2015

Refining the Search and Recovery Process: A Predictive Model for Vehicle Repossessions

Vijay Sachdev
University of Redlands

University of Redlands

Refining the Search and Recovery Process: A Predictive Model for Vehicle Repossessions

A Major Individual Project submitted in partial satisfaction of the requirements for the degree of Master of Science in Geographic Information Systems

by

Vijay Sachdev

Fang Ren, Ph.D., Committee Chair

Douglas Flewelling, Ph.D.

August 2015

Refining the Search and Recovery Process: A Predictive Model for Vehicle Repossessions

by

Vijay Sachdev

The report of Vijay Sachdev is approved.

Douglas Flewelling, Ph.D.

Fang Ren, Ph.D., Committee Chair

August 2015

Acknowledgements

The Major Individual Project was challenging in all of its facets. Completing the project would not have been possible without key contributors who provided me with the support and resources necessary to move forward throughout its various stages.

I first and foremost would like to recognize my advisor, Fang Ren, for her assistance on the project. Fang helped me navigate the various uncertainties involved in statistical analysis while offering her expertise in the field to help me develop a method that produced the strongest results.

I would also like to thank Doug Flewelling for his guidance in the conceptualization of the project. From data models to logic, Doug challenged me to assess the components of my analysis through a critical lens to identify potential errors in reasoning that could skew results.

In addition, I would like to thank Nate Strout for his assistance in the design of the prioritization tool developed for this project and Lisa Benvenuti for her guidance to obtain and mine the parcel dataset. Finally, I would like to thank Lynn Flewelling for editing the report and Mark Kumler for reviewing its content.

Abstract

Refining the Search and Recovery Process: A Predictive Model for Vehicle Repossessions

by

Vijay Sachdev

The automotive repossession industry has been transitioning to locationally aware hardware that permits the use of geographic data to enhance business intelligence operations. GIS initiatives in this industry are relatively new meaning that geospatial trends in repossession data have not been studied in an academic context. This project focuses on identifying trends in license plate recognition scan data collected from high-resolution cameras in Houston, Texas in order to predict future repossession locations. Through the use of opportunity terrain modeling with spatial statistics, a prediction surface was generated that accurately described the habitat of debt by combining seven opportunity variables that were significantly correlated with repossession densities. The findings can be used to narrow the search for vehicles by targeting high-opportunity locations for scanning and recovery.

Table of Contents

Table of Figures

List of Tables

List of Acronyms and Definitions

KDE – Kernel Density Estimation
LPR – License Plate Recognition
MAUP – Modifiable Areal Unit Problem
OLS – Ordinary Least Squares
OTM – opportunity terrain modeling
RTM – risk terrain modeling
VIF – variance influencing factor

Chapter 1 – Introduction

GIS is revolutionizing the automotive repossession industry, which has adopted advanced technology like License Plate Recognition (LPR) to narrow the search for vehicles. LPR utilizes a sophisticated process of text extraction coupled with data processing to match license plate scans to vehicle debtor information. The drivers who operate LPR vehicles and the routes that they drive ultimately determine the number of scans and subsequent repossessions that occur.

This project addresses a repossession company's desire to advance intelligence in its LPR department through predictive analysis. Since their drivers did not operate under any particular structure when scanning, the result was a series of arbitrary driving routes that were not based on empirical studies of repossession patterns or debtor habitats (the places where debtors live, work, and visit). This is partially because the spatial patterns of automotive repossessions have not been studied in an academic context but more importantly because the company was lacking the necessary data and tools to implicate demographic, financial, and geographic indicators as potential contributing factors to automotive delinquency. A better understanding of key variables that act as precursors to debtor habitats would help target opportune areas for scanning. This project served as a framework for this geospatial initiative.

1.1 Client

The client for the project was the License Plate Recognition (LPR) department of a major vehicle repossession company. The company provided approximately 180 LPR cameras to partner agents around the country and was planning on distributing 400 additional cameras at the time of project initiation. These partner agents employed drivers to operate vehicles onto which LPR cameras were attached. Each car collected roughly 10,000 scans per day. Each scan was sent to the company's database in real-time for further processing.

The client uses the term 'hit' to describe a scan that matches a vehicle in the open accounts database. Scans are all reads taken by LPR vehicles out in the field, and an open account refers to a vehicle out for repossession. Each open account has a set of addresses associated to the debtor. It is important to note that scans that do not register as a hit are not necessarily useless. Historical data generated from scans are essential to the daily operations of the client. These data provide leads to the locations of future debtors.

Through their rich endeavors, the repossession company and their partner agents developed a need for a stronger understanding of their historical repossession data in order to target areas where debtors lived, worked, and visited.

1.2 Problem Statement

The client's major problem was that License Plate Recognition (LPR) drivers in Houston drove along arbitrary routes based on personal preference. In one case, an attempt was made to route a driver using *Kernel Density Estimation* (KDE) on open account data. These locations were imported into a routing system where open account addresses located in hotspots were selected, optimized, and then pushed out to a GPS device to route the driver.

While this method proved to be effective at first, the number of hits and repossessions per scan decreased after a couple of months. One explanation for this is that the open account data were exhaustive, pulling associated addresses from each account from multiple years in the past. These addresses could be outdated, and there was no way to validate their accuracy. Secondly, the majority of these addresses included home addresses of debtors, and as most know, people do not stay at home all day. Lastly, this method had very little to do with prediction, so while it may have provided a general understanding of where debtors live, worked, and visited, it said little about the environments that were conducive for successful recoveries.

1.3 Proposed Solution

In order to address the repossession company's needs, a predictive surface generated with an opportunity terrain modeling (OTM) approach and a prioritization tool were proposed. The predictive surface would represent the likelihood of recovering a vehicle at any location in the study area while the prioritization tool would be used to select locations to route to based on the number of open accounts within a specified distance of each location of interest. The deliverables were created in ArcGIS 10.2.2 for Desktop backed by spatial statistical tests conducted in SPSS and GeoDa.

1.3.1 Goals and Objectives

The ability to generate and disseminate knowledge about debtor locations was the primary goal of this project. Increased knowledge would allow routing personnel to effectively direct drivers on their daily routes. While this knowledge would garner more hits and subsequent repossessions in the LPR department, the company also wanted to increase scan volumes in areas that future debtors were likely to inhabit and visit. Increasing scans in high-opportunity areas would provide the company with better leads.

Two objectives accompanied the primary goal. One was to examine and draw conclusions about the characteristics of areas with high rates of recovery. The second objective was to create an opportunity surface to predict the likelihood of finding current and future debtors at any given location in the study area.

1.3.2 Scope

The study area for the project was the Houston, Texas metropolitan area, the largest repossession market in the nation. Since LPR drivers are not confined to the boundaries of the city, all of Harris County (Figure 1-1) was considered in the analysis and output. The only historical data utilized for analysis were repossessions from a two-year period as they provided the most accurate information about debtor habitats with each repossession representing a parked car that was successfully recovered. A final validation on the prediction surface was administered using repossession points collected for three months after the date of the last repossession in the historical dataset.

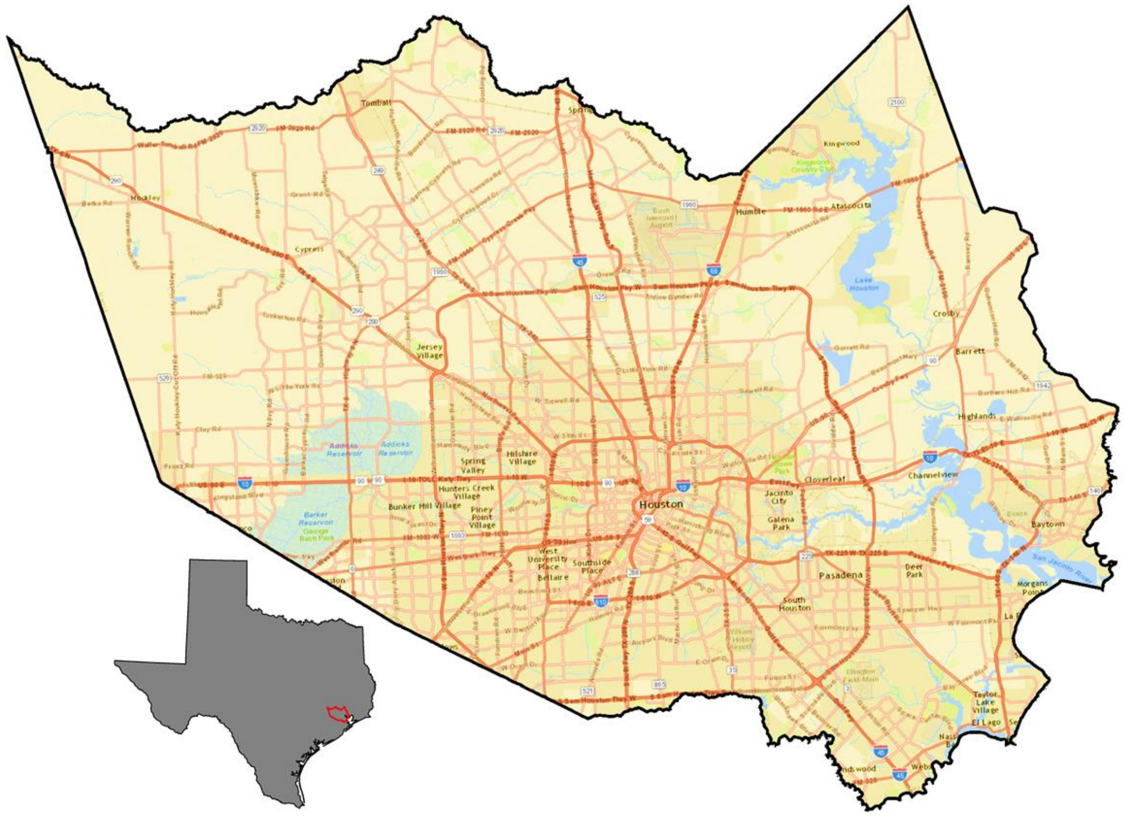

Figure 1-1: Harris County, Texas

The project was managed such that it could be easily replicated in other markets of interest. The deliverables were a predictive surface and a prioritization tool, all of which were delivered to the client for future work and reference in a Geodatabase.

1.3.3 Methods

The prioritization tool was hosted in a File Geodatabase and created with ModelBuilder. The tool was used to summarize repossessions by building types, which will be discussed in detail in Chapter 5.

The workflow used to create the predictive surface followed the steps outlined in Caplan and Kennedy (2010) with the addition of advanced spatial statistical methods to select and test variables for significance prior to the creation of the composite surface. In the final step of the analysis, a cross-tab Chi-square test was used to validate the model.

1.4 Audience

The intended audience for this report are GIS professionals interested in predictive modeling in spatial studies. The reader should possess knowledge of spatial statistics and analysis tools in ArcMap as well as general familiarity with SPSS and GeoDa for statistical analysis.

1.5 Overview of the Rest of this Report

This report provides a detailed summary of the project including background research, system design, database design, statistical analysis, and model validation. The literature review in Chapter 2 will cover hotspot analysis methods and opportunity variables associated with debt. The system design covered in Chapter 3 discusses the requirements of the project as well as the project plan. Each subsequent chapter is devoted to an individual step in the workflow of this project starting with building the database in Chapter 4 and ending with model validation in Chapter 8. The last chapter will summarize conclusions and provide ideas for future work.

Chapter 2 – Background and Literature Review

Although researchers have thoroughly explored key variables that affect the cost of purchasing a vehicle and loan lending practices, the spatial aspect of vehicle loan delinquency and repossessions has yet to be studied academically (Charles, Hurst, & Stephens, 2008). While the latter remains understudied, issues of foreclosure and subprime lending have been under question by several researchers and can provide insight to analyze repossession patterns in this project (Li, 2011; Immergluck & Smith, 2005; Schintler et al., 2010). The topic of crime hotspot analysis and police patrolling has also gained recognition in spatial studies, which has a close analogy to the issue of vehicle recovery. Triangulating these topics can assist in discovering effective methodologies and potential socio-environmental indicators of debtor habitats.

Since locating and recovering vehicles out for repossession can be observed at the crossroads of the aforementioned areas of research, this chapter reviews pertinent literature around topics of loaning practices, debt, and crime analysis. Section 2.1 reviews methods used for hotspot and cluster analysis while section 2.2 investigates potential indicators of debt and loan delinquency.

2.1 Crime Hotspot Analysis and Risk Terrain Modeling

Police patrolling and resource distribution is a major topic of discussion in crime prevention. Hotspot analysis has played a primary role in crime prevention and reduction strategies (Ratcliffe, 2004). Crime hotspots are defined as geographic areas with a significantly higher density of crime relative to comparable areas (Ratcliffe & McCullagh, 2001; Grubesic, 2006; Ratcliffe, 2004). Identifying crime hotspots allows personnel to better allocate resources to areas that are prone to crime rather than patrolling evenly across city boundaries (Braga, 2001). This tactic is called place-oriented or problem-oriented crime prevention, operating under the understanding that crime is not evenly distributed across space (Braga, 2001; Ratcliffe, 2004).

While hotspot analysis and problem-oriented patrolling techniques have proven to be effective in reducing crime, the general methodology of hotspot analysis is often misused (Grubesic, 2006). This is due to a general lack of guidance when it comes to selecting hierarchical or non-hierarchical partitioning methods and the fact that results vastly differ based on employed clustering methods (Grubesic, 2006). Ratcliffe (2004) categorized hotspot analysis of crime into three spatial patterns: dispersed patterns, clustered patterns, and hotpoints. There are several methods used to determine hotspots in a geographic region including cluster analysis, *k* nearest neighbor, and local indicators of spatial association (LISA) including Getis Ord *Gi* and Moran's Local Index (Ratcliffe, 2004, Ratcliffe & McCullagh, 2001). These methods are subject to the modifiable areal unit problem (MAUP) since aggregated data are joined to arbitrary boundaries causing the results to shift as the boundaries change (e.g., from census block groups to census tracts) (Ratcliffe, 2004; Grubesic, 2006). Non-hierarchal partitioning methods such as kernel density analysis, on the other hand, can mitigate the effect of the MAUP as each point is assigned to only one cluster (Grubesic, 2006).

Crime researchers have been interested in quantifying the efficacy of employing hotspot techniques to patrolling. In a review of nine law enforcement strategies implemented in four major cities in the United States and one suburb in Australia to target crime hotspots, Braga (2001) found that seven of nine programs resulted in a significant reduction in crime rates. The resulting crime reduction may be explained by theories in criminology like the deterrent effect (Akers & Sellers, 2013). The basic principle states that individuals make decisions based on the perceived consequences of their actions. In Braga's review, an increase in police visibility likely contributed to an overall higher perceived risk for those with intent to commit crime.

A particularly appealing case Braga reviewed was that of the Kansas City, Missouri Gun Project (Sherman and Rogan, 1995). This study is important because it relates to the daily patrols of LPR drivers with an objective of seizing contraband. Researchers from the University of Maryland collaborated with the Kansas City Police Department on a project design. Together, they chose one target police beat area and one control police beat area for comparison. The project produced statistically significant results in the seizure of guns through proactive patrols including door-to-door investigations, training police to recognize signals related to gun carrying, and "field interrogations in gun crime hotspots" within the police beat as determined by computer analytics at the University of Maryland (Sherman and Rogan, 1995, p. 677). The 29-week study resulted in a statistically significant increase in seizures by 65 percent and a statistically significant decrease of gun crimes by 49 percent.

Groff, et al. (2014) worked with the Philadelphia Police Department to isolate 81 hotspot boundaries based on a map of 2009 violent crime data. The mean area of the boundaries was .044 square miles, each including an average of three miles of streets. Through methods of foot patrolling, problem-oriented policing, and offender-focused policing, their study showed that the latter was the only approach that yielded a statistically significant reduction in crime by 42 percent. This approach identified violent offenders who lived in the target areas while criminal intelligence analysts determined which repeat violent offenders lived a criminal lifestyle and then focused on patrolling the areas they inhabited.

Although methods used in crime hotspot analysis can be a useful reference for conducting a cluster analysis of vehicle repossessions, the innate goals of these two phenomena are very different. Whereas crime prevention focuses on reducing overall crime rates, the repossession industry is not interested in reducing the number of delinquent debtors as this would eliminate the need for the industry altogether. Rather, it focuses on finding the best strategy to locate debtors in both the present and future. To address this issue, historical repossession datasets could be used to analyze social and environmental characteristics of the areas where clusters of repossessions are observed. This information could then be used to predict locations where debtors live and travel. In this regard, risk terrain modeling (RTM) (Caplan, & Kennedy, 2010) in criminology provides a useful approach to study the nature of vehicle repossessions.

Researchers generally accept that hotspot policing (problem-oriented policing) provides the most effective method for crime reduction (National Research Council Committee to Review Research on Police Policy and Practices, 2004). In traditional hotspot analysis, police personnel conceptualize crime in a static environment meaning that locations of crime do not change over time (Caplan, Kennedy, & Miller, 2011).

Caplan and Kennedy (2010) created RTM as an alternative method to predict events of interest. The goal was to equip police with the knowledge to patrol areas based on environmental precursors that create opportunity for crime. Focusing efforts on hotspots yields positive results by virtue, but overtime these results diminish due to the situational fallacy inherent to retrospective mapping (Caplan and Kennedy, 2010). The traditional hotspot approach ignores the dynamic nature of crime, which is more effectively addressed through prediction.

RTM operationalizes risk factors that may present an opportunity for crime or any other event to occur. The combination of these risk factors can lead to hotspots adapted to a changing urban environment. In their case study of Irvington, New Jersey, Caplan et al. (2011) presented the efficacy of RTM in predicting future shooting related crimes. They compiled three risk factors including the known location of gang members, drug arrests, and businesses like bars, strip clubs, check cashing outlets, pawnshops, fast food restaurants, and liquor stores. These risk factors were selected based on empirical evidence and general knowledge about gun related crime prevention. The results showed that the predicted crime locations based on RTM aligned with future shootings twice as well as retrospective hotspot policing would have.

The idea of finding variables that are related to an event of interest and combining them to create a terrain that points to high-risk areas will be a useful method for this project. Caplan and Kennedy (2010) note that these variables can be selected by means of empirical evidence, previous research, or knowledge in the field of interest; however, because there is a lack of knowledge about spatial repossession patterns, this project will empirically study relevant variables in order to confirm their significance to observed repossessions. Furthermore, since vehicle recovery and the presence of debtors is not a case of risk in the repossession industry but rather of opportunity, the method used hereon will be referred to as opportunity terrain modeling (OTM).

2.2 Loan Delinquency Analysis

The identification of hotspots for a phenomenon of interest often prompts an investigation into these spatial clusters. While social and geographic variables that affect the spatial distribution of crime have been well studied (Grubesic, 2006), a more relevant case to vehicle repossessions is focused on foreclosure analysis where researchers have worked to understand the environmental and social factors that determine foreclosure susceptibility in neighborhoods.

Due to the racial and socioeconomic segmentation of the housing market, regressions on foreclosure rates often include the following demographic variables: total and mean population, number of households, unemployment rate, income, race, education, median house value, and median household income (Immergluck & Smith, 2005; Rugh & Massey, 2010; Li, 2011). For example, Schintler, Istrate, Pelletiere, and Kulkarni (2010) used multivariate statistical methods at the neighborhood level to identify the socio-demographic indicators of foreclosure rates in four New England states. Their study defined hotspots as high foreclosure neighborhood tracts that were adjacent to other high foreclosure neighborhood tracts and described coldspots as low foreclosure neighborhood tracts adjacent to other low foreclosure neighborhood tracts. The data from the 2000 Census showed a positive relationship between low-income renter populations that were predominantly black and Hispanic and foreclosure rates, particularly in multiunit properties.

Other studies have bolstered the use of loan data to explain the spatial aspects of foreclosures. The theory holds that even though there may be a significant relationship between the percentage of black and Hispanic populations in a given neighborhood and foreclosure rates, access to certain types of loans can describe the crisis with more certainty. While it is not a main topic of this project, it should be noted that a legacy of discrimination in housing including redlining, a practice used to deny prime loans to borrowers of color, has created a culture of high-risk lending services concentrated in low-income, people of color neighborhoods in the present.

Immergluck and Smith (2005) conceptualized the foreclosure crisis as a result of uneven subprime lending practices, often occurring in racial minority neighborhoods. They tested several variables including demographic variables (unemployment rate, median income, population, median home value, and percentage of black and Hispanic populations), and conventional loan types (home purchase, home improvement, and refinance). They found that subprime loans had a large impact on foreclosure rates at a statistically significant level; however, when they included both demographic and loan variables in the regression, the demographic data became weaker although the percentage of black populations remained statistically significant.

Li (2011) tested the relationship between subprime lending and foreclosures with demographic, housing, and land use data in one county of South Florida and its respective census block groups. Their foreclosure data were calculated as a percentage of Certificate of Titles issued by the court divided by the total number of housing units. They found that the percentage of white populations accounted for 66.86 percent of the variation in subprime loan rates (a negative relationship) and the percentage of population with a college education accounted for 5.23 percent. In the regression of foreclosure rates, they found that the percentage of subprime loans accounted for 33 percent of variation in foreclosures while, similar to Immergluck and Smith (2005), demographic variables around race and education became less significant. Furthermore, their subprime regression model with an R-square value of 0.89 fit much better than their foreclosure regression model (R-square of 0.42). This implies that key variables were omitted from the foreclosure analysis further demonstrating how explaining trends in these data is extremely difficult when controlling for subprime lending. This is likely due to the interdependence that exists between several variables that are tested in foreclosure studies such as race, income, and subprime loans. Li noted that collinearity was not tested for given limitations in the dataset used.

While it is useful to identify key explanatory variables related to foreclosures and mortgages, it should be clear that there are limitations in using these types of studies alone to identify potential indicators of vehicle debt. This is because the practice of lending in the case of mortgages includes a locational analysis of the house being purchased whereas a vehicle loan is assessed solely on characteristics of the purchaser. For this reason, the characteristics of vehicle lending practices is more pertinent to this project.

Unlike describing the characteristics of foreclosures and subprime loans, few have paid attention to the trends of household spending on vehicles. As noted by Charles, Hurst, and Stephens (2008), households are likely to purchase several vehicles over the period that they own a home, and vehicle loans come from two main sources: banks or vehicle manufacturers. In their study, Charles et al. extracted data from 1992 to 2002 waves of the Survey of Consumer Finances to compare discrepancies in vehicle purchases and loans between black and white households. They found that black borrowers had interest rates of 10.6 percent on average versus 9.6 percent for white borrowers, which was a statistically significant difference. Furthermore, black purchasers were more likely to get a loan from a vehicle finance company in which case they were more likely to pay higher interest rates than white borrowers were. In addition, their study showed that significant differences existed between the creditworthiness of black and white borrowers with black borrowers being more likely to have been turned down for a loan or to have been late paying bills by over two months. Therefore, it seems that demographic and socio-economic variables would have important impacts on vehicle loan delinquency.

2.3 Summary

While there are few studies that address the spatial aspects of vehicle repossessions, studies focused on crime and the characteristics of loans and delinquency can augment the lack of knowledge around the topic. Targeting crime hotspots and predicting crime with RTM are proven techniques to reduce and identify criminal activity. These methods shed light on the techniques used in this project to define debtor habitats as places with high-opportunity to recover a vehicle. For example, non-hierarchical partitioning methods like kernel density estimation will help mitigate the effect of the MAUP encountered in this project when analyzing data at different levels of geography while an RTM framework used to model opportunity with OTM will accommodate the multivariate analysis to describe debtor habitats and the likelihood of recovering a vehicle at any given location.

The review of literature on loan lending practices explains that there are clear racial undertones influencing the unbalanced reality of the loan market in both mortgages and vehicle loans. Certain types of loans also have higher associated risks and interest rates (e.g. loans from a car dealership), which helps to understand the behavior of borrowers in their ability to make consistent payments. In addition, demographic and housing variables found to be related to foreclosures and subprime lending can be used to explain the behavior of these types of delinquencies in which case previous research provides a frame of reference to select potential explanatory variables for vehicle loan delinquency and repossession patterns.

Chapter 3 – Systems Analysis and Design

This chapter outlines the major components of the project that were considered prior to the data collection and analysis phases. Section 3.1 provides a description of the client's problem followed by an analysis of project requirements in Section 3.2. Section 3.3 details the system design, and Section 3.4 walks through the project plan.

3.1 Problem Statement

Due to the License Plate Recognition department's lack of knowledge around the nature of debt and repossessions, the routes driven by LPR drivers were arbitrary and based on personal preference. This led to an inefficient use of the company's resources. The client was unsure of the areas that had high debtor densities and was therefore unable to direct its drivers to those areas.

3.2 Requirements Analysis

One of the major deliverables in this project was a predictive surface indicating the chance of finding vehicles out for repossession at any given location. Another was a prioritization tool to help the client optimize locations for routing. The requirements for the project were categorized as functional or nonfunctional.

3.2.1 Functional Requirements

Six functional requirements were defined for the opportunity surface and prioritization tool. These requirements and a description for the surface are listed in Table 3-1 and for the tool in Table 3-2.

Table 3-1: Functional Requirements for Opportunity Surface

Requirement	Description
Opportunity Terrain Modeling	Process should align with guidelines recommended by RTM
Evidence Based	Output includes factors that are significantly correlated with repossessions
Causal Factors	Explains cause and effect of opportunity factors

Opportunity terrain modeling (OTM) was chosen as the approach to define debtor habitats. OTM was modeled off risk terrain modeling (RTM), and the methodology includes several steps outlined by Caplan and Kennedy (2010). These steps were used as a guide to ensure the proper execution of the final surface with the addition of more advanced spatial statistical methods not currently outlined in the RTM manual. Using the chosen statistical approach satisfied the requirement to consider only those explanatory variables that were significantly related to repossessions in order to produce a predictive surface based on empirical evidence rather than conjecture.

Table 3-2: Functional Requirements for Prioritization Tool

Requirement	Description
Input	Allows users to enter points to join to locations of interest
Search Radius	Allows users to enter a search radius for the near tool
Sum	Should sum number of points near input locations

The prioritization tool was built in order to allow the user to join points of interest to a set of target locations to count the number of points that fall 'near' each target location. The tool would also allow the user to specify a search radius when they do not want any points to be considered that are outside a certain distance from each target feature. While the tool was created for route prioritization as a complement to the predictive surface, it was also of use for the analysis of opportunity at the parcel level, which will be discussed in Chapter 5.

3.2.2 Non-Functional Requirements

Five non-functional requirements were defined for the deliverables of the project. A description of each is provided in Table 3-3.

Table 3-3: Nonfunctional Requirements

Requirement	Description
Replicable	Process should be replicable in other markets
ArcGIS 10.2.2	Surface and tool produced with ArcGIS 10.2.2
SPSS	Narrow explanatory variables in SPSS
GeoDa	Regression completed in GeoDa
Extensions	Utilizes Esri's Business Analyst and Spatial Analyst

It was important for the project's workflow to be replicable in other markets around the nation in order to understand the characteristics of vehicle delinquency across diverse geographic contexts. As such, the surface and the prioritization tool were created using ArcGIS 10.2.2 with a Business Analyst extension enabled for data acquisition while SPSS and GeoDa were used for statistical testing of opportunity variables. The steps and tools used in each type of software is outlined throughout this document, providing the client with a guide to repeat the process in new areas of interest.

3.3 System Design

The system design determined the major components of the project as well as the configuration among the components as shown in Figure 3-1. A File Geodatabase called StudyArea.gdb was created to host all inputs, outputs, and the tool. ModelBuilder was used to create the prioritization tool and OTM was applied to generate the opportunity surface. As will be shown in Chapter 4, the area considered for the analysis in StudyArea.gdb is smaller than the final predictive surface stored in HarrisCounty.gdb. The two deliverables were stored in HarrisCounty.gdb and were given to the client along with all data used for the analysis stored in StudyArea.gdb

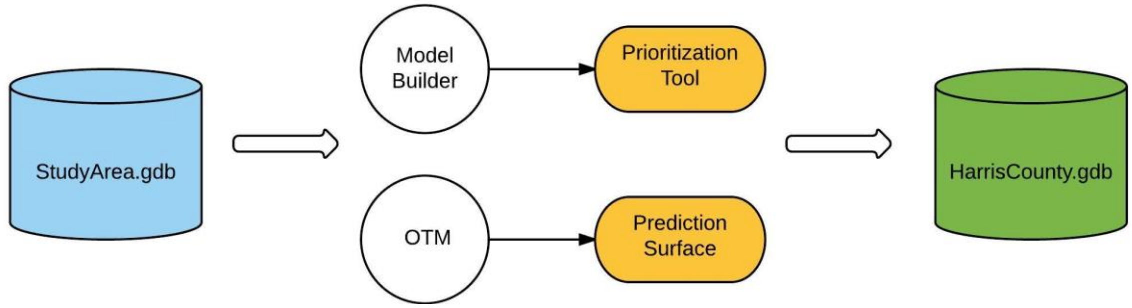

Figure 3-1: System Design

3.4 Project Plan

The project plan shifted as limitations in the client's data were discovered along with new methods of hotspot analysis including OTM. The initial goal was to create a web application for LPR drivers to access and view historical repossession hotspots. However, the client was working on creating a dashboard at the time of this initiative, which needed to be developed in .NET and C# in accordance with IT standards. This was out of scope for this project. Because the client needed information to support better routing decisions, it was decided that a surface displaying the likelihood of recovering vehicles out for repossession would be the most valuable product.

The project followed the workflow detailed in Figure 3-2. Each column represents a step in the OTM process and corresponds to each subsequent chapter in this paper.

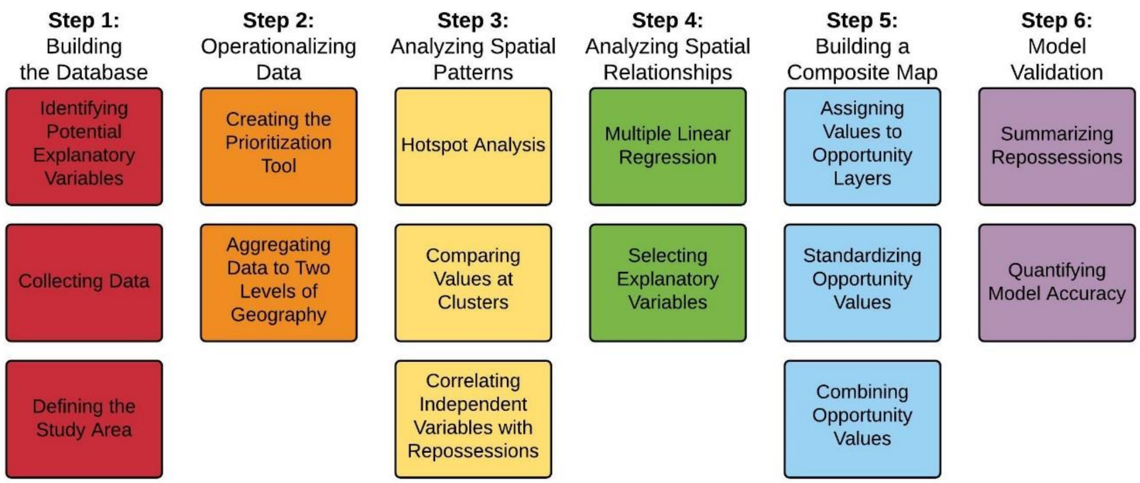

Figure 3-2: Project Plan

The first step was to build the Geodatabase with opportunity factors identified from the literature review and collected from the Business Analyst dataset. These data were clipped to the defined study area.

The second step included the creation of the prioritization tool, which assisted in summarizing repossessions at the parcel level. Repossessions were also aggregated to the census block group level during this step.

The third step was the first analysis phase of the project. Here, a hotspot analysis was completed to compare census block group opportunity values at hotspots versus coldspots. A correlation analysis was also run to select significant opportunity values to include in the regression analysis.

The fourth step included a multiple linear regression on census block group level opportunity values in relation to repossession densities. In this step, final opportunity values were selected to include in the composite map.

The fifth step included the standardization of values for each opportunity layer, which were then combined to form a predictive surface.

The final step in the project, step six, was to validate the model. Here, repossessions collected from a three-month testing period were compared to the predictive surface to quantify the model's accuracy and significance.

3.5 Summary

The system analysis and design helped move the project forward into implementation. Defining the requirements of the project provided a point of reference to evaluate the success and ultimate completion of the project while the system design further defined the major components needed to complete the project at a satisfactory level. Defining the system as a function of requirements and design components created a strong foundation to carry out the steps defined in the project plan.

Chapter 4 – Database Development

This chapter describes the process of building the database to satisfy the client's requirements. The main data components and relationships will be explained in the following sections: Section 4.1 provides a conceptual model of the project, Section 4.2 outlines how the data were organized and stored, Section 4.3 lists the data sources, Section 4.4 explains how data were collected, and Section 4.5 describes the process of creating spatial representations of the collected data. The chapter ends with a summary.

4.1 Conceptual Data Model

The conceptual model provides context about the main components involved in repossessing a car. It also references the entities used to generate a repossession opportunity surface.

Figure 4-1 describes the daily operations of License Plate Recognition (LPR) drivers. LPR drivers scan all plates throughout the routes that they drive. The goal of the LPR driver is to locate vehicles that match delinquent accounts.

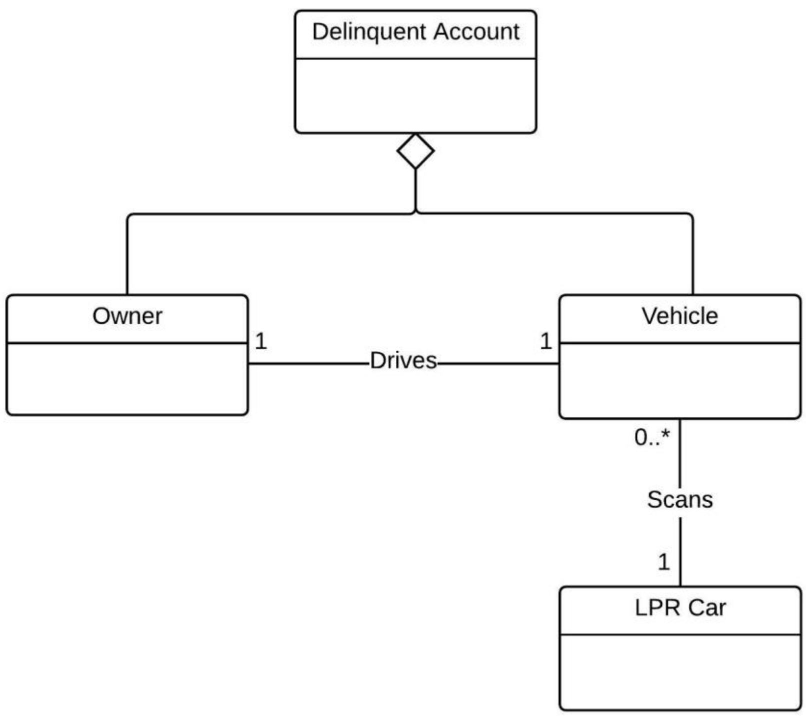

Figure 4-1: Scanning Process

These scans extract text from the license plate, which is sent to the client's server. Scans register as a hit when the text matches a delinquent account's license plate number as shown in Figure 4-2. From here, a notice is sent to a member of the dispatch team who either verifies the scan as a match or declares it a mismatch. There are several reasons why a dispatcher may not create an order for repossession after a hit occurs. For instance, if a car out for repossession is scanned while in motion (e.g. on a highway or main street), it would not make sense to send a tow truck to that location. Furthermore, hits are often invalid when, for example, the plate number matches a delinquent account but the state names are different. Although hit data may provide a detailed view of the places that delinquent debtors have been seen, the accuracy of the hit data is low since several scans that register as hits are invalid and not removed from the client's database. Repossession data, on the other hand, provide locations of successful recoveries and were, thus, more useful for this project.

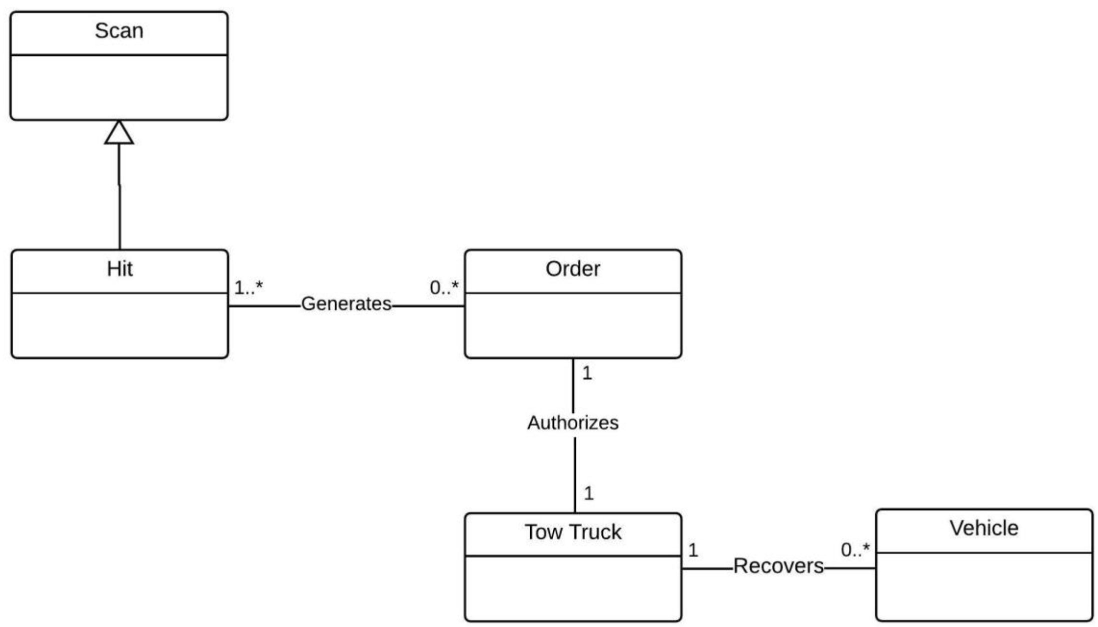

Figure 4-2: Recovery Process

The last conceptual model shown in Figure 4-3 describes the process of defining opportunity at any given location. Here, opportunity is measured through the inheritance of variables at the census block group and parcel level. The main objective of the project was to describe locational trends in repossession data in order to determine the key precursory environmental contexts that make any given area more opportune for scanning and recovery. Since vehicles are driven by owners (Figure 4-1), the idea was to describe the vehicle using the owner as a surrogate and then to describe the owner using census block group variables as a proxy. The parcel data were then used to refine the geography at a granular level such that opportunity would be represented at the parcel level while maintaining opportunity values from census block groups.

In order to test several variables for significance at the census block group level, a kernel density estimation was created for both repossessions and delinquent addresses. These density values were then aggregated to the census block group level such that each block group had a measure of density for both repossessions and delinquent addresses. This process is described in detail in Chapter 5.

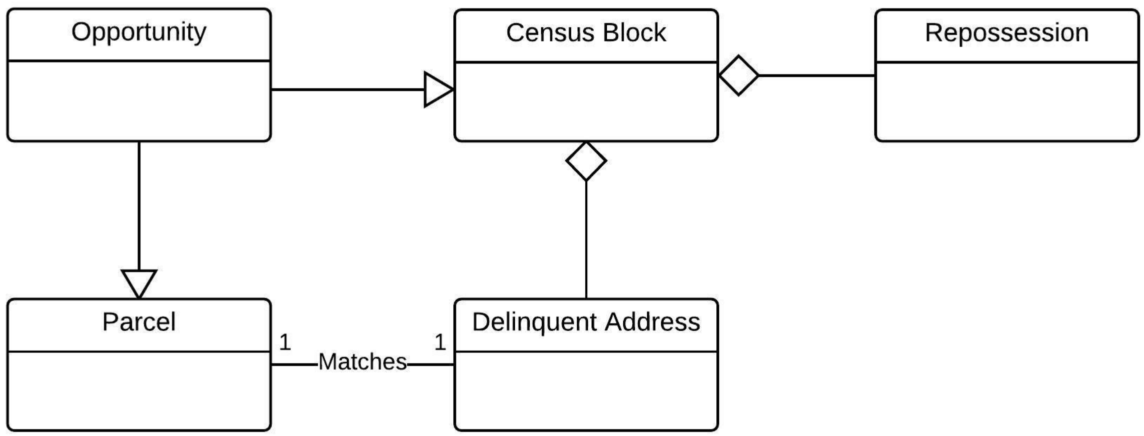

Figure 4-3: Components of Opportunity

4.2 Database Development

The final opportunity surface was generated based on variables that were significantly correlated with repossessions as identified through a multivariate linear regression. Two databases were maintained to store the entities that describe repossessions including a temporary and final Geodatabase.

The temporary file Geodatabase stored all potential opportunity factors for correlation analysis, the parcel layer, and company data including repossessions and open accounts. Potential demographic and financial opportunity factors were selected based on the literature review and examined at the census block group level (Table 4-1). Open account addresses were all of the associated addresses for each assignment (delinquent account) on any given day generated over the month of September 2014.

Table 4-1: Potential Opportunity Factors

Census Block Group
Renter Occupied Household Units (%)
Population density
Average Household Size
White (%)
Diversity Index
Higher Education (%)
Median Age
Median Household Income
Median Net Worth
Median Home Value
Median Disposable Income
Have Auto Loan (%)
Purchased Vehicle: Dealer Financing (%)
Purchased Vehicle: Manufacturer Loan (%)
Purchased Vehicle: Bank or Credit Union Loan (%)

The final file Geodatabase included all opportunity factors used to generate the final opportunity surface for Harris County including demographic, financial, and parcel layers. The final opportunity surface was also stored here along with the prioritization tool developed for the project.

4.3 Data Sources and Collection

All repossession and open account data were collected from the client. These data did not come with metadata. Repossessions were pulled from November 2012 to February 2015 with data from November 2012 to November 2014 used for correlation analysis and data from November 2014 to February 2015 used for model validation.

Demographic and financial data are available from Esri's Business Analyst for Desktop 2014 dataset. These came with detailed documentation including attribute descriptions and collection methods. These data were appended to the study area at the block group level using the *Business Analyst* Toolbar. Finally, the parcel data came from the City of Houston's GIS department. All metadata for these fields are available online with descriptions of each coded value and table. The building types associated with parcels were downloaded separately as text files and joined to the parcel layer.

4.4 Data Scrubbing and Loading

All repossession data came in address form and required geocoding. The client's database changed in 2014, which introduced an inconsistent entry method allowing users to enter repossession locations as single line text as opposed to the original database, which separated repossession locations by address, city, state, and zip code. The latter were easily geocoded using the *Business Analyst* geocoder. The former required the use of

Google's Geocoding API in Excel to process all repossessions for the entire country as there was no easy way to distinguish those that occurred in Harris County from the single line text. These were then added to ArcMap as *XY Events* where all repossessions outside of Harris County were deleted in an editing session.

The Business Analyst data were compiled using the *Append* tool in the *Business Analyst* Toolbar. This tool provided a way to select financial and demographic data of interest and add them as attributes to the study area. Several variables had to be normalized due to variations in populations and households (see table 4-2).

Table 4-2: Normalizing Demographic and Financial Variables

Variable	Normalized By
Renter Occupied Household Units	No. of Household Units
Black Population	Total Population
White Population	Total Population
Age 25+ with Bachelors or Graduate Degree	Total Population Age 21+
Have an Auto Loan	Total Population Age 18+
Purchased Vehicle: Dealer Financing	Total Population Age 18+
Purchased Vehicle: Manufacturer Loan	Total Population Age 18+
Purchased Vehicle: Bank or Credit Union Loan	Total Population Age 18+

The parcel data were processed with Microsoft Access. Building information was downloaded as two text files (one as residential buildings and one as commercial buildings). The text files were then imported into Microsoft Access, which included all of the headings for each text file. Once added to Microsoft Access, the tables were brought into ArcMap, joined to the parcel shapefile, and exported to the temporary Geodatabase.

4.4.1 Creating the Study Area

All of the opportunity layers were clipped to the defined study area. The area of interest for the project was Houston, Texas, but since the drivers of License Plate Recognition Vehicles do not strictly scan within the city boundaries, all of Harris County was initially considered.

Since census block groups were the largest geography of opportunity used in this project, the census block groups that intersected the maximum extent of repossession points from November 2012 to November 2014 (245 points) defined the study area. In order to do this, a convex hull was created from the repossession points, which was then used to select intersecting census block groups. The final study area shown in Figure 4-4 consists of the dissolved version of all census block groups considered for analysis.

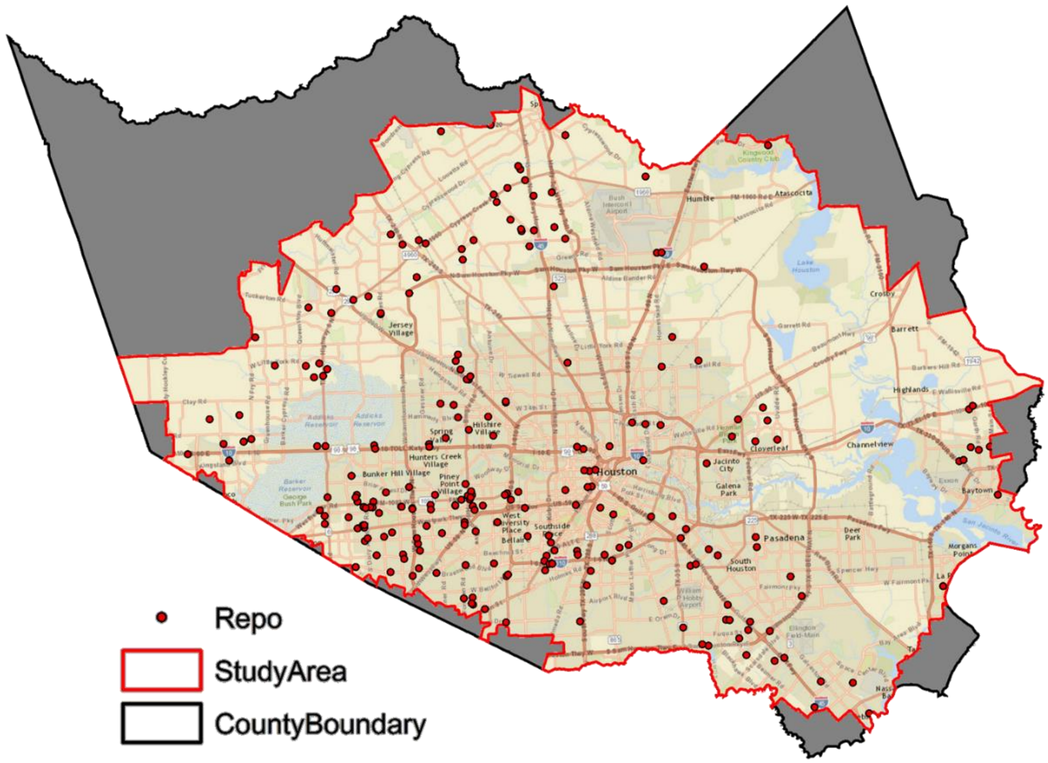

Figure 4-4: Study Area

Eliminating extraneous land reduced uncertainty about where drivers had not scanned in the past, which could skew the results of the analysis. This does not mean that the discarded land was unsuitable for repossessions. Thus, while the analysis only considered the study area, the final prediction surface was applied to all of Harris County.

4.5 Summary

This chapter discussed the main entities of the project including their relationships to one another as they related to the final opportunity surface. The availability and quality of the data from the repossession company largely shaped the scope of this project while access to third party data led the analysis through to completion.

Chapter 5 – Operationalizing Opportunity Factors

It was necessary to operationalize opportunity factors identified through the literature review prior to building the repossession opportunity surface. Opportunity was conceptualized as a combination of three levels of geography including census block groups, parcels, and addresses. Census block groups were used for analyzing demographic and financial variables, parcels included lots and building types, and addresses included associated delinquent addresses. This chapter describes the processing of repossession data to prepare for a statistical analysis of these indicators.

5.1 Categorizing Parcels by Building Type

The first step in preparing repossession data for analysis was to categorize each recovery by the type of building it occurred at (residential, commercial, or agricultural). This was accomplished by joining repossessions to their associated parcel using the prioritization tool created as a requirement for this project.

Since repossessions were geocoded at the street level, an intersect could not complete this type of join since repossessions often fell in-between parcel polygons. Since it was likely for multiple repossessions to occur in the same parcel, a spatial join was also unsuitable for this type of analysis since this tool only allows a user to aggregate join features to target features using the 'intersect', 'distance', or 'closest' functions. The 'closest' function only considers one join feature (repossession) for every target feature (parcel). The prioritization tool remedies this by facilitating a 'near' function along with a 'summary statistics' calculation to append a count of the number of repossessions that occurred at each parcel. Figure 5-1 shows the design of the tool created in ModelBuilder.

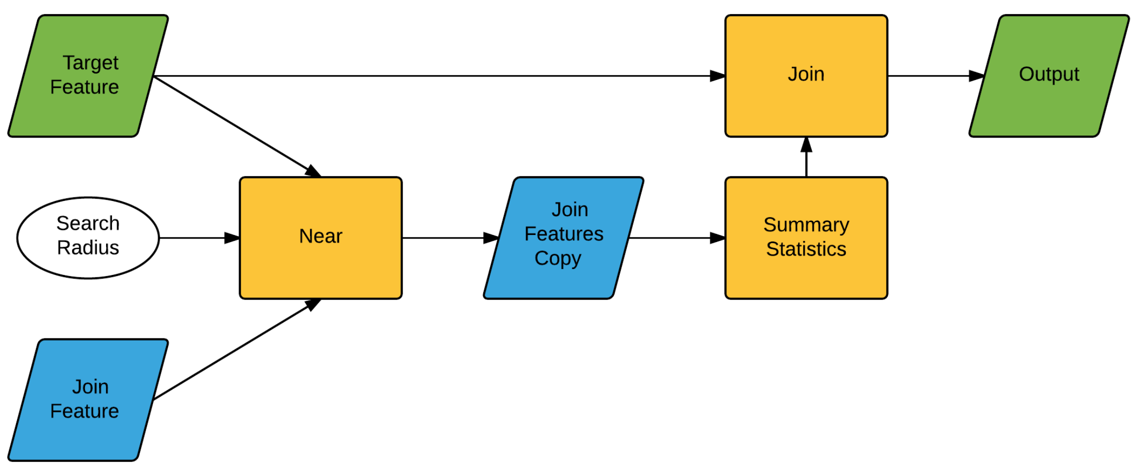

Figure 5-1: Prioritization Tool

The tool is set up with similar parameters as a spatial join asking the user for target features and join features. The tool then generates a new layer with all of the join features (repossessions) and appends an attribute for the Object-ID of the nearest feature (parcel) and another attribute for the distance between the two features. The layer is then summarized using *Summary Statistics* to count the frequency of repossessions that were associated with each parcel. Finally, the tool joins the generated summary table back to a copy of the original parcel layer to include the count of repossessions per record. This tool permits the effective categorization of repossessions by building type at the parcel level, which provided a finer level of granularity to assess opportunity in the final prediction model.

While the result from joining repossessions to parcels will be discussed in detail in Chapter 8, Figure 5-2a, 5-2b, and 5-2c show a generalized distribution of the type of buildings where repossessions occurred.

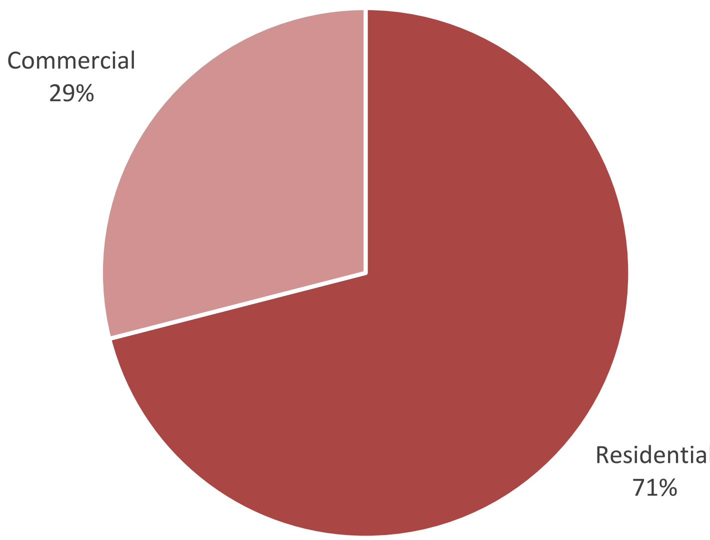

Figure 5-2a: Repossessions by Building Type

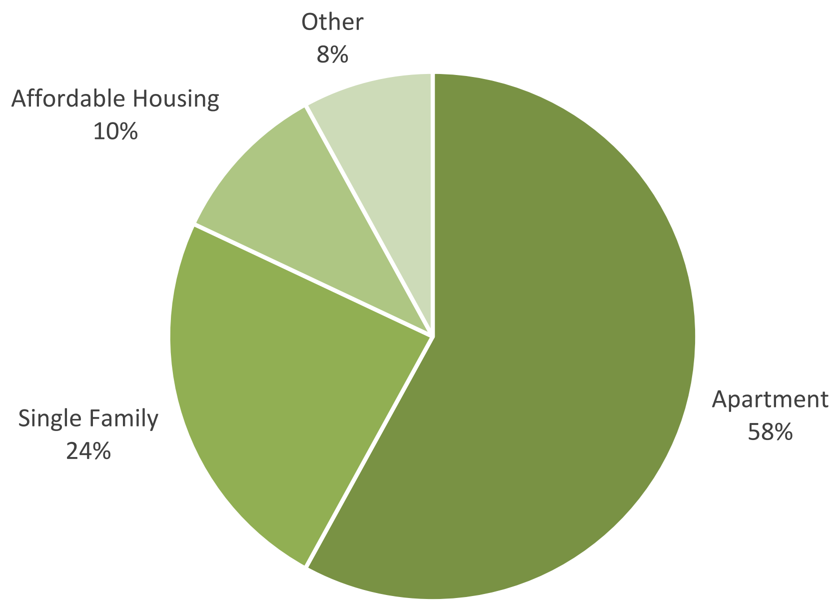

Figure 5-2b: Residential Repossessions

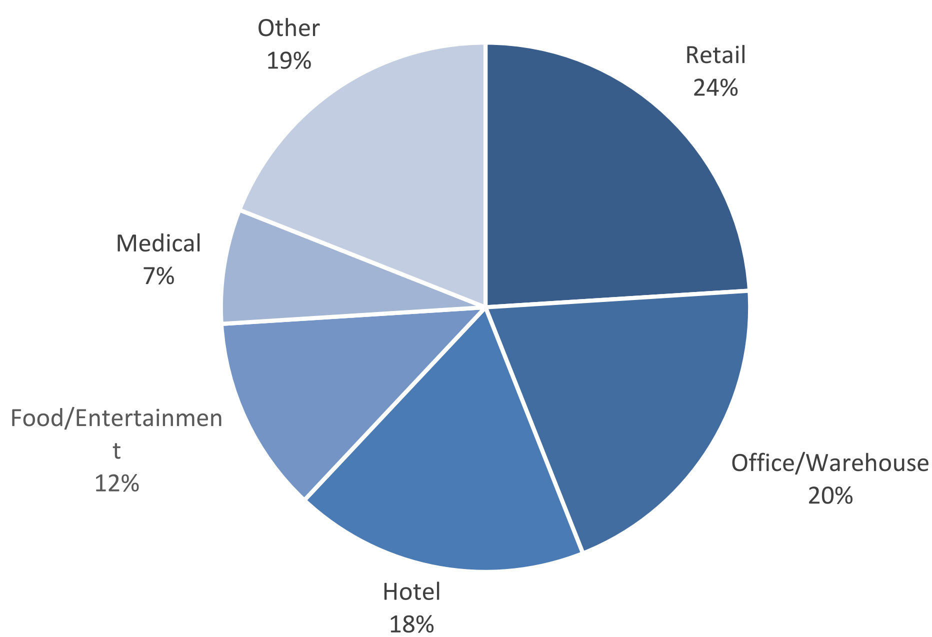

Figure 5-2c: Commercial Repossessions

The results from the prioritization tool show that repossessions occur at certain types of buildings more frequently than others with the distribution heavily concentrated at residential units in comparison to commercial buildings. Over half of residential repossessions occurred at an apartment complex while over half of commercial repossessions occurred at a retail store, office or warehouse, or a hotel. This information will be factored into the final prediction surface as it is apparent that certain types of buildings or residential units present more opportunity for recovery than others.

5.2 Demographic and Financial Factors of Census block groups

Since demographic and financial variables at the census block group level were aggregated from individual household data, the analysis of these variables in relation to repossessions ran the risk of being skewed if residential and commercial repossessions were treated the same. This is because types of businesses and the people who visit them are not necessarily indicative of the profile of people who live near those businesses. Figure 5-3 shows, however, that when categorizing by land use, commercial repossessions occur very close to residential repossessions meaning that the places where debtors live and visit are relatively similar. The conclusion made was that cars recovered at commercial buildings belonged to debtors who lived in a similar type of census block group, and as such, repossessions were combined for the remainder of the analysis.

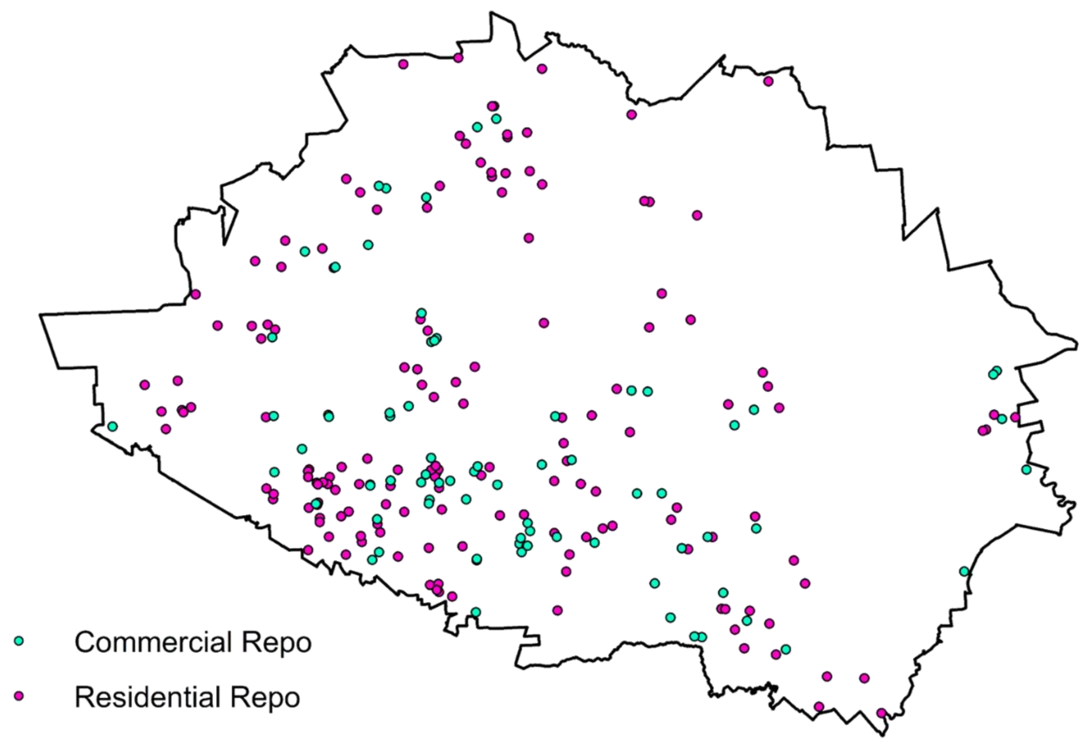

Figure 5-3: Repossessions by Land Use

In order to represent repossessions in a meaningful way with demographic and financial variables at the census block group level, the points were converted into densities and then aggregated to census block groups. *Kernel Density Estimation* (KDE) was used as a non-hierarchical partitioning method to smooth the points shown in Figure 5-4. This method was used to mitigate the effect of the Modifiable Areal Unit Problem (MAUP), which would be more severe if repossessions were aggregated using a simple spatial join. The search radius used for the KDE was approximately twice the value of the average nearest neighbor distance for repossessions.

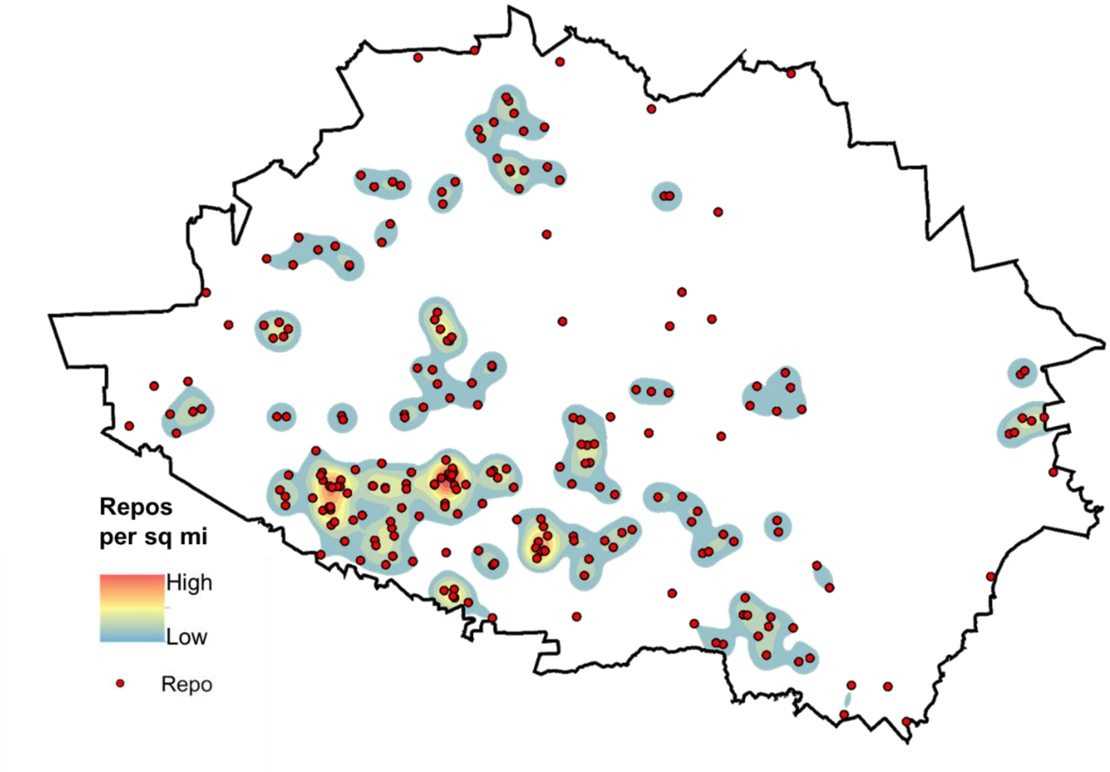

Figure 5-4: KDE on Repossessions

After running the KDE on repossessions, the density was aggregated to the census block group level using the maximum value of the cells within each block as it provided the most variation between block values. The aggregation was accomplished by creating a table with the *Zonal Statistics* tool and joining it to the census block group layer. The output is shown in Figure 5-5, which consisted of the census block group layer with all of the demographic and financial opportunity attributes including a field for repossession density values.

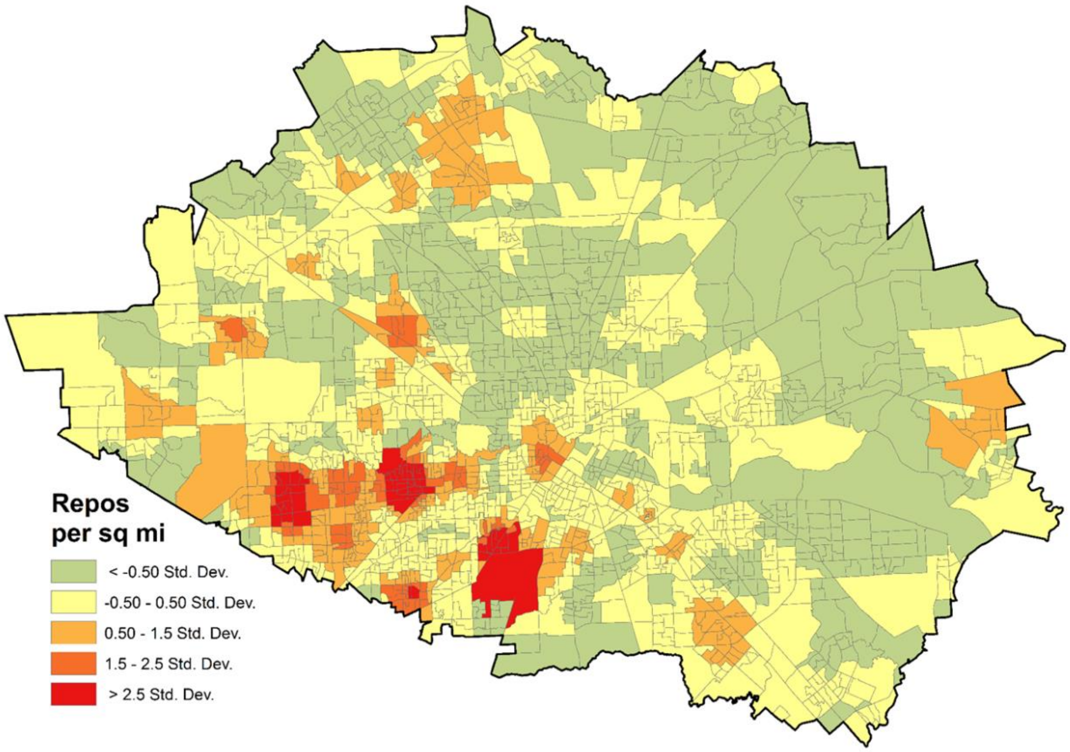

Figure 5-5: Aggregating Repossession Density

5.3 Open Accounts

The final opportunity layer to operationalize was open accounts. Intuitively, one would think that open accounts could accurately predict the whereabouts of vehicle debtors. In reality, the open account dataset was very expansive, covering a large geographic area as shown in Figure 5-6. While these data alone are not suitable for prediction, it was important to test the relationship between associated addresses of debtors and repossession densities to add potential strength to the opportunity surface.

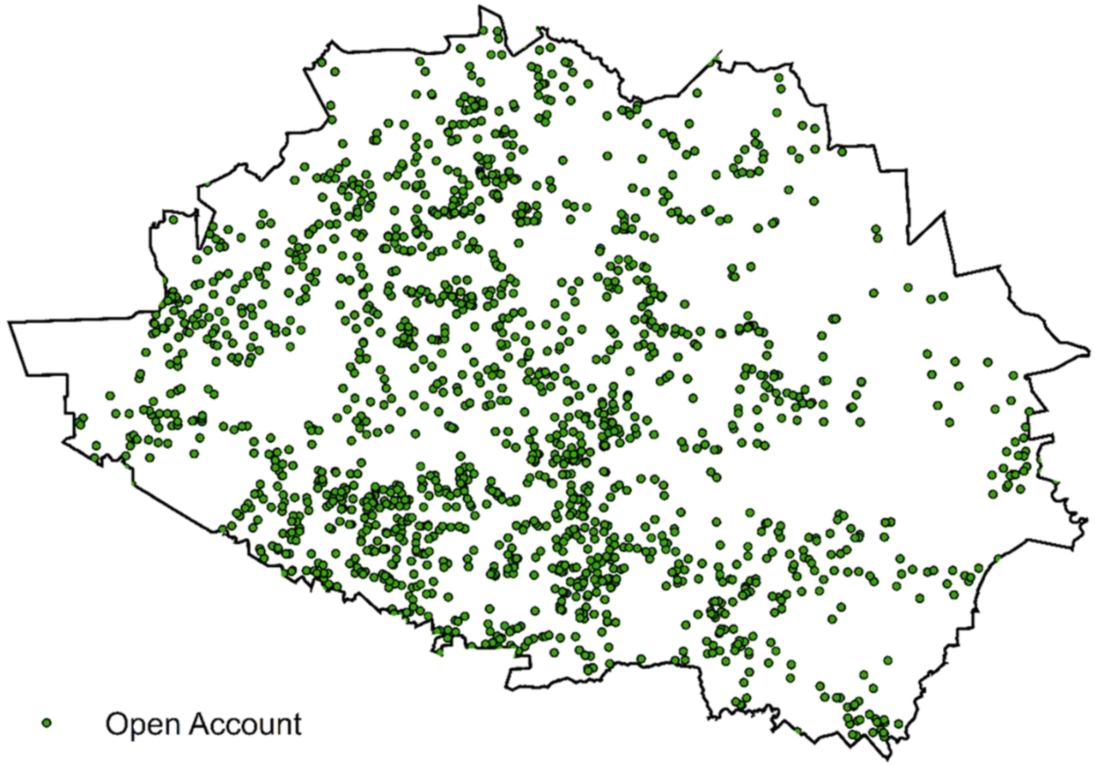

Figure 5-6: Open Accounts September 2014

Both open accounts and repossessions needed to be presented in a similar fashion in order to permit correlation testing between the two. The solution was to smooth open accounts using KDE as shown in Figure 5-7 and aggregate the density values to census block groups using the same technique as repossessions as shown in Figure 5-8.

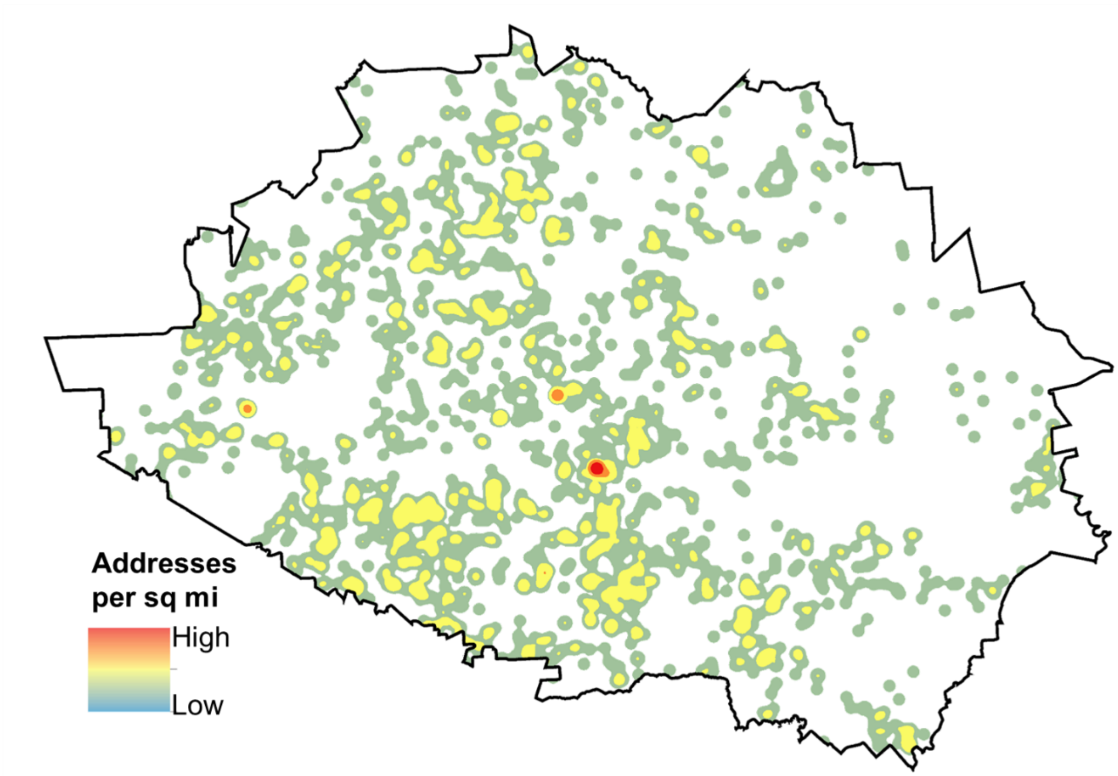

Figure 5-7: KDE on Open Accounts

The search radius used to generate the KDE was approximately twice the value of the average nearest neighbor distance of open accounts. The generated density layer was aggregated to the census block group layer using the *Zonal Statistics* tool such that demographic and financial variables, open accounts, and repossessions were all represented at the same level of geography.

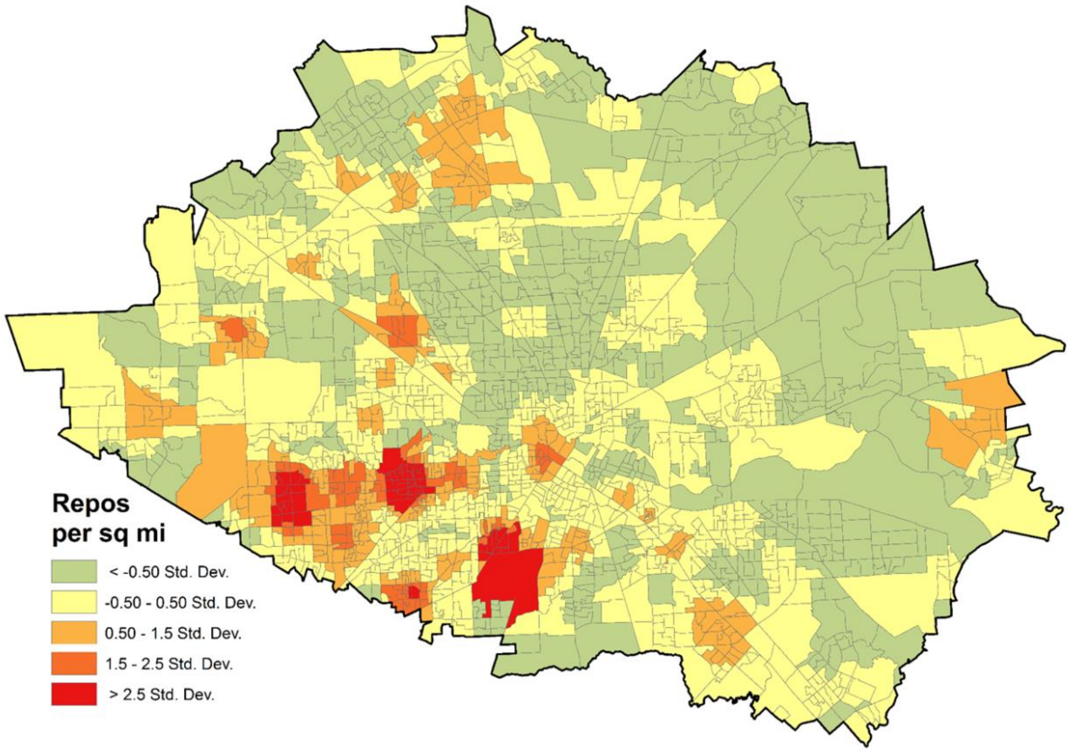

Figure 5-8: KDE on Open Accounts

5.4 Summary

Three categories of opportunity were operationalized to permit statistical testing on their relationship with repossessions. The resulting layers consisted of two levels of geography including census block groups and parcels. The final census block group layer contained attributes for demographic, financial, and repossession variables (repossession density and open account density) while the parcel layer contained an attribute for the number of repossessions that happened at each building type.

Chapter 6 – Analyzing Spatial Patterns

With all of the data operationalized, the next step was to examine the spatial distribution of repossessions to identify significant clusters and to analyze significant demographic, financial, and repossession variables at these clusters. This chapter discusses the cluster analysis process along with the method used to identify important opportunity factors.

6.1 Hotspot Analysis of Repossession Densities

The *Hotspot Analysis* tool with Getis-Ord *Gi* statistics was used to identify significant clusters of census block groups with high and low repossessions densities. As discussed in Chapter 5, repossessions were aggregated to the census block groups using *Kernel Density Estimation* (KDE) and *Zonal Statistics* in order to observe the correlation between demographic and financial variables, and repossession densities. As shown in Figure 6-1, there are clusters of both types within the study area at a confidence level of 90 percent. The next task was to determine the demographic and financial data that were significantly different at hotspots versus coldspots. The census block group data were exported to Excel using the *Conversion* toolset in order to create a dataset in SPSS for analysis.

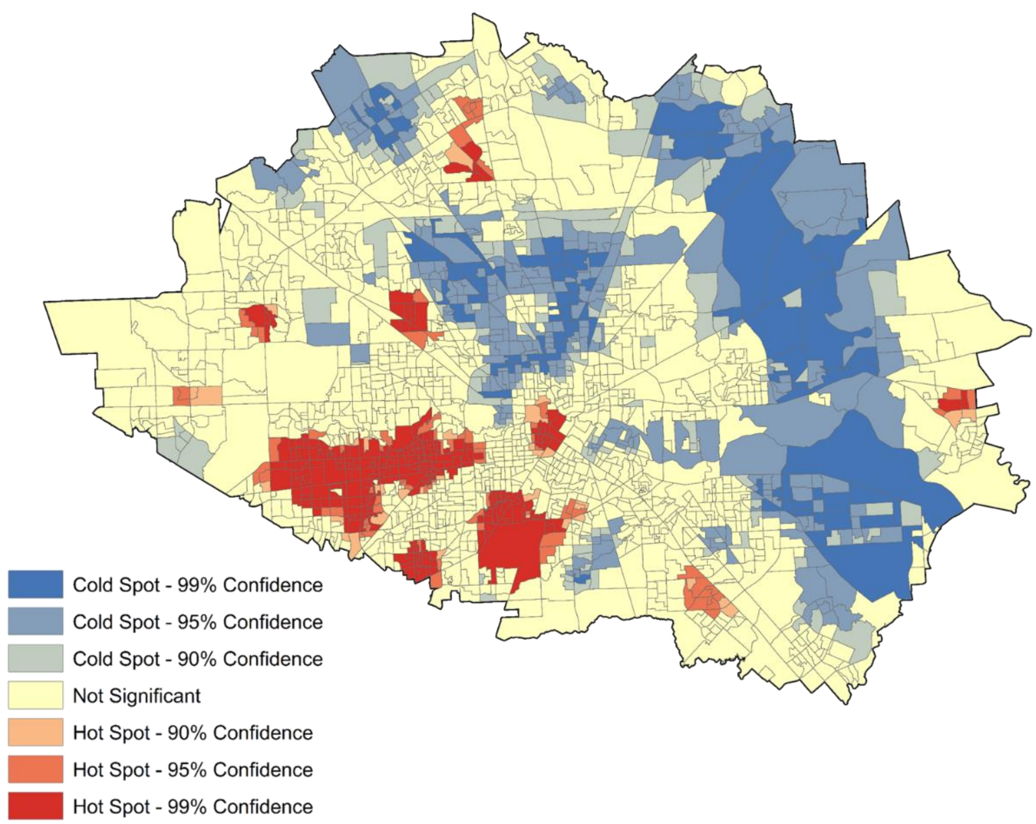

Figure 6-1: Hotspot Analysis on Repossession Density

6.2 Independent Sample T-Test

After compiling an Excel spreadsheet with all census block groups from the hotspot analysis, a two sample T-Test was completed in SPSS to determine the variables that had significantly different mean values at hotspots versus coldspots. Since the two sample T-Test assumes samples to be independent, 100 samples were randomly chosen from both hotspots and coldspots with at least a 90 percent confidence level in order to remove samples that contributed similar information. The 100 samples were grouped by hotspots and coldspots to compare the demographic, financial, and open account data. Table 6-1 shows that, out of the 17 variables tested, 11 were significant.

Table 6-1: Two Sample T-Test Results

Variable	Probability
Population Density	$p<.01$
Average Household Size	$p<.01$
Median Age	Not Significant
Median Household Income	$p<.05$
Diversity Index	$p<.01$
Median Disposable Income	$p<.05$
Median Net Worth	$p<.01$
Median Home Value	Not Significant
Unemployment Rate	Not Significant
Renter Household Units (%)	$p<.01$
White (%)	Not Significant
Have Auto Loan (%)	Not Significant
Higher Education (%)	$p<.01$
Dealer Financing (%)	$p<.01$
Bank Credit Union Loan (%)	$p<.01$
Manufacturer Loan (%)	Not Significant
Open Accounts Density	$p<.01$

Compared to coldspots, hotspots had significantly higher values for population density, median household income, diversity index, the percentage of renter household units, the percentage of people with a bachelor's degree or higher (Higher Education), the percentage of people who purchased their last vehicle with dealer financing, the percentage of people who purchased their last vehicle with a bank or credit union loan, and the density of open accounts. The measure for median household income and the percentage of people with a bachelor's degree or higher was unexpected based on the literature review and required closer observation.

At the same time, hotspots had significantly lower values for average household size, median disposable income, and median net worth. These results were expected based on the literature review.

Finally, among the variables tested, the values for median age, median home value, unemployment rates, the percentage of white people, the percentage of people with an auto loan, and the percentage of people who purchased their last vehicle with a vehicle manufacturer loan were not significantly different at hotspots versus coldspots. These variables were disregarded for the remainder of the analysis.

6.3 Bivariate Correlation Test

After the potential indicators of debt were reduced to those demographic, financial, and open account data that were significantly different at hot and cold clusters, the next task was to run a correlation test between each opportunity factor and repossession density. This was accomplished using a Bivariate Correlation test in SPSS to confirm the relevance of the opportunity variables across the study area.

The test of bivariate correlation coefficient requires samples to be independent of each other, and as such, a stratified sample was collected using 50 percent of records from each of seven categories of census block groups. The seven categories included hotspots at a 90 percent, 95 percent, and 99 percent confidence, coldspots at a 90 percent, 95 percent, and 99 percent confidence, and those that were insignificant in the hotspot analysis.

The results in Table 6-2 show that all variables that remained significant after the two-sample T-Test were significantly correlated with repossessions. Of these variables, population density, diversity index, renter household units, higher education, dealer financing, bank/credit union loan, and open accounts were positively related to repossessions while average household size, median household income, median disposable income, and median net worth were negatively related to repossessions. Except for the variable for higher education, all variables showed the expected relationship to repossessions based on the literature review.

Table 6-2: Bivariate Correlation Analysis

Variable	Pearson Correlation	Probability
Population Density	0.273	$p<.01$
Average Household Size	-0.303	$p<.01$
Median Household Income	-0.058	$p<.01$
Diversity Index	0.111	$p<.01$
Median Disposable Income	-0.057	$p<.05$
Median Net Worth	-0.104	$p<.01$
Renter Household Units (%)	0.329	$p<.01$
Higher Education (%)	0.178	$p<.01$
Dealer Financing (%)	0.179	$p<.01$
Bank Credit Union Loan (%)	0.201	$p<.01$
Open Accounts	0.209	$p<.01$

6.4 Summary

This chapter described the methods used to analyze the spatial distribution of repossession densities in order to eliminate opportunity layers that were insignificant. Among 17 variables, 11 were significantly different at hotspots versus coldspots. Correlation testing on all census block groups further confirmed the opportunity factors that were significantly related to repossessions overall. Of the 11 variables used in the bivariate correlation test, all were significant. The next step in the project was to run a multivariate analysis on the variables against repossession densities to select opportunity variables to include in the final prediction surface.

Chapter 7 – Analyzing Spatial Relationships

Once the variables significantly correlated with repossessions were found, the next step in the analysis phase was to run a multivariate analysis between explanatory variables and repossessions to quantify their relationship. This was accomplished with multilinear regression techniques discussed throughout this chapter.

7.1 Multiple Linear Regression

A multiple linear regression was completed using the *Ordinary Least Squares* (OLS) tool in ArcMap to determine the relationship between the explanatory (opportunity) variables and the dependent variable (repossession densities) from this study.

The 11 variables from Table 6-2 were included in the first OLS analysis. The results from this model, shown in Table 7-1, include a coefficient to measure the relationship between independent variables and repossessions with a probability value based on the T-Test statistic and a variance influencing factor (VIF), which measures the severity of multicollinearity.

Table 7-1: OLS Trial #1

Variable	Coefficient	Probability	VIF
Population Density	0.000016	p<.01	1.31
Average Household Size	-0.005272	Not Significant	3.56
Median Household Income	-0.000003	Not Significant	200.71
Diversity Index	0.004664	p<.01	1.77
Median Disposable Income	-0.000002	Not Significant	199.26
Median Net Worth	0.000001	p<.01	7.13
Renter Household Units (%)	0.612741	p<.01	3.38
Higher Education (%)	1.030357	p<.01	5.31
Dealer Financing (%)	-1.601339	p<.05	4.5
Bank Credit Union Loan (%)	3.582886	p<.01	3
Open Accounts	0.00984	p<.01	1.08

F-Statistic = 63.512221, p<.01, R-Square = 0.26

The VIF statistics point to a multicollinearity issue between median household income and median disposable income. This is logical as one would expect a household's disposable income to decrease or increase as the income decreases or increases respectively. Along with Average Household Size, these two variables did not happen to be significant and were excluded from the second regression. Table 7-2 provides the results from the second OLS test in which case all controlled variables were significant. The model is valid given the significant F test result; however, the fitness of the model is not desirable (R-Square = 0.26).

Table 7-2: OLS Trial #2

Variable	Coefficient	Probability	VIF
Population Density	0.000016	p<.01	1.29
Median Household Income	-0.000004	p<.01	7.88
Diversity Index	0.004606	p<.01	1.56
Median Net Worth	0.000001	p<.01	7.11
Renter Household Units (%)	0.619739	p<.01	2.71
Higher Education (%)	1.037238	p<.01	4.17
Dealer Financing (%)	-1.589858	p<.05	4.35
Bank Credit Union Loan (%)	3.600852	p<.01	2.84
Open Accounts	0.009883	p<.01	1.07

F-Statistic = 85.57, p<.01, R-Square = 0.26

Examining the residuals from OLS is an important step in regression analysis. The OLS tool in ArcMap creates an output layer with residual values for each observation. The output map from OLS results is illustrated in Figure 7-1. The map identifies census block groups that have relatively high and low predicted repossession values in comparison to the observed value.

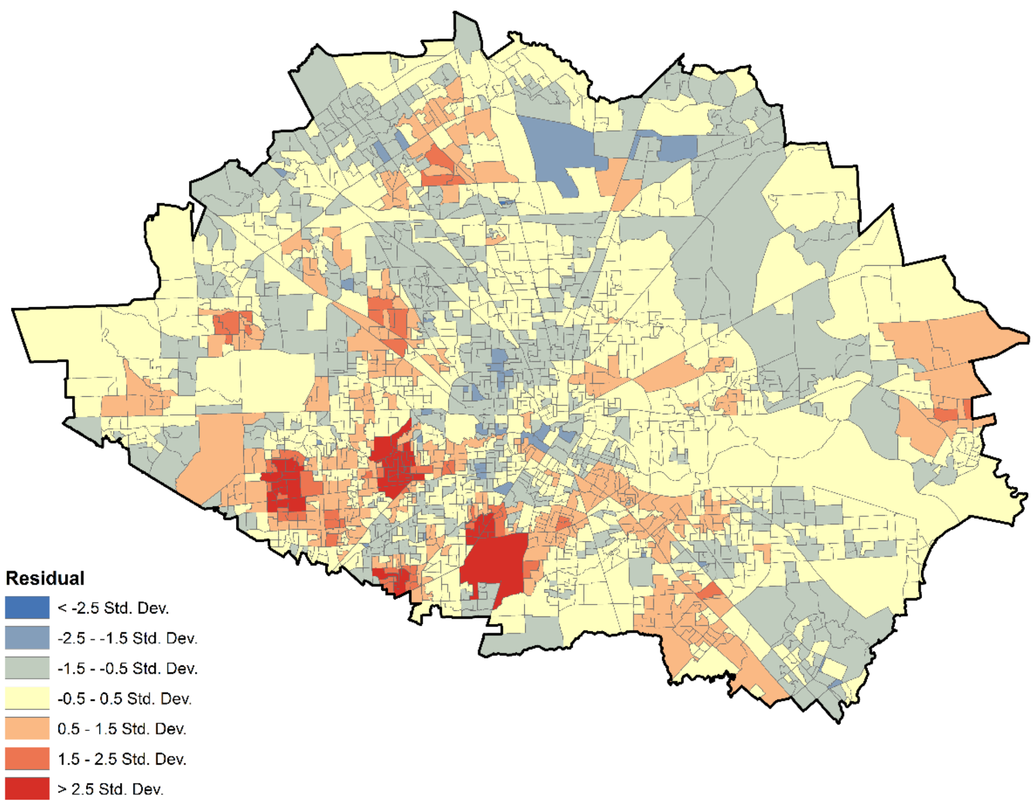

Figure 7-1: OLS Residual Map

In order to accept the model completed through the OLS test, a Global Moran's I was generated for the mapped residuals with a first order queen weights matrix to identify a potential issue with spatial autocorrelation. A returned Moran's Index of 0.7116 with p<.01 indicated that the residuals were autocorrelated, showing that the model from OLS on repossessions could not be trusted.

7.2 Spatial Regression Analysis

Due to problems with spatial autocorrelation, OLS results could not be considered for the prediction surface. As a result, a spatial regression was completed on demographic, financial, and open account variables as an alternative approach. There are two types of spatial regression models including the Spatial Lag model and the Spatial Error model. An OLS regression on the same nine variables from Table 7-2 was completed in GeoDa with the addition of a first order queen contiguity weights matrix in order to generate Lagrange Multiplier statistics as shown in Table 7-3.

Table 7-3: Autocorrelation Statistics from OLS Output

Statistic	Value	Probability
Lagrange Multiplier (lag)	3249.2266	0.00
Robust LM (lag)	366.1371	0.00
Lagrange Multiplier (error)	2910.8484	0.00
Robust LM (error)	27.7589	0.00

The GeoDa workbook created by Anselin (2006) was used to determine the best spatial regression model through the comparison of these statistics. Because all of the statistics had equally significant probabilities, a lag model was chosen as the most suitable due to the larger generated value of the robust lag statistic in comparison to the robust error statistic.

Generating a Spatial Lag regression was very similar to OLS. The dependent variable was the repossession density and the independent variables included those that were significant from the OLS analysis in Table 7-2. A first order queen contiguity weights matrix was applied to the regression. The output from the Spatial Lag regression, shown in Table 7-4, included the coefficient of each variable in addition to a lag coefficient.

Table 7-4: Spatial Lag Model Trial #1

Variable	Coefficient	Probability
Lag Coefficient	0.9557	p<.01
Population Density	-2.68E-06	p<.01
Median Household Income	-4.78E-07	Not significant
Diversity Index	0.0007	p<.05
Median Net Worth	1.14E-07	Not significant
Renter Household Units (%)	0.1286	p<.01
Higher Education (%)	0.0478	Not significant
Dealer Financing (%)	0.1866	Not significant
Bank Credit Union Loan (%)	0.2155	Not significant
Open Accounts	0.0031	p<.01

The output from the Spatial Lag model showed that several independent variables were no longer significant when controlling for spatial autocorrelation. However, when comparing the Log Likelihood, Akaike Info Criterion, and Schwarz Criterion between the OLS and Lag results, there were drastic improvements shown in Table 7-5.

Table 7-5: Model Improvement

	OLS	Lag
Log Likelihood	-1466.67	302.516
Akaike Info Criterion	2953.34	-583.032
Schwarz Criterion	3009.34	-521.428

After removing the insignificant independent variables and rerunning the regression, the variables with corresponding values and significance levels shown in Table 7-6 remained. Based on the two-sample T-Test and bivariate correlation analysis completed in Chapter 6, the direction of the relationship between diversity, renter household units, higher education, bank/credit union loan, and open accounts to repossessions matched the expected outcome. Population density, however, became negatively related to repossessions. This was likely due to the fact that population density tends to be highly autocorrelated, and when controlling for autocorrelation of repossession densities in the Spatial Lag model, its strength was reduced. For the purpose of creating a predictive surface, population density was included as a contributing opportunity factor with a positive relationship to recoveries since the surface does not include a variable to account for the lag coefficient.

Table 7-6: Spatial Lag Trial #2

Variable	Coefficient	Probability
Lag Coefficient	0.9564	p<.01
Population Density	-2.57E-06	p<.01
Diversity (Index)	0.0006	p<.05
Renter Household Units (%)	0.1173	p<.01
Higher Education (%)	0.0406	p<.10
Bank Credit Union Loan (%)	0.3545	p<.05
Open Accounts	0.0032	p<.01

A Moran's scatter plot was generated for the Spatial Lag model's residual values displayed in Figure 7-2. The model was affirmed with a Moran's Index very close to zero and an insignificant p-value meaning that the issue of autocorrelation from OLS results was, in fact, corrected.

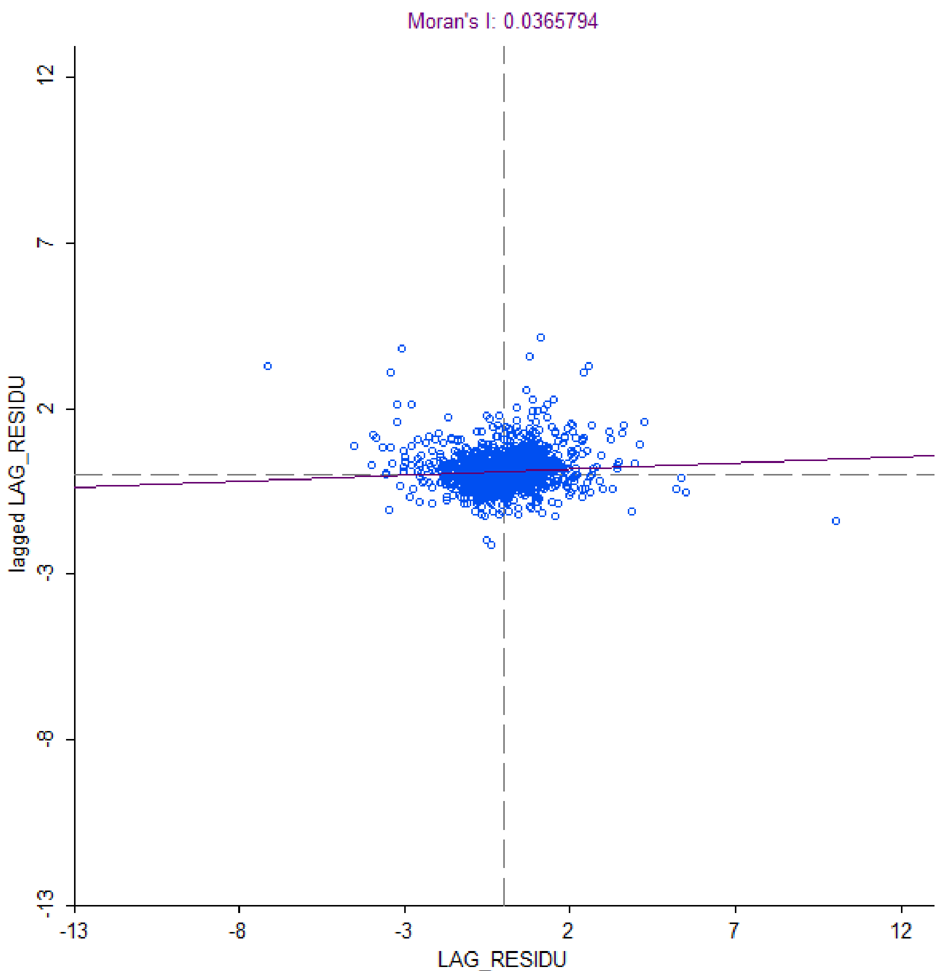

Figure 7-2: Moran's Scatter Plot of Lag Residuals

The last task was to analyze the relative strength of each significant opportunity factor on repossession rates. Table 7-7 shows a Spatial Lag model run with z-score values for each variable. Since all variables were standardized, the coefficients indicate the strength of their effect on repossession densities relative to one another. The results show that all of the explanatory variables, excluding the lag coefficient that was not considered for the final output, had a similar level of effect on repossessions densities. Thus, it would not be appropriate to apply weights to each opportunity layer for the final output.

Table 7-7: Spatial Lag Z-Score Values

Z-Score Variable	Coefficient	Probability
Lag Coefficient	0.9558	p<.01
Population Density	-0.0233	p<.01
Diversity (Index)	0.0181	p<.05
Renter Household Units (%)	0.0461	p<.01
Higher Education (%)	0.0157	p<.10
Bank Credit Union Loan (%)	0.0183	p<.05
Open Accounts	0.0411	p<.01

7.3 Summary

A multiple linear regression was completed using a Spatial Lag model to correct issues of autocorrelation that were present in the OLS regression. This method identified six significant independent variables that explained repossession densities in the study area and were used for the final prediction surface. These variables included population density, diversity, renter household units, higher education, vehicle/bank loans, and open accounts, which all had a positive relationship with repossession rates.

Chapter 8 – Building a Composite Map

With all opportunity layers operationalized and summarized in terms of their contribution to repossession rates, the next step was to create a visual representation of repossession opportunity for Harris County and validate the result with repossessions that occurred between November 2014 and February 2015. This chapter will lay out each step involved in building the opportunity surface and validating the model.

8.1 Predicting Vehicle Repossession Likelihood

The likelihood of recovering debtors' vehicles is represented by an opportunity surface that considers the various opportunity factors associated with vehicle repossessions. In this project, seven opportunity factors were identified and operationalized including demographic, financial, and building type. These variables were standardized and then combined to create the final composite opportunity map.

8.1.1 Standardizing Opportunity Factors

Six variables—including population density, diversity, renter household units, bank/credit union loans, and open accounts—were considered contributing (positive) opportunity factors at the census block group level. Each block group was assigned a standard value based on its values for each opportunity factor using Eq. 8-1 to generate a number between zero and one with zero representing no opportunity and one representing the highest opportunity. This equation is only appropriate for opportunity factors with a positive relationship to repossessions. The result was six new attributes representing standardized opportunity scores for each contributing factor.

$$\frac{\text{value - min}}{\text{max - min}}$$

Eq. 8-1

A closer look at the frequency of repossessions that happened in each parcel, which was achieved using the prioritization tool discussed in Section 5.1, showed that all 245 repossessions happened at a commercial or residential building. This could be due to the fact that the client has to associate an address with each repossession. Without any evidence that repossessions happened in lots without buildings, the assumption made was that cars are only picked up at lots with an address.

Each parcel was assigned an opportunity score based on its corresponding building type. Each building type was then summarized by the number of repossessions that occurred. The result showed that a total of 43 types of buildings had at least one repossession in past two years. The summary statistics in Table 8-1 shows each type of building with its number of observed repossessions. Building types with a null value were parcels with no building while building types labeled 'other' were all building types with no repossessions. The column called 'Count' was the total number of buildings of each type, 'Repos' were the number of recoveries that happened at each type, and the 'Opportunity' was the standardized value created for each type.

The opportunity value describes the probability of recovering a vehicle at any given building type. It was calculated by the ratio of repossessions at each building type to all repossessions divided by the ratio of the number of each building type to the total number of buildings. For example, even though only three recoveries were made at cinemas, the opportunity associated with this type of building was very high given that there were only 15 buildings of this type in the study area.

This value was converted to standard values using Eq. 8-1, with one representing buildings associated with the highest opportunity and zero representing buildings with the lowest. Parcels without a building maintained a null opportunity value as the potential of recovering a vehicle on an agricultural field, for example, is virtually nonexistent without any evidence of this type of recovery occurring over the past two years.

Table 8-1: Summary Statistics by Building Type

Building Type	Count	Repos	Opportunity
Cinema/Theater (Multi-Screen)	15	3	1
Apartment High Rise (13+ Stories)	13	1	0.3846
Extended Stay Hotels/Motels	95	5	0.2632
Apartment Structure	66	3	0.2273
Apartment - Tax Credit	237	11	0.2321
Apartment Garden (1 to 3 Stories)	2270	93	0.2048
Subsidized Housing	289	7	0.1211
Hotel/Motel, Low-Rise (1 to 3 Stories)	528	8	0.0758
Community Shopping Center	136	2	0.0735
Auto Dealer Full Service	239	3	0.0628
Recreational/Health	197	2	0.0508
Nursing Home	101	1	0.0495
Retirement Home	106	1	0.0472
Neighborhood Shopping Center	546	5	0.0458
Residential Fourplex	120	1	0.0417
Drugstore (Freestanding)	249	2	0.0402
Service Center Warehouse	568	4	0.0352
Parking Garage	158	1	0.0316
Supermarket	181	1	0.0276
Light Industrial - Non Metallic	237	1	0.0211
Office Bldgs. Hi-Rise (5+ Stories)	516	2	0.0194
Car Wash (Manual)	338	1	0.0148
Medical Condominium	393	1	0.0127
Retail Condominium	402	1	0.0124
Club House	438	1	0.0114

Building Type	Count	Repos	Opportunity
Fast Food	1436	3	0.0104
Bar/Lounge	554	1	0.0090
Medical Office	1186	2	0.0084
Office - Warehouse	609	1	0.0082
Restaurant	1240	2	0.0081
Strip Shopping Center	2151	3	0.0070
Apartment Struct. 4-20 Units	2985	3	0.0050
Distribution Warehouse	1044	1	0.0048
Retail Multi-Occupancy	2235	2	0.0045
Service Station (Self)	1282	1	0.0039
Warehouse - Metallic	12407	6	0.0024
Retail Single-Occupancy	4570	2	0.0022
Office Bldgs. Low-Rise (1 to 4 Stories)	2646	1	0.0019
Residential Duplex	7959	3	0.0019
Residential Condo	63345	10	0.0008
Residential Townhome	21641	1	0.0002
Residential Single Family	971977	42	0.0002
Other	35013	0	0
<Null>	199454	0	<Null>
Total	**1342172**	**244**	

8.1.2 Combining Opportunity Layers

Each opportunity factor from the census block group and parcel layer was combined by completing an intersect on the data. The output consisted of the shared boundaries of the two layers with seven standardized opportunity scores. These scores were combined using Eq. 8-2 where variable X_i represents each opportunity layer's standardized value and $\bar{X}$ represents the final opportunity output into a new field called 'Opportunity.' This equation normalizes opportunity on a scale from zero to one with equal weight assigned to each input factor.

$$\bar{X} = \sum_{i=1}^{7} \frac{x_i}{7}$$

Eq. 8-2

The resulting layer was the composite map displayed in Figure 8-1a, which describes the likelihood of recovery based on seven opportunity factors. A larger scale is used in Figure 8-1b to show how the output consists of a vector layer representing parcels with contributing information from seven opportunity variables. Opportunity is represented at the parcel-building level since non-building parcels were left with null values, thus excluding these areas from the final output.

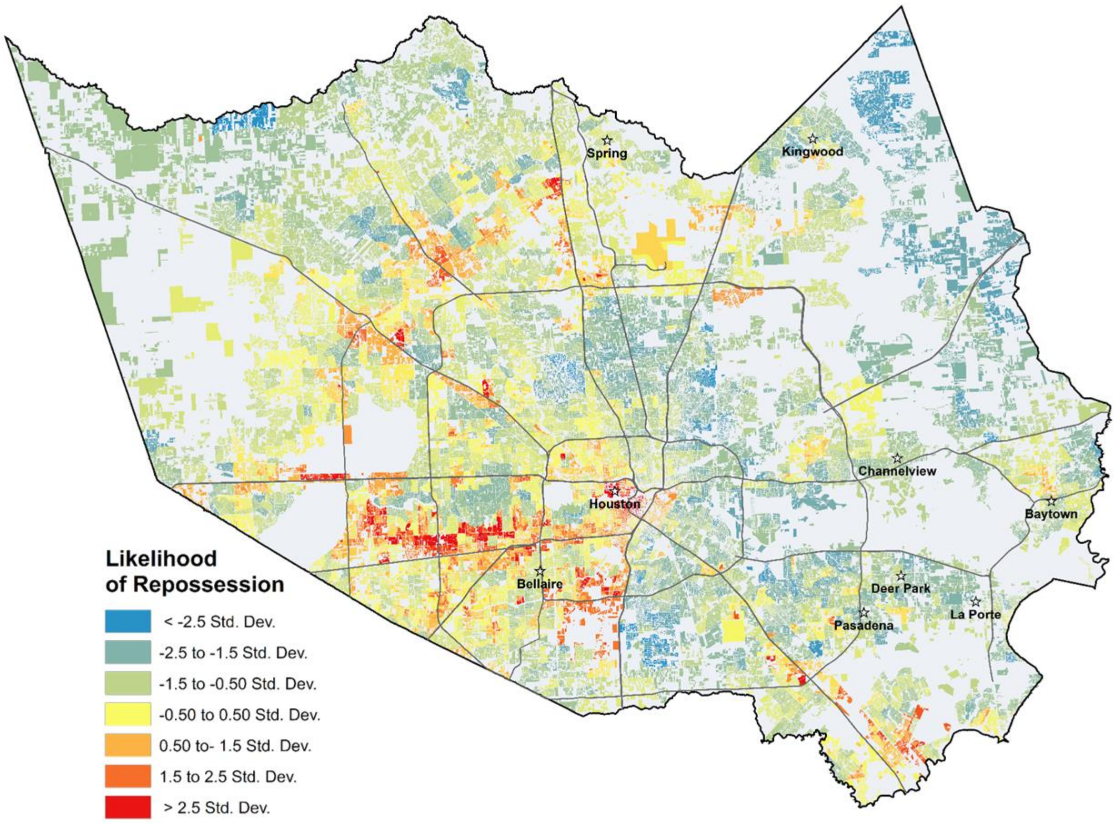

Figure 8-1a: Predictive Surface

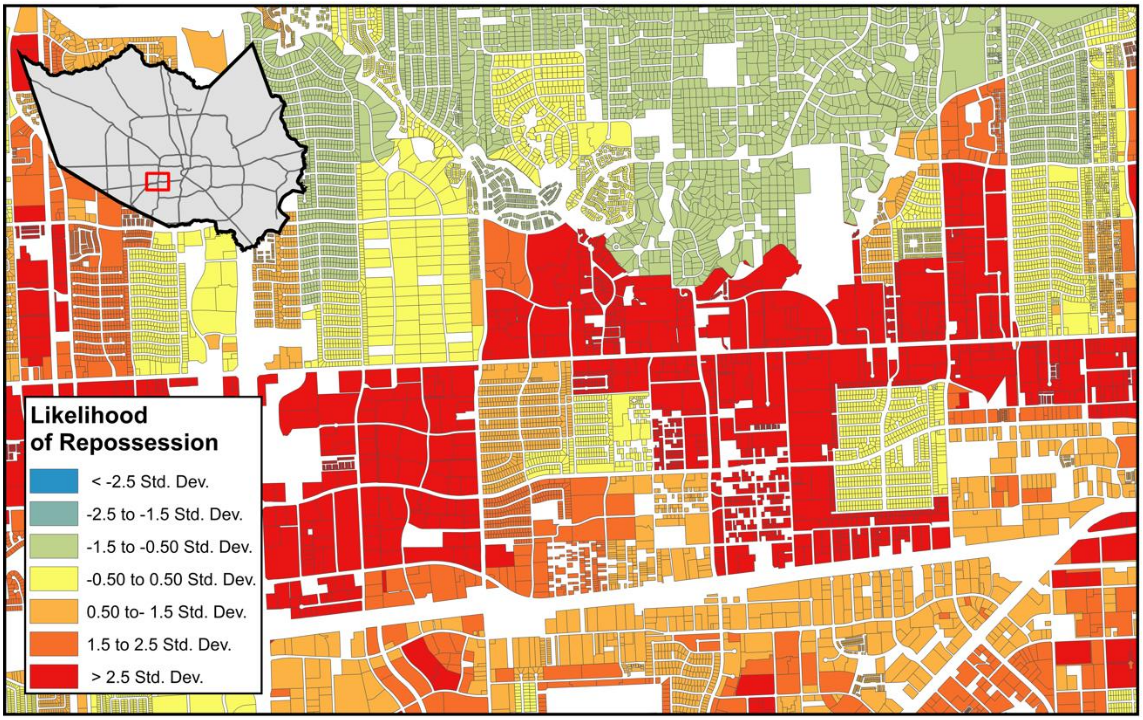

Figure 8-1b: Predictive Surface

The potential opportunity ranges from zero to one; however, the actual predicted opportunity runs from .08 to .59 because no single polygon fulfilled all seven demographic, financial, open account and building type requirements that would yield the highest opportunity. According to the surface, it is apparent that repossession opportunities are not randomly distributed across the study area with higher opportunity concentrated in the southwest corner of Harris County.

8.2 Model Validation

The last step in the project was to validate the model by examining whether the final opportunity surface accurately predicted repossession locations in a three-month period between November 12, 2014 and February 12, 2015.

Figure 8-3 shows an apparent relationship between the prediction surface and repossession locations; however, to confirm there was a significant relationship between the prediction surface and observed repossession patterns, a cross-tab Chi-square Test was used.

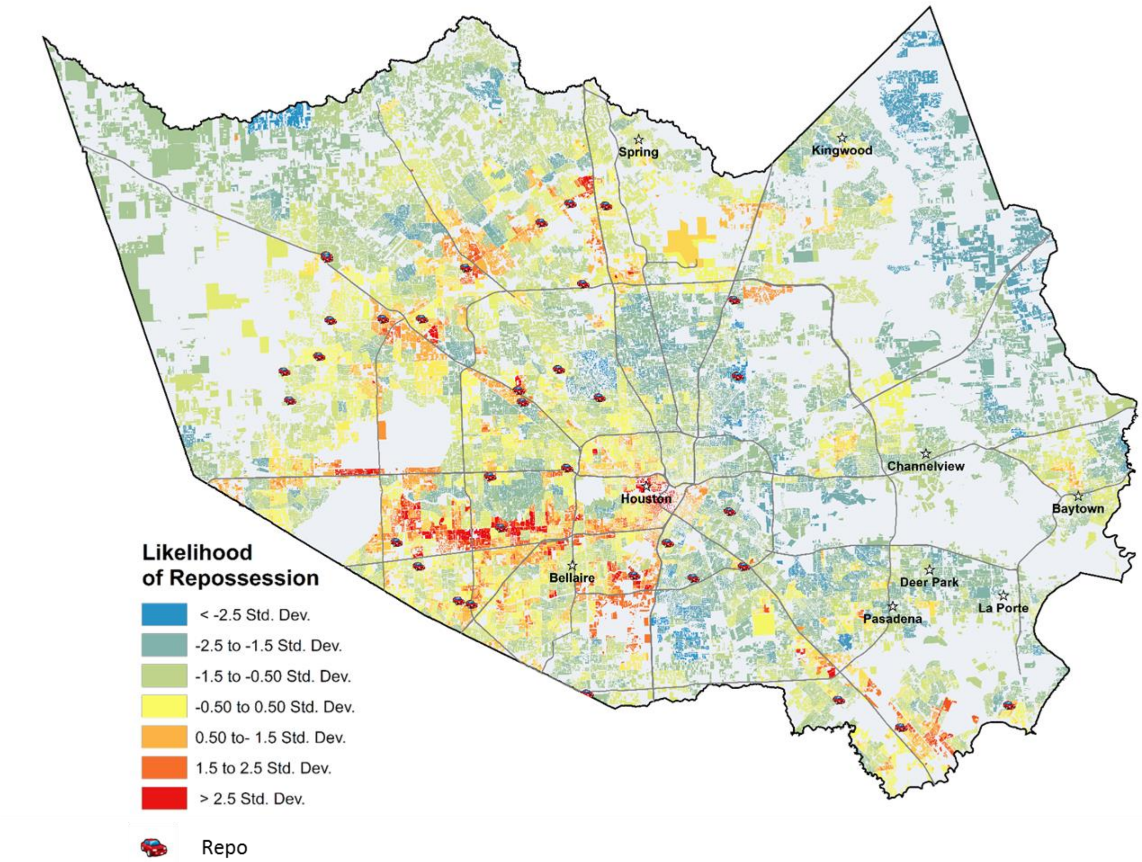

Figure 8-2: Repossessions from November 2014 to February 2015

The first task in the model validation was to join repossession points from the three-month period to the parcel layer. This was completed using the prioritization tool created for this project discussed in Section 5.1. Table 8-2 shows the sum of repossessions by building type.

Table 8-2: Repossessions by Building Type

Building	Repos
Residential Single Family	11
Apartment Garden (1 to 3 Stories)	8
Community Shopping Center	3
Residential Townhome	2
Apartment - Tax Credit	2
Retail Power Center	2
Residential Duplex	1
Apartment Structure (4-20 Units)	1
Subsidized Housing	1
Strip Shopping Center	1
Supermarket	1
Surgery Center	1
Hospitals	1
<Null>	1
Total	36

There were a total number of 36 repossessions from November 2015 to February 2015, which were distributed among thirteen building types. One repossession occurred in a parcel without a building or address. Because all repossessions from two years prior to the validation period occurred at a building, it is likely that this is an error due to missing information in the parcel layer or to an error in the recorded repossession address. For the purposes of this project, it was acceptable to assume that this repossession simply did not occur within a high-opportunity parcel.

8.2.1 Extracting Opportunity Values to Repossession Points

A new set of points was generated with the *Polygon to Point* tool to display all repossessions from the test period at the centroid of the parcels they belonged to. This was done in order to prepare data for value extraction such that all repossessions would properly intersect the corresponding parcel where the recovery occurred instead of the street, for example, which has a null value. Values were extracted to the adjusted repossession points using the *Intersect* tool with the opportunity surface. The output was a layer of repossession points located inside of the parcel they belonged to with a corresponding opportunity value extracted from the opportunity layer.

8.2.2 Significance Testing

A cross-tab Chi-square Test was completed to quantify the validity of the composite map. The predicted opportunity value at each repossession location from the testing period was compared to the mean opportunity value from the prediction surface (mean opportunity = 0.27), in which values above the mean were considered high-opportunity while values below the mean were considered low-opportunity. Of 36 repossessions, 11 occurred in low-opportunity polygons while 25 occurred in high-opportunity polygons. Similarly, all polygons in the composite map, or the predictive surface, were classified as either high-opportunity or low-opportunity by comparing their value to the mean.

A cross-tab was then created to describe high and low-opportunity polygons in terms of the sum of polygons that did and did not have repossessions shown in Table 8-3. The Chi-square test returned a significant p-value meaning that the occurrence of repossessions was not independent of predicted opportunities with more repossessions taking place in high-opportunity polygons. This result showed that the predicted opportunity surface can be used to guide the future search efforts for vehicle recoveries.

Table 8-3: Chi-square Test on Cross-Tab Values

	No recovery	Recovery	Total
Low-Opportunity	614,431	11	614,442
High-Opportunity	538,654	25	538,679
Total	1,153,085	36	1,153,121

$p < .01$

8.3 Summary

Opportunity values were generated and standardized for census block group attributes and parcels. Only parcels with buildings were considered for the final opportunity in order to eliminate the areas where debtors and people in general do not park. Each opportunity variable was combined to create parcel level prediction for Harris County. The composite map was compiled using opportunity values from significant demographic, financial, and open account variables at the census block group level and building types at the parcel level. The output excluded parcels with no buildings to create such that every building in in Harris County received an opportunity score. A cross-tab Chi-square Test of repossessions at high-opportunity and low-opportunity buildings confirmed that the final opportunity surface accurately predicted future repossession locations. This meant that the surface was suitable for future search and recovery endeavors of the client’s License Plate Recognition department.

Chapter 9 – Conclusions and Future Work

This project was designed to develop unprecedented research on the habitat of vehicle debt and recovery in order to narrow the search for vehicles out for repossession by an asset recovery company. The project successfully accomplished the client's goals with the creation of a prioritization tool, which also served as a summarization tool to join points to a target layer with the use of a 'near' function. It produced a highly significant opportunity surface using statistical methods, including two-sample T-Tests, bivariate correlation analysis, and multiple linear regression. The opportunity surface was successful as it was able to predict future repossession locations with highly significant accuracy.

The next step for a future project considering repossession prediction could be to replicate the predictive model in other markets around the country to draw comparisons between precursory environmental characteristics that influence repossession patterns in different geographic contexts. This could incorporate automated processes to generate opportunity surfaces, allowing users to input the opportunity layers found to be significantly related to repossessions to produce a composite, predictive surface.

Future work could also obtain accurate data about 'hits' on vehicles out for repossession to describe opportunity in terms of the locations that delinquent debtor vehicles are observed rather than successfully recovered. This could help identify building types and geographies that are unsuitable for recoveries thereby strengthening the level of prediction that was produced by this project.

Furthermore, additional work might take the findings from the prediction surface to develop a system to route License Plate Recognition vehicles between high-opportunity buildings. This would require more research about Traveling Sales Person diagnostics and routing heuristics to direct LPR vehicles using several waypoints. This could be taken to a step further by generating scanning zones to further optimize the search for vehicles out for repossession.

Works Cited

Akers, A. L., Seller, C. S. (2013). *Criminological Theories: Introduction, Evaluation, and Application* (6th ed.). New York, NY, USA: Oxford University Press.

Anslein, L. (2006). *Exploring Spatial Data with GeoDa: A Workbook.* Urbana, IL: University of Illinois, Urbana-Champaign.

Braga, A. A. (2001). The Effects of Hot Spots Policing on Crime. *American Academy of Political and Social Science*, 578, 104-125.

Caplan, J. M., Kennedy, L. W. (2010). *Risk Terrain Modeling Manual.* Newark, NJ: Rutgers Center on Public Security.

Caplan, J. M., Kennedy, L. W., Miller, J. (2011). Risk Terrain Modeling: Brokering Criminological Theory and GIS Methods for Crime Forecasting. *Justice Quarterly,* 28(2), 360-381.

Charles, K. F. Hurst, E. Stephens, M. (2008). Rates for Vehicle Loans: Race and Loan Source. *American Economic Review,* 98(2), 315-320.

Groff, E. R., Ratcliffe, J. H., Haberman, C. P., Sorg, E. T., Joyce, N. M., Taylor, R. B. (2014). Does what Police do at Hot Spots Matter? The Philadelphia Policing Tactics Experiment. *Criminology*, 53(1), 23-53.

Grubesic, T. (2006). On The Application of Fuzzy Clustering for Crime Hot Spot Detection. *Journal of Quantitative Criminology,* 22(1), 77-105.

Grubesic, T., Mack, E. (2008). Spatio-Temporal Interaction of Urban Crime. *Journal of Quantitative Criminology,* 24(3), 285-306.

Immergluck, D., Smith, G. (2005) Measuring the Effect of Subprime Lending on Neighborhood Foreclosures: Evidence from Chicago. *Urban Affairs Review,* 40, 362-389.

Li, Y. (2011). Geography of Opportunity and Residential Mortgage Foreclosure: A Spatial Analysis of a U.S. Housing Market. *Journal of Urban and Regional Analysis,* 3(2), 195-214.

National Research Council Committee to Review Research on Police Policy and Practices. (2004). *Fairness and Effectiveness in Policing: The Evidence,* eds. Wesley Skogan and Kathleen Frydl. Washington, DC: The National Academies Press.

Ratcliffe, J. H. (2004). The Hotspot Matrix: A Framework for the Spatio-Temporal Targeting of Crime Reduction, *Police Practice and Research,* 5(1), 5-23.

Ratcliffe, J. H., McCullage, M. J., (2001). Chasing Ghosts? Police Perception of High Crime Areas. *The Centre for Crime and Justice Studies,* 41, 330-341.

Rugh, J. S., Massey, D. S. (2010). Racial Segregation and the American Foreclosure Crisis. *American Sociological Review,* 75(5), 629-651.

Schintler, L., Istrate, E. Pelletiere, D. Kulkarni, R. (2010). The Spatial Aspects of the Foreclosure Crisis: A look at the New England Region. *School of Public Policy, George Mason University,* research paper 2010-28, 1-23.

Sherman, L. W., Rogan, D. P. (1995). Effects of Gun Seizures on Gun Violence: “Hot Spots” Patrol in Kansas City. *Justice Quarterly,* 12(4), 673-693.