Predicting Parcel Scale Redevelopment within The Berkeley Neighborhood in Denver Colorado Using Linear and Logistic Regression

by

Lisa Cherry

A thesis submitted to the Faculty and the Board of Trustees of the Colorado School of Mines in partial fulfillment of the requirements for the degree of Master of Science (Environmental Engineering Science).

Golden, Colorado

Date ______________________

Signed: ___________________________
Lisa Cherry

Signed: ___________________________
Dr. John McCray
Thesis Advisor

Golden, Colorado

Date ______________________

Signed: ___________________________
Dr. John McCray
Professor and
Department Head of CEE

ABSTRACT

Many water resource issues associated with urban development result from increased impervious cover. As impervious cover increases, rainwater infiltration decreases leading to increased flows and potentially higher pollutant loads in the runoff. Most of the prior research on this topic investigates the increase of impervious cover through the transformation of undeveloped to developed regions, or the expansion of dense urban development into outlying suburban areas. A topic that is not as widely understood is the impact of infill redevelopment on stormwater runoff. Infill redevelopment is rapidly occurring in many Denver neighborhoods, where previously developed properties with low-density structures are being replaced by larger, higher density units. Regardless of impervious cover increase, these redevelopment projects are only required to incorporate stormwater detention and treatment systems if they are greater than one acre. Due to most of the redevelopment in Denver (86%) occurring on sites less than one acre, the burden of stormwater treatment and detention ultimately falls on the city.

This study focuses on modeling the spatial distribution of infill re-development on a parcel scale to investigate its cumulative impacts on stormwater quality and quantity for near-term and future conditions. Future redevelopment and imperviousness is determined by distributing a “business as usual” linear growth scenario to the parcels with the greatest probability of future redevelopment. Then, a logistic regression model is used to determine the parcels that will be redeveloped. Results indicate that building cover change within study site from 2004 – 2014 followed a linear pattern ($R^2 = 0.98$). During this period the total building cover increased by 17% or 1.7 % per year on average. The logistic regression model determined the total value, year built, percent difference between current and max building cover, the current use classification: rowhomes, and current use classification apartments to be the greatest

predictors of redevelopment, resulting in a model that was 81 % accurate. The "Building to Land Area Ratio" variable was found to be highly correlated with the “Improvement to Land Value Ratio”. However, the “Building to Land Area Ratio” was found to be a better predictor of redevelopment. The final model forecasted an increase of 820,498 sq. ft. (18.8 acres) in building coverage between 2014 and 2024, resulting in a 14% increase in impervious area due to building coverage. This method will provide municipalities with a tool that can be used to estimate parcel scale impervious cover growth from publicly available planning data resulting in more informed urban watershed planning and policy development.

TABLE OF CONTENTS

LIST OF FIGURES

List of Tables

Acknowledgements

Without the support and assistance of countless individuals affiliated with both the ReNUWIt research center and the City and County of Denver this research would not have been made possible. I owe an enormous thank you to my thesis advisor, John McCray, for his role in my success. The wealth of knowledge and genuine interest in my research that he provided was an endless source of support and encouragement. Secondly, I would like to thank Darren Mollendor, my supervisor at the City and County of Denver. His mentorship has vastly expanded my perspective on urban storm water management and has greatly impacted my research. I would also like to thank my committee members, Terri Hogue and Bill Eisenstein, for the time they invested in me and their insightful recommendations throughout this process.

Finally, I must express gratitude to my family and friends for providing me with boundless support and encouragement throughout this journey. I especially need to thank my mother and grandmother, Ruthann Geary and Ruth Weber, who taught me to be strong and hardworking, even in the face of adversity, my younger sister and brother, Kelly and Sean Geary, who bring out the best in me, and my partner, Ryan Tyndall, who is my emotional rock. Without these individuals I would not be the person I am today.

CHAPTER 1
INTRODUCTION

Statement of Problem

Urban land use varies in both its type and intensity of use. A parcel, or properties, designated land use type can be classified by many terms including, but not limited to, single-family residential, industrial, commercial, and open space. The intensity of land use can be described by classifications such as vacant, undevelopable, or developed. In urbanized areas where there is limited vacant land available for new development, infill and redevelopment strategies are often employed to increase the utility of the land and promote economic growth. The land can either be completely redeveloped or the utilization of a parcel can increase due to additions to the previously developed area. Infill redevelopment is the process of renovating or restoring a previously developed parcel through demolition and construction activities resulting in higher density land use.

Infill redevelopment is rapidly occurring in many urban areas. In many cases it involves low density single-family properties being redeveloped with larger, higher density multi-family units, high-density commercial units, or mixed uses. The City and County of Denver is expected to grow by more than 100,000 people in the next 10 years, and 50% growth is projected through 2050. Subsequently, population increases will be accompanied by new development, much of which will be in the form of infill redevelopment.

Many urban planners advocate for infill redevelopment because of its many benefits. It increases walkability and access to public transportation while decreasing (or decreases to stay consistent with increases) urban sprawl, dependency on the automobile, and per capita carbon emissions. Furthermore, the beneficial use of the additional stormwater produced from re-development occurring in many Denver neighborhoods could be a novel solution to relieve some pressure on future water supply needs. Even without beneficial re-use, most cities are interested in implementing green infrastructure to help mitigate water quality deterioration from additional development.

However, there are also many challenges that result from this unique type of rapid redevelopment. As natural landscapes become covered by concrete and other construction materials, the amount of impervious cover within a lot can increase significantly. Considering the growing frequency of this occurrence it can have a significant impact on neighborhoods and municipalities, especially in areas that are experiencing rapid population growth.

The increase of developed land leads to a decrease in rainwater infiltration, thus a greater amount of stormwater runoff. Many water resource issues have been associated with population growth and urban development. Pollutant concentrations in storm-water runoff are likely to increase when population and impervious area increases. The pollutants that accumulate on the impervious surfaces are deposited by the stormwater runoff into nearby streams potentially causing impairments.

While the link between impervious cover and stormwater pollution has been demonstrated, the discipline of stormwater management is young. The first large stormwater regulation was enacted in 1990 when the EPA created the Municipal Separate Storm Sewer Systems (MS4) permit program. Through this program, municipalities were required to create personalized stormwater management programs (SWMP). The SWMP details how the permit holder will decrease the pollutants entering the local waterway through its storm drainage systems. Located in every SWMP are eight key subparts including:

1. Construction Site Runoff Control Illicit
2. Discharge Detection and Elimination

3. Pollution Prevention/Good Housekeeping
4. Post-Construction Runoff Control
5. Public Education and Outreach
6. Public Involvement/Participation
7. Program Effectiveness
8. Total Maximum Daily Loads (TMDL's) [1].

To mitigate water quality and quantity damage done after rapid redevelopment has occurred, post-construction stormwater management is needed. MS4 permit holders, like the City and County of Denver, manage post-construction runoff and other sections of their Storm Water Master Plan (SWMP) through the installation of Best Management Practices (BMP). Some examples of BMPs that are currently utilized are wetlands, retention ponds, and vegetated swales. These natural systems help treat the stormwater before it enters the storm sewer network. To ensure that these systems are strategically placed in order to maximize the flood risk mitigation, reduce pollutant discharge, and increase over dollar efficiency, decision support tools are used.

Every five years the City and County of Denver evaluates its drainage basins to obtain up to date information on hydrologic conditions and determine areas that need improvement. This information is then organized into a document called the Storm Drainage Master Plan (SDMP). The most recent update was created with the collaboration of six consulting firms: Enginuity Engineering Solutions, Icon Engineering Inc., Matrix Design Group, Olsson Associates, San Engineering, and Zoeller Consulting. They did so using information from Urban Drainage and Flood Control District's Stormwater Management Model, the City of Denver's Blueprint Denver Report, LiDAR topography maps, Potential Inundation Areas, storm drainage network updates, as well as many other inputs [2]. Based on the information presented in the document, the city determined areas in the city that require capital investments and have prioritized those sites accordingly.

In addition to the SDMP, the city is developing many other initiatives that will help achieve their goal of optimal stormwater management and remain in compliance with their MS4 permit. One project that the city is currently working on with Matrix Design Group is The Stormwater Quality Prioritization and BMP Analysis Plan. This project will supplement the efforts of the Storm Drainage Master Plan by expanding the scope to include a water quality component of watershed evaluation. The tool was created in November 2014 and is only currently available in draft form but the main intention of the plan, as the name implies, is to create a prioritization tool for stormwater capital improvement projects. From this analysis, municipalities can construct BMPs in locations that will provide the largest benefit and help watersheds that need it the most. The prioritization process relies on scoring the watersheds based on eight primary categories: existing TMDL's, 303(d) listed waterbodies, wet weather pollutant loading, dry weather pollutant loading, disconnected impervious area, redevelopment potential, impervious area within the right of way, and existing treatment.

These strategies are just a few tools that the City of Denver uses to ensure that they are making informed choices about where they are investing in stormwater management technologies. However, all of these strategies rely on Blueprint Denver's "Areas of Change", which were developed in the year 2000 (16 years ago), for their future land use condition inputs.

City of Denver Research Request to Colorado School of Mines

This thesis research addresses the need for an urban land-use forecasting tool that can be applied at the neighborhood scale for stormwater modeling of future conditions and is part of a larger research project funded jointly by ReNUWIt research center at Mines and the City and County of Denver. The larger project encompasses innovative stormwater management in a west-Denver neighborhood, called the Berkeley neighborhood, undergoing the pressures of redevelopment. In this section, the overall research project is described, followed by the objective of this thesis research.

The rapid redevelopment within the neighborhood caught the City and County of Denver's attention and they decided it would be beneficial to evaluate The Willis Case Golf Course's potential to be converted into a stormwater treatment facility (Figure 1.1). Thus, in September of 2014 the city entered a collaborative effort with a Colorado School of Mines and ReNUWIt to complete a feasibility study on using the Willis Case golf course and/or the Berkeley Lake City Park for a regional stormwater treatment facility. The study became known as the Berkeley Neighborhood Stormwater Feasibility Study. It evaluates the legal, economic, and technical feasibility for the beneficial use of non-potable stormwater in basin by performing four tasks:

1. Analyze re-development in Berkley neighborhood and predict near and longer term land use trends (the topic of this thesis).
2. Determine BMP performance needed to achieve water quality standards for using reclaimed water for urban irrigation

3. Analyze the legal, policy and economic barriers to using stormwater for beneficial use for newly generated stormwater in storm sheds where streams do not have alluvial aquifers and determine strategies for overcoming these barriers
4. Complete a 30% design for implementing the appropriate BMPs on the Willis Case Golf Course

The optimal future irrigation goal of the project is 100% non-potable treated stormwater from the tributary drainage area. However, the most probable outcome will be a combination of irrigation sources ranging from treated stormwater, raw water from Rocky Mountain Ditch Share, and Denver Water with consumptive use augmentation from variable sources. An acceptable, but less preferred potential outcomes would be to have stormwater quality treatment on the city parks properties to enhance future stormwater quality entering the local stream Clear Creek. Clear Creek is a tributary of the South Platte River, which is the primary source of drinking water for Denver.

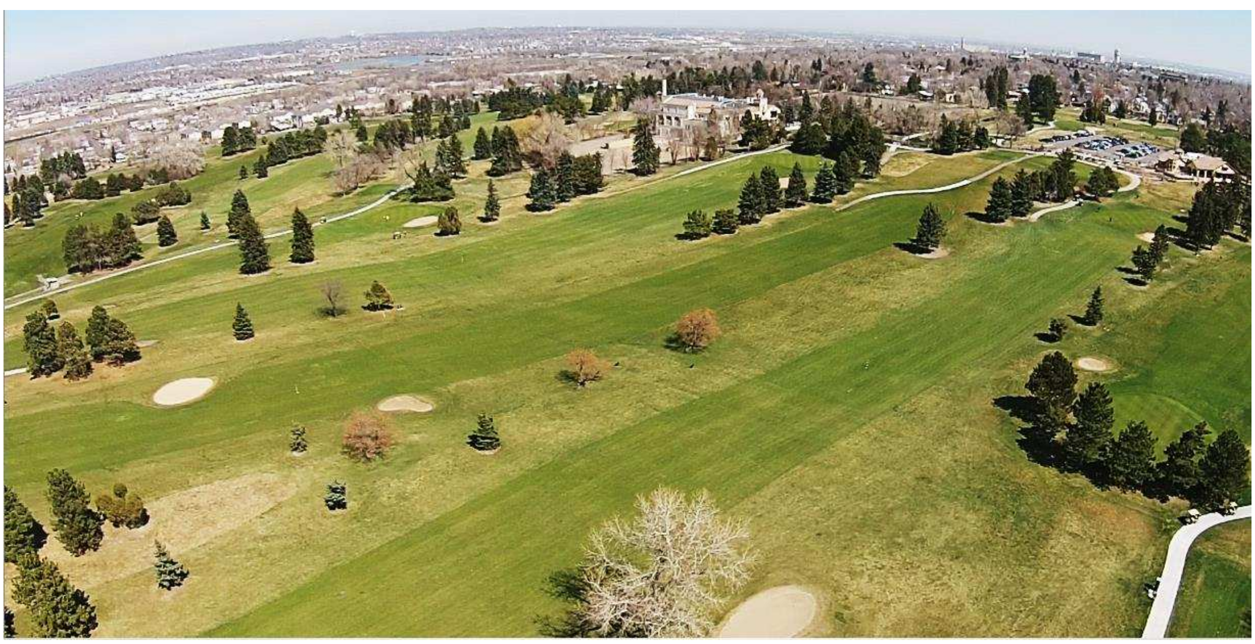

Figure 1.1 Willis Case Golf Course

The project promotes the initiatives of several Colorado agencies including the City and County of Denver (CCD) - Wastewater Management, the ReNUWIt research center, CCD Department of Parks and Recreation, Colorado Department of Transportation, The Colorado Water Conservation Board, Denver Water, and Urban Drainage and Flood Control District. Specifically, a regional stormwater treatment system located on Willis Case Golf Course that could treat the entire Berkeley neighborhood drainage basin would provide water conservation, environmental protection, drainage and flood protection, infrastructure research and development opportunities, as well as, education and outreach. Each of the stakeholders involved would directly benefit from these outcomes.

Research Objective

The overarching goal of this thesis research is to evaluate the current and future redevelopment within the Berkeley Neighborhood. The primary hypothesis is that future impervious cover in parcels (i.e., properties) with urban infill development in west Denver can be predicted using a combination of linear and logistic regression, where the independent variables in logistic regression are factors found in County assessor's data. This research is intended to develop a reproducible method that can be used by municipalities to forecast spatial and temporal change in impervious cover at the parcel scale for urban neighborhoods experiencing infill redevelopment using public available urban planning data. Results from this research will be used as inputs for other aspects of the larger project that evaluate the impacts of urbanization on stormwater quality and quantity within the Berkeley Neighborhood drainage basin.

CHAPTER 2
LITERATURE REVIEW

Urban Land Use Change Models

There are many models available to assist researchers and professionals in investigating land use change within the urban environment. According to Agarwal et al. all land use change models can be categorized according to five attributes: spatial, temporal, human decision making, scale, and complexity [3]. Finer scale models evaluate small areas of land, like a watershed or sub watershed and shorter time steps, while broader-scale models evaluate areas larger in size, like an entire region or continent and over longer periods of time[3]. As the model complexity increases the model considers more interactions.

Infill development in Denver, and probably most cities, occurs at the neighborhood scale. For example, in Denver this type of development has occurred in separate neighborhoods across the city at different times (e.g., Highlands was one of the first neighborhoods in West Denver where infill redevelopment started to increase rapidly in the early 2000's). Thus, simulating infill redevelopment a fine spatial scale, where redevelopment can be predicted in an area the size of a couple square miles or less (the Berkeley neighborhood has an area of less than 2 square miles), is ideal.

However, all commonly used urban land use planning models are intended for implementation at larger scales. The US Forest Service reviewed 19 of the most representative land use change models, including the California Urban Futures model (CUF) and the Land Use Change Analysis System (LUCA). Among these models, the minimum spatial scale was around 10,000 ha or about 40 mi^2 [3]. This is appropriate for small to large cities, but not for urban neighborhoods. As a result most land use change models are not suitable for infill redevelopment modeling since their spatial extent is much broader than what is required for redevelopment forecasting at the neighborhood scale.

One land use change model with a parcel scale resolution and the ability to simulate infill redevelopment processes is “Urbansim” [4]. The model has a land development tool that models developer behavior and can estimate the amount of vacant or developed land converted to a greater density. This tool completes this task by evaluating a parcels profitability and constraints. Additionally Urbansim model has been accepted as one of the best land use change models available [5]. However, the model is limited in its implementation by its large amount of model inputs. In fact when conversing with the developer of Urbansim about the applications of the model for this project he suggested that the model would not be appropriate for neighborhood scale infill redevelopment (Paul Waddell, 2015, verbal communication).

Municipality Tools

The high spatial resolution required by redevelopment models and the limited resources for data collection and synthesis limit the options that are available to municipalities. As a result, the majority of municipalities and city planning groups have created their own methods of redevelopment capacity estimation and forecasting [6–14]. Three strategies commonly incorporated in the municipality studies are the analysis of a parcel’s floor area ratio (FAR), its improvement to land-value ratio (ILV), and its residual land value (RLV). The FAR is the ratio of building area to total land area. The ILV is the ratio of improvement value of a parcel to its land value, where a parcel’s assessed building value is divided by the assessed land value. The RLV is the value of completed development subtracted by the cost of development and profit earned by the developer [15] .

The FAR can help municipalities determine if a parcel is being utilized to its full potential (regarding spatial utilization and taxation potential). A FAR value of one (1) indicates that the building floor area is equal to the parcel area, a FAR value that is greater than one (>1) means that the floor space is greater than the parcel, thus a multiple story building. A FAR value that is less than one (<1) would imply that the building is smaller than the total parcel area [16]. The maximum FAR value is the same for all parcels within a zoning type. In a rapidly developing neighborhood such as Berkeley, parcels with FAR significantly less than 1 are logically candidates for near-term redevelopment.

During the ILV evaluation, all the developed land is ranked per its improvement to land value ratio. For an example, City of Bellingham, Washington sets the ILV threshold at 10%, whereby if the structural value is less than or equal to 10% of the total land value then redevelopment will occur within 20 years [14]. Parcels with low ILV ratios are considered "underutilized". The city of Portland states that a parcel is considered underutilized if the ILV is 0.5 or less [8]. Eagle Point, Oregon has an ILV redevelopment threshold of anything less than 1. Eugene, Oregon uses a method that determines the low, moderate, and high development potential using ILV: low redevelopment potential exhibits an ILV between 0.3 and 0 .5 with the building coverage less than 20% of the land area; moderate redevelopment potential is characterized by an ILR of 0.3-0.5 with building coverage accounting for less than 10% of the total parcel; high redevelopment potential as any ILV less than 0.3 regardless of the building utilization [17].

The RLV Method evaluates redevelopment likelihood based on a residual land value analysis. This method compares the market price of the parcel after redevelopment to the residual land value, which is the purchased value of the land minus the redevelopment costs and desired profit [10]. This method requires knowledge of construction costs, projected profits, and selling price after redevelopment. The method was designed and tested on two Pierce County, Washington cities, Tacoma and Puyallup. They found that the model was successful and only needed adjust baseline calculations for a few “unique” parcels [10].

Many municipalities use these three techniques as a way of locating potential locations for future redevelopment. While these methods are a great way to get a basic understanding of parcels with a potential for redevelopment, the literature regarding their accuracy is limited (particularly the peer-reviewed literature). The methods currently employed by many municipalities use specific planning variables that intuitively are related to development, and then associate somewhat arbitrary values for these variables with somewhat arbitrary time periods for redevelopment.

CHAPTER 3
METHODS

Overview

As previously stated, this research's intent is to produce a model that can be used by municipalities to predict the future redevelopment of parcels at the scale of an urban neighborhood. To achieve this objective a combination of linear and logistic regression was utilized, using open source information obtained from the City and County of Denver's Assessors Office.

The fitted linear regression model was then used to predict the future building cover growth within the neighborhood. Once the projected building cover increase was determined, a logistic regression model was developed using 2004 parcel attributes as dependent variables (e.g., building value, year built, building type etc.) to predict which parcels would redevelop between 2004 to 2014. 2004 was used as a starting period because this is about the time that rapid infill redevelopment started in Berkeley. 2014 is the most recent year of data. The projected building cover increase was then distributed to the parcels that were predicted to be redeveloped in the future (using logistic regression) resulting in an impervious cover forecast for the Berkeley neighborhood that is deemed appropriate for the year 2022. Details on the linear regression model, including justification for the time period used, are provided below.

Site Description

The Berkeley Neighborhood is in the far northwestern section of the City of Denver's boundary (Figure 3.1). Currently, the stormwater that runs off the Berkeley Neighborhood is largely captured in storm drains, and delivered untreated to Clear Creek, a tributary of the South Platte River (the primary drinking water source for Denver). The total area of the urban drainage basin is 1,114 acres (1.74 sq. mi.). The impervious cover in 2014 was about 46%. The neighborhood contains 4,710 parcels.

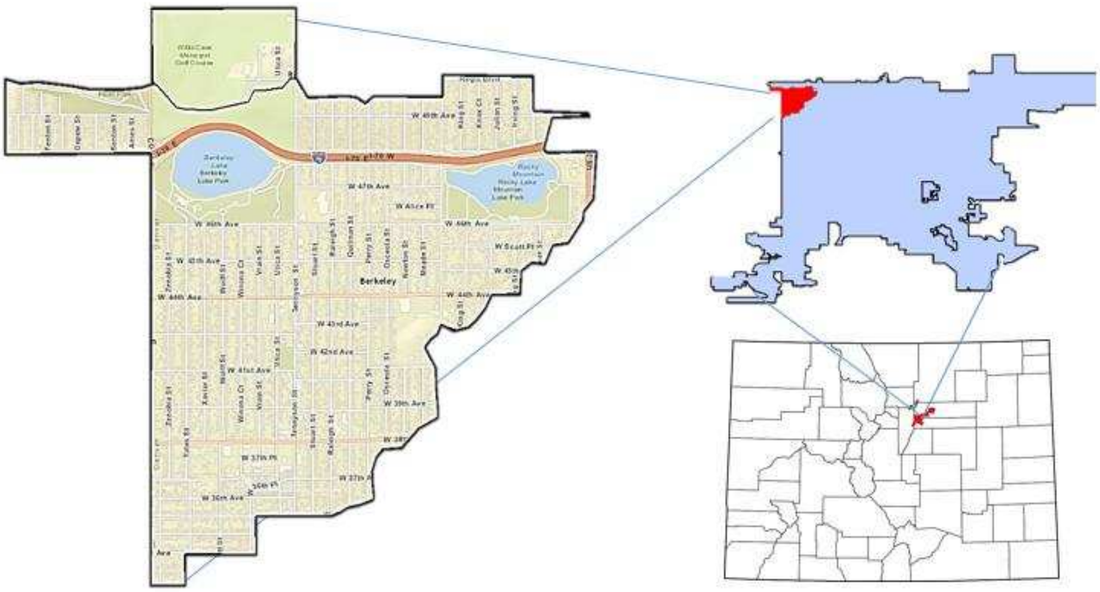

Figure 3.1 The Berkeley Neighborhood Drainage Basin

The dominant land use in the Berkeley neighborhood is low-density residential where single-family homes (41.2%) reside on 6,000 square foot lots (Figure 3.2). However, zoning allows a range of uses, from single unit to mixed use, with the highest density development occurring closest to Tennyson Street.

In 1892 the Berkeley neighborhood had only 450 residents. Between the years 1926 and 1950, the number of properties grew by 50% [18]. At present, the neighborhood is in a state of rapid infill redevelopment. Properties that once held small single family cottages are being replaced by larger single and multifamily units (Figure 3.3). Over the past ten years, urban growth has begun to convert the neighborhood into higher density apartments, row homes, and condominiums that can house multiply families on one parcel. Large projects currently in the pipeline include a Natural Grocer, and two large apartment complexes with over 50 apartments per building [18,19].

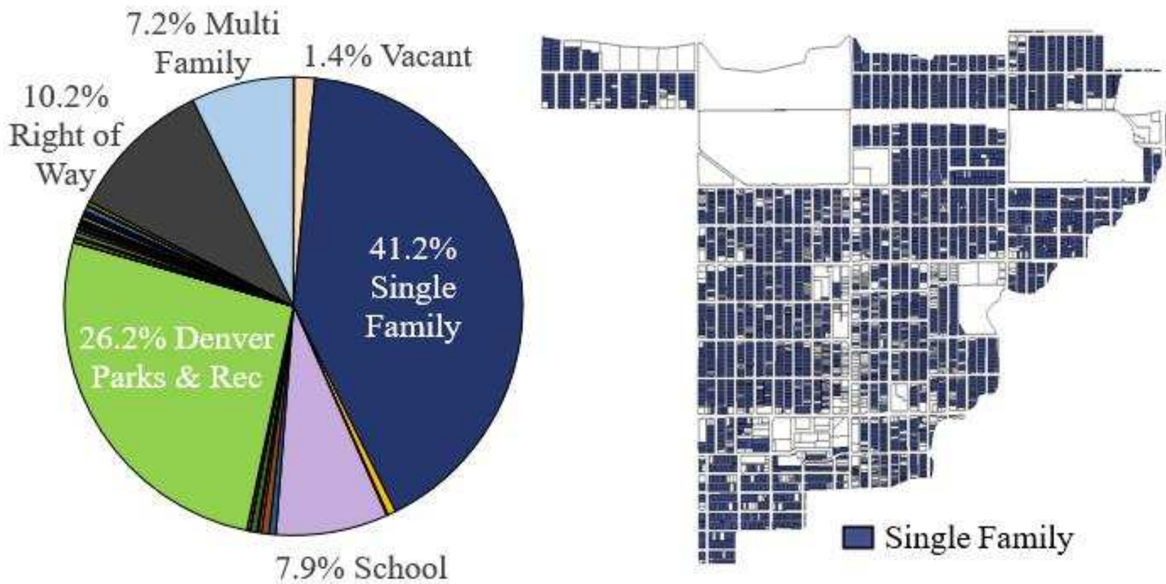

Figure 3.2 The Berkeley Neighborhood 2014 Land Use

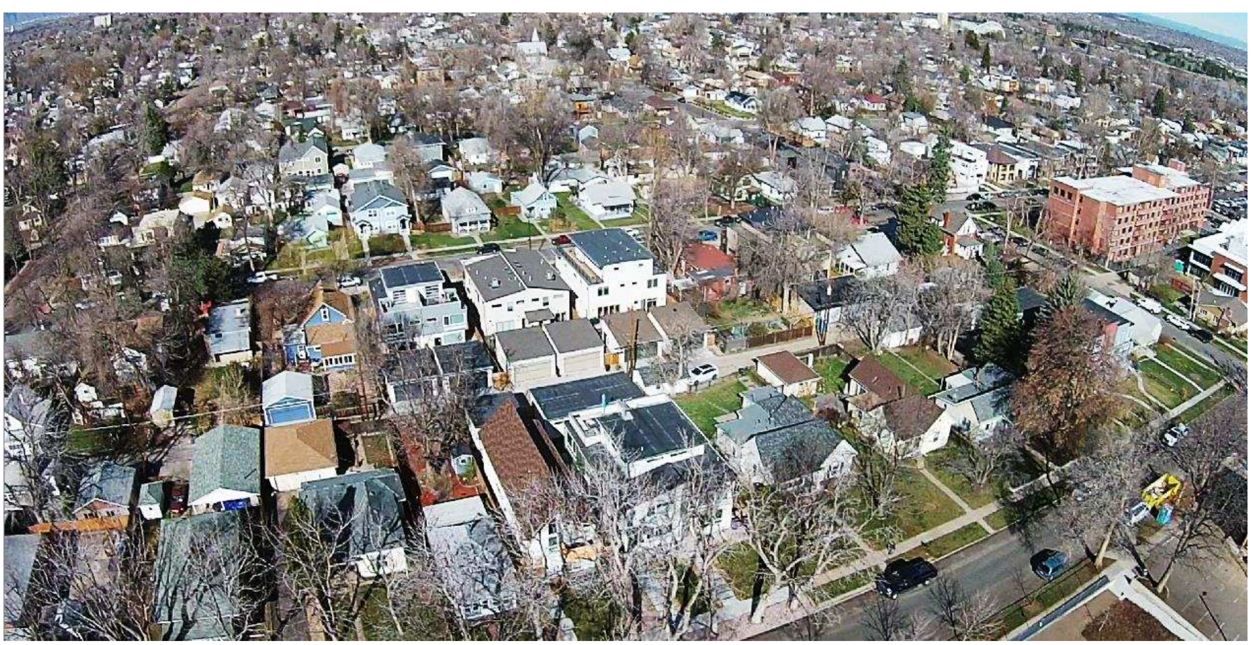

Figure 3.3 Berkeley Neighborhood Aerial

Data Organization

The best available data source for parcel scale modeling in the Berkeley neighborhood was determined to be the City of Denver Assessor's Office parcel data. The Assessor's Office parcel data included information such as:

1. Improvement to Land Value Ratio (ILV)
2. Parcel Area
3. Building Area
4. Max Building Area (according to zoning code)
5. Year Built
6. Percent Difference Between Current and Max Building Coverage
7. Building to Land Area Ratio
8. Total Value
9. Building Value
10. Land Value
11. Current Use Classifications: Vacant, Single Family, Row house, Other, Condo, Business, and Apartment

It is useful to note that the "percent difference between current and max building coverage" was not directly available in the Assessor's Office parcel data, but was calculated from current coverage and maximum coverage.

The Assessor's Office parcel data is released every two years and was available from 2004 to 2014. In order to prepare this data for the linear and logistic regression modeling considerable quality assurance and control needed to be performed.

The 2014 data was missing building coverage for all apartments. As a result, the building coverage had to be calculated manually using google earth's calculated geometry feature. While this task was tedious, it was necessary in order to include the apartments within the Berkeley Neighborhood into the analysis.

Additionally, the parcels that were classified as condos were individual units within the condominium building. Since the goal was to determine redevelopment on a parcel scale, the individual condo units had to be joined into one observation for each condominium building. Since all of the condo units shared the same address and year built these features were used to group them. Once the condo units were grouped, they were combined into a single parcel through the addition of the individual unit's variable values.

Another data organization process that needed to be performed was the grouping of the parcels "Current Use" categories. The current use describes a parcels primary function. Some of the most common uses within the Berkeley neighborhood are single family and row homes. Less common uses are churches, car washes, and schools. Since the original data had over 45 different use categories sometimes with only one parcel in some categories they were grouped into larger more encompassing categories that would allow for better model fitting and prediction. The resulting categories were vacant, single family, business, apartment, other, row house, and condo. Each current use category was then made into its own variable and dichotomized or turned into a "dummy" variable for ease of model interpretation.

The parcel redevelopment indicator that was used as the outcome variable in the logistic regression model (discussed more below) was determined by evaluating a parcel's change in the building coverage between 2004 - 2014. In order to track this change, the parcels schedule number was used, which is an ID that is unique for each parcel. Parcels that had a change in building coverage between the study period were labeled as 1 for redeveloped and 0 for not redeveloped.

Over the ten year period that was being investigated 4% of the parcels were subdivided. When the subdivision occurred the parcels schedule number was either changed into two completely new schedule numbers or one was arbitrarily kept as the old number while the other was given a new schedule number. In order to accurately track the building cover change resulting from a subdivision, all of the parcels resulting from the spilt had to be connected with the original 2004 parcel. This task was completed by spatially joining the parcels in 2004 with the parcels in 2014 and determining where the 2004 parcel overlapped with the new subdivided 2014 so that the 2004 and 2014 total building change could be calculated.

After performing quality assurance and control on the raw data in the assessor data base the linear and logistic regression modeling could occur.

Linear Regression

The first step to quantifying the spatial extent of redevelopment within the Berkeley neighborhood is to create an informed growth scenario. This process was completed by applying a linear regression model to the building coverage between 2004 and 2014 and using the equation to predict the total building cover at a future point in time. Building cover was deemed the most appropriate variable to use because it is an actual indicator of redevelopment and is also a good measure of increased impervious area. Sheds, garages, or driveways were not included in this calculation because the lack of available data for these structures over time.

As it will be shown in the Results section, the linear regression model provided a good fit with the observed building cover over time, despite the 2008-2010 recession period. Much thought was given on what years to consider for the growth analysis, given that new data becomes available every two years. The year 2004 was selected as the starting date, because of data availability and that rapid infill-redevelopment began in Berkeley neighborhood in 2003-2004, and 2014 was selected as the end date because it is the most recent data available.

Logistic Regression

After developing a growth scenario that could be used to determine future building cover increase, the next step in the process was to determine where the expected growth would occur, in other words, what parcels would be redeveloped in the future. This was completed using logistic regression. Logistic regression is a statistical modeling method, used for forecasting in the engineering and sciences, that can balance model complexity with user friendliness by measuring the relationship between an outcome of interest (i.e., a dependent variable, in this case whether or not a parcel will re-develop) and one or more independent variables (i.e., urban planning variables from assessor database).

Logistic regression has promise to improve urban land-cover forecasting because of its ability to evaluate the effects of multiple variables on land cover change due to redevelopment, distinguish which of these variables has the greatest influence on redevelopment, and finally to predict the probability that a parcel will be redeveloped in the future. Thus, this method was selected as the most suitable tool for this research because of its ability to incorporate complexity (i.e., include numerous redevelopment drivers), maintain relatively user friendliness, while meeting the primary goal to predict future redevelopment.

To fit a logistic regression model, statistical software, like R and SAS, estimates the equation intercept and variable coefficients, where α represents the equation intercept and β represents the variable coefficients, through a process called maximum likelihood estimation (Equation 3.1-3.3). The likelihood function selects the coefficient values that maximize the probability of observing the data given the parameters. Once the model is fit using some known set of data, the log odds of the outcome occurring can be calculated for observations that are unknown. The term “log odds” is not intuitive, but is a probability of the desired outcome (in this case, the probability of re-development) with a value between zero and one. The R computer program written to implement the logistic regression modeling is included in Appendix B.

The logistic regression equations can manipulated to determine the probability of each parcel developing, which could be used to rate a parcel’s affinity for development (i.e., low, medium, high, etc.) as has been done by other methods (Equation 3). However, for this research, we want to forecast whether a parcel will develop (yes or no) so that the model results can be compared to the observed outcomes. Thus, this probability will eventually need to be dichotomized to express whether a parcel had been redeveloped (Y = 1) or not redeveloped (Y = 0) between the years 2004 and 2014.

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_n X_n \quad (3.1)$$

$$\left(\frac{p}{1-p}\right) = \exp(\alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_n X_n) \quad (3.2)$$

$$p = \frac{1}{1+\exp(\alpha+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3+\beta_4 X_4+\beta_5 X_5+\beta_n X_n)} \quad (3.3)$$

The general procedure for the logistic regression model is to use assessor data (independent variables) from one year (we chose 2004) along with outcome data from a future year (i.e., developed or not) to ”calibrate” the model (i.e., optimize the values of the coefficients in Equation 1, which relate the relative importance of each independent variable on the outcome.) A standard approach used in the statistics literature is to fit, or calibrate, the logistic regression equation to one set of data (often called training data set), and then evaluate the model performance for its predictive ability using another data set (typically called the testing data set) [20–22].

As with the linear regression model, the most challenging aspect of the methodology was to select a future year for both the training and testing, acknowledging that assessor data (the basis for both the independent and dependent variables) is available every two years. Ideally, one might use one period (for example 2004 with a model calibration target of development 2010) to determine regression-model coefficients, and a different period (2010-2014) for the testing. However, the timing of the recession (2008-2010) complicated the choices. As shown in the Results section, while a good linear trend in building coverage existed for the entire period between 2004 and 2014, the period between 2008 and 2010 demonstrated a clear slowdown of development. In addition, the rate of development between 2004 and 2008 was similar but somewhat less than between 2010 and 2014 and we desired for the model to be applicable to as long a period as possible. Given the impacts of the recession, a regression model "calibrated" for any single period may not be appropriate for an entire 10-year period. Moreover, using the entire 10 year period (2004-2014) for the calibration (training) procedure was though to better encompass reality (i.e., a 2-year recession within a 10-year period).

Thus, years 2004 and 2014 were used for both the training and testing of the regression model, where the independent (assessor's data) and outcome data (developed or not) was spilt into two groups: a training set and a testing set. 75 % of the 2004 parcel data was randomly selected in R to be used for the training set and the other 25 % was used to compose the testing data set. Using this procedure, we anticipate that the developed model could be used to forecast which parcels will be developed over a future 10-year period (i.e., 2014 to 2024). Then, the predicted building coverage (from the linear regression) will be assigned to specific parcels using a method descried below.

Model Training

The first step in the logistic regression modeling process is fitting the model using the independent variables from the 2004 parcel training data. To determine the best subset of variables from this dataset each variable was tested for collinearity. Once the variables were narrowed down to a smaller subset, all model combinations were used to fit a logistic regression model using the R package *'Glmulti'* to select the final variables that would be used for model predictions.

Collinearity Including variables that are correlated with one another into the logistic regression model is unnecessary because they provide redundant information with regard to their relationship to the dependent variable, redevelopment. Moreover, including two correlated variables in a model can result in a model calibration (training) producing non-unique results that are can preclude prediction [23]. Including unnecessary variables can also unnecessarily complicate the model thus each variable in the 2004 parcel data was evaluated individually for its correlation with the other independent variables.

Spearman's rank correlation coefficient (r_s), which is a measure of correlation between non-normally distributed continuous variables, and the chi square values (*X*), for categorical variables, was used. If two variables were highly correlated (i.e., $r_s > 0.6$ or $X < 0.05$), only one was included in the final analysis. The variable that was selected was the one with the greatest correlation with the dependent variable.

To evaluate the strength of correlation between each independent variable and the dependent variable the R squared value was used. Variables that were not found to have a reasonably significant R value (i.e., $R^2 < 0.5$) when fitted to the outcome of redevelopment on their own were not considered in the multinomial logistic regression model. This process resulted in 7 of the variables being eliminated, which is explained in the Results section.

Multivariate Analysis Once the potential issue of collinearity was addressed, the variables were further evaluated to determine which combination of variables created the best model. The goal of this process was to find the simplest model with the least number of variables (to enhance user friendliness) that could accurately predict the parcels that had redeveloped between 2004 and 2014.

However, there are too many possible models that running each in the statistical software separately would be a time prohibitive endeavor. For example, a model with just 10 variables has 2^{10}, or 1,024, different candidate models that could be used. Because it would be extremely tedious to manually test each one of these combinations, the '*Glmulti'* R software was utilized to select the top models for analysis and prediction.

The software produces all possible variable combinations, fits them with logistic regression, and returns the best models based on the Akaike Information Criterion, which measures the quality of the logistic regression model relative to the other candidate models [24]. This was performed as part of the model training (sometimes called calibration). After this process was completed, the highest rated model was selected and used in the model testing process (More details are provided in the Results section).

Model Testing

To ensure that the model selected could accurate predict redevelopment, its predictive ability was evaluated using the 2004 testing data variables. The predictive ability of a logistic regression model is typically evaluated based on its sensitivity, and specificity, and overall accuracy (Equation 4-6)[25–27]. In the logistic-regression literature, model *sensitivity* evaluates a model's ability to correctly classify the outcome or the proportion of observed positives that were predicted to be positive. Model *specificity* quantifies the true negative rate or the proportion of observed negatives that were predicted to be negatives. The sensitivity and specificity of a model are inversely related to one another; as one increases the other decreases. Model accuracy considers both the true positive and true negative classification ability of the model, and is often considered the best overall measure of logistic regression model accuracy. Specificity or sensitivity may also be considered depending on the goals of the model. For example, model sensitivity might be be chosen as the primary measure if it is highly important to correctly classify the positive outcomes, even it results in a number of false positives (i.e., predicts parcels will develop that don't actually develop). Specificity might be considered the primary measure it if it more important to correctly identify false negatives (i.e., correctly predicting the properties that do not develop). The equations associated with these measures are provided below where the terms a, b, c, and d are defined in Table 3.1 below (Equation 3.4-3.6).

$$\text{Sensitivity} = d / (c + d) \qquad (3.4)$$

$$\text{Specificity} = a / (a + b) \qquad (3.5)$$

$$\text{Model Accuracy} = (a + d) / (a + b + c + d) \qquad (3.6)$$

Table 3.1 Confusion Matrix

		Predicted Outcome	
True Outcome		No	Yes
	No	True Negative (a)	False Positive (b)
	Yes	False Negative (c.)	True Positive (d)

Since the logistic regression model predicts the probability of redevelopment for each parcel, to determine the models predictive ability these probabilities needed to be dichotomized so that the predicted vales could be compared to the true outcomes, which were only available is a binary form. This was completed by either rounding the predicted probability up or down depending on a certain cutoff value [28–30]. The cutoff value that is selected can either maximize the sensitivity or specificity of the model. To determine the best cutoff all cutoffs between 0.1 and 0.9, at 0.1 increments, were tested and evaluated. The final model cutoff that was selected was one that maximized the model accuracy, while balancing sensitivity and specificity.

Model Forecast

Following the model training and testing processes the parcels expected to be redeveloped in the future needed to be determined. To complete this task the top fitted model was applied to the 2014 parcel data and the probabilities of future redevelopment for each parcel were produced. The probabilities were then dichotomized using the selected cutoff point as explained above. This step resulted in each of the 2014 parcels being classified as either candidates for future redevelopment or not. The 2014 parcels that were predicted to be candidates for future redevelopment then could then be incorporated into the business as usual growth scenario.

Combining Linear and Logistic Models

Once the parcels expected to have redevelopment in the future were determined the total expected increase in building cover would need to be distributed to these parcels. However, the total increase in building cover depends on the future year. Since the logistic regression model does not have a time component, which is required to determine the value to plug into the linear regression equation, the future year associated with the forecast was determined by using the annual parcel redevelopment rate per year. The temporal evolution of building area, assumed to be a good proxy for impervious cover, was determined by calculating a growth rate using the number of parcels that experience redevelopment between 2004 and 2014. This information gave insight to the rate and extent that redevelopment had occurred in the past and was used to create a scenario for future growth using a "business as usual" growth scenario.

To complete this step, the number of parcels that had been redeveloped between the years 2004-2014 was calculated and converted to a percent of the total parcels in the neighborhood. The percent was then divided by ten, resulting in the percent of the total parcels in the neighborhood that redeveloped historically per year. The annual parcel redevelopment rate was used to determine the number of years into the future the logistic regression was for. For instance, if the rate of redevelopment historically was 1 % and the percent of total neighborhood parcels that the 2014 model predicted was 10 % then we assumed that the model was for a ten year prediction. Once the temporal scale of the logistic regression prediction was determined, the linear regression for building cover could be used to determine the building coverage for the expected year.

The projected additional building coverage then need to be distributed to the predicted parcels. This task has not been previously conducted, based on the literature review, and the task is not straight forward. As a result several approaches were considered:

1. Add building coverage to the parcel with the highest fractional probability of re-development until the max zoning coverage is achieved, and then repeat this process for the parcel with the next highest probability until the total projected building coverage (820,498 sq. ft) is reached. This would be consistent with the observation of redevelopment resulting in large multi-use buildings constructed as close to property lines as allowed. However, a non-negligible portion of the redevelopment is remodeled homes with some pervious area.
2. Distribute building coverage equally across all properties above a certain probability fraction until the additional building coverage is reached. This approach is appealing for its simplicity, but is not consistent with observations of current development.
3. Distribute building coverage according to the predicted probability of redevelopment (i.e., properties with a larger fractional probability of redevelopment receive a larger fractional portion of the increased building coverage). This approach is realistic in that the properties most likely to redevelop are often the candidates for razing the current building (usually small with low value) and building a larger multi-use or multi-family structure or single-family home. The properties with lower fractional probability are most likely properties that already have moderate size building on them, and thus are more likely to be remodeled (adding rooms, etc.), and thus building coverage is a

lower percentage of the property area. In addition, this scenario conceptually incorporates the likelihood that the City of Denver will implement stormwater quality controls for lots smaller than 1 acre, which would result in less development for those lots developed later in the time period (presumably the parcels with the lower fractional probability).

Neither of the methods above are without flaws, and all require limiting assumptions. However, option 3 above is more generally consistent with observations of actual redevelopment in the Berkeley neighborhood, thus this method was selected. It is likely that either of the three options would not produce significantly different results for future stormwater generation, although this would need to be tested using a hydrologic model.

Therefore, using option 3, we distributed the expected growth of building coverage among the different parcels based on each parcel's fractional probability of redevelopment (between 0 and 1). The parcels predicted to redevelop were assigned additional building coverage (beyond that existing in 2014) that was proportional to their percentage of the sum of the probabilities. If a parcel reached the max development based upon the zoning requirements for the parcel, the balance was redistributed amongst parcels that had available development capacity. This process was completed until all the projected building cover was distributed to the predicted parcels resulting in the final model output, expected impervious growth due to redevelopment for each parcel.

Assumptions

There are a number of assumptions, uncertainties, and limitations with the method presented above that require stating. One limitation, associated with the linear regression model, is that the linear trend in building coverage was determined using few data points. Additionally, this model assumes that a projected building cover increase will result in an equal impervious area increase without taking into consideration the existing imperviousness. For example, it does not account for the difference in imperviousness resulting from properties that redevelop with new building coverage over previous imperviousness like patios, outbuildings, and within-lot paved pathways. We project the contribution of this type of imperviousness per parcel to be between 10%-20% based on experience, but that this should be evaluated in more detail in future research. Finally, while the cutoff points in the logistic regression model predictions were based on careful reasoning it was still somewhat arbitrary and different priorities of the user could result in a different cutoff, which could result in a somewhat different development scenario. For future research, different cutoffs could be evaluated depending on the end goal (i.e., future development, future impervious area, future infrastructure needs, etc.).

CHAPTER 4
RESULTS AND DISCUSSION

Linear Regression

The total building cover from 2004 to 2014 was plotted and fit with a linear regression trend line (Figure 4.2). Regardless of the 2008 recession, the total neighborhood building cover steadily increased over the ten year period with a total increase of 17 % or 820,498 square feet (19 acres). The fitted equation had an R^2 value of 0.98 (Equation 4.1). Unlike previously assumed, the spatial distribution was fairly uniform across the entire neighborhood (Figure 4.1).

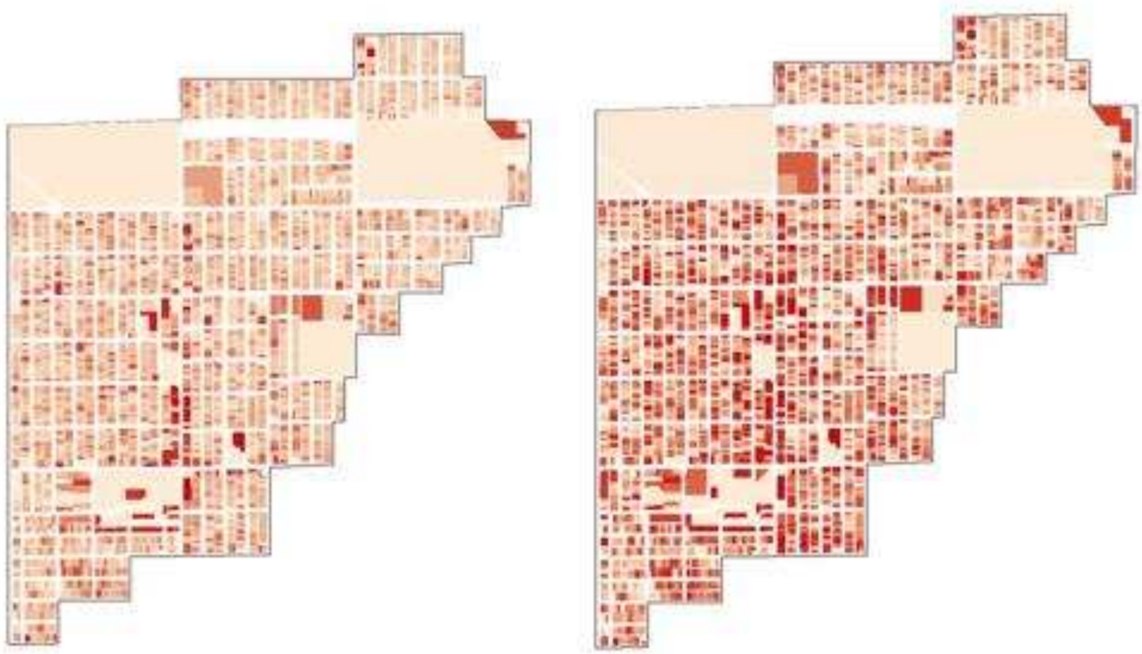

Figure 4.1 Building to Land Area Ratio 2004 (left) and 2014 (right)

Using the linear regression equation and the max allowable building coverage per zoning (18,783,107 sq. ft.) it was determined that, at the current rate of development, max build out of the neighborhood would not occur over the next ten years, In addition, City zoning is likely to change to limit complete buildout. For example, the City is currently considering implementing a requirement for water quality treatment for infill-redevelopment (Darren Mollendor, City and County of Denver, personal communication, 2016). If such a rule, it is likely to encourage installation of green infrastructure, which could reduce development space compared to that, allowed by current zoning. It is probably reasonable, however, to assume that max buildout would not occur before 2050.

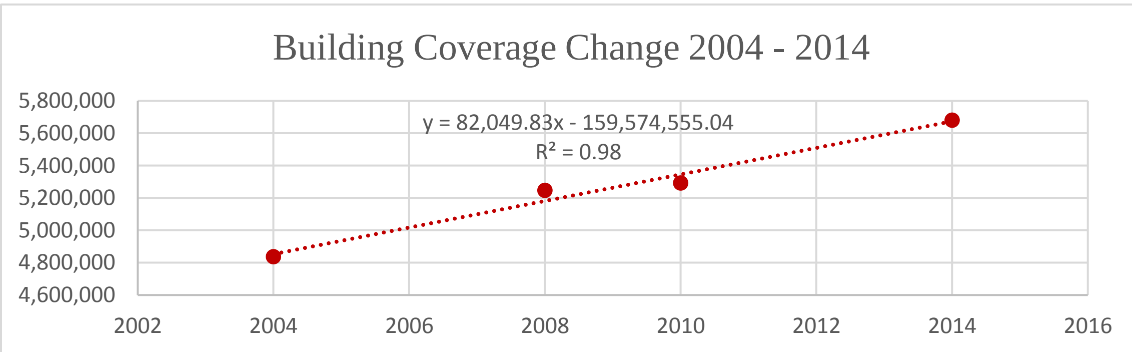

Figure 4.2 Linear Regression of Building Coverage Change 2004-2014

$$y = 82{,}049.83x - 159{,}574{,}555.04 \tag{4.1}$$

Logistic Regression

The logistic regression process was then conducted to determine the probability of redevelopment for future parcels. This process, as mentioned above, includes the model testing, training, and future prediction.

Model Training

Training the model using 2004 assessor's office data was the first step in this process which included testing for collinearity and multivariate analysis.

Collinearity Analysis The collinearity analysis of the 2004 parcel data gave great insight to the variables that needed to be discarded from the model fitting process. Results indicate that the following variable pairs were correlated with each other: total parcel value- improvement value; building coverage - land cover type; and

Improvement to Land Value Ratio (ILV) - Building to Land Area Ratio (BLA) (Appendix A). Of these pairs, we eliminated the variable that showed the less strong correlation to development during the univariate model runs, leaving the following variables: the total parcel value, building coverage, and the BLA (Appendix A). Additionally, the parcels current use classification of vacant, other, condo, and businesses did not show statistical significance with redevelopment. Based on this information 9 variables remained for the logistic regression testing and training. These variables include the following:

1. Parcel area
2. BLA
3. Max coverage allowed by zoning
4. Percent difference between current and max building coverage
5. Total monetary value of the land and improvements,
6. Year built of a structure on the parcel
7. Parcel use classification -apartment
8. Parcel use classification - row house
9. Parcel use classification - single family unit

Multivariate Analysis Using the 2004 parcel training data and the variable set selected above, the '*Glmulti*' package determined the top model from the 512 (or 2^9) candidate models. This model included the following 5 independent variables:

1. Total value
2. Year built

3. Percent difference between current and max building cover (PD)
4. Current use classification: rowhomes (RH)
5. Current use classification apartments (A)

Equation 4.2 displays the intercept and coefficient values for this model. By taking the natural log of this equation the odds ratio can be determined for each variable coefficient to understand how to odds of redeveloped are influenced by a one unit change in the variables. Evaluation of the coefficient, in the form of the odds ratio, for each variable indicate that as the total value, year built and percent difference increase so does the odds of a parcel being redeveloped (Table 2). A one unit increase percent difference between the existing and max cover, or the amount of land available for development, increase the odds of redeveloped by 2.27. This indicates that parcels with small footprints on large lots are desirable candidates for redevelopment. Similarly, a parcel with a use classification of row home or apartment has a higher likelihood of redevelopment than the other use classifications in the neighborhood. Regardless of the magnitude of influence the variables selected in the final model have with the dependent variable of redevelopment, they all have a p value below 0.05 (Table 4.1).

$$\ln\left(\frac{p}{1-p}\right) = 27.91 + (1.88\times 10^{-6})(Total\ Value) + (-1.57\times 10^{-2})(Year\ Built) + (0.82)(PD) + (1.33)(RH) + (5.59)(A) \tag{4.2}$$

Table 4.1 Fitted Model Results

Variable	β	e^β	P Value
Intercept	27.91	1.32178E+12	6.55E-09
Total Value	1.88E-06	1.00000188	0.000683
Year Built	-0.0157	0.984422603	5.33E-10
PD	0.82	2.270499838	0.005288
Row House	1.33	3.781043388	3.07E-07
Apartment	5.59	267.7356197	2E-16

Model Testing

The fitted logistic regression equation was then used to test the models predictive ability. As shown in Table 4.2, the model accuracy remains stable (around 84 %) for all cutoff values above 0.5, which indicates that the selected model variables can predict redevelopment well, regardless of the cut off used. It is not until the cutoff value is set below 0.3 that the model accuracy falls below 80 %. The model sensitivity was below 50 % for all cutoff values greater than 0.2. Conversely, the model specificity remained high (above 80 %) for all cutoff values above 0.2. The predicted number of parcels remained low (<9%) for all cutoff values greater than 0.4. The cut off that best represented actual % of parcels redevelop (22%) was between 0.2 and 0.3. Based on these considerations, it was determined that a cutoff between 0.2 and 0.3 would be appropriate because a natural break is evident in model accuracy between 0.2 and 0.3.

Thus, a cutoff value of 0.25 was selected as the model performance evaluation criteria. This model cutoff displayed 81% accuracy with a specificity of 92% and a sensitivity of 39% (Table 4.3). This optimized the specificity without causing a large reduction in accuracy or sensitivity, and best matched the observed parcel development rate during that time period. This cutoff is used in the logistic regression model to forecast parcels redevelopment from 2014 into the future.

Table 4.2 Cutoff Value confusion matrix results

Cutoff									
	0.9	**0.8**	**0.7**	**0.6**	**0.5**	**0.4**	**0.3**	**0.2**	**0.1**
# of parcels predicted to be redeveloped in testing data	55	58	58	61	66	82	108	237	858
% of total parcels in testing data	6	6	6	7	7	9	12	25	92
Model Accuracy	0.84	0.84	0.84	0.84	0.84	0.83	0.82	0.75	0.27
Sensitivity	0.26	0.28	0.28	0.28	0.29	0.31	0.35	0.50	0.94
Specificity	1.00	1.00	1.00	0.99	0.99	0.97	0.95	0.81	0.08

Table 4.3 Fitted Model Performance

Cutoff	0.25
Model Accuracy	0.81
Specificity	0.92
Sensitivity	0.39

Model Forecast

The fitted logistic regression equation with the 0.25 cutoff was applied to the 2014 parcel data set to predict redevelopment in the future at the parcel scale. The model predicted that 135 parcels, or 15 percent of the total neighborhood parcels, were likely to be redevelopment at some time in the future (Table 4.4) (Figure 4.3). This is slightly less than the 22% of total neighborhood (818) that redeveloped between 2004 and 2014.

Similar to the historic redevelopment, the parcels that were predicted for redevelopment were distributed evenly across the entire neighborhood. Out of the parcels that were projected for redevelopment the average size of the parcel was 103,412 sq. ft. or 0.24 acres, which is much less than the current 1 acre post construction stormwater facility requirement in the City and County of Denver.

Table 4.4 2014 redevelopment forecast

Number of Parcels Predicted for Redevelopment	135
% of Total Neighborhood Parcels	15

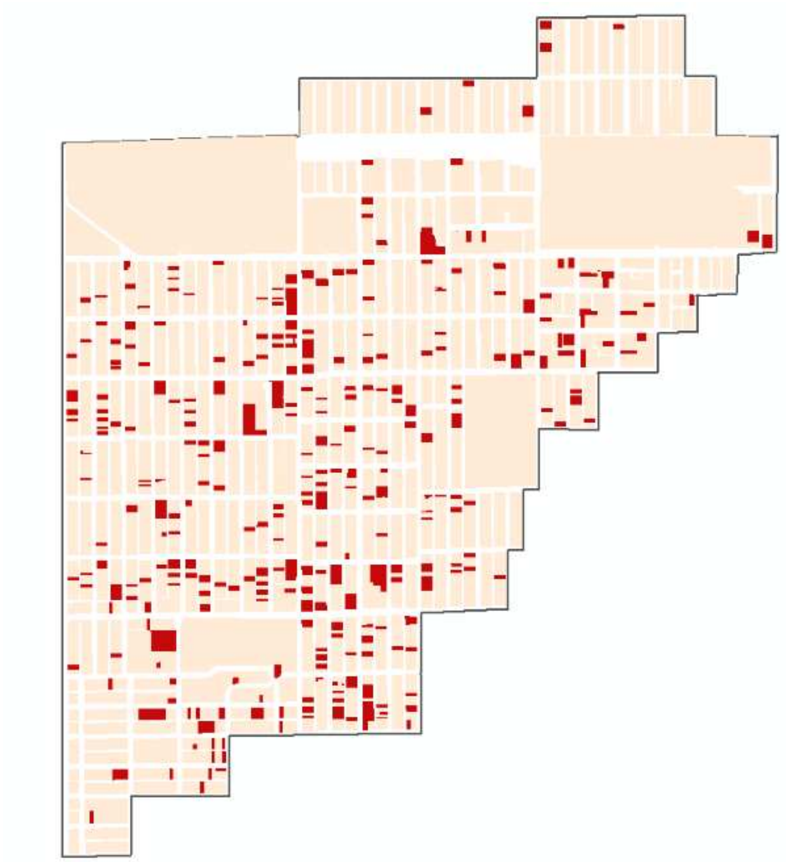

Figure 4.3 2014 Parcels Predicted for Future Redevelopment

The predicted probability of redevelopment for each parcel is presented in Figure 4.4 below and in Appendix C. This value can be used by stormwater modelers to create more informed future subbasin imperviousness scenarios. Most of the predicted probabilities of redevelopment prior to the dichotomization process were below .25 which explains the high specificity of the model (Figure 4.5).

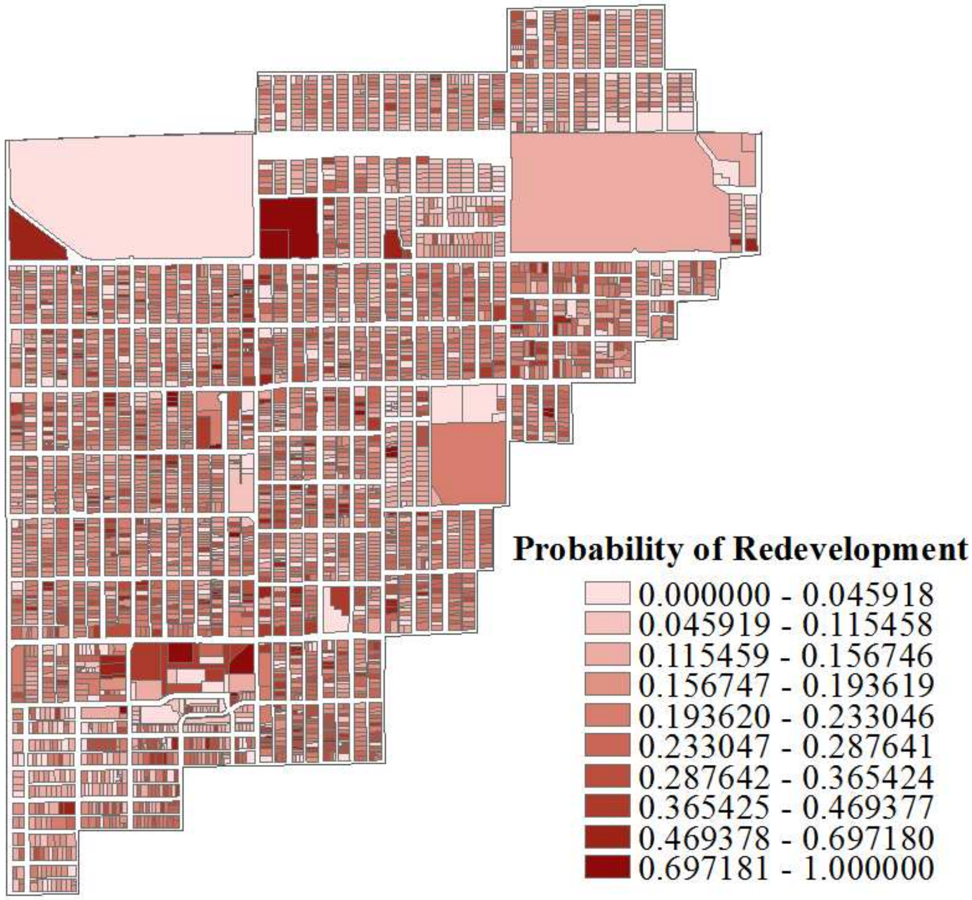

Figure 4.4 Probability of Redevelopment

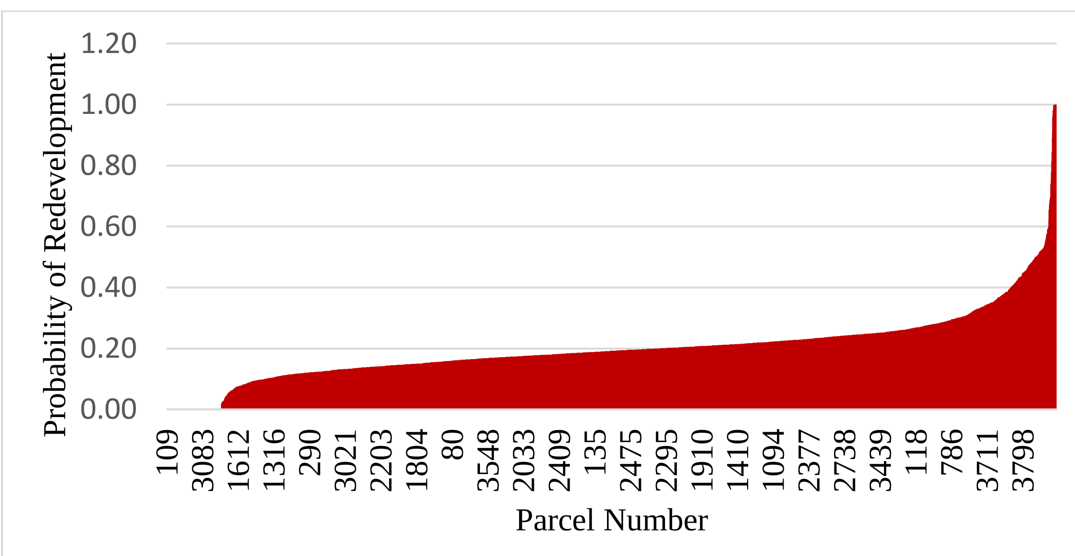

Figure 4.5 Redevelopment Probability Distribution by Parcel

Combining Linear and Logistic Models

Between 2014 and 2024 it was estimated that there would be about 820,498 sq. ft. (18.8 acres) new building square footage within the watershed that needed to be distributed to the predicted parcels. This equates to a 14 % increase in building imperviousness.

Figure 4.6 illustrates the forecast of future impervious area for the Berkeley neighborhood in 2024 by combining the logistic regression model to identify parcels that will redevelop, and linear regression models for building coverage and parcel development rate.

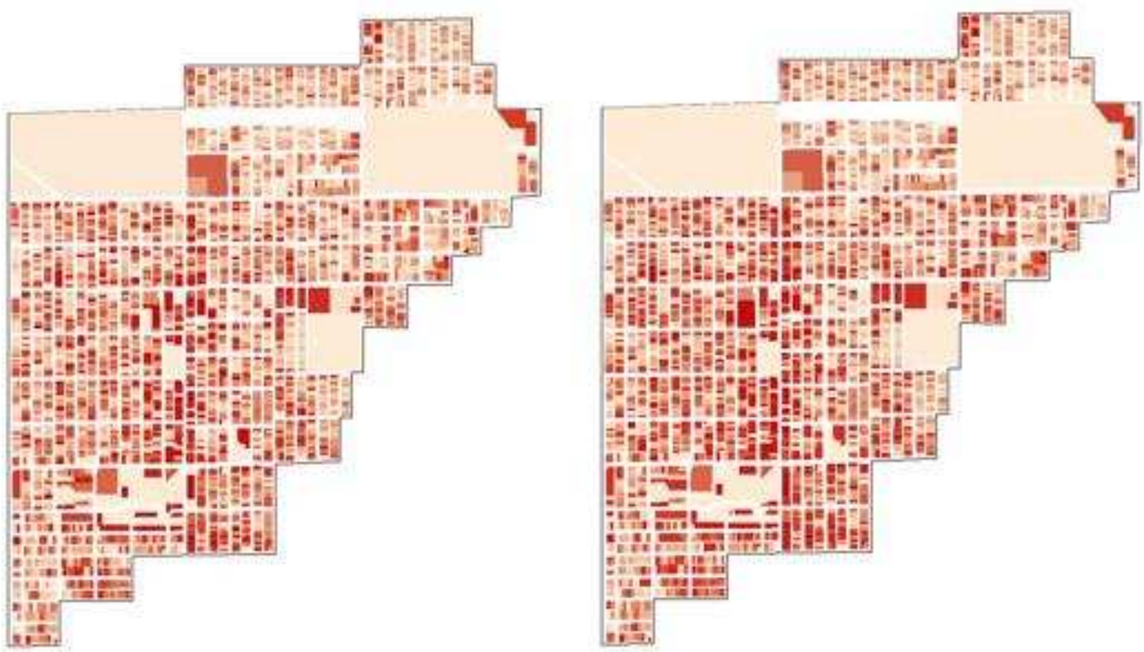

Figure 4.6 Building to Land Area Ratio 2014 (left) and 2024- Predicted (right)

CHAPTER 5
CONCLUSION

The results presented above indicate that building cover change within study site from 2004 – 2014 followed a linear pattern. During this period the total building cover increased by 17% or 1.7 % per year on average. The logistic regression model determined the total value, year built, percent difference between current and max building cover, the current use classification: rowhomes, and current use classification apartments to be the greatest predictors of redevelopment, resulting in a model that was 81 % accurate. The BLA variable was found to be highly correlated with the ILV and according to univariate analysis, the BLA was found to be a better predictor of redevelopment. This is an important finding because it suggests that municipalities may be able to use the variable BLA instead of ILV.

The final model forecasted an increase of 820,498 sq. ft. (18.8 acres) in building coverage between 2014 and 2024. The average size of the parcels expected to be redeveloped in the future was 0.24 acres which is far less than the required 1 acre threshold for the implementation of a permanent stormwater detention and treatment facility.

Both the linear and logistic regression model can be used as a starting point to understand the extent and location of redevelopment on a parcel scale. The logistic regression model had high accuracy and specificity. However, since the models were only fit using ten years of historic data, the presented methods may be better used for the planning and evaluation of watershed projects within 10-year period.

This information can be used by the City and County of Denver to estimate locations of likely future development using open source easily obtainable data, which can help the City plan for infrastructure upgrades, including road improvements, bicycle lanes, public transportation routes, and location and sizing of grey or green infrastructure for storm water peak flow reduction or storm water quality improvement. The tool can also be applied by the city to justify a change in there post construction stormwater management requirements since the average area of parcel redevelopment is less than their current 1 acre minimum. Additionally since many of the models used by the City’s Capital Projects Management group, like EPA SWMM and CUHP 2.0, offer a future imperviousness scenario, this tool will be highly useful for their Storm Drainage Master Planning updates.

While the model performed fairly well, there are many areas that would be appropriate for future evaluation. Some of which include expanding the analysis to other neighborhoods within the City, incorporating more variables into the analysis, and a more in depth validation process using current data as it becomes available. By expanding the analysis to additional neighborhoods and comparing the best models that were selected trends could be established regarding the variables of influence. Incorporating more variables into the analysis, like economic indicators and population data, may result in a greater model accuracy. Lastly, validating the model forecast using current data as it becomes available would provide insight to the exact temporal scale of the redevelopment forecast.

REFERENCES

[1] United States Environmental Protection Agency, U.S. Environmental Protection Agency NPDES Permit Writers' Manual, 2010.

[2] Enginuity Engineering Solutions, Icon Engineering Inc, Matrix Design Group, Olsson Associates, San Engineering LLC, Zoeller Consulting LLC, City and County of Denver Storm Drainage Master Plan, 2014.

[3] C. Agarwal, G.M. Green, J.M. Grove, T.P. Evans, C.M. Schweik, A Review and Assessment of Land-Use Change Models: Dynamics of Space, Time, and Human Choice, Apollo Int. Mag. Art Antiq. (2002) 62. doi:10.1289/ehp.6514.

[4] P. Waddell, UrbanSim: Modeling urban development for land use, transportation, and environmental planning, J. Am. Plan. Assoc. 68 (2002) 297–314. doi:10.1080/01944360208976274.

[5] F. Zhao, S. Chung, A Study of Alternative Land Use Forecasting Models, (2006). http://www.wgianalytics.com/landuse/docs/florida/FDOT_BD015_10_rpt.pdf (accessed October 26, 2015).

[6] B.J. Son, F. County, Residential Build-Out Analysis Based on Zoning Meeting infrastructure demand using GIS, Analysis. (2009) 28–29.

[7] Maryland Department of Planning, Estimating Residential Development Capacity: A Guidebook for Analysis and Implementation in Maryland, 2005.

[8] Portland Bureau of Planning and Sustainability, City of Portland Development Capacity Analysis, Portland, 2010. papers://ef64220a-a077-48ec-ae81-be13b32d2073/Paper/p227.

[9] E. Janowicz, Predicting Redevelopment : Modeling Building Demolition, in: 2010.

[10] Community Attributes, Identifying Redevelopable Lands: Application of a Land Value Potential (LVP) approach in Urban Centers, 2009.

[11] Tashman Associates Planners and Policy Consultants, Infill and Redevelopment Strategies, (1994).

[12] T. Watt, Solimar Research Group, California Infill Estimation Methodology Project, 2004.

[13] Community Planning Workshop, Eagle Point Buildable Lands Analysis, 2001.

[14] City of Bellingham, Land Supply Methodology, 2006.

[15] P.M. Torrens, How Land-use-transportation Models Work, 2000.

[16] American Society of Planning Officials, Floor Area Ratio, Chicago: The Society, 1958.

[17] Envision Eugene Technical Resource Group, Commercial Redevelopment Methodology, 2011.

[18] B. Rolett, New Apartments Headed to Tennyson, Bus. Den. (2015).

[19] B. Rolett, Tennyson apartment rush adding $ 23M project, Bus. Den. (2015).

[20] E.P. Xing, M.I. Jordan, R.M. Karp, Feature Selection for High-Dimensional Genomic Microarray Data, (n.d.).

[21] E.W. Steyerberg, M.J. Eijkemans, F.E. Harrell Jr., J.D. Habbema, Prognostic modeling with logistic regression analysis: in search of a sensible strategy in small data sets, Med.Decis.Making. 21 (2001) 45–56. doi:10.1177/0272989X0102100106.

[22] B. Pradhan, S. Lee, Landslide susceptibility assessment and factor effect analysis: backpropagation artificial neural networks and their comparison with frequency ratio and bivariate logistic regression modelling, Environ. Model. Softw. 25 (2010) 747–759. doi:10.1016/j.envsoft.2009.10.016.

[23] M. Geza, K. Murray, J. McCray, Watershed-Scale Impacts of Nitrogen from On-Site Wastewater Systems: Parameter Sensitivity and Model Calibration, J. Environ. Eng. 136 (2010).

[24] Vincent Calcagno, glmulti: An R Package for Easy Automated Model Selection with (Generalized) Linear Models, (n.d.).

[25] S.S. Coughlin, B. Trock, M.H. Criqui, L.W. Pickle, D. Browner, M.C. Tefft, The logistic modeling of sensitivity, specificity, and predictive value of a diagnostic test, J. Clin. Epidemiol. 45 (1992) 1–7. doi:10.1016/0895-4356(92)90180-U.

[26] J. Pearce, S. Ferrier, Evaluating the predictive performance of habitat models developed using logistic regression, Ecol. Modell. 133 (2000) 225–245. doi:10.1016/S0304-3800(00)00322-7.

[27] J. Blacher, R. Asmar, S. Djane, M. London, M.E. Safar, Aortic Pulse Wave Velocity as a Marker of Cardiovascular Risk in Hypertensive Patients, (1999).

[28] L.H.J. Eberhart, G. Geldner, P. Kranke, A.M. Morin, A. Scha, H. Treiber, H. Wulf, The Development and Validation of a Risk Score to Predict the Probability of Postoperative Vomiting in Pediatric Patients, (2004) 1630–1637. doi:10.1213/01.ANE.0000135639.57715.6C.

[29] F. Barkhof, M. Filippi, D.H. Miller, P. Scheltens, A. Campi, C.H. Polman, G. Comi, H.J. Adèr, N. Losseff, J. Valk, Comparison of MRI criteria at first presentation to predict conversion to clinically definite multiple sclerosis, Brain. 120 (1997) 2059–2069.

[30] D.P. Devanand, X. Liu, M.H. Tabert, G. Pradhaban, K. Cuasay, K. Bell, M.J. de Leon, R.L. Doty, Y. Stern, G.H. Pelton, Combining Early Markers Strongly Predicts Conversion from Mild Cognitive Impairment to Alzheimer's Disease, Biol. Psychiatry. 64 (2008) 871–879. doi:10.1016/j.biopsych.2008.06.020.

APPENDIX A

Table A-1 Variable Correlation with the Dependent Variable

Predictor	**p value**
ILV	0.00348
Land Cover	0.0106
Building Coverage	0.000131
Max Building Coverage	0.043
Year Built	0.596
% Difference Between Current and Max	0.0733
BLA	1.45E-08
Total Value	1.88E-08
Building Value	2.65E-09
Land Value	0.00394
Vacant	0.309
Single Family	2E-16
Row House	0.000855
Other	0.0854
Condo	0.952
Business	0.12
Apartment	2E-16

Table A-2 Variable Spearman's Correlation Values

Spearman's correlation										
	LANDVAL	IMPROV-EMENT VALUE	TOTAL VALUE	LANDCOV	MAXCOV	YRBLT	BLDCOV	PD	ILV	BLA
LANDVAL	1.000	0.079	0.362	0.895	-0.236	0.139	0.268	0.186	-0.559	-0.345
IMPROVEMENTVAL	0.079	1.000	0.930	0.103	0.123	0.222	0.670	-0.442	0.687	0.584
TOTALVAL	0.362	0.930	1.000	0.357	0.052	0.253	0.721	-0.319	0.421	0.411
LANDCOV	0.895	0.103	0.357	1.000	-0.045	0.151	0.324	0.255	-0.479	-0.312
MAXCOV	-0.236	0.123	0.052	-0.045	1.000	0.154	0.221	0.421	0.278	0.271
YRBLT	0.139	0.222	0.253	0.151	0.154	1.000	0.167	0.073	0.087	0.063
BLDCOV	0.268	0.670	0.721	0.324	0.221	0.167	1.000	-0.471	0.307	0.693
PD	0.186	-0.442	-0.319	0.255	0.421	0.073	-0.471	1.000	-0.504	-0.718
ILV	-0.559	0.687	0.421	-0.479	0.278	0.087	0.307	-0.504	1.000	0.736
BLA	-0.345	0.584	0.411	-0.312	0.271	0.063	0.693	-0.718	0.736	1.000

Table A-3 Variable Chi Square Correlation Values

Chi Square			
Predictor 1	**Predictor 2**	**Chi.Square**	**p.value**
VACANT	SINGLEFAMILY	343.712	0
VACANT	BUSINESS	1.02	0.313
VACANT	APARTMENT	2.985	0.084
VACANT	OTHER	0.007	0.931
VACANT	ROWHOUSE	0.909	0.34
VACANT	CONDO	0	1
SINGLEFAMILY	BUSINESS	669.038	0
SINGLEFAMILY	APARTMENT	1395.379	0
SINGLEFAMILY	OTHER	196.34	0
SINGLEFAMILY	ROWHOUSE	626.317	0
SINGLEFAMILY	CONDO	33.836	0
BUSINESS	APARTMENT	6.643	0.01
BUSINESS	OTHER	0.317	0.573
BUSINESS	ROWHOUSE	2.497	0.114
BUSINESS	CONDO	0	1
APARTMENT	OTHER	1.364	0.243
APARTMENT	ROWHOUSE	6.161	0.013
APARTMENT	CONDO	0	1
OTHER	ROWHOUSE	0.264	0.607
OTHER	CONDO	0	1
ROWHOUSE	CONDO	0	1

APPENDIX B

```
######################################################################
################ Logistic Regression Model R Code ##################
################       By : Lisa Cherry            ##################
################     November 12th, 2016           ##################
######################################################################

############### load parcel data
setwd("C:/Users/Lisa/Documents/Berkeley Backup/Thesis/R")
data04 <- read.csv("data04.csv")
data10 <- read.csv("data10.csv")
data14 <- read.csv("data14.csv")

library(glmulti)
library(ROCR)
library(caret)

###############Organize the data 2004

#Spilt data into Training (75%) and Validation (25%) datasets
## 75% of the sample size
smp_size04 <- floor(0.75 * nrow(data04))

## set the seed to make your partition reproductible
set.seed(123)# fix random seed to make this report reproducible!
train_ind04 <- sample(seq_len(nrow(data04)), size = smp_size04)

train04 <- data04[train_ind04, ]
test04 <- data04[-train_ind04, ]

#################### Run Logistic Regression 2004-2014
glmulti.04to14.out <- glmulti(RD14 ~ TOTALVAL + LANDCOV + YRBLT + BLA +
MAXCOV + PD + ROWHOUSE + APARTMENT + SINGLEFAMILY, data=train04,
                       level=1,              # No interaction considered
                       method="h",            # Exhaustive approach
                       crit = "aicc",           # AIC as criteria
                       confsetsize = 10,        # Keep top 100 models
                       plotty = F, report=F,   # No plot or interim reports
                       fitfunction= "glm",    # glm function
                       na.action= na.exclude,# omitting NA's
                       family="binomial")    # binomial family for logistic
regression

# a brief summary of the results
print(glmulti.04to14.out)

# obtain a plot of the AIC values for all of the models
plot(glmulti.04to14.out)

#show results for top 8 models
top04.14 <- weightable(glmulti.04to14.out)
top04.14 <- top04.14[top04.14$aic <= min(top04.14$aic) + 2,]
top04.14

#run top 8 models
```

```
glm1 <- glm(RD14 ~ 1 + TOTALVAL + YRBLT + PD + ROWHOUSE + APARTMENT,
data=train04)
glm2 <- glm(RD14 ~ 1 + TOTALVAL + YRBLT + MAXCOV + ROWHOUSE + APARTMENT,
data=train04)
glm3 <- glm(RD14 ~ 1 + TOTALVAL + YRBLT + PD + ROWHOUSE + APARTMENT +
SINGLEFAMILY, data=train04)
glm4 <- glm(RD14 ~ 1 + TOTALVAL + YRBLT + BLA + MAXCOV + ROWHOUSE +
APARTMENT, data=train04)
glm5 <- glm(RD14 ~ 1 + TOTALVAL + YRBLT + MAXCOV + PD + ROWHOUSE + APARTMENT,
data=train04)
glm6 <- glm(RD14 ~ 1 + TOTALVAL + YRBLT + BLA + PD + ROWHOUSE + APARTMENT,
data=train04)
glm7 <- glm(RD14 ~ 1 + TOTALVAL + YRBLT + MAXCOV + ROWHOUSE + APARTMENT +
SINGLEFAMILY, data=train04)
glm8 <- glm(RD14 ~ 1 + TOTALVAL + LANDCOV + YRBLT + PD + ROWHOUSE +
APARTMENT, data=train04)


#average the top 8 models together
library(MuMIn)
glm.ave <- model.avg(glm1, glm2, glm3, glm4, glm5, glm6, glm7, glm8)
summary(glm.ave)

## Show result for the best model
glmtop04.14 <- glmulti.04to14.out@objects[[1]]
summary(glmtop04.14)

#evaluate variable importance
plot(glmulti.04to14.out, type="s")

##calculating multimodel inference
#registering getfit method
setOldClass("rma.uni")
setMethod('getfit', 'rma.uni',function(object, ...) {
  if (object$knha==FALSE) {
   cbind(estimate=coef(object), se=sqrt(diag(vcov(object))), df=100000)
  } else {
   cbind(estimate=coef(object), se=sqrt(diag(vcov(object))), df=object$k-
object$p)
  }
})
#carrying out multimodel inference calculations
round(coef(glmulti.04to14.out, type="s"),4)


################### Final Model Evaluation and Diagnostics

## calculate the odds ratios for each preditor
exp(coef(glmtop04.14))

###################### use the model parameters to predict the value of the
target
##################### variable in a completely new set of observations using
best model

pred.prob.04.14 <- predict(glmulti.04to14.out@objects[[1]], newdata=test04 ,
type="response")
```

```
binary.pred04.14 <- ifelse(pred.prob.04.14 > 0.3,1,0)
write.table(pred.prob.04.14, "pred.prob.04.14.txt", sep="\t")
write.table(binary.pred04.14, "binary.pred04.14.txt", sep="\t")
confusionMatrix(binary.pred04.14, test04$RD14,positive= "1")
confusionMatrix(binary.pred04.14, test04$RD08,positive= "1")

pred.prob.10.14 <- predict(glmulti.04to14.out@objects[[1]], newdata=data10 ,
type="response")
binary.pred10.14 <- ifelse(pred.prob.10.14 > 0.3,1,0)
write.table(pred.prob.10.14, "pred.prob.10.14.txt", sep="\t")
write.table(binary.pred10.14, "binary.pred10.14.txt", sep="\t")
confusionMatrix(binary.pred10.14, data10$RD14,positive= "1")

#selecting cutoff based on accuracy
acc.perf = performance(pred, measure = "acc")
plot(acc.perf)
ind = which.max( slot(acc.perf, "y.values")[[1]] )
acc = slot(acc.perf, "y.values")[[1]][ind]
cutoff = slot(acc.perf, "x.values")[[1]][ind]
print(c(accuracy= acc, cutoff = cutoff))

### Calculate ROC and AUC
pred04.14 <- prediction(pred.prob.04.14, test04$RD14)
perf04.14 <- performance(pred04.14, "tpr", "fpr")
plot(perf04.14)
abline(a=0,b=1)
performance(prediction(pred.prob.04.14, test04$RD14), "auc")@y.values[[1]]

#selecting cutoff based on accuracy
acc.perf = performance(pred04.14, measure = "acc")
plot(acc.perf)
ind = which.max( slot(acc.perf, "y.values")[[1]] )
acc = slot(acc.perf, "y.values")[[1]][ind]
cutoff = slot(acc.perf, "x.values")[[1]][ind]
print(c(accuracy= acc, cutoff = cutoff))


#selecting the cutoff by maximizing sensitivity and specificity
opt.cut04.14 = function(perf04.14, pred04.14){
  cut.ind = mapply(FUN=function(x, y, p){
   d = (x - 0)^2 + (y-1)^2
   ind = which(d == min(d))
   c(sensitivity = y[[ind]], specificity = 1-x[[ind]],
     cutoff = p[[ind]])
  }, perf04.14@x.values, perf04.14@y.values, pred04.14@cutoffs)
}
print(opt.cut04.14(perf04.14, pred04.14))


#### Predict 2014 Parcel Redevelopment

pred.prob14 <- predict(glmulti.04to14.out@objects[[1]], newdata=data14 ,
type="response",na.action= na.exclude)
binary.pred14 <- ifelse(pred.prob14 > 0.3,1,0)
write.table(pred.prob14, "pred.prob14.txt", sep="\t")
write.table(binary.pred14, "binary.pred14.txt", sep="\t")
```

APPENDIX C

A list of the 2014 parcels and their predicted probability of redevelopment can be located in the following electronic spreadsheet:

https://docs.google.com/spreadsheets/d/1r0FXuNXVVzGi-Fd011BQfB4zslySb9tUMbhoPk7mDCM/edit?usp=sharing