The Report Committee for Natalia Zuniga Garcia
certifies that this is the approved version of the following report:

Spatial Pricing Empirical Evaluation of Ride-Sourcing Trips Using the Graph-Fussed Lasso for Total Variation Denoising

Committee:

James G. Scott, Supervisor

Randy B. Machemehl, Co-Supervisor

Spatial Pricing Empirical Evaluation of Ride-Sourcing Trips Using the Graph-Fussed Lasso for Total Variation Denoising

by

Natalia Zuniga Garcia

Report

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Science in Statistics

The University of Texas at Austin

May 2018

To my parents.

Acknowledgments

I would like to express my sincere gratitude to Dr. Randy B. Machemehl, my Ph.D. advisor, for his unconditional support and guidance in this project and during my doctoral studies. Dr. Machemehl has been an inspiration for me, and I am very grateful for the opportunity he gave me to work with him.

I also would like to thank Dr. James G. Scott, my master's supervisor, for giving me the chance to collaborate with him. I sincerely appreciate his constant guidance, support, and motivation.

I would like to acknowledge Mauricio Garcia-Tec for his valuable contributions to this project. Working with Mauricio has been one of the best experience during my graduate school. He has provided me with so much knowledge and guidance that more than a classmate he has become a teacher.

I also want to acknowledge Wesley S. Tansey for providing us with the code for his algorithm.

Additionally, I want to thank my friends and family in Costa Rica, for showing their support from the distance and for visiting and cheering me up always. I am also grateful to all my new friends in Austin because it wouldn't be the same without them.

Finally, I want to thank my parents and brother for their love and support. I owe all my accomplishments in life to them.

Spatial Pricing Empirical Evaluation of Ride-Sourcing Trips Using the Graph-Fussed Lasso for Total Variation Denoising

Natalia Zuniga Garcia, M.S.Stat.
The University of Texas at Austin, 2018

Supervisor: James G. Scott
Co-Supervisor: Randy B. Machemehl

This study explores the spatial pricing discrimination of ride-sourcing trips using empirical data. We use information from more than 1.1 million rides in Austin, Texas, provided by a non-profit transportation network company from a period where the main companies were out of the city. We base the analysis on operational variables such as the waiting or idle time between trips, reaching time, and trip distance. Also, we estimate three different productivity measures to evaluate the impact of the trip destination on the driver continuation payoff.

We propose the application of a total variation denoising method that enhances the spatial data interpretation. The selected methodology, known as the

graph-fussed lasso (GFL), uses an ℓ_1-norm penalty term that presents a variety of benefits to the denoising process. Specifically, this approach provides local adaptivity; it can adapt to inhomogeneity in the level of smoothness across the graph. Solving the GFL smoothing problem involves convex-optimization methods, we make use of a fast and flexible algorithm that presents scalability and high computational efficiency.

The principal contributions of this research effort include a temporal and spatial evaluation of different ride-sourcing productivity measures in the Austin area, an analysis of ride-sourcing trip pricing and its effect on driver equity, and a description of the principal ride-sourcing travel patterns in the city of Austin. The main results suggest that drivers with rides ending in the central area present favorable spatial differences in productivity when including the revenue of two consecutive trips. However, the time effect was more contrasting. Weekend rides tend to provide better driver productivity measures.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Ride-sourcing companies, also known as transportation network companies (TNCs) or ride-hailing, provide pre-arranged or on-demand transportation service for compensation [Shaheen et al., 2016]. They operate as a two-side market connecting drivers of personal vehicles with passengers. Drivers work as independent contractors, with the flexibility to drive on their own schedule. Some TNCs allow ride-sharing trips in their platforms, where multiple passengers share a ride with a similar destination at a lower cost. For example, Uber and Lyft provide services called UberPOOL and Lyft Line, respectively, explicitly used for ride-share trips.

TNCs have been involved in controversies in different cities around the world due to multiple factors, such as taxi service unfair competition, and lack of regulation of their pricing system and driver selection. Specifically, the pricing strategy has been criticized due to concerns for the welfare of providers and consumers [Cachon et al., 2017]. The price setting in ride-sourcing markets is important because it has a crucial impact on the availability of resources (drivers) and the demand from passengers. A known challenge is the incentive equity across drivers. Trips may be mispriced relative to other trip opportunities [Ma et al., 2018], which leads to a concern for fairness among drivers. For this reason, TNCs implemented different

compensation or subsidization policies – such as reduced commissions, or bonuses for meeting a certain number of rides [Bogage, 2016] – to avoid supply shortage and beat the competition among platforms.

The design of pricing strategies in the ride-sourcing market is challenging. Driver revenue is determined by a fare scheme that is similar to the taxi market, with a base fare and price variating based on the time and distance of the trip. However, unlike taxi driver, ride-sourcing driver revenue variates based on factors like pooled or shared-rides, company compensation policies, and the price surge factor multiplier, determined based on temporal and spatial supply-demand unbalance.

Recent ride-sourcing research has been focusing on analysis of the surge price [Zha et al., 2017a, Wang et al., 2016, Banerjee et al., 2015] and its effect on labor supply [Chen and Sheldon, 2016, Sheldon, 2015]. The spatial analysis research is limited. Recent evaluations include spatial pricing analysis [Ma et al., 2018, Bimpikis et al., 2016, He et al., 2018]. However, the main limitation is the accessibility to empirical data. Also, there is a lack of studies evaluating the drivers perspective. Therefore, the present research intends to contribute to the current ride-sourcing literature providing a spatial pricing evaluation of trips, using empirical data from an Austin-based TNC, with emphasis on the equity among drivers.

1.1 Goal and Objectives

The principal objective of this research is to evaluate the spatial pricing discrimination of ride-sourcing trips and its effect on the competitive equilibrium among drivers. Specifically, we analyze the spatial impact on the drivers' continuation payoff, evaluating productivity factors such as waiting time, reaching time, and trip longitude. We use the data that a non-profit Austin-based ride-sourcing company made available, including trips during the period that Uber and Lyft were temporar-

ily out of the city [1].

1.2 Methodology

The first stage of the methodology includes data processing and mining. We selected different measures as indicators of drivers productivity to evaluate its spatial effect. Also, we discretized in *time* based on weekday peak and off-peak hours and weekends, and in *space* using the traffic analysis zones (TAZ), the unit of geography most commonly used in conventional transportation planning models.

Then, we smoothed (denoised) the data over the physical distances, using a *graph total variation (TV) denoising* technique, to compensate for inherent sampling noise and enhance the interpretability. The data presented diverse complexity including sparsity and a heterogeneous inherent space, because of the distribution of high and low trip demand zones. The city of Austin presents areas with high trip density – such as downtown – and inter-connected zones with low demand. Thus, a global smoothing approach is desirable. These methods include the long-range dependencies, while local procedures, such as the *Gaussian Kernel* or *K nearest neighbor*, tend to smooth over a specific window and do not account for the complete available data.

We selected the *graph-fussed lasso (GFL)* method for the TV denoising. This technique allows one to globally smooth anisotropic and discrete areas using an ℓ_1-norm penalty term[2]. The effect of the ℓ_1 penalty is that it enables local adaptivity, i.e. it can adapt to inhomogeneity in the level of smoothness of an observed signal across the graph. It can set many high-order differences to zero exactly and leave others at large nonzero values [Wang et al., 2015]. Thus, the approximated "true"

[1] Uber and Lyft left the city from May 2016 to May 2017 after the Austin City Council passed an ordinance requiring ride-hailing companies to perform fingerprint background checks on drivers, a stipulation that already applies to Austin taxi companies [Samuels, 2017].

[2] Formally, the ℓ_p-norm of x is defined as: $\|x\|_p = \sqrt[p]{\sum_i |x_i|^p}$, where $p \in \mathbb{R}$.

denoised signal can simultaneously be smooth in some parts of the graph, and wiggly in others. In contrast to other global methods, such as the *Laplacian*, which uses an ℓ_2 penalty and enforces a much heavier contrast, thus tends to estimate either smooth or else wiggly throughout [Wang et al., 2015]. A relevant analogy for the differences between ℓ_1 and ℓ_2 regularizations is the comparison between the lasso and ridge regression.

The main disadvantage of global approaches is that they are slower than the local ones and fail to scale to large graphs. Specifically, solving the GFL smoothing problem involves convex-optimization methods that tend to require a high computational cost. We applied a *fast and flexible algorithm*, developed by Tansey and Scott [2015], that is scalable and highly efficient. The algorithm decomposes the graph into a set of trails which can each be solved efficiently using techniques for the ordinary (1D) fused lasso.

The denoised information is then analyzed to evaluate the space and time effect on driver productivity and provide insights of the principal operational variables in ride-sourcing trips in Austin.

1.3 Contributions

The main contributions of this research effort include: (1) temporal and spatial evaluation of different ride-sourcing productivity measures in the Austin area, including waiting time, and reaching time, among others; (2) analysis of ride-sourcing trip pricing and its effect on driver equity; (3) description of the principal ride-sourcing travel patterns in the city of Austin; and (4) verification of the usefulness of big data analytics in transportation problems.

1.4 Outline

The subsequent sections of the paper are organized as follow. Section 2 provides a literature review of the principal aspects of pricing, labor supply, and spatial analysis in ride-sourcing markets. Section 3 presents an introduction to graph smoothing techniques and a description of the selected graph fused lasso solver algorithm. Section 4 describes the dataset and includes the methodology. Section 5 presents the principal results. Finally, Section 6 contains conclusions and final remarks.

Chapter 2

Literature Review

This chapter presents a literature review of the research focused on (i) pricing strategies, and (ii) spatial pricing, for ride-sourcing trips.

2.1 Price Strategies

The popularity of ride-sourcing platforms relies not only on the advanced technology of connecting users and providers through cell phone applications but also on the pricing strategies. The price-setting analysis is complex because, as a two-sided market, it requires economic models that capture the incentives of both driver and passengers [Banerjee et al., 2015]. A key tool used by TNCs is the dynamic (or surge) price, that help in managing both supply and demand. The surge-price can be defined as the output of an algorithm which automatically raises the cost of a trip when demand outstrips supply within a fixed geographic area [Chen and Sheldon, 2016]. The literature in price strategies encompasses the analysis of surge prices and its effect on labor supply.

2.1.1 Surge Price

The recent research in price strategies has been focused on analyzing the surge price. Banerjee et al. [2015] studied the optimal pricing strategies using a queuing model to explore the effect of surge pricing on the platform performance, defined by throughput and revenue. Among the principal conclusions, the authors found that the surge price is more robust to fluctuations on the system parameters such as arrival rates, service rates, and preference distributions of passengers and drivers, compared to the static pricing. Castillo et al. [2017] studied the motivations of the surge price. They found that when prices are too low a "perverse equilibrium" – called the *wild goose chase* – arises. In this equilibrium, drivers spend too much time picking up passengers instead of driving them or waiting to be matched, resulting in a low number of idle drivers and thus completing a vicious circle. The authors found that the surge price is a natural tool to avoid this issue. Cachon et al. [2017] analyzed the role of surge pricing of self-scheduling platforms and found that both, providers and consumers, benefit from the surge price because providers are better utilized and passengers benefit from lower rates during regular hours.

2.1.2 Labor Supply

A broader research area has been dedicated to studying pricing policies and its effect on the labor supply. Specifically, some authors focused on evaluating two labor supply theories: neo-classic and income-target. The neo-classic labor supply theory predicts a positive response to transitory changes in wages. Thus, a driver is expected to work more extended hours as average hourly earnings get higher in shorter horizons. Camerer et al. [1997] introduced the income-target theory based on an empirical study of New York taxi drivers. The authors found negative elasticities – where, as earnings increase, the total driving time decreases – and argued that taxi drivers have a reference point of income that influences the daily labor supply

decision.

In the context of ride-sourcing trips, labor supply elasticities have substantial implications on the effectiveness of surge pricing, because the temporary increase in wages can have an immediate effect on whether or not drivers continue working. Recent empirical studies have demonstrated that ride-sourcing trips are mainly influenced by the neo-classic labor supply theory. For instance, Chen and Sheldon [2016] found that Uber drivers drive more at times when earning are high, and flexibly adjust to operating more at top surge times. The authors found that surge price significantly increases the supply of rides on the Uber system. Sheldon [2015] found that the driver elasticities increase with experience using empirical data from a peer-to-peer ride-sharing firm. His results suggest that income-targeting behavior, if present, is only temporary.

Despite the empirical evidence found, the research in this area is limited, and there is not a substantial agreement on which theory would outperform. Some authors opted to include both approaches in their analysis, for instance, Zha et al. [2017a] investigated the performance of surge pricing using a time-expanded network, and proposed different formulations for the labor supply models that included both theories.

2.2 Spatial Pricing

The research that incorporates the spatial distribution of the demand and supply is limited. The existing literature mainly addresses the problem of dealing with temporal demand fluctuations at a given location. Recent studies focused on evaluating the spatial pricing of ride-sourcing.

Ma et al. [2018] proposed a Spatio-Temporal Pricing (STP) mechanism based on a driver dispatching methodology. The model included a limited drivers supply that is kept constant during the analysis period. Also, they considered multiple

locations and time periods, with rider demand, willingness to pay, and driver supply varying over space and time. The main motivation is based on the different market failures of current ride-sourcing systems, as explained by the authors: (1) incorrect spatial pricing, caused by prices substantially higher than adjacent locations that will cause drivers to "chase the surge"; and (2) incorrect temporal pricing, where drivers anticipate that prices will increase at certain times (e.g. at the end of a major sport event) and decline trips in anticipation of the surge. Thus, prices need to be appropriately "smooth" in space and time, the STP intends to provide this with a drivers' competitive equilibrium. The authors probed with a simulation that this mechanism provides higher social welfare than the myopic origin-based pricing scheme.

Bimpikis et al. [2016] explored the spatial price discrimination in a network of locations. Unlike the Ma et al. [2018] study, their model included unlimited driver supply, a continuum of potential riders that have heterogeneous willingness to pay, and drivers that endogenously determine whether to provide service and where to relocate themselves with the objective of maximizing their earnings. Using simulations, the authors found that if the demand pattern is not"balanced", the platform can benefit substantially from pricing rides differently depending on the location they originated from. Besides, they also found benefits when pricing rides based on both origin and destination, but the improvements were less significant. Similarly, Castro et al. [2018] propose a framework where a platform chooses prices for the different locations, and drivers respond by deciding where to relocate based on rates, travel costs, and driver congestion levels. They analyzed the spatial pricing problem specifically in the short-term supply and demand imbalance.

Zha et al. [2017b] investigated the effect of spatial pricing on ride-sourcing markets using a model with a discrete-time geometric-matching framework. Their model includes frictions in the matching and meeting process, i.e., the waiting time

for both drivers and customers calibrated using empirical data from Didi Chunxing (a Chinese TNC). The primary results suggest that the platform and drivers are better off under revenue-maximizing spatial pricing, while the effect on costumers may vary.

Buchholz [2015] analyzed the spatial equilibrium in the taxi industry using empirical data from New York City yellow medallion taxis. They proposed a spatial equilibrium model to understand the welfare cost of taxi fare regulations. The authors found that allowing tariffs to vary by time, location or distance can enhance allocation efficiency given the presence of search frictions.

2.3 Contributions to Current Literature

Based on the previous literature review, we found a gap in empirical evidence of the spatial pricing discrimination. Therefore, the contributions of the present research are focused on providing empirical evidence of the spatial unbalance of trip pricing, and its effect on driver equity.

Chapter 3

Graph Spatial Smoothing

The present research effort makes use of the graph-fussed lasso (GFL) technique to perform total variation (TV) denoising. This chapter presents an introduction to graph spatial smoothing. The first section corresponds to a background of the use and the most important aspects. The second section expands on the inherent statistical model and introduces the most relevant techniques: kernel smoothing, Laplace smoothing, and GFL. The last section presents details of the selected approach to solve the GFL smoothing problem, which is based on the work developed by Tansey and Scott [2015].

3.1 Background

The main purpose of the smoothing process is to increase the signal-to-noise ratio. Spatial smoothing techniques are typically used for a wide range of applications. For example, in the image processing field, smoothing approaches are used for image denoising [Chambolle, 2004]; in computational geometry and object modeling, to reconstruct surfaces [Yu and Turk, 2013, Tasdizen et al., 2002]; in machine learning, to impute missing values [Compton et al., 2014]. Other applications include spatial

statistical analysis, where data is smoothed over a physical distance to compensate for inherent sampling noise [Gelfand et al., 2010]. Examples include predicting crime hotspots by smoothing incident report locations [McLafferty et al., 2000], detecting crash hotspots using historical crash data [Thakali et al., 2015], or event detection in taxi trips [Wang et al., 2015].

Graphs can be continuous or discrete. Smoothing techniques for the continuous case include *Gaussian process* and *continuous random fields*, while for the discrete case some methods include *kernel smoothing* and *Laplacian smoothing*. This review is focused on techniques applied to the discrete case.

Graph-smoothing techniques can be classified into local and global approaches [Tansey, 2017]. Local approaches smooth only a local window around each point, such as neighboring pixels in an image, while global methods typically define an objective function over the entire graph and simultaneously optimize the entire set of points. The most simplistic local approaches simply replace each point with the average or median of the points in its window [Tansey, 2017].

An important aspect of using spatial data is the specification of the covariance function of the random field. Data can be isotropic, meaning that the spatial dependence does not depend on the direction of the spatial separation between sampling locations [Weller et al., 2016]. Methods such as the *Gaussian kernel* assume isotropy. However, this assumption is often violated by real-world data, where arbitrary discontinuities can be present in the graph. In some cases, it is more appropriate to rely on anisotropic smoothing techniques.

Anisotropic local methods include the *bilateral filter* and *guided filter*, used to preserve the edge. Global techniques for anisotropy include the *Markov random fields* (MRFs). This method defines a joint distribution over a graph via a product of exponentiated potential functions over cliques, or as a conditional autoregressive (CAR) model where each nodes unnormalized likelihood is written conditioned on

all other nodes in the graph [Tansey, 2017].

An alternative to MRFs for global smoothing is the graph-based trend filtering (GTF) [Wang et al., 2015], which is a special case of the generalized lasso [Tibshirani, 2011]. GTF applies an ℓ_1 penalty to the vector of $(k+1)^{st}$-order differences, where the integer $k \geq 0$ is a hyperparameter. While global approaches like MRFs and GTF typically yield better results, they often fail to scale to large graphs due to every node being dependent on the rest of the graph [Tansey, 2017]. One exception is a special case of the GTF with k = 0, also known as *graph-based total variation* denoising, also called the *graph-fussed lasso (GFL)*.

3.2 Statistical Model

This section presents the statistical model and expands on the smoothing techniques for discrete graphs. For the model, let's say that we have observations y_i, each associated with a vertex $s_i \in \mathcal{S}$ in an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with node set $\mathcal{V}$ and edge set $\mathcal{E}$. The edge set tells which sites are neighbors on the graph. In spatial smoothing, the underlying statistical model can be represented as shown in Equation 3.1 [Scott, 2017].

$$y(s_i) = x(s_i) + \epsilon(s_i)\,, \quad i = 1, \ldots, n, \tag{3.1}$$

where s_i is the spatial location of the ith data point, x is the "true" denoised signal, y is data, and ϵ is mean-zero error. The goal is to estimate x_i in a way that leverages the assumption of spatial smoothness over the underlying graph. The next subsections provide details of three different techniques to find x_i, including kernel smoothing, Laplacian smoothing, and GFL.

3.2.1 Kernel Smoothing

The simplest technique for denoising a spatial signal is called *kernel smoothing*. Suppose we want to estimate $x(s)$ at some target location s. The kernel-smoothing estimate takes the form of a weighted average of all the points in the dataset.

$$x(s) = \frac{\sum_{i=1}^{n} w_i(s, s_i) y(s_i)}{\sum_{i=1}^{n} w_i(s, s_i)}, \tag{3.2}$$

where $w(s, s_i)$ is a weighting function that gets smaller as s and s_i get further apart.

The selection of the weights can lead to different kernel methods. A common approach is called "K nearest neighbors," where the weight function $w(s, s_i)$ takes the value $1/K$ if s_i is one of the K closest points to s, and 0 otherwise. Another common approach is to define the weights in terms of a kernel function. One example is the Gaussian kernel:

$$w_i(s, s_i; b) = \frac{1}{b} \exp\left\{ -\frac{(s - s_i)^2}{2b^2} \right\}, \tag{3.3}$$

which depends upon a bandwidth parameter b. The bandwidth will determine the spread of kernel weights. Figure 3.1 illustrates Gaussian kernel smoothing in a brain structure image. Another example is the quadratic or "Epanechnikov" kernel, which has the advantage that it decays to zero beyond a certain distance.

$$w_i(s, s_i; b) = \max\left(0, \frac{3}{4b}\left[1 - \frac{(s - s_i)^2}{b^2}\right]\right), \tag{3.4}$$

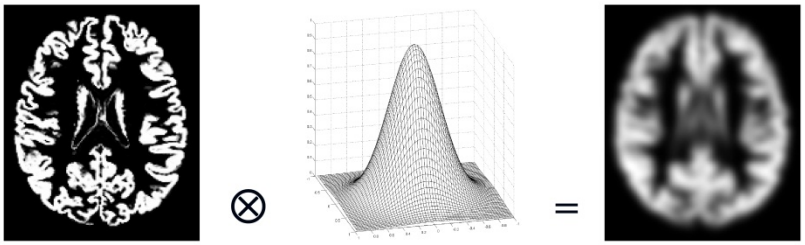


Figure 3.1: Denoising of an image. From left to right: original image, 2D Gaussian kernel and smoothed image. [García-Martí et al., 2013]

3.2.2 Laplacian Smoothing

In image processing, electronics, and spatial statistics, a common approach is *Laplacian smoothing*. The Laplacian smoothing problem can be expressed as follow,

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \quad \frac{1}{2}\|y-x\|_2^2 + \frac{\lambda}{2}x^T L x\,, \tag{3.5}$$

where L is the graph Laplacian, and $\lambda > 0$.

The Laplacian matrix has the alternate representation $L = D^T D$, therefore the penalty term $x^T L x$ can be rewritten as $x^T L x = \|Dx\|_2^2$, known as the ℓ_2 penalty, where D is the oriented edge matrix of the graph [1]. The solution of Equation 3.5 can be written as,

$$\hat{x} = (I + \lambda L)^{-1} y \tag{3.6}$$

which is a linear system that can be solved using (1) a direct solver that uses a sparse matrix factorization, (2) the Gauss-Seidel method, (3) the Jacobi iterative method, among others.

[1] Letting $m = |\mathcal{E}|$ be the size of the edge set, D is the $m \times n$ matrix defined as follows. If $(j,k), j < k$ is the ith edge in $\mathcal{E}$, then the ith row of D has a 1 in position j, a -1 in position k, and a 0 everywhere else. Thus the vector Dx encodes the set of pairwise first differences across the edges of the graph.

3.2.3 Graph-Fused Lasso

As mentioned previously, there are cases where the isotropic assumption is violated, and it is necessary to apply anisotropic methods. An example of this type of approach is the GFL, which consider a version of the spatial smoothing problem where we change the ℓ_2 penalty to an ℓ_1 penalty. The penalty term rewards the solution for having small absolute first differences across the edges in the graph. Figure 3.2 illustrates the GFL technique. The GFL smoothing problem can be represented as,

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \quad \frac{1}{2}\|y - x\|_2^2 + \lambda \|Dx\|_1 \,, \tag{3.7}$$

where D is the oriented edge matrix of the graph, and $\lambda > 0$ is the regularization parameter.

The Equation 3.7 does not have a closed-form solution. Therefore, convex optimization approaches – such as the *alternating direction method of multipliers* (ADMM) [2] [Boyd et al., 2011] – are required. Many efficient, specialized approaches using ADMM have been developed, e.g. Wahlberg et al. [2012], Barbero and Sra [2014], and Tansey and Scott [2015]. Specifically, the Tansey and Scott [2015] method lead to an efficient approach that presents a fast solution and is also scalable. The next section provides a brief description of this method.

3.3 A Fast and Flexible Algorithm for the GFL

Tansey and Scott [2015] proposed an ADMM approach to solving the GFL, where the key insight is to decompose the graph into a set of trails which can then each be solved efficiently using techniques for the ordinary (1D) fused lasso. "The resulting technique is both faster than previous GFL methods and more flexible in the choice

[2] The ADMM is an algorithm that solves convex optimization problems by breaking them into smaller pieces, each of which is then easier to handle.

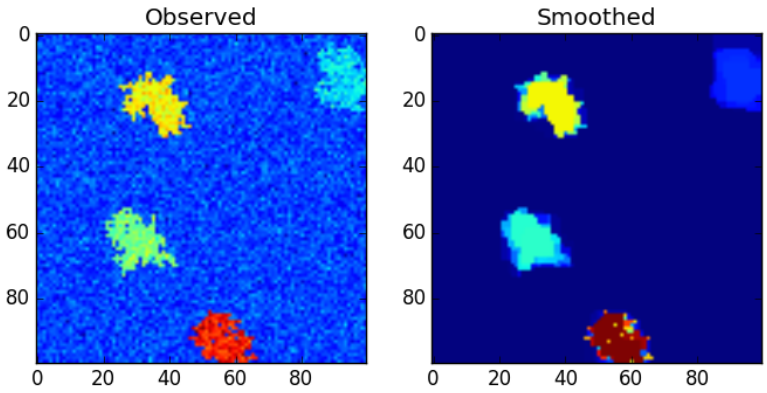


Figure 3.2: TV denoising of an image using the GFL [Tansey, 2017]

.

of loss function and graph structure" [Tansey and Scott, 2015]. This section provides a summary of the method.

First, let's rewrite Equation 3.7 for a more general case,

$$\underset{\mathbf{x}\in\mathbb{R}^n}{\text{minimize}} \quad \ell(\mathbf{y},\mathbf{x}) + \lambda \sum_{(r,s)\in\mathcal{E}} |x_r - x_s|, \tag{3.8}$$

where ℓ is a smooth convex loss function. For the case of Equation 3.7, the loss function corresponded to the squared-loss error. Solving for a general case is an advantage of the method, which is flexible enough to handle any smooth convex loss function and any generic graph.

The core idea of the algorithm is to decompose a graph into a set of trails. Recall the undirected graph definition, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with node set $\mathcal{V}$ and edge set $\mathcal{E}$. Tansey and Scott [2015] denote two preliminaries:

- Every graph has an even number of odd-degree vertices.
- if $\mathcal{G}$ is not connected, then the objective function is separable across the connected components of $\mathcal{G}$, each of which can be solved independently.

Theorem 3.3.1 *(Tansey and Scott [2015], Theorem 1) The edges of a connected graph with exactly $2k$ odd-degree vertices can be partitioned into k trails if $k \geq 0$. If $k = 0$, there is an Eulerian tour. Furthermore, the minimum number of trails that can partition the graph is $max(1, k)$.*

The Theorem 3.3.1 reassures that any connected graph can be decomposed into a set of trails $\mathcal{T} = \{t_1, t_2, ..., t_k\}$ on which the optimization algorithm can operate, and allows one to rewrite the penalty function as:

$$\sum_{(r,s)\in\mathcal{E}} |x_r - x_s| = \sum_{t\in\mathcal{T}} \sum_{(r,s)\in t} |x_r - x_s| \tag{3.9}$$

The updated penalty function allows proposing an efficient ADMM algorithm. The next sections present details of the updated optimization method and the trial decomposition approaches suggested by the authors.

3.3.1 Optimization via ADMM

The objective function, shown in Equation 3.8, can be rewritten using the updated penalty function. In addition, for each trail t (where $|t| = m$), we introduce $m + 1$ slack variables [3], one for each vertex along the trail. Multiple slack variables are introduced if a vertex is visited more than once in a trail. Equation 3.8 is rewritten as:

$$\begin{aligned} \underset{\mathbf{x}\in\mathbb{R}^n}{\text{minimize}} \quad & \ell(\mathbf{y}, \mathbf{x}) + \lambda \sum_{t\in\mathcal{T}} \sum_{(r,s)\in t} |x_r - x_s| \\ \text{subject to} \quad & x_r = z_r \\ & x_s = z_s \end{aligned} \tag{3.10}$$

This problem can be solved using the ADMM algortihm [Boyd et al., 2011]

[3] In an optimization problem, a slack variable is a variable that is added to an inequality constraint to transform it into an equality.

based on the following updates:

$$\mathbf{x}^{k+1} = \underset{\mathbf{x}}{\operatorname{argmin}} \left(\ell(\mathbf{y}, \mathbf{x}) + \frac{\alpha}{2} \|\mathbf{A}\mathbf{x} - \mathbf{z}^k + \mathbf{u}^k\|^2 \right) \tag{3.11}$$

$$\mathbf{z}_t^{k+1} = \underset{\mathbf{z}}{\operatorname{argmin}} \left(w \sum_{(r \in t} (\tilde{y}_r - z_r)^2 + \sum_{(r,s) \in t} |z_r - z_s| \right), t \in \mathcal{T} \tag{3.12}$$

$$\mathbf{u}^{k+1} = \mathbf{u}^k + \mathbf{A}\mathbf{x}^{k+1} - \mathbf{z}^{k+1} \tag{3.13}$$

where u is the scaled dual variable, α is the scalar penalty parameter, $w = \frac{\alpha}{2}$, $\tilde{y}_r = x_r - u_r$ and A is a sparse binary matrix used to encode the appropriate x_i for each z_j . Here t is used to denote both the vertices and edges along trail t.

For the squared-loss function, as used in Equation 3.7, the x updates have the simple closed-form solution:

$$x_i^{k+1} = \frac{2y_i + \alpha \sum_{j \in \mathcal{J}} (z_j - u_j)}{2 + \alpha |\mathcal{J}|}, \tag{3.14}$$

where $\mathcal{J}$ is the set of dual variable indices that map to x_j. Crucially, the trail decomposition approach means that each trail's z update in Equation 3.12 is a one-dimensional fused lasso problem which can be solved in linear time via an efficient dynamic programming routine.

3.3.2 Trail decomposition

Tansey and Scott [2015] proposed two approaches for the trail decomposition summarized as follow – for a broader explanation see Tansey and Scott [2015]:

1. Create k "pseudoedges" connecting the 2k odd-degree vertices, and then find an Eulerian tour on the surgically altered graph. To decompose the graph into trails, we then walk along the tour (which by construction enumerates

every edge in the original graph exactly once). Every time a pseudo-edge is encountered, we mark the start of a new trail.

2. Iteratively choose a pair of odd-degree vertices and remove a shortest path connecting them. Any component that is disconnected from the graph then has an Eulerian tour and can be appended onto the trail at the point of disconnection.

Chapter 4

Methodology

This section describes the methodology used for the analysis of the ride-sourcing trips. The first part presents a description of the data used, including the mining process and the variable definition. The second part explains the graph-fussed lasso (GFL) total variation (TV) denoising process applied to the specific set of data.

4.1 Data

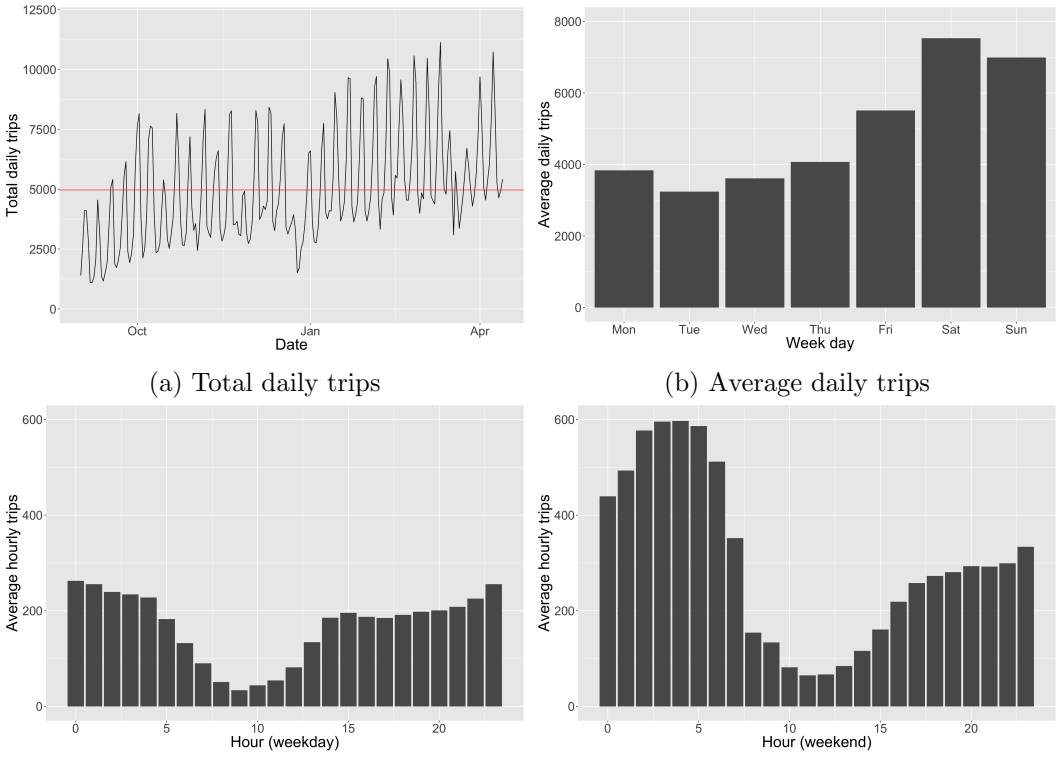
In this study, we used the data that a non-profit Austin-based TNC – known as *Ride Austin* – made available in early 2017[1]. The dataset consisted of 1,494,125 rides between June 2nd, 2016 and April 13th, 2017. Each trip corresponds to a row in the database. Also, the dataset provides a description of the trip, rider, and driver (anonymized), payment, cost, and weather.

Since the demand during the first month was limited, we focused our analysis on data from September 1st, 2016 to April 13th, 2017. We selected rides with the origin and destination coordinates within the traffic analysis zones (TAZs) [2]. Besides, our analysis only includes regular car category trips; we omitted trips by

[1]Available through the website https://data.world/ride-austin

[2]Defined by the Capital Area Metropolitan Planning Organization (CAMPO)

sport-utility vehicles (SUV), premium, and luxury categories because the fare rate is different in each case. Similarly, we only analyzed flat-rated trips, i.e., trips that do not include any surge price. The total number of rides examined based on the previous restrictions is 1,117,943 rides, with approximately 5,000 average daily trips. Figure 4.1 provides a summary description of the evaluated trips including the total daily trips, the average daily trips per week day, and the average hourly trips during weekdays and weekends. We can observe that the majority of trips are concentrated on weekends, mainly during the morning hours.



(a) Total daily trips

(b) Average daily trips

(c) Average hourly trips during weekdays

(d) Average hourly trips during weekends

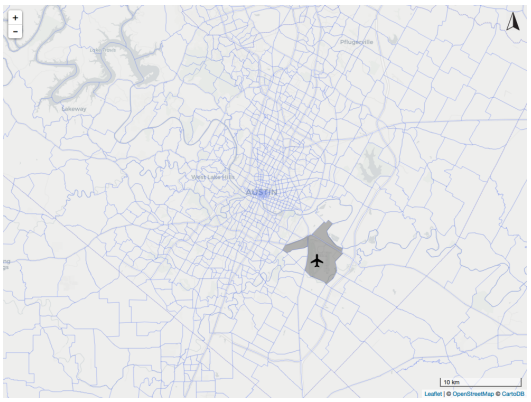
Figure 4.1: Description of evaluated trips

4.1.1 Space Discretization

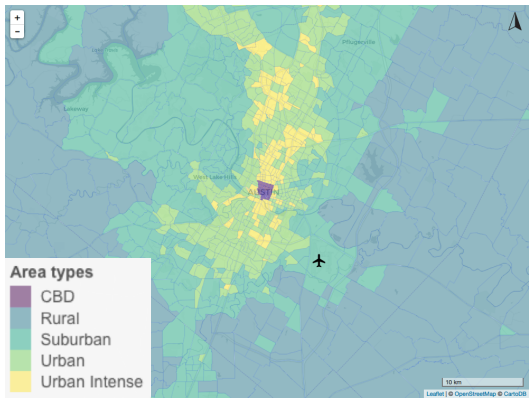
The space discretization consists of summarizing the trip variables within the origin or destination TAZ using the average value. We matched the trips pick-up (origin) and drop-off (destination) longitude and latitude coordinates with the corresponding TAZ location. Figure 4.2 (a) presents the TAZs in the Austin area and provides the location of the TAZs corresponding to the Austin-Bergstrom International Airport (ABIA) area. Also, as a specific case study, we selected trips with origin within the central business district (CBD), located in Austin downtown, using the area type TAZ classification. CBD trips correspond to a total of 176,219 trips, approximately 16% of the total evaluated rides. Figure 4.2 (b) provides a spatial description of the area types in Austin, and Figure 4.2 (c) presents a detailed view of the downtown area.

4.1.2 Time Discretization

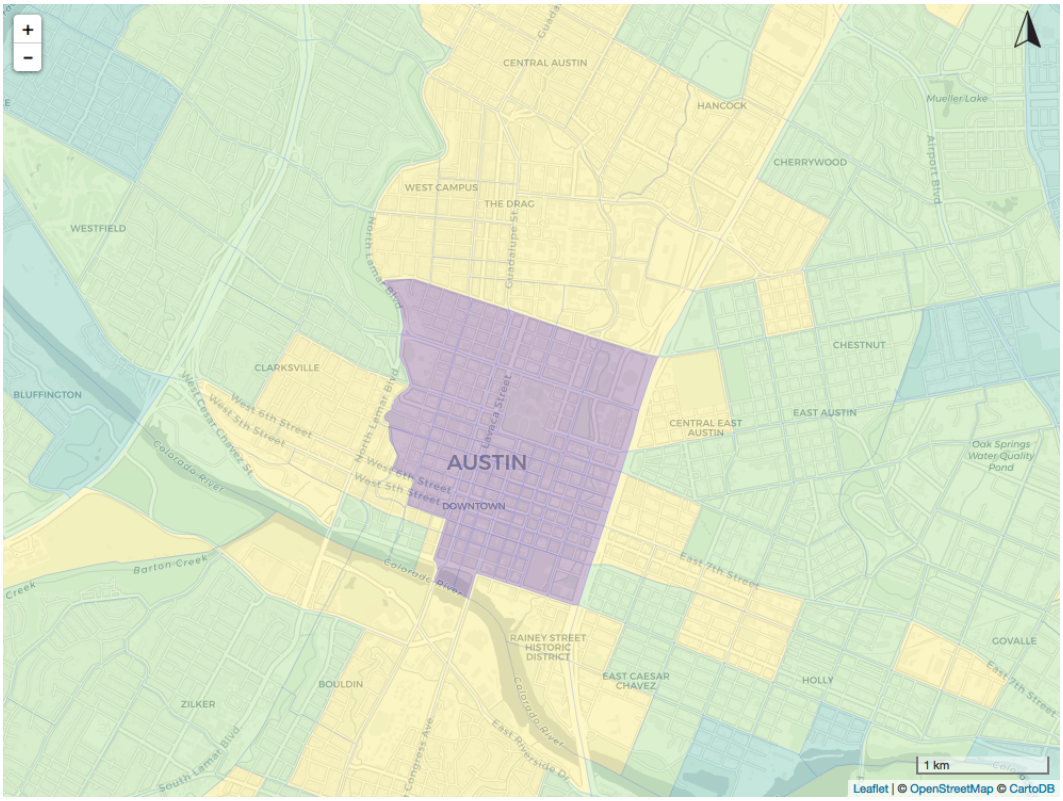
The time discretization is based on the system peak hours because these correspond to higher travel time and delays. We used four different time classifications. First, we divided the trips into weekdays and weekend trips. Then, the weekday trips were divided into AM-peak (from 6 to 9 AM), PM-peak (from 4 to 7 PM), and off-peak hours. Table 4.1 provides the number of rides per each time classification. Figure 4.3 presents the total trip count per TAZ for each of the time frames evaluated based on both origin and destination.



(a) TAZs in Austin (airport TAZs shaded)



(b) Area types TAZs



(c) Area types TAZs - Downtown

Figure 4.2: Description of TAZs

Table 4.1: Number of trips evaluated

Time discretization		CBD-origin	System-wide
Weekday	AM-peak	16,371	49,430
	PM-peak	15,661	122,653
	off-peak	69,748	480,900
Weekend		74,439	464,960
Total		***176,219***	***1,117,943***

4.1.3 Description of Variables

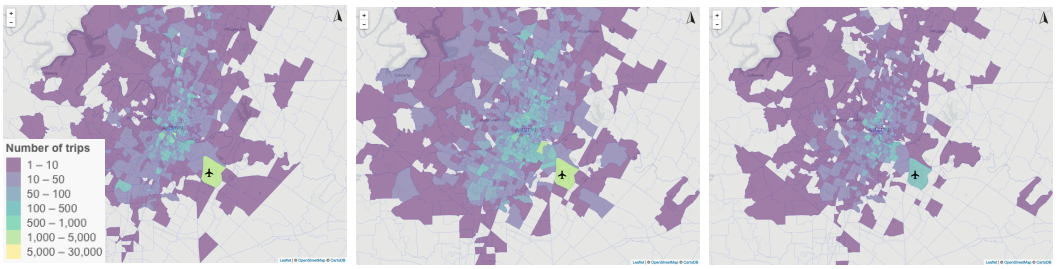
We selected different measures as indicators of driver productivity. Specifically, we focused on operational variables such as the trip fare, trip distance, waiting or idle time, and reaching time. Additionally, we estimated three different productivity variables based on trip fare and driver time.

Operational Variables

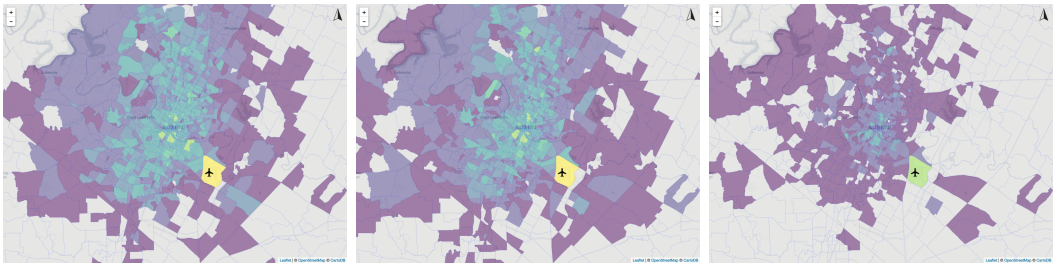
Among the operational variables, we selected the *trip fare*, corresponding to the total passenger cost except for the tip, roundup amount[3], and other operational fees, such as booking and airport fees. The trip fare consists of the sum of base fare, time rate, and distance rate, and the minimum fare is $4 (see Equation 4.1), based on a regular car category.

$$fare_{trip} = max(base\ fare + distance\ rate + time\ rate, 4) \tag{4.1}$$

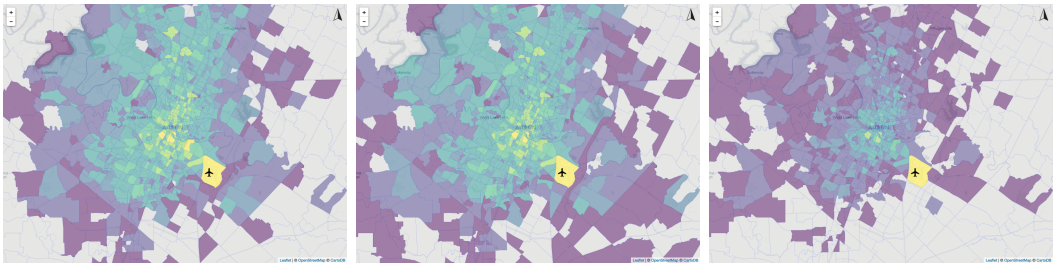
[3]Ride Austin allows riders to round up the total fare to the nearest dollar and designate it to a local charity.



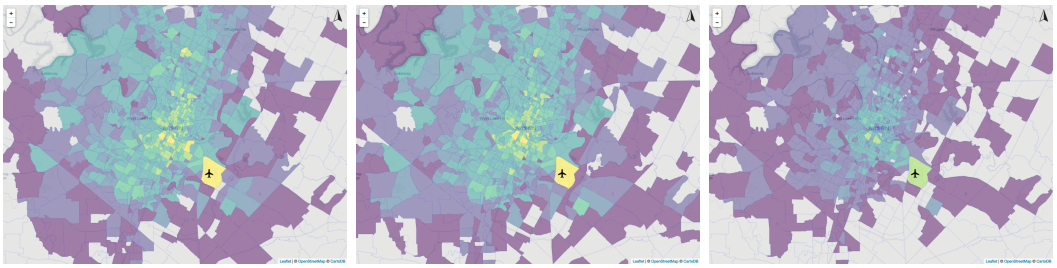
(a) Origin (AM-peak) (b) Dest. (AM-peak) (c) Dest. (AM-peak) CBD



(d) Origin (PM-peak) (e) Dest. (PM-peak) (f) Dest. (PM-peak) CBD



(g) Origin (off-peak) (h) Dest. (off-peak) (i) Dest. (off-peak) CBD



(j) Origin (weekend) (k) Dest. (weekend) (l) Dest. (weekend) CBD

Figure 4.3: Total count number of trips per TAZ origin and destination

The *trip distance* used represents the distance from the pick-up to the drop-off location, in kilometers, provided in the database. The *reach time* corresponds to the time it took the driver to reach the rider since the trip was assigned to him, also provided in the database. In addition, the *idle time* was estimated based on the driver unique identification information. This time correspond to the time it took the driver from the previous trip drop-off time to the next trip pick-up time. We only considered idle time lower than 60 minutes in the analysis. Figure 4.4 provides a driver time diagram with a graphical representation of these variables. From the figure we can describe the driver time values as follow:

- t_0 : time $trip_1$ started at the pick-up location
- t_1 : time $trip_1$ finished at the drop-off location
- t_2 : time $trip_2$ is assigned to the driver
- t_3 : time $trip_2$ started at the pick-up location
- t_4 : time $trip_2$ finished at the drop-off location

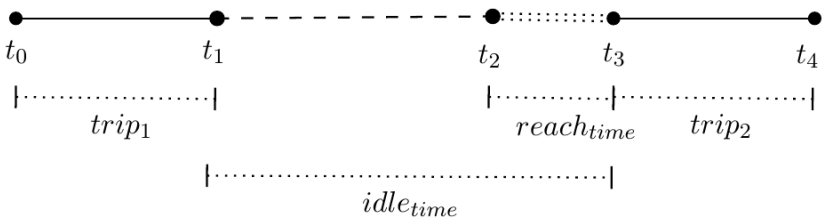


Figure 4.4: Driver time diagram

.

Using the previous driver time description, we can estimate the idle time and the reach time using Equations 4.3 and 4.2, receptively.

$$reach_{time} = t_3 - t_2 \tag{4.2}$$

$$idle_{time} = t_3 - t_1 \tag{4.3}$$

Productivity Variables

We estimated three different productivity variables based on the driver time diagram. *Productivity A* corresponds to the throughput of the $trip_1$ in dollars per hour, estimated using Equation 4.4. *Productivity B* refers to the productivity of the $trip_1$ including the idle time after finishing that trip, estimated using Equation 4.5. Finally, *Productivity C* corresponds to the productivity of two consecutive trips including the idle time between them, as an indirect measure of the continuation payoff, it was estimated using Equation 4.6. The main objective of estimating these variables is to analyze the impact of the trip destination on the driver productivity.

Productivity A only captures the revenue of a trip divided by its duration, so it is not expected to capture the effect that ending a trip on a region of low density may have. Productivity B is adding the idle time after the end trip to the formula; in this way, penalizing for traveling to low-demand zones. Productivity C also includes the revenue and duration of the following trip, seeking to capture further spatial dynamics for ending in a specific region.

Table 4.2 provides summary statistics of the seven variables described in this section.

$$Productivity\ \mathbf{A} = \frac{fare_{trip_1}}{t_1} \tag{4.4}$$

$$Productivity\ \mathbf{B} = \frac{fare_{trip_1}}{t_3} \tag{4.5}$$

$$Productivity\ \mathbf{C} = \frac{fare_{trip_1} + fare_{trip_2}}{t_4} \tag{4.6}$$

Table 4.2: Summary statistics of the analyzed variables

Variable	Min.	Max.	Mean	Median	Std. Dev.
Operational					
Trip fare ($)	4.0	57.9	10.4	8.3	6.5
Trip distance (km)	0.1	50.0	9.0	6.2	7.8
Idle time (min)	1.0	60.0	18.0	13.3	13.7
Reach time (min)	1.0	20.0	6.3	5.6	3.5
Productivity					
Productivity A ($/hr)	15.7	100.0	48.2	46.7	11.0
Productivity B ($/hr)	1.0	60.0	17.9	17.7	10.5
Productivity C ($/hr)	1.1	60.0	24.8	25.8	11.5

4.2 GFL TV Denoising

We used the GFL for TV denoising of the variables described previously. This section describes the method used in the current dataset.

4.2.1 Penalized Weighted Least Squares

We selected a penalized weighted least square loss function (Equation 4.7) to take into account the differences in the number of observations with each zone. Let us denote by η_i the count of trips observed within the i-th TAZ, then the objective function takes the form:

$$\underset{\mathbf{x}\in\mathbb{R}^n}{\text{minimize}} \quad \sum_{i=1}^{n} \frac{\eta_i}{2}(y_i - x_i)^2 + \lambda \sum_{(r,s)\in\mathcal{E}} |x_r - x_s| \tag{4.7}$$

The justification for this set of weights is the following. If $(y_{i,1}..., y_{i,\eta_i})$ are the observations in the i-th TAZ, $(x_{i,1}..., x_{i,\eta_i})$ the predicted values, then the squared error of the full model would be:

$$\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{\eta_i}(y_{ij} - x_{ij})^2$$

And if in the above formula we replace all the y_{ij} and x_{ij} with their means y_i and x_i, then we recover the squared error term of (4.7).

4.2.2 Graph Definition

The edges for joining the TAZ nodes where chosen according to a k-nearest neighbors principle. The location of a TAZ was computed as the mean longitude and latitude of all the points observed in that region. The exact assigned location differed slightly depending on whether the data were being classified by start or end location. Once the node locations were calculated, then for each node r, an edge (r, s) was added for all s within its k-nearest neighbors. We used $k = 4$ so that the graph represented spatial adjacency. We remark that there was little variation in the final results for other close values of k.

4.2.3 Choosing the Regularization Parameter

The optimal regularization parameter λ depended on each variable and was selected by splitting the data in a training and test set. The metric used is the root mean square error (RMSE) at the individual level, shown in Equation 4.8. Figure 4.5 presents an example of the RMSE obtained using different λ values, the optimal λ is the value that minimizes the RMSE.

$$\widehat{RMSE} = \left(\frac{1}{\sum_{i=1}^{n}\eta_i}\sum_{i=1}^{n}\sum_{j=1}^{\eta_i}(y_{ij} - \hat{x}_i)^2\right)^{1/2}, \tag{4.8}$$

where n and η_i are the number of TAZ regions and counts for the i-th taz from the test set, and $\hat{x}_i$ is the prediction for the i-th TAZ obtained from the training data.

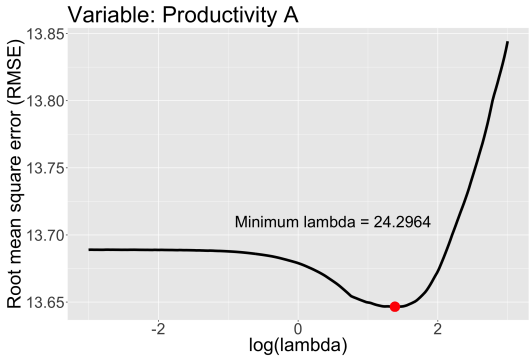


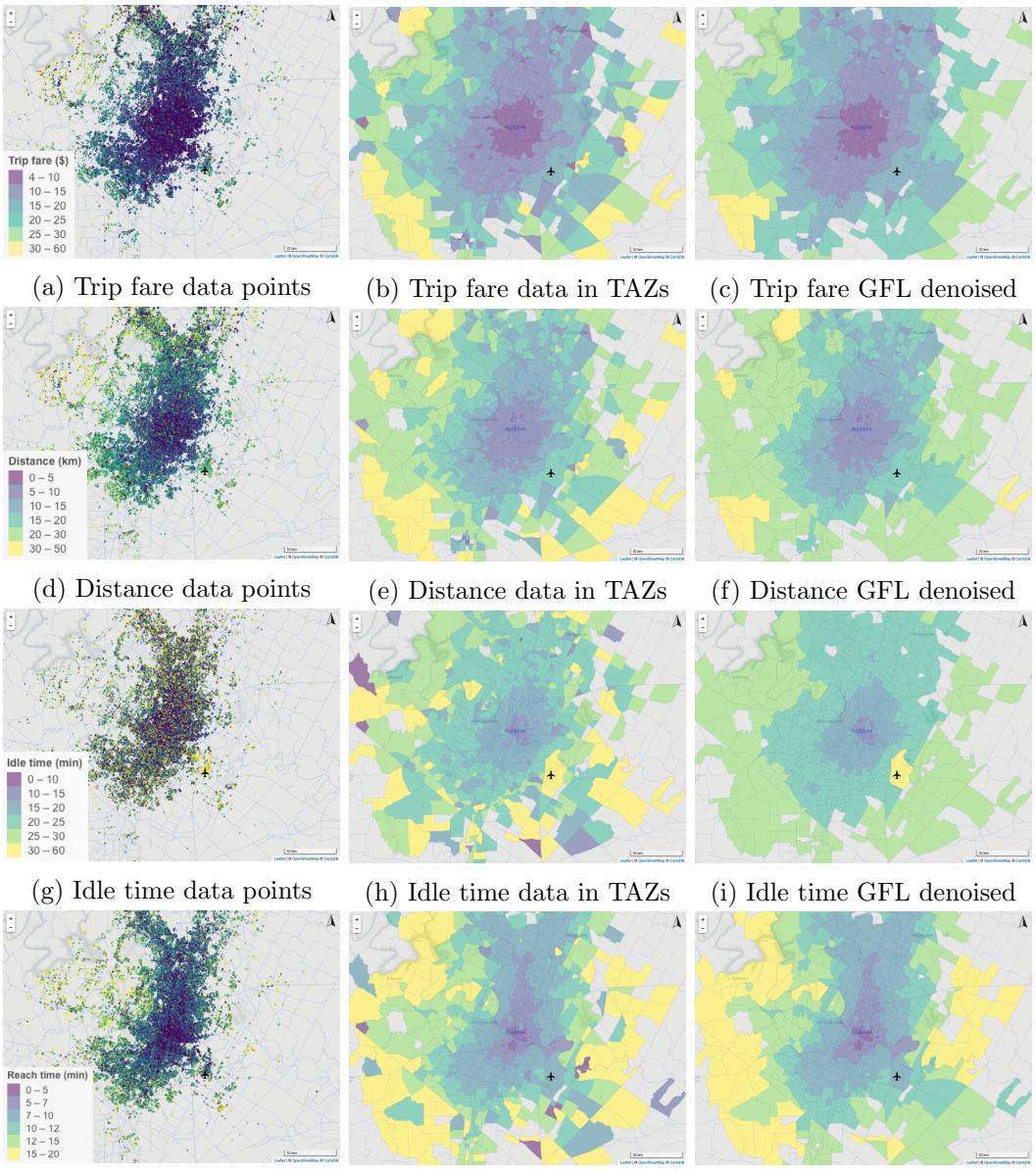
Figure 4.5: RMSE to find the optimal regularization parameter

.

4.2.4 GFL TV Denoising Examples

We provide examples of the application of the GFL TV denoising to the variables of interest in Figure 4.6. The first image presents the raw data points. The next image provides the information summarized per TAZ. Finally, the third image presents the denoised graph.

The denoised image allows a better interpretation of the spatial distribution of the variables. Also, we can observe a clear example of the ℓ_1 penalty benefits in the airport area of Figure 4.6 (i), corresponded to the idle time variable. The GFL is able to preserve the high contrast of the values in this area and keep it independent of the surrounding area values.



(a) Trip fare data points (b) Trip fare data in TAZs (c) Trip fare GFL denoised

(d) Distance data points (e) Distance data in TAZs (f) Distance GFL denoised

(g) Idle time data points (h) Idle time data in TAZs (i) Idle time GFL denoised

(j) Reach time data points (k) Reach time data in TAZs (l) Reach time GFL denoised

Figure 4.6: GFL TV denoising examples (system-wide weekend origin trips)

Chapter 5

Results and Discussion

This section presents the results of the total variation (TV) denoising process using the graph-fussed lasso (GFL) for the variables described in the methodology section. We analyze and discuss significant findings based on the denoised results.

5.1 Operational Variables

The operational variables are analyzed based on the origin (pick-up) traffic analysis zone (TAZ) of the trips. The results are shown in Figure 5.1.

5.1.1 Distance

The resulting maps of *trip distance* spatial distribution show that the majority of short trips are concentrated in the central area, while longer trips are originated from the urban sectors. Additionally, trip distances do not show significant changes across the time frames analyzed. The main difference is found during the morning peak hour. We observed areas with a trip length lower than 5 km in the downtown sector, while the airport presented long trips. Thus, it is possible that users prefer to use ride-sourcing trips for commute inside the central area and airport trips during

the critical morning hours.

5.1.2 Idle Time

The *idle time* provides an indirect measure of the trip supply and demand relation. It is based on the driver time between the end of a trip and the beginning of the next one (see Figure 4.4). In this case, the results are shown based on the origin of the trips and correspond to the idle time prior to the rider pick-up at that specific location. Downtown trips present lower driver idle time than the periphery-area trips. Short idle time means that the ratio supply-demand is near to one (the demand is similar to the supply). High idle time can be related to low rides' demand or high drivers' supply, thus a supply-demand ratio higher than one.

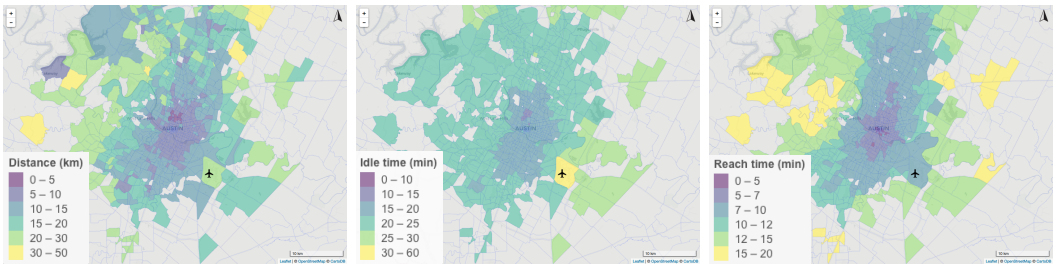
The results showed a marked difference between the airport trips and the other TAZs trips, and this difference is constant over the time. Specifically, rides beginning at the airport showed higher idle time prior to the rider pick-up. This result suggests that there could be an excessive driver supply in the area, which causes drivers to wait longer until the next trip. However, the airport-area presented a high number of trips. Thus it is possible that drivers are aware of this and prefer drive there to warranty trips.

We observed a high contrast of idle time at the AM and PM-peaks. Weekday evening rides tend to have a greater idle time compared to weekday mornings. This result suggests that the supply and demand interaction of these hour frames is different. PM-peak trips present higher supply-demand ratio compared to AM-Peak. The number of rides during weekday evenings is significantly higher to morning trips. Thus, we can conclude that there is a considerably higher supply of drivers during the PM-Peak compared to AM-Peak.

5.1.3 Reach Time

The *reach time* consists of the driver time since the trip is assigned until the pick-up moment, as shown in Figure 4.4. This variable is an indirect measure of the driver supply in the area. Low reach-time regions have more drivers available nearby than high reach-time zones. However, it is also affected by the driving speed and accessibility. In general, the reaching time is less than 10 min for most of the central area, including the airport zone.

The results present a notable difference between north-south and east-west reaching time, which can be related to the accessibility provided by the main north-south corridors including the Interstate Highway (IH) 35, State Highway Loop 1, and other Arterial corridors (e.g., Lamar and Guadalupe). Among the time discretization, the morning peak-hour shows the higher variation. The north-south pattern is lesser than other time slots and the low reaching-times are limited to the center of the city. This result suggests that the during AM-peak hours the accessibility to the north and south areas is limited, probably caused by the delays due to rush hour.



(a) Distance AM-peak (b) Idle time AM-peak (c) Reach time AM-peak

(d) Distance PM-peak (e) Idle time PM-peak (f) Reach time PM-peak

(g) Distance off-peak (h) Idle time off-peak (i) Reach time off-peak

(j) Distance weekend (k) Idle time weekend (l) Reach time weekend

Figure 5.1: Operational variables comparison for system-wide trips (trip origin)

5.2 Productivity Variables

The productivity variables are analyzed based on the destination (drop-off) TAZ of the trips, which allows us to investigate the impact of the trip destination on the drivers' productivity. We provide a natural experiment where we select only trips with origin in the central business district (CBD)– refer to Figure 4.2 for more details – these trips represent 16% of the total rides evaluated. We estimated three different productivity measures to capture the main variables influencing drivers' effective profit. *Productivity A* measures the throughput of the first trip. *Productivity B* includes the idle time to take into account the trade-off between trip benefit (profit) and ending zone operational condition. Finally, *Productivity C* provides a measure of the continuation payoff by taking into account also the following trip profit.

5.2.1 Trips with CBD Origin

Productivity A results show an interesting relation between very short (less than 0.8 km) and long trips (more than 25 km). For instance, based on the AM-peak outputs, we can observe a small area of high throughput ($60/hr-$80/hr) within the downtown zone, given by short rides. Longer trips provide lower measures in the range of $45/hr to $55/hr, but as trips become further away from the CBD, the productivity raises again to high values. This "donut-like" effect is also present for the PM-peak, off-peak, and weekend rides, but with less contrast than AM-peak results. This result can be related to the base fare, which warranties a minimum fix amount for very short trips. Thus, the driver productivity for a single journey is comparable between trips lower than 0.8 km and trip longer than 25 km, approximately.

Using the measure of Productivity B, we want to capture the effect of the ending-zone idle time in the driver productivity. The results show an interesting spatio-temporal dynamic. First, the spatial contrast is lesser. For instance, AM

peak shows a result variation of only $10/hr (from $15/hr to $25/hr). Also, the central area tends to show lower values than the periphery area. Second, the time effect is significant. For example, we can observe that Productivity B values variate significantly from the AM-Peak to the PM-peak and weekend results. This result suggests that it is a spatial influence in the driver productivity but the time effect has a more significant impact.

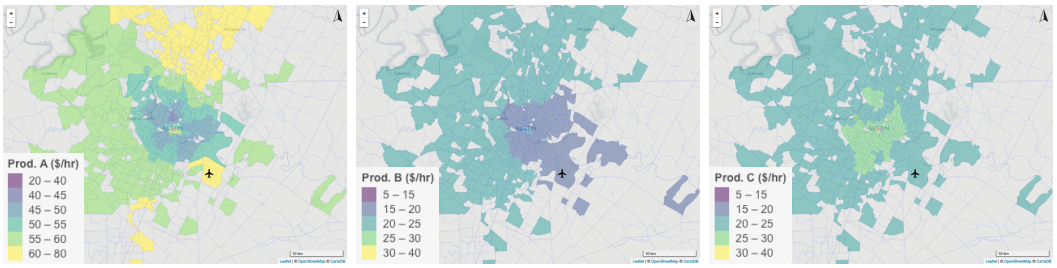
Productivity C provides insights on the productivity of two consecutive trips and takes into account the ending-zone idle time between them. In this case, results showed a lower spatial impact compared to Productivity B. Mainly for off-peak and weekend trips where the majority of the area present similar results. Am-peak presented the most favorable productivity measure located in the central region. In general, AM-peak rides results indicate that drivers that stayed in the central area ended their second trip with higher productivity compared to those who made longer trips. Regarding the time effect, weekend trips are more favorable for drivers.

5.2.2 System-Wide Trips

The system wide results provide a generalization of the CBD rides. The results are presented in Figure 5.3. In this case we want to generalize the idea of the trip destination effect on the drivers' equity. Productivity A show that the revenue for a single ride is higher in the periphery area of the city, meaning that trips ending in zones distant from the city provide higher revenue in general. The AM-peak presents greater productivity values than other time frames. The idle time of the ending zone provides measures of Productivity B with low spatial contrast. However we can observe that PM-peak trips present lower values compared to other times of the day.

Productivity C during weekday peak hours provides valuable information. Rides ending in downtown result with higher productivity compared to those ending

in the peripheral area. Specifically, AM-peak rides present the highest value in downtown and airport zones. During off-peak and weekend hours we still have the central area as the most desirable. However, we can observe that this area not only comprises the central zone but also south-west area, corresponding to suburban developments. This result is contra-intuitive, because the area does not present high intensity of trip demand. The main reason of this result can be related with the higher uncertainty of the data points in the zone, the amount of data is more limited and the underlying spatial adjacencies provide a GFL best guest estimate comparable with the downtown area. Figure 5.4 presents the trip count and the Productivity C before and after the GFL denoising process. We can confirm that the south-west area presents low trip count and high variability on the productivity values.



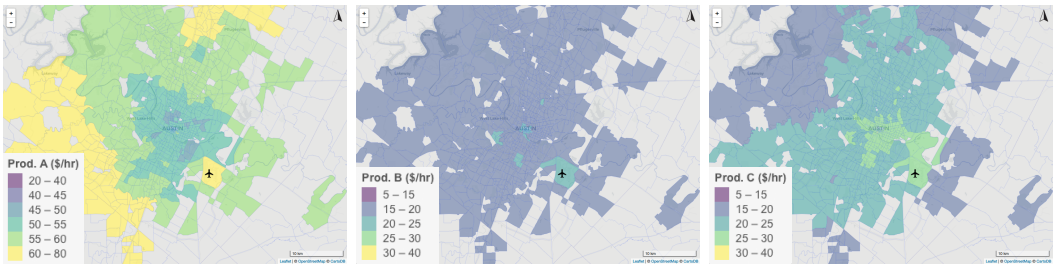
(a) Productivity A AM-peak (b) Productivity B AM-peak (c) Productivity C AM-peak

(d) Productivity A PM-peak (e) Productivity B PM-peak (f) Productivity C PM-peak

(g) Productivity A off-peak (h) Productivity B off-peak (i) Productivity C off-peak

(j) Productivity A weekend (k) Productivity B weekend (l) Productivity C weekend

Figure 5.2: Productivity comparison for CBD trips (trip destination)



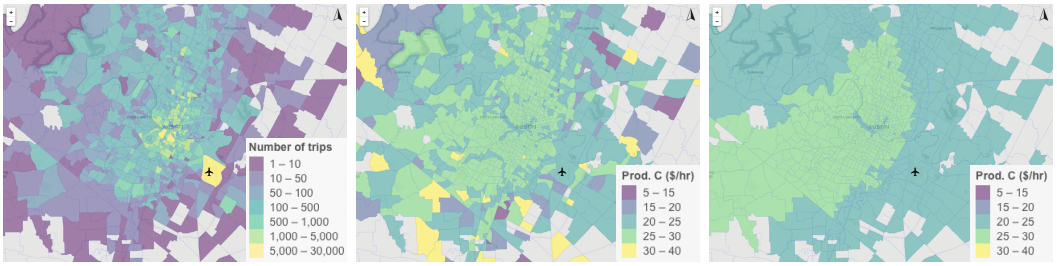
(a) Productivity A AM-peak (b) Productivity B AM-peak (c) Productivity C AM-peak

(d) Productivity A PM-peak (e) Productivity B PM-peak (f) Productivity C PM-peak

(g) Productivity A off-peak (h) Productivity B off-peak (i) Productivity C off-peak

(j) Productivity A weekend (k) Productivity B weekend (l) Productivity C weekend

Figure 5.3: Productivity comparison for system-wide trips (trip destination)



(a) Trip count in destination (b) Productivity C original (c) Productivity C denoised

Figure 5.4: Productivity C comparison for weekend trips (destination)

Chapter 6

Conclusions

This study explored the spatial pricing discrimination of ride-sourcing trips using empirical data. We used information from more than 1.1 million rides in the Austin area, provided by a non-profit TNC from a period where the leading companies were out of the city. We based our analysis on operational variables such as the waiting or idle time between trips and reaching time. Also, we estimated three different productivity measures to evaluate the impact of the trip destination on the driver continuation payoff.

The analysis of the operational variables provided insights about the ride-sourcing travel patterns and the balance between supply and demand across space and time. We found that during weekday AM-peak hours (from 6 to 9 AM), riders within the central area prefer shorter trips, compared to the other time frames. Additionally, PM-peak hours (from 4 to 7 PM) tend to have significantly higher driver supply which causes a greater supply-demand ratio, compared to AM-peak. Also, Airport-area trips showed a marked difference compared to other Austin areas in term of operational variables, including a significantly higher amount of origin and destination rides. The results suggest that drivers prefer to drive there to warranty rides.

Furthermore, the evaluation of different productivity variables allowed the investigation of the effect of the trip destination on driver productivity. Based on the results, the productivity of a single trip for rides shorter than 0.8 km is comparable to rides longer than 25 km, approximately. Regarding spatial effects, drivers with rides ending in the central area presented favorable spatial differences in productivity when including the revenue of two consecutive trips for AM-peak rides. However, the other time slots evaluated did not show significant differences. The time effect was more contrasting than the spatial effect. Weekend rides tend to provide better driver productivity measures.

The results and methods provided in this study can serve multiple purposes. First, from a driver and operator point of view, we identified the spatial and temporal distribution of the principal operational and productivity variables, which can lead to a more efficient driver supply method. Second, from the planners and engineers' perspective, we provided insights on the ride-sourcing travel patterns in the Austin area that can help to understand the main characteristics of this type of service. Third, we provide empirical evidence of the driver productivity inequality due to spatial and temporal factors. This evaluation can lead to policies that warranty fair driver-earning conditions. Finally, our results may also have relevance in the field of transportation research. We provide an application of spatial smoothing approaches to a transportation problem. Furthermore, we specifically focused on a method that can be solved with a highly efficient and scalable algorithm that can be used with evaluations that include big-data.

Bibliography

Siddhartha Banerjee, Carlos Riquelme, and Ramesh Johari. Pricing in ride-share platforms: A queueing-theoretic approach. 2015.

Alvaro Barbero and Suvrit Sra. Modular proximal optimization for multidimensional total-variation regularization. *arXiv preprint arXiv:1411.0589*, 2014.

Kostas Bimpikis, Ozan Candogan, and Saban Daniela. Spatial pricing in ride-sharing networks. 2016.

J Bogage. Uber's controversial strategy to finally defeat lyft. *The Washington Post*, 2016.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

Nicholas Buchholz. Spatial equilibrium, search frictions and efficient regulation in the taxi industry. Technical report, Technical report, University of Texas at Austin, 2015.

Gerard P Cachon, Kaitlin M Daniels, and Ruben Lobel. The role of surge pricing on a service platform with self-scheduling capacity. *Manufacturing & Service Operations Management*, 19(3):368–384, 2017.

Colin Camerer, Linda Babcock, George Loewenstein, and Richard Thaler. Labor supply of new york city cabdrivers: One day at a time. *The Quarterly Journal of Economics*, 112(2):407–441, 1997.

Juan Camilo Castillo, Dan Knoepfle, and Glen Weyl. Surge pricing solves the wild goose chase. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, pages 241–242. ACM, 2017.

Francisco Castro, Omar Besbes, and Ilan Lobel. Surge pricing and its spatial supply response. 2018.

Antonin Chambolle. An algorithm for total variation minimization and applications. *Journal of Mathematical imaging and vision*, 20(1-2):89–97, 2004.

M Keith Chen and Michael Sheldon. Dynamic pricing in a labor market: Surge pricing and flexible work on the uber platform. In *EC*, page 455, 2016.

Ryan Compton, David Jurgens, and David Allen. Geotagging one hundred million twitter accounts with total variation minimization. In *Big Data (Big Data), 2014 IEEE International Conference on*, pages 393–401. IEEE, 2014.

G García-Martí, A Alberich-Bayarri, and L Martí-Bonmatí. Brain structure mr imaging methods: morphometry and tractography. In *Novel Frontiers of Advanced Neuroimaging*. InTech, 2013.

Alan E Gelfand, Peter Diggle, Peter Guttorp, and Montserrat Fuentes. *Handbook of spatial statistics*. CRC press, 2010.

Fang He, Xiaolei Wang, Xi Lin, and Xindi Tang. Pricing and penalty/compensation strategies of a taxi-hailing platform. *Transportation Research Part C: Emerging Technologies*, 86:263–279, 2018.

Hongyao Ma, Fei Fang, and David C Parkes. Spatio-temporal pricing for ridesharing platforms. *arXiv preprint arXiv:1801.04015*, 2018.

Sara McLafferty, Doug Williamson, and PG McGuire. Identifying crime hot spots using kernel smoothing. *V. Goldsmith. PO McGuire, JH Mollenkopf and TA Ross CRIME MAPPING AND THE TRAINING NEEDS OF LAW ENFORCEMENT*, 127, 2000.

Alex Samuels. Uber, lyft returning to austin on monday. *Texas Tribune*, May 2017. URL https://www.texastribune.org/2017/05/25/uber-lyft-returning-austin-monday/.

James Scott. Spatial smoothing at scale. University Lecture, 2017.

Susan Shaheen, Adam Cohen, and Ismail Zohdy. Shared mobility: current practices and guiding principles. Technical Report FHWA-HOP-16-022, U.S. Depatment of Transportation, 2016. URL https://ops.fhwa.dot.gov/publications/fhwahop16022/index.htm.

Michael Sheldon. Income targeting and the ridesharing market. *Work. Pap., Univ. Chicago*, 2015.

Wesley Tansey. *Scalable smoothing algorithms for massive graph-structured data*. PhD thesis, University of Texas at Austin, 2017.

Wesley Tansey and James G Scott. A fast and flexible algorithm for the graph-fused lasso. *arXiv preprint arXiv:1505.06475*, 2015.

Tolga Tasdizen, Ross Whitaker, Paul Burchard, and Stanley Osher. Geometric surface smoothing via anisotropic diffusion of normals. In *Proceedings of the conference on Visualization'02*, pages 125–132. IEEE Computer Society, 2002.

Lalita Thakali, Tae J Kwon, and Liping Fu. Identification of crash hotspots using kernel density estimation and kriging methods: a comparison. *Journal of Modern Transportation*, 23(2):93–106, 2015.

Robert Tibshirani. Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3): 273–282, 2011.

Bo Wahlberg, Stephen Boyd, Mariette Annergren, and Yang Wang. An admm algorithm for a class of total variation regularized estimation problems. *IFAC Proceedings Volumes*, 45(16):83–88, 2012.

Xiaolei Wang, Fang He, Hai Yang, and H Oliver Gao. Pricing strategies for a taxi-hailing platform. *Transportation Research Part E: Logistics and Transportation Review*, 93:212–231, 2016.

Yu-Xiang Wang, James Sharpnack, Alex Smola, and Ryan Tibshirani. Trend filtering on graphs. In *Artificial Intelligence and Statistics*, pages 1042–1050, 2015.

Zachary D Weller, Jennifer A Hoeting, et al. A review of nonparametric hypothesis tests of isotropy properties in spatial data. *Statistical Science*, 31(3):305–324, 2016.

Jihun Yu and Greg Turk. Reconstructing surfaces of particle-based fluids using anisotropic kernels. *ACM Transactions on Graphics (TOG)*, 32(1):5, 2013.

Liteng Zha, Yafeng Yin, and Yuchuan Du. Surge pricing and labor supply in the ride-sourcing market. *Transportation Research Procedia*, 23:2–21, 2017a.

Liteng Zha, Yafeng Yin, and Zhengtian Xu. Geometric matching and spatial pricing in ride-sourcing markets. 2017b.