshields' Relation is $q(k) = k\left(1 - \frac{k}{k_J}\right)$, where k_J is jam density. The CTM uses the flow-density relationship to construct "sending" and "receiving" functions, which at each time step determine the maximum population which may leave or enter a cell, respectively. These sending and receiving functions are combined to determine the outflow from each cell at each time step.

Under Greenshields' Relation capacity flow and jam density are linearly related as $q_m = \frac{1}{4} v_0 k_J$. Thus, to reduce the capacity within a cell in the CTM, we simply reduce

the jam density by the same proportion in that cell. This enables us in a simple way to adapt the CTM to allow for the introduction of incidents, by reducing the capacity of a cell if an incident is in progress at that cell.

A problem arises regarding a direct implementation of the incident rate model we developed and its implementation into the CTM, since the incident rate model only allows for incidents to occur at the 13 detector locations along the corridor at each aggregated 5-minute interval, yet the space-time discretization in the CTM requires a much finer grid of cells. A naive implementation of the incident rate model to every cell in the CTM results in strongly inflated incident occurrence, since the calibration of the incident rate model is not on the same scale as the CTM cell discretization. To overcome this difficulty in the simulation study we proceed as follows: Choose ten equally spaced points along the corridor at which incidents may occur, and only allow the possible occurrence of an incident every 0.25 time-units (equivalent of 5 minutes).

5.5.3 Simulation Study

We conducted three simulation studies to determine how the introduction of stochastic incidents distorts traffic flow dynamics in the Single-Entry Corridor Problem. In the three studies we employed three different corridor inflow rates, a capacity inflow rate, the UO inflow rate and the SO inflow rate. The corridor ranged from $x = 0$ to $x = 1$, and we only allowed the possibility of incidents to occur at at 10 equally spaced points from $x = 0.5$ to $x = 9.5$. At each of these 10 points and every 0.25 time-units, if an incident was not already in progress at that point, then we used Model (5.2) calibrated with the fitted values in Table 5.4 to determine the occurrence of an incident. In Model (5.2), since the downstream-incident indicator variable was calibrated to check for an incident at a downstream location occurring 30 minutes into the past, for the simulation

study we converted this time to 30 minutes $= 1.5$ time-units into the past. During the simulation study, if an incident occurred at one of the 10 points, then using the current normalized density at that cell-point (normalized by jam density) we stochastically determined the duration and capacity reduction of the incident from the linear regression models for incident duration and capacity reduction fitted in the previous section. We subsequently decreased the jam-density at that cell-point by the same proportion as the determined capacity-reduction, and maintained the decreased jam-density for the determined duration of the incident.

In summary, we allow an incident to occur at only 10 points in the corridor, and model the occurrence of an incident by reducing the jam-density in the corresponding cell for the duration of the incident.

In the Single-Entry Corridor Problem time is normalized so that the first vehicle inflow into the corridor occurs at time $t = 0$, and as a consequence the length of the rush-hour, $\bar{t}$, is endogeneous and is determined by the time of the final vehicle arrival at the CBD. We adopted the same approach during our simulation study, where for each simulation the length of the rush-hour, $\bar{t}$, is determined by the time of the final vehicle arrival at the CBD. If an individual commuter enters the corridor at time t_d and arrives at the CBD at time t_a, then the travel time cost incurred is $t_a - t_d$ (since $\alpha_1 = 1$ in normalized cost units), and the schedule delay cost incurred is $\alpha_2 (\bar{t} - t_a)$, where in the simulation study we chose $\alpha_2 = \frac{1}{2}$.

For each of the three inflow rates (capacity, UO and SO inflow rates), we performed 10,000 simulations. In Table 5.5 we present a summary of the incidents which randomly occurred over the 10,000 simulations. Note that in the SO, for which there is less congestion, we observe fewer observed incidents.

	Capacity	**UO**	**SO**
# incidents	4786	4476	3865
Mean Incident Capacity	0.552	0.549	0.560
Mean Incident Duration	2.94	2.95	3.35
Mean Occupancy Rate at Incident Start	0.406	0.409	0.331

Table 5.5: Summary of incidents occurring in the three simulation studies which used three different inflow rates, capacity inflow, UO inflow and SO inflow. Each simulation study consisted of 10,000 simulations.

In Figures 5.14, 5.15 and 5.16 we present graphs showing the unit travel time cost, unit schedule delay cost and total unit trip cost as functions of departure time for the three inflow rates considered in the simulation study. In these graphs the solid lines show the unit trip costs with no incidents, and the dashed lines show the unit trip costs with the introduction of stochastic incidents, averaged over the 10,000 simulations. In these graphs we observe that vehicles departing at the start of the rush-hour do not experience increased travel times, since they do not experience congestion and, thus, on average do not encounter incidents. In the UO solution, for which total unit trip cost is constant for all departure times, we observe that with the introduction of incidents the average total unit trip cost is larger for vehicles departing later, since these vehicles have a higher likelihood for encountering an incident.

The aggregate trip cost is the sum of the unit trip costs over the entire population, for which the SO solution is a minimum. In Table 5.6 we present the aggregate travel time cost, aggregate schedule delay cost and aggregate total trip cost for the three inflow rates considered in the simulation study. Here we also show each aggregate cost in the absence of incidents, so that we may compare how the introduction of stochastic incidents increases the aggregate costs. Note that with the SO inflow, the aggregate

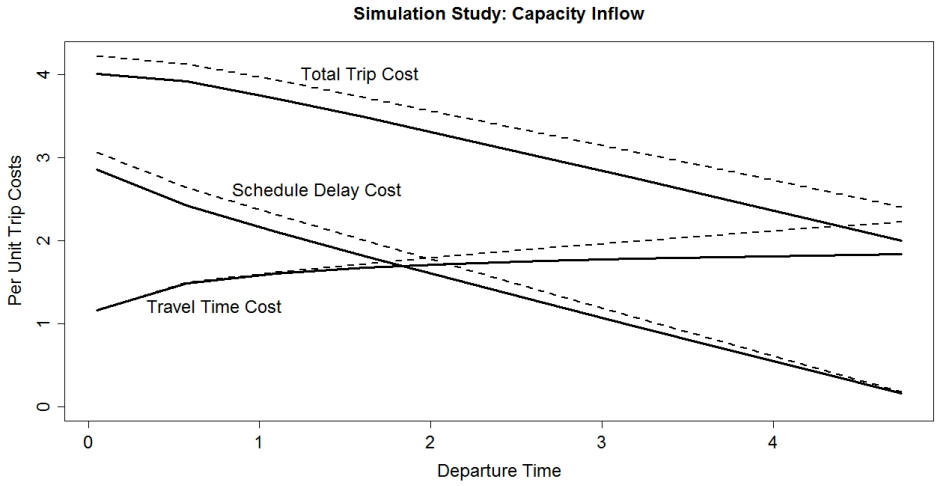

Figure 5.14: Unit travel time, schedule delay, and total trip costs as a function of departure time for capacity inflow into the corridor. The solid lines are the unit trip costs in the absence of incidents, and the dashed lines are the unit trip costs with the introduction of incidents, averaged over the 10,000 trials in the simulation study.

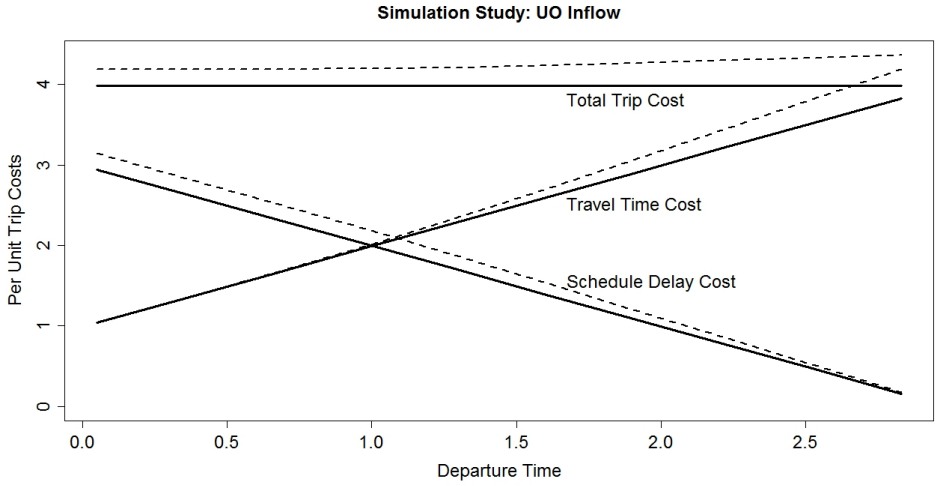

Figure 5.15: Unit travel time, schedule delay, and total trip costs as a function of departure time for UO inflow into the corridor. The solid lines are the unit trip costs in the absence of incidents, and the dashed lines are the unit trip costs with the introduction of incidents, averaged over the 10,000 trials in the simulation study.

costs are less affected with the introduction of stochastic incidents than with the UO inflow or the capacity inflow. We attribute this to the decreased congestion occurring in the SO solution, which leads to a decreased likelihood of incident occurrence.

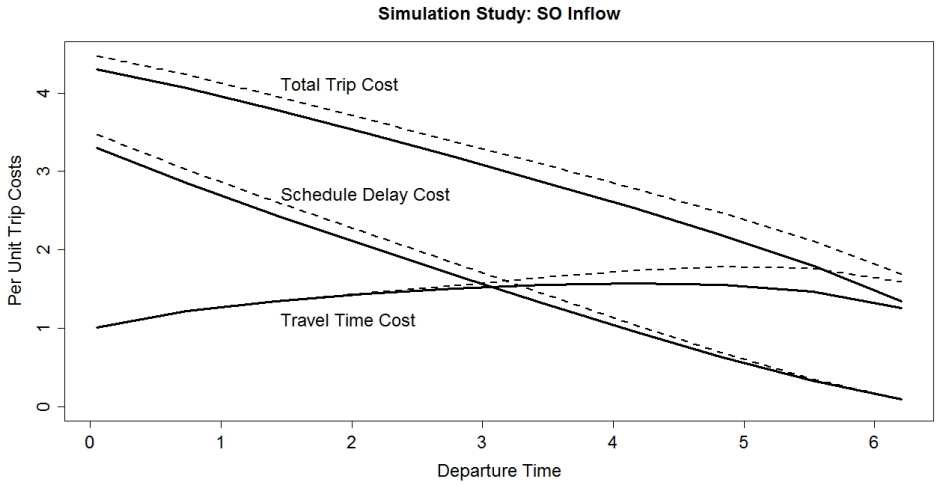

Figure 5.16: Unit travel time, schedule delay, and total trip costs as a function of departure time for SO inflow into the corridor. The solid lines are the unit trip costs in the absence of incidents, and the dashed lines are the unit trip costs with the introduction of incidents, averaged over the 10,000 trials in the simulation study.

	Capacity	**UO**	**SO**
Aggregate Travel Time Cost	$8.44 \rightarrow 9.24$	$13.23 \rightarrow 13.96$	$7.23 \rightarrow 7.81$
Aggregate Schedule Delay Cost	$6.77 \rightarrow 7.43$	$6.69 \rightarrow 7.31$	$7.30 \rightarrow 7.87$
Aggregate Total Trip Cost	$15.22 \rightarrow 16.67$	$19.92 \rightarrow 21.27$	$14.53 \rightarrow 15.68$

Table 5.6: Aggregate travel time, schedule delay, and total trip cost for the three inflow rates considered in the simulation study. The first value is the aggregate cost in the absence of incidents, and the second value is the aggregate cost with stochastic incidents, averaged over the 10,000 trials in each simulation study.

5.5.3.1 Remark Regarding Late Arrivals

The occurrence of an incident may lead to huge delays, potentially resulting in a commuter arriving late to their work-destination. Understanding how commuters alter their departure times to account for severe incidents may be important in understanding rush-hour traffic dynamics. Therefore, to accurately model the occurrence of incidents in the Corridor Problem requires that we allow for late arrivals. The most significant deficiency in our simulation study is that we endogenously determine the work-start time at the CBD to coincide with the final vehicle arrival at the CBD, so that late arrivals

do not occur. Correcting this deficiency requires that we determine UO and SO inflow rates in a Single-Entry Corridor Problem model which allows for late arrivals, and then perform a simulation study of the effect of stochastic incidents in the case when late arrivals are permitted. Although we do not allow for late arrivals in our current study, our work here is a first step towards this goal.

5.6 Concluding Remarks

5.6.1 Directions for Future Research

The Corridor Problem is a very general model of morning rush-hour traffic dynamics, and the Single-Entry Corridor Problem is a simplification of this general model which requires all traffic to enter the corridor at a single location (also known as the "Morning Commute" model). DePalma and Arnott, working paper, have completely characterized the UO and SO solutions in the case of no late arrivals, which is useful to a governing agency seeking to employ a tolling policy so as to minimize congestion on roadways.

A more realistic model of traffic flow would allow for late arrivals at the CBD, and the development of UO and SO solutions for the Single-Entry Corridor Problem in the case when late arrivals are permitted would comprise a complete generalization of the "Bottleneck Model" (Arnott and Lindsey, 1990), allowing for a more realistic model of traffic flow congestion (i.e., kinematic wave dynamics instead of queueing dynamics). It would be straightforward to apply the incident rate model developed in this paper to a simulation study of the UO and SO solutions of the Single-Entry Corridor Problem model which allows for late arrivals, and, thus, provide a more realistic understanding of how incidents affect morning commute patterns. An end goal would be to not only

understand how the introduction of incidents distorts the UO and SO solutions, but to develop equilibrium solutions which account for the occurrence of stochastic incidents. However, before reaching this goal we must first develop UO and SO solutions to the Single-Entry Corridor Problem model for the case in which late arrivals are permitted.

5.6.2 Conclusion

This work has combined several different active research areas, including automatic incident detection, incident rate models, HGLM's and the Corridor Problem. We have succeeded in modifying and employing a recently developed adaptive thresholding, change-point detection algorithm to PeMS traffic flow data as an automatic incident detection algorithm, followed by incorporating the detected incidents into a stochastic incident rate model, fitted using techniques from the theory of HGLM's. We adapted the Cell Transmission Model to allow for stochastic incidents, and then implemented our fitted incident rate model into a simulation study to determine how the introduction of stochastic incidents distorts previously determined economic equilibrium solutions to the Single-Entry Corridor Problem.

Based on the results of our simulation study we observed that commuters departing later face, on average, increased trip costs due to a higher likelihood of incident occurrence, suggesting that commuters will shift their departure patterns to depart earlier so as to avoid potential incidents. Furthermore, we observed that the SO inflow is the less affected by the introduction of incidents than the UO or capacity inflows, due to the decreased congestion in the SO solution. This conclusion suggests that the benefits of a tolling policy to shift traffic dynamics to the SO results in not only decreased, overall traffic congestion, but also results in fewer traffic incidents as well.

References

Ahmed, S. and Cook, A. (1982). Application of time-series analysis techniques to freeway incident detection. *Transportation Research Record*, 841:19–21.

Arnott, R. and Lindsey, R. (1990). Economics of a bottleneck. *Journal of Urban Economics*, 27(1):111–130.

Arnott, R. and DePalma, E. (2011). The corridor problem: preliminary results on the no-toll equilibrium. *Transportation Research Part B*, 45, 743–768.

Byrdia, R., Johnson, J. and Balke, K. (2005). An investigation into the evaluation and optimization of the automatic incident detection algorithm used in TxDOT traffic management systems. Texas Transportation Institute, Report 0-4770-1, Project 0-4770.

Carvell, J., Balke, K., Ullman, J., Fitzpatrick, K., Nowlin, L. and Brehmer, C. (1997) Freeway management handbook. Report No. FHWA-SA-97-064. U.S. Department of Transportation, FWHA, Washington D.C.

Chang, E. and Wang, S. (1994). Improved freeway incident detection using fuzzy set theory. *Transportation Research Record*, 1453:75–82.

Cook, A. and Cleveland, D. (1974). Detection of freeway capacity-reducing incidents by traffic-stream measurements. *Transportation Research Record*, 495:1–11.

Daganzo, C. (1995). A finite difference approximation of the kinematic wave model of traffic flow. *Transportation Research Part B*, 29(4):261–276.

DePalma, E. and Arnott, R. working paper. Morning commute in a single-entry traffic corridor with no late arrivals. Submitted to *Transportation Research Part B*.

Giuliano, G. (1989). Incident characteristics, frequency, and duration on a high volume urban freeway. *Transportation Research Part A*, 23A(5):387–396.

Golob, T., Recker, W. and Alvarez, V. (2004) Freeway safety as a function of traffic flow. *Accident Analysis and Prevention*, 36:933-946.

Golob, T., Recker, W. and Pavlis, Y. (2008). Probabilistic models of freeway safety performance using traffic flow data as predictors. *Safety Science*, 46:1306–1333.

Henderson, C. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*, 31:423–447.

Lambert, D. and Liu, C. (2006). Adaptive thresholds: monitoring streams of network counts. *Journal of the American Statistical Association*, 101:78–88.

Lee, Y. and Nelder, J. (1996). Hierarchical generalized linear models (with discussion). *J. Roy. Statist. Soc. B*, 58:619–678.

Lee, Y. and Nelder, J. (2001). Hierarchical generalized linear models: a synthesis of generalized linear models, random-effect models and structured dispersion. *Biometrika*, 88:987–1006.

Lee, Y. and Nelder, J. (2001). Modelling and analysing correlated non-normal data. *Statistical Modelling*, 1:3–16.

Lee, Y. and Nelder, J. (2005). Likelihood for random-effect models. *SORT*, 29(2):141–164.

Lee, Y., Nelder, J. and Pawitan, Y. (2006). Generalized linear models with random effects: unified analysis via h-likelihood. *Chapman and Hall, CRC*, Boca Raton, FL USA.

Lighthill, M.H. and Whitham, G.B. (1955). On kinematic waves II: a theory of traffic flow on long crowded roads. *Proc. Roy. Soc. (Lond.)*, 229(1178), 317–345.

Lord, D. and Mannering, F. (2010). The statistical analysis of crash-frequency data: a review of methodological alternatives. *Transportation Research Part A*, 44:291–305.

Meng, X. (2009). Decoding the h-likelihood. *Statistical Science*, 24(3):280–293.

Montes De Oca, V., Jeske, D., Zhang, Q., Rendon, C. and Marvasti, M. (2010). A cusum change-point detection algorithm for non-stationary sequences with application to data network surveillance. *The Journal of Systems and Software*, 83(7):1288–1297.

Nelder, J. and Lee, Y. (1991). Generalized linear models for the analysis of Taguchi-type experiments. *Appl. Stochastic Models Data Anal.*, 7:107–120.

Newell, G.F. (1988). Traffic flow for the morning commute. *Transportation Science*, 22(1), 47–58.

Oh, J., Oh C., Ritchie, S. and Chang, M. (2005) Real-time estimation of accident likelihood for safety enhancement. *Journal of Transportation Engineering*, 131(5):358–363.

Payne, H. and Tignor, S. (1978). Freeway incident-detection algorithms based on decision trees with states. *Transportation Research Record*, 682:30–37.

Parkany, E. (2005). A complete review of incident detection algorithms and their deployment: what works and what doesn't. New England Transportation Consortium, Project No. 00-7.

PB Farradyne, Inc. (2010) Traffic incident management handbook. Federal Highway Administration, Office of Travel Management, Washington D.C., 20590 USA.

Persuad, B., Hall, F. and Hall, L. (1990). Congestion identification aspects of the McMaster incident detection algorithm. *Transportation Research Record*, 1287:167–175.

Petty, K. (1997). Incidents on the freeway: detection and management. P.h.D. dissertation, Department of Engineering, University of California, Berkeley.

Richards, P.I. (1956). Shock waves on the highways. *Operations Research*, 4(1), 42–51.

Sinha, P., Hadi, M. and Wang, A. (2006). Modeling reductions in freeway capacity due to incidents in microscopic simulation models. *Transportation Research Board*, 62–68.

Small, K., Winston, C. and Yan, J. (2005). Uncovering the distribution of motorists' preferences for travel time and reliability. *Econometrica*, 73(4):1367–1382.

Smith, B., ASCE, M., Qin, L. and Venkatanarayana, R. (2003). Characterization of freeway capacity reduction resulting from traffic accidents. *Journal of Transportation Engineering*, 129(4):362–368.

Stephanedes, Y. and Liu, X. (1995). Artificial neural networks for freeway incident detection. *Transportation Research Record*, 1494:91–97.

Urban Crossroads, Inc. (2006). PeMS data extraction methodology and execution technical memorandum for the southern California association of goverments.

Wang, W., Chen, H. and Bell, M. (2005). A review of traffic incident duration analysis. *Journal of Transportation Systems Engineering and Information Technology*, 5(3):127–140.

Wolfinger, R., Tobias, R. and Sall, J. (1994). Computing gaussian likelihoods and their derivatives for general linear mixed models. *SIAM J. Sci. Comput.*, 15(6):1294–1310.

Appendix

5.6.3 SAS Code

The following SAS code, described in Section 5.4.2, was used to fit logistic models with random detector effects to small subsets of the adjusted data set:

```
proc glimmix data=SubsetData;
       class detector;
       model incident(event=last) = norm_flow norm_flow2 downstream_incident
                                    / s  dist=binary;
       random detector / s ;
run;
```

5.6.4 R Code

The following R code, described in Section 5.4.4, fits the incident rate model to the adjusted data set using HGLM fitting techniques. The code repeatedly calls *gamma.glm.fit.R*, which fits an intercept-only gamma GLM (with log-link function) using IWLS.

5.6.4.1 HGLM Fit

```
setwd("C:\\Users\\Elijah DePalma\\Documents\\Incident Model -
Dissertation\\HGLM Model")

library(DAAG)
rm(list=ls())

###################################
#Format data

in.data <- as.matrix(read.table(
               file="IncidentData_Adjusted3.txt", header=TRUE))
#"detector" "day" "time" "norm.occ" "norm.occ2" "norm.flow"
#     1  2   3   4   5   6
# "norm.flow2" "downstream.incident" "incident"
#        7   8   9
#Incident = 1 corresponds to a pre-alarm point
#Downstream.incident = 1 if an incident
# immediately downstream within 1/2 hr
```

```
data <- in.data

########################################
#Joint Mean-Dispersion Fitting Procedure for Augmented-GLM
# and Dispersion-GLM

#Response
Y <- data[,9]

#Design matrices
X <- cbind(
           rep(1,nrow(data))   #intercept
           ,data[,6]           #flow
           ,data[,7]           #flow2
           ,data[,8]           #downstream.incident
           )

frequencies <- numeric(13) #13 detector random effects
for (i in 10:22){ frequencies[i-9] <- length(which(data[,1]==i)) }
Z <- c( rep(1,frequencies[1]), rep(0,nrow(X)-frequencies[1]) )
zeros <- frequencies[1]
for (i in 2:12){
    Z <- cbind(Z, c( rep(0,zeros),rep(1,frequencies[i]),
                     rep(0,nrow(X)-zeros-frequencies[i]) ) )
    zeros <- zeros + frequencies[i]
}
Z <- cbind(Z, c(rep(0,zeros), rep(1,frequencies[13])))

#Construct Augmented-GLM Model
#psi.M is mean/expected value of random effects
psi.M <- rep(0,ncol(Z))
Y.aug = c( Y, psi.M )

T <- rbind(cbind(X, Z),
           cbind( matrix(nrow=ncol(Z),ncol=ncol(X),
                         rep(0,ncol(Z)*ncol(X))),
                  diag(rep(1,ncol(Z))) ) )

#GLM: Bernoulli responses with log-link
variance <- function(mu){ mu * (1-mu) }
linkfun <- function(mu){ log( mu / (1-mu) ) }
linkinv <- function(eta){ pmax(exp(eta), .Machine$double.eps) /
                               (1+ pmax(exp(eta), .Machine$double.eps))}
deviances.scaled <- function(y,mu){
        2*(ifelse(y==0,log(1/(1-mu)),log(1/mu)))}
mu.eta <- function(eta){  pmax(exp(eta), .Machine$double.eps) /
        ( 1+pmax(exp(eta), .Machine$double.eps) )^2 }
```

```
#GLMM: Identity-Normal random effects
variance.M <- function(u){ 0*u + sig2r.actual }
linkfun.M <- function(u){ u }
linkinv.M <- function(r){ r }
deviances.scaled.M <- function(y,mu){ (y-mu)^2 }
u.r.M <- function(r){ 0*r + 1 }

#Initialize response means, .25 for 0 and .75 for 1
mu <- (Y+.5)/2
eta <- linkfun(mu)
#Initialize covariance parameter using fit from gamma-GLM
source("gamma.glm.fit.R")
dispersion.fit <- gamma.glm.fit( deviances.scaled(Y,mu),
         rep(1/2,length(Y)) )
sig2r.old <- exp( dispersion.fit$coefficients )
sig2r <- sig2r.old

############################
#Alternate between Steps (i),(iii) until covariance parameter fitted
for (outer.iter in 1L:100){

    #Step (i): Given covariance parameter, fit
    # Augmented-GLM for beta, r
    devold <- sum( deviances.scaled(Y,mu) )
    for (iter in 1L:25){
        mu.eta.val <- mu.eta(eta)
        z <- eta + (Y-mu)/mu.eta.val
        z.M <- psi.M
        z.aug <- c( z,z.M )

        varmu <- variance(mu)
        w <- (mu.eta.val^2)/varmu
        w.M <- rep( 1/sig2r, ncol(Z) )
        w.aug <- c( w,w.M )
        Sigma <- 1/w.aug

        nobs <- NROW(z.aug); nvars <- ncol(T)
        fit <- .Fortran("dqrls", qr = sqrt(1/Sigma) * T, n = nobs,
                        p = nvars, y = sqrt(1/Sigma) * z.aug, ny = 1L,
                        tol = min(1e-12),
                        coefficients = double(nvars),
                        residuals = double(nobs),
                        effects = double(nobs),
                        rank = integer(1L), pivot = 1L:nvars,
                        qraux = double(nvars),
                        work = double(2 * nvars),
                        PACKAGE = "base")

        omega <- fit$coefficients
```

```
        beta <- omega[1:ncol(X)]
        r <- omega[(ncol(X)+1):(ncol(X)+ncol(Z))]
        eta.aug <- drop( T %*% omega )
        eta <- eta.aug[1:(nrow(X))]
        mu <- linkinv(eta)

        deviances <- deviances.scaled(Y,mu)
        deviances.M <- deviances.scaled.M(psi.M,r)
        dev <- sum( c(deviances,deviances.M) )
        if (abs(dev - devold)/(0.1 + abs(dev)) < 1e-8) { break }
        else { devold <- dev }
    }

    #Step (iii): Use deviances to fit gamma-GLM
    # for covariance parameter
    #Leverages from last fit above
    qrs <- structure(fit[c("qr","qraux","pivot","tol","rank")],
             class="qr")
    q <- hat(qrs,intercept=FALSE)
    q.M <- q[(nrow(X)+1):(nrow(X)+ncol(Z))]

    #Responses for gamma-GLM model
    dstar.M <- deviances.M/(1-q.M)
    dispersion.fit <- gamma.glm.fit( dstar.M, weights=(1-q.M)/2)

    #Updated covariance parameter
    sig2r <- exp( dispersion.fit$coefficients )

    #Track inner-iterations per each outer-iteration
    if (outer.iter==1) { inner.iterations <- c( iter ) }
    if (outer.iter>1) { inner.iterations <- c( inner.iterations,
                                                    iter ) }

    if (abs(sig2r - sig2r.old)/abs(sig2r) < 1e-4 ) { break }
    else { sig2r.old <- sig2r }

}

########################################
#Iterative fitting procedure completed
# Now calculate standard errors

#Calculate standard errors of omega: diagonals of
# sqrt( (T' Sigma^(-1) T)^(-1) )
covmat.unscaled <- chol2inv(fit$qr[1L:nvars, 1L:nvars, drop = FALSE])
sd.omega <- sqrt( diag( covmat.unscaled ) )
sd.beta=sd.omega[1:ncol(X)]
sd.r=sd.omega[(ncol(X)+1):(ncol(X)+ncol(Z))]
```

```
#Calculate standard error of dispersion parameter estimate, sig2r,
# using aphl

A11 <- matrix(nrow=ncol(X),ncol=ncol(X))
for (l in 1:ncol(X)){
    for (lp in 1:ncol(X)){
        A11[l,lp] <- sum( mu*(1-mu)*X[,l]*X[,lp] )
    }
}

A22<- matrix(nrow=ncol(Z),ncol=ncol(Z))
for (m in 1:ncol(Z)){
    for (mp in 1:ncol(Z)){
        A22[m,mp] <- sum( mu*(1-mu)*Z[,m]*Z[,mp] )
    }
}
A22 <- A22 + 1/sig2r*diag( ncol(Z) )

A12 <- matrix(nrow=ncol(X),ncol=ncol(Z))
for (l in 1:ncol(X)){
    for (m in 1:ncol(Z)){
        A12[l,m] <- sum( mu*(1-mu)*X[,l]*Z[,m] )
    }
}

A21 <- t(A12)

D <- rbind( cbind( A11, A12 ),
            cbind( A21, A22 ) )

A11.inv <- solve(A11)
A22.inv <- solve(A22)
C1 <- A11 - A12 %*% A22.inv %*%A21
C2 <- A22 - A21 %*%A11.inv %*%A12
C1.inv <- solve(C1)
C2.inv <- solve(C2)

D.inv <- rbind( cbind( C1.inv, -A11.inv %*% A12 %*% C2.inv ),
                cbind( -C2.inv %*% A21 %*% A11.inv, C2.inv ) )

dD.dsig2r <- -1/sig2r^2 *
    rbind( matrix(nrow=ncol(X),ncol=(ncol(X)+ncol(Z)),
                  rep(0,ncol(X)*(ncol(X)+ncol(Z)))),
           cbind( matrix(nrow=ncol(Z),ncol=ncol(X),
                    rep(0,ncol(X)*ncol(Z))),
                  diag(ncol(Z)) ) )
d2D.dsig2r2 <- 2/sig2r^3 *
    rbind( matrix(nrow=ncol(X),ncol=(ncol(X)+ncol(Z)),
                  rep(0,ncol(X)*(ncol(X)+ncol(Z)))),
```

```
          cbind( matrix(nrow=ncol(Z),ncol=ncol(X),
                    rep(0,ncol(X)*ncol(Z))),
                 diag(ncol(Z)) ) )

sd.sig2r <- 1 / sqrt( sum(r^2)/sig2r^3 +
                      1/2*sum(diag( D.inv%*%d2D.dsig2r2 -
                                    D.inv%*%dD.dsig2r%*%
                                    D.inv%*%dD.dsig2r )) )

#Model results showing # of iterations,
# fitted fixed coefficients and their
# standard errors, fitted random effects and their standard errors
sig2r.fit <- sig2r; sd.sig2r.fit <- sd.sig2r
beta.fit <- beta; sd.beta.fit <- sd.beta
r.fit <- r; sd.r.fit <- sd.r
Results <- list(outer.iterations=outer.iter,
                inner.iterations=inner.iterations,
                sig2r=cbind(sig2r.fit, sd.sig2r.fit),
                beta=cbind(beta.fit, sd.beta.fit),
                r=cbind(r.fit, sd.r.fit))
print(Results)
```

5.6.4.2 Gamma GLM Fit

The R-code for *gamma.glm.fit.R*, which fits an intercept-only gamma GLM model (with log-link), and which is repeatedly called by the HGLM fitting algorithm whose code is given in the previous section.

```
#Fit an intercept only gamma-GLM with log-link and with prior weights

gamma.glm.fit <-
    function (y, weights=NULL)
{
    x <- as.matrix( rep(1, NROW(y)) ) #Intercept only model

    #GLM: Gamma responses with log-link
    variance <- function(mu){ mu^2 }
    linkfun <- function(mu){ log(mu) }
    linkinv <- function(eta){ pmax(exp(eta), .Machine$double.eps) }
    deviances.scaled <- function(y,mu){
        2 * (y * log(ifelse(y == 0, 1, y/mu)) - (y - mu)/mu) }
    mu.eta <- function(eta){ pmax(exp(eta), .Machine$double.eps) }

    #Initialize means using observed values
    mu <- y
```

```
    eta <- linkfun(mu)

    #Begin IWLS iterative procedure
    if (is.null(weights)){ weights <- rep.int(1, nobs) }
    devold <- sum( deviances.scaled(y, mu) )
    for (iter in 1L:25) {
        mu.eta.val <- mu.eta(eta)
        z <- eta + (y - mu)/mu.eta.val
        varmu <- variance(mu)
        w <- (mu.eta.val^2)/varmu
        Sigma <- 1/weights * 1/w

        nobs <- NROW(y); nvars <- ncol(as.matrix(x))
        fit <- .Fortran("dqrls", qr = sqrt(1/Sigma) * x, n = nobs,
                        p = nvars, y = sqrt(1/Sigma) * z, ny = 1L,
                        tol = min(1e-8),
                        coefficients = double(nvars),
                        residuals = double(nobs),
                        effects = double(nobs),
                        rank = integer(1L), pivot = 1L:nvars,
                        qraux = double(nvars),
                        work = double(2 * nvars),
                        PACKAGE = "base")

        beta <- fit$coefficients
        eta <- drop( as.matrix(x) %*% beta )
        mu <- linkinv(eta)

        dev <- sum( deviances.scaled(y,mu) )
        if (abs(dev - devold)/(0.1 + abs(dev)) < 1e-8) { break }
        else { devold <- dev }
    }


    list(coefficients = beta)
}
```

Chapter 6

Conclusions

† In Aqua Sanitas (In Water Is Health)

An Avocado Farmer's Solution

"He crisply drove through the orchard, closely monitoring his smartphone to provide him with the GPS coordinates of the next tree to sample. The smartphone application he used required him to input only the GPS coordinates of the corners of his orchard, and using this information it provided him with a sequential sampling strategy, at each stage indicating the GPS coordinates of the next tree to be sampled. The smartphone application indicated how many leaves from each tree must be sampled, and after inputing the sampling results from a tree it provided confidence limits for the mean pest density, as well as providing a suggestion to either treat the orchard, not treat the orchard, or continue sampling. He smiled as he drove, knowing that within just a few short hours he would have reliable information to use regarding a treatment decision."

A Free-Flowing Morning Commute

"Whistling with the radio, she drove steadily along the roadway at a smooth pace. Her clock read 8:20am, and she smiled knowing that she would arrive to work within 20 minutes. Traffic had been much smoother ever since the introduction of a morning toll to enter the roadway, and the toll for her to enter by 8:10am was only $1, which was negligibly small compared to the $10 toll required to enter the roadway at 8:30am. Nowadays she rarely encountered traffic congestion, and the occurence of an accident was a truly rare event. Plus, the additional 20 minutes she now had before beginning her workday gave her just enough time to enjoy a cup of tea with her co-workers and settle into a beautiful day. "

The research presented in Ch. 2 is the first statistical application of spatial analyses coupled with sequential sampling for the development of a sampling plan for pest management. Our proposed presence-absence sampling methodology for *O. perseae* evaluates a sequential hypothesis test of pest population densities which, 1) accounts for aggregation of pest populations on individual trees, and 2) mitigates spatial correlation of pest populations on adjacent trees using a tree-selection rule which sequentially selects trees to be maximally spaced from all other previously selected trees. Although the results of our simulation study demonstrate the effectiveness of our presence-absence sampling methodology for parameter estimates relevant to *O. perseae*, the methodology can easily be applied to other pests, and even other non-pest spatial sampling situations.

With further research involving field validation, our sampling methodology has the potential to be customized as a reliable decision-making tool for pest control advisers and growers to use for control of the *O. perseae* mite in commercial avocado orchards. To meet this goal, software would be needed to help a pest manager with tree selection and with evaluating the treatment decision hypothesis test at each sequential step. With the widespread ownership and use of smart phones, sampling programs like that developed here could be made available as a downloadable "app." Because the popularity of smart phone apps is increasing, a well-developed app that is attractive in appearance and easy to use may help greatly with the adoption of sampling plans, like that developed here for *O. perseae*.

Future work might include a more sophisticated per tree sampling cost which varies during the sequential sampling process to account for both the distance and the land topography between subsequently sampled trees, which may be of interest to a pest manager seeking to minimize their distance traveled and seeking to avoid sampling from trees which are difficult to reach (e.g., trees on steep hillsides). Additionally, the spatial

GLMM model of pest populations which we employed assumes that pest individuals are distributed randomly within a tree, and that correlations of pest populations on adjacent trees are spatially symmetric. Future research on sequential sampling with spatial components which extends beyond these model assumptions may address issues pertaining to pest populations that are systematically distributed within trees, and may include anisotropic (i.e., asymmetric) correlation structures of pest populations, allowing for stronger correlation along orchard edges or within orchard rows.

The work in Ch. 3 takes a preliminary step in developing a theory of the spatial dynamics of metropolitan traffic congestion. Even though the no-toll equilibrium Corridor Problem assumes away many features of a much more complex reality, it appears very difficult to solve. We have not yet succeeded in obtaining a complete solution to the problem, and this work presented preliminary results. We have succeeded in providing a solution for the no-toll equilibrium for a restricted family of population distributions along the traffic corridor. What remains to be done is to obtain a complete solution for an arbitrary population distribution. Either an equilibrium does not exist for arbitrary population distributions, or our specification of the problem is incomplete, or our solution method somehow overconstrains the problem. We have been unable to uncover the root of the difficulty, which is why we are unable to provide a complete solution.

In Ch. 4 we build upon earlier work by Newell which replaces a bottleneck with a single-entry corridor of uniform width that is subject to LWR flow congestion, for which the bottleneck is a limiting case. Our work considered a special case of Newell's model in which local velocity is a negative linear function of local density, and all commuters have a common desired arrival time at the central business district. These simplifying assumptions permitted a complete closed-form solution for the social optimum and a quasi-analytical solution for the user optimum. Providing detailed derivations and ex-

ploring the model's economic properties added insight into rush-hour traffic dynamics with this form of congestion, and into how the dynamics differ from those of the bottleneck model. Future work may treat late, as well as early, arrival, a road of non-uniform width, merges and capacity constraints at the CBD. Once the non-uniform road and merge problems have been solved, the stage will be set to tackle the difficult problems of determining the SO and UO with LWR flow congestion on a general corridor and a general network.

In Ch. 5 we modify a recently developed, online change-point detection algorithm used to detect unusual decreases in the number of users on a network server, to an automatic incident detection algorithm to detect the occurrence of incidents (i.e., unusual congestion spikes) in freeway traffic data. We apply the incident detection algorithm to morning, freeway traffic data for a San Diego freeway, and then use the detected incidents to calibrate an incident rate model using fitting techniques from the theory of Hierarchical Generalized Linear Models. Implementing the algorithms for fitting the incident rate model is a great success, and the same algorithms may be applied to fit other statistical models. Lastly, we implemented the calibrated incident rate model into a simulation study (via the Cell Transmission Model) of economic equilibrium solutions to the Single-Entry Corridor Problem. Our simulation results confirm what may be theoretically anticipated, namely, that the introduction of stochastic incidents leads to increased average trip costs for individual users departing later in the rush-hour, and that the occurrence of incidents is less frequent in less congested traffic situations. However, to provide a proper study of the effect of stochastic incidents on traffic commute patterns requires that we allow for late arrivals, and we strongly recommend that future work in this area focus on developing equilibrium solutions to the Corridor Problem which allows for late arrivals.