An Econometric Analysis of Paid Sewer Backup Damage Claims In Seattle

Joshua Simpson

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Science

University of Washington

2013

Committee:

Sergey Rabotyagov

Robert Halvorsen

John Perez-Garcia

Terry Martin

Program Authorized to Offer Degree:

School of Environment and Forest Sciences

University of Washington

Abstract

An Econometric Analysis of Paid Sewer Backup Damage Claims In Seattle

Joshua Simpson

Chair of the Supervisory Committee:

Assistant Professor Sergey Rabotyagov

School of Environment and Forest Sciences

Seattle is known for its high occurrence of rainfall events but most of them are low intensity events. However, when it rains heavily, sewer pipes can reach capacity and sewer backups may result. Damage claims are filed by the parties with sewer backup damage incurred on their property and, in some cases, the city will pay a damage claim amount to cover the amount of damage. The dataset used in this project contains sewer backups that caused a total of $8 million damage from August 2004 to March 2011. Nearly half of the damage claims in the dataset were due to three major storms that occurred within that timeline.

Meteorological, demographic, environmental and structural variables that explain the damage caused by those three storms are analyzed using a rare events logistic regression model. Sewer backups are rare events in Seattle since the highest claim-producing storm induced 147 claims in Seattle, a city with over 180,000 parcels. The model uses the claims from a particular storm and a random stratified citywide sample of parcels (stratified by neighborhood) to examine the

explanatory variables that explain the occurrence of backups. A conditional backup probability is calculated for each sample parcel.

A spatial econometric model is used to measure the effect of explanatory variables that explain various levels of sewer backup damage while accounting for spatial effects of clustered claims. The results of the model are used to calculate potential damage for each sample parcel. The probability and potential damage calculations are multiplied together to produce an expected sewer backup damage (ESBD) amount for the sample parcels. These calculations were used to create three maps that represent probabilities of backups (conditional on the occurrence of a claim-producing storm), potential damage and ESBD.

These maps and the data that makes up the map can be used to prioritize preventative maintenance before a storm season. There are many other risks that face utility customers in Seattle but focusing on sewer backup risk allows for the application of two econometric models to better assess this specific risk. Such analysis has not been utilized to analyze the occurrence of sewer backups to date. Given the results of Salathe et al. (2010) and Zhu (2012) that suggest that higher frequency and higher intensity storms will affect the Puget Sound area, the accumulation of data and the use of the best information can efficiently mitigate damage caused by future storms.

Acknowledgements

In my two and a half years at the University of Washington, I was basically learning economics in my own program so I did not have a community of students that I associated with. However, there are a lot of individuals that I need to thank and recognize for helping me with this project.

I would like to thank my academic advisor, Dr. Sergey Rabotyagov, was immensely helpful in helping me with the research design of this project as well as the projects before this one that were scraped for one reason or another. The models that I used were not his expertise yet he was able to associate the data and the problem I was trying to explore with the correct model to use. He was my advisor from the beginning and helped me craft my custom track in applied economics. I am very grateful for the advice and encouragement I got from him during my time at UW.

I would like to thank the other members of my committee. They are Dr. John Perez-Garcia, Dr. Robert Halvorsen and Terry Martin. Dr. Perez Garcia and Dr. Halvorsen agreed to be on my committee when I was exploring carbon cap and dividend schemes but they are well-respected economists and it was an honor to have them on my committee. Terry Martin, the director of Asset Management and Economic Services group at Seattle Public Utilities, was very helpful in providing information and advice for my project when I was an intern at Seattle Public Utilities. When I left the utility, he was helpful in translating my research to other employees at Seattle Public Utilities.

When I was exploring the data and research design for this project, I interviewed many employees at Seattle Public Utilities and they were all helpful in providing information. These individuals are Brian Morgenroth, Mark Benson, Jason Sharpley, James Rufo-Hill, Holly

McCracken, Albert Ponio, George Sidles, Scott Reese and anyone else that helped me out. While this project was not funded by Seattle Public Utilities, I was allotted time to conduct these exploratory interviews and gather data whilst interning at Seattle Public Utilities.

Next, I would like to thank Dr. Monika Moskal and the Remote Sensing Geospatial Analysis Lab for providing the tree density LIDAR image. It was high-quality data and it allowed me to accurately calculate tree density data in GIS, which was a useful exercise.

Finally, I would like to thank all of the people that worked in the Social Science lab when I was working in there.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Figure No. **Page**

LIST OF TABLES

Table No. **Page**

Chapter 1: Introduction

Seattle is known for its persistent rainy weather though most rain events in the Puget Sound area are low-intensity events. The rainfall can be described as a slow trickle that lasts seemingly for months. The drainage and waste water system is built to conduct the rain water that is not absorbed by the soil to a receiving water body or a treatment plant. The system is designed to handle most rain events but it is not equipped to handle high-intensity rain events (Martin, 2012). The evidence suggests that drainage infrastructure designed using mid-20th century rainfall records [in the Puget Sound] may be subject to a future rainfall regime that differs from current design standards (Rosenberg et al., 2010). When high-intensity rain events occur, issues in the form of floods and sewer backups result due to excess stormwater and wastewater flow exiting the system into customers' property.

A sewer backup occurs when the sewer and stormwater flow is blocked from continuing to its intended endpoint due to an obstruction within the sewer mainline. The flow enters a residential or commercial building through the side sewer infrastructure connected from the mainline to the plumbing in the building. When this process of flow is disrupted in the system and there is enough pressure cause by the blockage and the heavy flow of water, stormwater and wastewater enters a residential or a commercial building and causes damage.

If a residence or a business experiences property and/or personal damage as a result of sewer backup and flood damage, a claim can be filed with the city to receive reparations for the amount of the damage. The claim will go to litigation and, if the city is liable, the claimant will be awarded a damage payment in the amount decided in the litigation process. A potential claimant has three years to file a claim after the date of loss (Martin, 2012).

This thesis project aims to explain why sewer backups occurred where they did and why varying levels of damage occurred. The occurrence of sewer backups is modeled using a rare events logistic regression model. The parcels that received payments as a result of settled claims associated with a storm will be measured against parcels that did not. The largest storm in the dataset, which ranges from 2004 to 2011, led to 148 claims. Given the amount of parcels in Seattle (~ 183,000), these sewer backups are rare events and a model that corrects the estimates for the true population is necessary to measure why backups occurred where they did.

This thesis also aims to explain various levels of sewer backup damage. Potential property damage is modeled using a spatial econometric model. Given the clustered nature of the spatial distribution of claims, spatial dependence may need to be accounted for. Spatial dependence reflects a situation where values, observed at one location or region, depend on the values of neighboring observations at nearby locations (LeSage and Pace, 2009). A spatial econometric model adds a spatial weights matrix to the dependent variable or the error term of an ordinary least squares (OLS) regression model and controls for spatial dependence.

The results of the rare events logistic regression model produce a probability of a positive sewer backup claim occurring on a parcel given the rainfall variables associated with the storm that caused them. The spatial econometric model produces coefficients that attach a potential damage amount to a parcel. The combination of the probability and the damage amount produce an expected damage amount for each parcel estimated.

Known Causes of Damage

The three main factors that cause backups in SPU's sewer system are rainfall intensity, tree roots and grease (Martin, 2012). Rainfall is ostensibly the primary cause of system issues. The other

two factors, tree roots and grease, exacerbate the rainfall intake in the system by robbing the system of the designed capacity. Tree roots infiltrate into the pipe via natural processes and through cracks in the pipe. Grease enters the system via disposal into the drains connected to the system. The grease coats the inside walls of the pipes and can coat the tree roots already in the system. The loss of capacity combined with the large amounts of flow and the speed of the flow entering the system leads to sewer backups.

Do these three variables tell the whole story? Do other factors explain why backups occur and explain the level of damage incurred on a residence or business? This thesis project postulates that these factors may explain much of the story but other factors such as demographic and structural variables may also explain the causes of backups and size of the damage claim amounts.

Paying out damage compensation via sewer claims that occur as a result of some of the explanatory variables that will be used in this thesis makes the occurrence of backups an economic problem for the utility. The factors that add to the cause of backups and the damage of sewer backups may be mitigated and the factors that prevent backups and backup damage may be further established to lower the amount paid out by backups. Many of the factors are static or exogenous and cannot be altered to prevent backups but the purpose of these models is to explain the cause and damage potential of backups within Seattle and estimate areas of highest risk where sewer backup prevention measures can be implemented.

Mitigating the risks of sewer backup occurrences should then translate into reduced costs of claim arbitration and payout for the utility and the claimants. This, in turn, will improve the level of service provided by the utility due to the absence of sewage causing damage or less damage in

some cases. This higher level of service would help the utility to improve social welfare for the customers in Seattle.

There is a possibility of an increase in significant storms due to global climate change which may mean more claim-producing storms for Seattle. Both simulations in the Salathé Jr. et al. (2010) yield an increase in the measures of extreme precipitation even though one simulation produced mostly reductions in total precipitation during winter and spring. Climate simulations by Zhu (2012) reveal that rainfall intensity differences from the historical runs and future simulations were significant at all intervals from two years to 100 years. Given these findings, mitigation should be a higher priority and this thesis provides tools to aid in proper mitigation of increased rainfall events and intensities.

Purpose of Study

This thesis estimates the expected damage from system overflow caused by heavy rainstorms. Rain is necessary for the flora, fauna and humans in the Puget Sound region and is considered a positive benefit for what it provides. However, when excess rain infiltrates the drainage and waste water system and causes damage to a residence or a business, that rainfall allocation provides negative benefit. Since rainfall is an exogenous atmospheric process, only damages can be mitigated. The purpose of the study is to find the factors that affect the damage. The variable factors can be changed to mitigate the damage and the fixed factors can be used as reference so that mitigation efforts can be concentrated to these areas before a future event occurs.

More extreme rainfall events may lead to more claims paid out by the city and more issues for utility customers. Knowing which areas could be impacted and what measures need to be carried

out for prevention of system issues can mitigate the damages of extreme rainfall. More future Pacific Northwest climate issues will be discussed later in the paper.

An econometric analysis of sewer backup claims has not been published to date, although flood claims and risk have been analyzed in a similar manner. This type of empirical analysis allows for the relative measurement of risk of excess rainfall and how other factors (demographic, environmental and structural) add or subtract to the occurrences of and damages from backups. This risk will always be present and the framework allows for the accumulation of information to better inform risk managers as time goes on.

This thesis is structured in the following manner: Chapter 2 covers the history of the sewer system and recent storms as well as a literature review of relevant topics, Chapter 3 explains the methodology of the data and the models, Chapter 4 includes results and discussion and Chapter 5 contains implications and conclusions.

Chapter 2: Background and Literature Review

The Sewer System

The areas in Seattle that were incorporated first are the central part of the city around downtown and the area in North Seattle around Green Lake in the mid-to-late 1800's (Phelps, 1978). The earliest existing sewer system was implemented in 1891 after the devastating fire where the city had to build over existing buildings and infrastructure (Phelps, 1978). These systems were combined sewer systems (storm water and sewer flow together) and most of these assets remain combined today.

Over the years, Seattle has incorporated areas of Seattle north and south of the existing incorporated sections of the city and, as a result, has inherited informal drainage systems (Martin, 2012). Different parts of the city's sewer systems were built differently with roughly one third of the city's infrastructure were built as combined, one third partially separated and the rest separated systems (Martin, 2012). The older infrastructure is downtown and was built as combined and the post-WWII construction in the north and in the south was decidedly built as separated when the city decided to connect the systems and treat wastewater (Martin, 2012).

This study focuses on combined and sanitary sewer assets as drainage-only pipes do not produce sewer backups, though they can produce flooding. The different types of sewer systems are displayed in Figure 1.

The Major Storms

The rare events logistic regression model (may also be referred to as the probability model) is used to model the location of sewer backup claims occurred where they did using parcels where

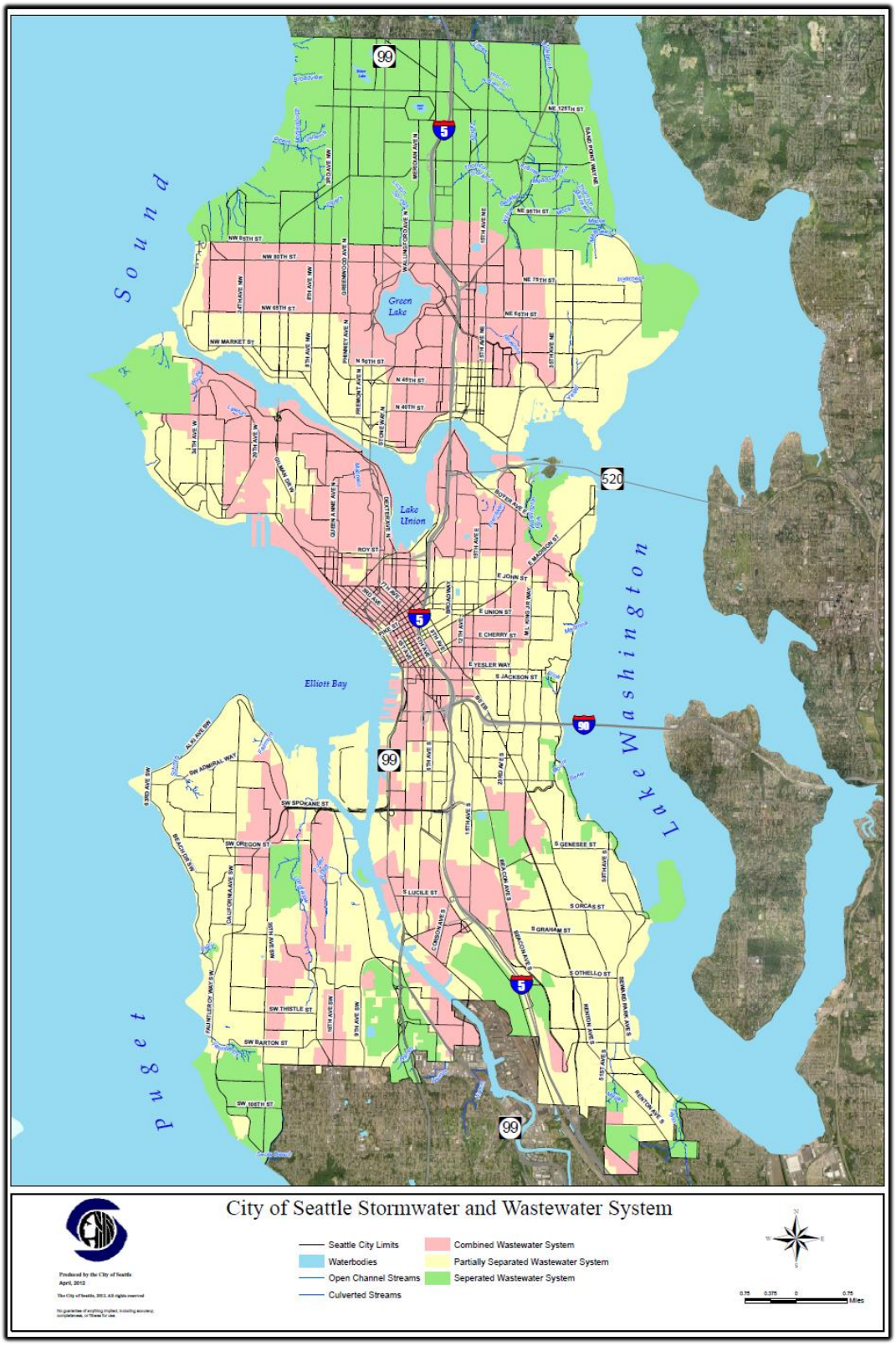


Figure 1: The City of Seattle Drainage and Wastewater System

claims occurred and parcels where no claims occurred. To do so, a specific storm must be selected based on the amount of claims resulting from the storm. The dataset contains backups that occurred as a result of many different various-sized storms. There were three storms in the dataset that produced a sizeable amount of claims to be used in the Rare Events logistic regression model.

The first storm took place on May 27th, 2006. This storm was a short storm (around 2 to 3 hours) that only produced substantial rainfall amounts in the area of three rain gauges (0.66 inches of rain in one hour and 1.1 inches of rain in three hours), yet the storm produced 32 sewer backup claims. The statistical characteristics of the rainfall variables attached to sampled parcels that pertain to the May 2006 storm are displayed in Figure A in Appendix A.

The next took place on December 14th, 2006. This storm can be described as a long rainstorm with a 2 to 3 hour spike of heavy rainfall, especially in the Madison Valley where the rain gauge in the area recorded 1.35 inches of rain from 2 p.m. to 5 p.m. Within that spike, the peak five and ten minute rainfall amounts were on average 0.14 and 0.22 inches, respectively. This storm produced 147 sewer backup claims and the summary statistics pertaining to this storm are displayed in Figure 7 in Chapter 4. This storm will be the focus of the analysis for this thesis.

The third storm took place on December 3rd, 2007. This storm began later in the day on December 1st according to most gauges but produced low, steady rainfall until the 3rd where heavier rainfall persisted for most of the day, consistently ranging from 0.2 to 0.45 inches per hour for much of the day. The rainfall intensity for the storm period measured, which began later in the day on the 2nd, was 0.22 inches per hour. That rainfall intensity value is twice as large as

what was measured in the other two storms. This storm produced 48 sewer backup claims and the summary statistics applicable to this storm are displayed in Figure B in Appendix B.

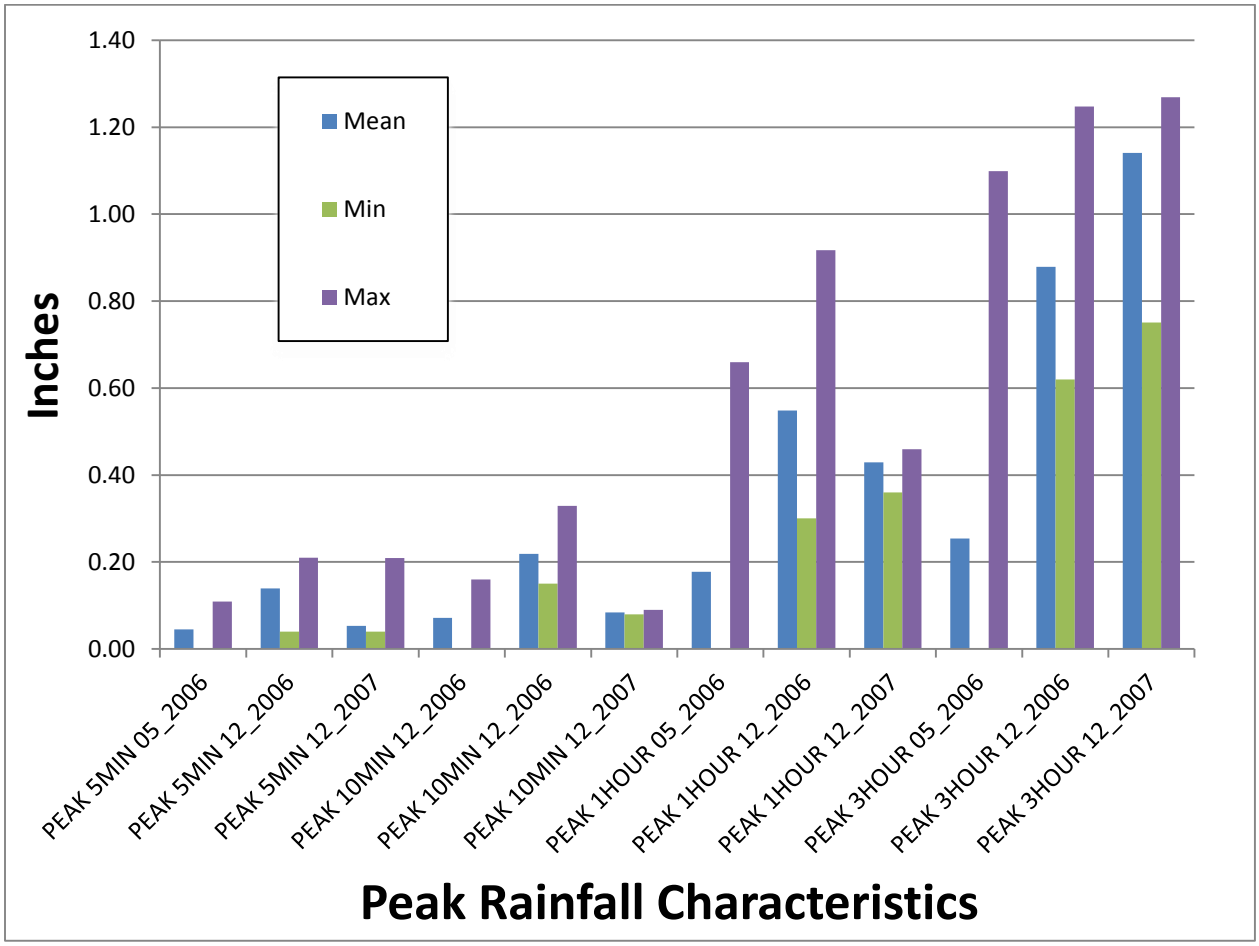


Figure 2: Peak rainfall amounts for the three major storms

The varying nature of these storms is fortuitous as different storms should reveal different factors that prove explanatory for why backups occurred. A future storm fitting the characteristics of one of the aforementioned storms can be used as a comparison to estimate the level of claim amounts that can be expected to be paid out as a result of the storm.

Sewer Backup Damage and Cause Assessment

No studies that examine the cause of and levels of damage from sewer backups using econometric models were found in the literature search for relevant studies. Related damage claim studies were relied upon for shaping the research design of this thesis.

Yoder (2002) examines wildlife-imposed crop damage claims filed by Wisconsin farmers using crop dummy, wildlife management and land size variables among others in a truncated regression model. Yoder (2002) does not analyze spatial effects for his study, though this project uses a truncated statewide sample. The single most valuable and inexpensive addition would be data on undamaged fields managed by the claimant, in addition to the data on damaged fields already collected (Yoder, 2002). This thesis uses data from parcels that did not submit claims in addition to parcels that submitted claims to predict conditional sewer backup probabilities and forecast potential damage on the non-claim parcels.

Van Tassel et al. (2000) examine the depredation claim process in Wyoming using a probit model based on a utility theory where landowner utility is a function of management and personal characteristics as well as depredation conditions. When various forms of wildlife cause damage to landowner property, the landowner can file a depredation claim and the state will reimburse the landowner for the amount of damage. The state of Wyoming has the program because landowners may be prone to retaliate against the wildlife in absence of the program.

While this project has some similar characteristics, much of this project is focused on determining whether or not a landowner will make a claim and assessing attitudes towards different kinds of wildlife. Claimant characteristics may be of interest for those involved in the sewer backup claim process but the particular specification Van Tassel et al. (2000) use is not applicable to this project.

Lubini and Fuamba (2011) model the deterioration timeline of sewers using linear and exponential regression models as well as neural networks and fuzzy set theory. Structural aspects of the sewer system including age, diameter, length, slope and material were included in the regression models. While measuring deterioration is not directly applicable to measuring aspects of sewer backup claims, this is one of the few examples of a study that uses regression models to measure aspects of a sewer system.

Hall et al. (2007) measure flood risk using hydrological models by analyzing the variance of parameters including pipe size, permeable areas and river bottom width. Their expected attributable damage function is a function of rainfall as well as loading and resistance (dike capabilities). Their measures of risk reinforced the use of explanatory variables and added conceptual ideas to this project related to risk such as risk being a common currency and attributing risk based on the capacity to reduce risk.

Spekkers et al. (2012) used rainfall intensities from 10 minutes to 4 hours to investigate whether high numbers of water-related damage claims from a Netherlands insurance database were associated with high rainfall intensities using logistic regression. They found that rainfall intensities were a significant damage predictor but much of the variance in damage amounts remained unexplained.

Zhou et al. (2011) use GIS to measure flood vulnerability and risk to calculate an Expected Annual Damage amount. With a given extreme external loading, the key principle is to assess and quantify the hazard and vulnerability characteristics of an area and then link both hazard and vulnerability information in a GIS-based risk model (Zhou et al., 2011). This thesis will apply

similar methods in terms of measuring sewer backup hazard and vulnerability as hazard and vulnerability are conceptually synonymous with probability and potential damage[1].

Spatial Econometric Model Applications

Spatial dependence reflects a situation where values, observed at one location or region, depend on the values of neighboring observations at nearby locations (LeSage and Pace, 2009). Many of the claims in the dataset were clustered and that suggests that many claims may be caused by the same problem within the system. *Introduction To Spatial Econometrics* by LeSage and Pace (2009) serves as a useful reference for implementing and interpreting spatial models and was referred to often during this project.

Two potential sources of spatial dependence can cause econometric problems: structural spatial dependencies across observations on the dependent variable and spatial dependence among the error terms (Bell and Bockstael, 2000). The former can be corrected for using a spatial lag model while the latter can be corrected for using a spatial error model. When a spatial relationship results from a biophysical process (e.g. water flowing across a landscape) or because of behavior of neighbors, there exists spatial lag dependence (Pattanayak and Butry, 2002). Due to the nature of the sewer system and the clustering of the claims, the spatial econometric specifications should be able to adequately account and adjust for spatial dependence accordingly.

---

[1] The MIKE URBAN hydrologic model that Zhou et al. (2011) use was investigated but cannot be applied to sewer backups as it is not setup to measure the proper dynamics of a sewer backup occurrence. Perhaps the volume of flow in the sewer can be measured within the model and used as a structural variable but this thesis project did not pursue the use of this model.

Correcting for spatial autocorrelation can also add to the goodness-of-fit of the model. Griffith and Layne (1999) report that exploiting [autocorrelation] tends to increase the R-squared value by about 5%, and obtaining 5% additional explanatory power in this way is much easier and more reliably available than getting it from collecting and cleaning additional data or from using different statistical methods. For this project, the goodness of fit for the OLS model used to determine which variables should be used in the spatial econometric model was low (~ 0.2). Applying the corrections for spatial autocorrelation was helpful in this project.

Spatial econometric models can be used in the estimation and inference regarding parameters, prediction, and out-of-sample forecasting as well as model comparison of alternative specifications to answer the question regarding observed spatial dependence in dependent variables from our models and residuals (LeSage and Pace, 2009). While the lack of studies that have applied econometric modeling to sewer backup or other storm-induced claims, it is believed that an OLS model will explain some of the variance of sewer backup claims and, if spatial autocorrelation is present, a spatial econometric model will correct for spatial autocorrelation and produce more robust results that can be used to attribute damage caused by other storms via out-of-sample forecasting.

Rare Events Logistic Regression Model Applications

King and Zeng (2001) wrote the explanation of the rare events logistic regression model specification and Tomz et al. (2003) wrote the code for use in Stata and Gauss. Imai et al. (2007) wrote the code for use in R. King's website describes the rare events specification as such: "How to save 99% of your data collection costs; bias corrections for logistic regression in estimating probabilities and causal effects in rare events data; estimating base probabilities or any quantity

from case-control data; automated coding of events (King, 2013).” These authors provide many applications and evidence for effective bias corrections.

King and Zeng applied the rare events logistic regression model specification to international relations including examining causes of interstate conflicts (2001a) as an application of the methods in King and Zeng (2001) and predicting state failure (2001b) using case-control data. Their methods in King and Zeng (2001a) allow them to examine interstate conflicts using conflicts that already occurred and a sample of dyadic (two country) relationships that did not result in conflicts. With all of the countries in the world, modeling relationships with every country and every other country created a large amount of dyadic relationships. In King and Zeng (2001b), different status measures including infant mortality, trade openness and level of democracy.

There are other applications of rare events logistic regression models besides international relations and many of them involve natural resource applications. Many studies examine factors that cause landslides in forests and greenspaces. These studies use LIDAR images and GIS data to examine the areas. Qualitative data seems to be prevalent in these studies. A good example of this is Guns and Vanacker (2012), who use Monte Carlo simulation to replicate landslide conditions in the Andes. These methods are helpful for examining characteristics that cause landslides in places that cannot be directly surveyed but the remote data exist such as slope, proximity to watercourses and land use changes.

Another study related to natural resources is Vospernick (2006), who uses Austrian National Forest Inventory data to examine which tree species red deer prefer to strip bark (with their antlers) in Austrian forests. This is another example of remotely examining a problem that is

occurring to provide an explanation for the problem. Vospernick (2006) use the weighted correction specification.

Many of the above studies had rare events but were not as rare as this study's rare events in terms of available ones (from the dependent variable perspective). King and Zeng (2001) define a rare event as an event with an occurrence rate under 5%. King and Zeng (2001b) explain that the rare events assumption states that τ (tau, the true frequency of the rare event) is arbitrarily small [while P(X) stays bounded away from zero, or instead that $\Pr(Y_i = 1 | X_j) \rightarrow 0$ for j = [0, l]. This assumption is not merely that cases are ''rare,'' but that they occur, at the limit, with zero probability (King and Zeng, 2001b).[2]

Such conditions allow for the use of risk ratios, odds ratios and first differences to analyze how changes in variable amounts affect changes in probabilities that the rare events models project. These methods are especially useful for estimating susceptibility to diseases and viruses by examining those characteristics. King and Zeng (2002) explore how these methods can be implemented in case control studies and such studies apply to the medical field.

Braitman and Rosenbaum (2002) look at how these ratios apply to the prediction of rare outcomes. In the medical field, a disease with a small probability of occurrence still has the potential to affect a large portion of the population when the probabilities apply on a continental or global scale. The way these problems are examined aided in the application of first differences and risk ratios in this thesis. The conclusions drawn from calculated probabilities and changes in the variables that induce estimated changes in probabilities should be closely analyzed for bias and context and such considerations were taken for this project.

---

[2] $\Pr(Y_i = 1)$ is the probability that the dependent variable, Y, equals one and X_j (where j = [0, l]) represents the change from one value of $X(0)$, an explanatory variable, to another (l).

The rare events assumption is logical when applied to sewer backups. Not all rainfall events produce sewer backups but, when rainfall events occur with higher rainfall intensity and volume than normal, backups can occur. Even the largest events still bring about a small amount of claims and the absolute probability of a backup occurring is still very low for most areas within the City of Seattle, rendering sewer backup occurrences rare events.

Finally, spatial effects for the use of rare events logistic regression model in this project were not examined beyond a Moran's I test. Robertson et al. (2009) investigated the influence of spatial effects and how to correct for it. They state that while spatial effects might be expected in the consideration of land use decisions as well as other discrete choice settings, how to go about specifying such effects in practice has been challenging (Robertson et al, 2009) The most prevalent workaround in the literature is the idea of reducing the data set by systematic geographic subsampling in hopes of avoiding spatial effects (Robertson et al, 2009).

The random stratified sample used in this project should eliminate significant clustering. The Moran's I test for spatial clustering on the sample parcels and the largest claim-producing storm allayed any concerns about spatial effects.

Pacific Northwest Climate Research

Since this thesis examines damage due to extreme rainfall amounts, it is worth examining future expectations for the frequency of extreme rainfall events. The design of stormwater infrastructure is based on an underlying assumption that the probability distribution of precipitation extremes is statistically stationary (Rosenberg et al., 2010). If the rainfall frequency changes, this may present capacity concerns for the combined portions of the sewer system in Seattle.

Forecasts of global climate change are largely driven by global climate circulation models. Global models do not represent local terrain and mesoscale weather systems well, owing to their coarse horizontal resolution (150– 300 km, Dulière et al., 2008). Dulière et al. (2008) examined whether or not regional climate models can represent the intensities and frequencies of extreme events in the Pacific Northwest at the scale important for climate impacts assessment.

They found that extreme precipitation can be adequately simulated in regional models given boundary conditions from the reanalysis. The large-scale conditions that control the spatial distribution of heavy precipitation are well represented by the reanalysis, and the regional models can simulate the local effects (such as orographic enhancement and mesoscale weather patterns) that produce heavy precipitation (Dulière et al., 2008).

Having these capabilities allows for the results of regional simulations to inform decision makers about the expectations of claim-producing storms occurring in the future. Both simulations in the Salathé Jr. et al. (2010) yielded an increase in the measures of extreme precipitation even though one simulation produced mostly reductions in total precipitation during winter and spring. Consistent with previous findings, these results suggest that extreme precipitation changes are more related to increased moisture availability in a warmer climate than to increases in climate-mean precipitation (Salathé Jr. et al., 2010).

Rosenberg et al. (2010) examined historical precipitation levels and simulations of future rainfall to evaluate past and prospective changes in the probability distributions of precipitation extremes across Washington State. In the Puget Sound region, statistically significant increases in annual maxima were observed at the 24 hour duration, which is the interval most frequently used for the design of stormwater infrastructure (Rosenberg et al., 2010). This suggests that the sewer system

in Seattle has in fact received increased inputs into the system. The system may already be under capacity to handle current rainfall inputs and may further be unable to handle higher rainfall intensities and amounts in the future.

Zhu (2012) examined historical and future scenario model runs at many different time intervals with updated information and calculated a 13% increase in rainfall intensity. Seattle was the only region (of the six US regions studied) with the average AF (Adjustment Factor, the difference between historical and future scenario model runs) being greater than one and its significance level being small (<0.02, Zhu, 2012). Rainfall intensity differences from the simulations Zhu (2012) used were significant at all intervals from two years to 100 years.

While a comparison between the HH (Hadley RCM and GCM used in this study) historical runs and the future scenario runs generally indicated a greater increase in magnitude at longer return periods, the percentage increase was similar for most return periods (Zhu, 2012). Such forecasts suggest that longer, more intense storms can be expected at all different return periods. This also implies that the return periods established may not be valid going forward into the future.

The timing and the importance of this research suggests that there will be more studies on future projections of Pacific Northwest rainfall intensity at the regional level. More knowledge on this subject may better inform decisions on the implementation of system upsizing and measures to prevent or mitigate issues resulting from high intensity rainfall events.

Chapter 3: Methodology

For the methodology section, the research questions will be stated, the data collection and processing methods will be explained, the dependent and explanatory variables will be explained, the specification of the models will be explained and the process that produced the model results will be explained.

These results can be calculated for each sample parcel and these parcels can be converted to raster files.[3] They can be used to extrapolate the values of explanatory variables as well as extrapolate sample measures of risk across the city of Seattle to attribute risk for the entire city and can be displayed on maps to be used as risk analysis tools.

Research Questions

Here are the research questions that this project sought to answer:

1. What meteorological, environmental, demographic and structural variables explain the cause of sewer backups that are settled with a positive claim being paid out by the City of Seattle?
2. What meteorological, environmental, demographic and structural variables explain the amounts of settled positive sewer backup claims?
3. If these variables explain the cause of sewer backups and amount of sewer backup claims, what is the expected damage of sewer backups at the parcel level and the neighborhood level in Seattle given the characteristics of prominent rain storms that occurred within the timeframe of the data set?

[3] Raster files are collections of pixels that represent different levels of spatial information.

Data Collection and Processing Methods

The claim dataset was procured from the claims manager at Seattle Public Utilities. This dataset was exported into a spreadsheet from a shapefile in GIS. All names and addresses were removed from the dataset and the claims were identified by the remaining characteristics in the database file, mostly from the claim number. The dataset contained sewer and flood claims and the flood claims as well as non-sewer backup claims were removed so that only definitive sewer backup claims (as defined by the brief description in the claims shapefile) with positive payout amounts remained.

The sewer backup claim dataset contained 459 positive sewer backup claims that were due to sewer backup damage that occurred between September 2004 and May 2011. While citizens have three years after the date of loss to file a claim with the city (Martin, 2012), the list contained no new claims within the aforementioned timeframe when updates were sought.

The meteorological data were gathered from the City of Seattle's rain gauges. Data were exported from the Hydstra (2012, the rain gauge database where rainfall information is imported) for the periods of the claim data set. The data were exported to spreadsheets at the following intervals available: 5 minute, 10 minute, hour and day. For the damage model, Thiessen polygons[4] were generated for the existing rain gauges in ArcGIS for every year in the dataset. These polygons were joined to the claims using the Identity tool[5]. The rain gauge domains were joined with the claims and imported to a spreadsheet. Special case Thiessen polygons were generated when a rain gauge generated bad results or was out of commission by omitting the rain

[4] Polygons that represent the spatial coverage of rain gauges given location of other rain gauges
[5] The Identity tool transfers information from one shapefile that is overlaid on another shapefile

gauge with erroneous results. The rain gauge domains were used to manually enter the rain gauge information into the claims spreadsheet.

The sole environmental variable was the tree density variable, though other variables such as soil corrosivity and soil type were considered but neither variables seemed to be applicable to the study. Soil type could be a good variable but Seattle is an urban area and much of the soil has been altered and replaced over the years it has been developed.

The tree density was calculated through the use of a raster file obtained from the 1-meter LIDAR and Imagery Land Use Land Change (LULC) raster dataset developed by Dr. Monika Moskal's Remote Sensing and Geospatial Analysis lab at the University of Washington. The raster was a binary dataset where a pixel is deemed as either containing a tree or trees or not containing a tree. The raster was converted to a polygon and then reconverted to a raster format (the original file was in TIFF format). Then, the raster values were aggregated to different sizes so that the values were percentages that represented tree densities. The different raster files were compared against an aerial image of Seattle and the chosen raster represented the tree density well and had a pixel size that was large enough to adequately represent the scale of tree density for an individual household and its connection to the system.

The structural characteristic data (sewer pipe data) were gathered from Drainage and Wastewater mainline shapefile in GIS. For the damage model, this data were manually attached to the claims by using the side sewer and laterals layer in GIS to see where the building or parcel connects to the system. The unique ID number of the sewer pipe was entered in the claim table so that the structural characteristics could be joined to the claim dataset.

For the probability model, the claims were added to a sample of non-claim parcels where the structural variables were joined using the Spatial Join tool in ArcGIS. This tool joins the parcel to the nearest pipe from the randomly generated point. Though this may not be as accurate as manually joining the claim to a sewer pipe, the manual process would be too time-consuming and most of the variables do not vary much from their connected neighbors due to the nature of sewer infrastructure installation.

The demographic variables consist of restaurant, household and population densities as well as parcel values. Restaurant density data, used as a proxy for potential grease in the sewer system, were collected by Seattle Public Utilities' F.O.G. (Fats, Oils and Greases) restaurant inspection program. This data were in raster format and were attributed to parcels and claim points in ArcGIS.

Parcel values and lot sizes were obtained from spreadsheets at the King County Assessor's Office website and joined to a parcel shapefile obtained from King County GIS Data Center (2013). Parcels that did not have values were manually obtained from the King County Assessor's Office.

Household and population data were obtained from King County GIS Data Center (2013) and were calculated from the 2010 Census and were conflated to census blocks by the King County GIS Data Center. The data were exported from the Data Center and imported into ArcGIS and joined with a census block layer (also obtained from the King County GIS Data Center). The joined data were then converted to raster files using the Feature To Raster tool in ArcGIS. This allows for more of a robust sampling of household and population densities than a direct transfer of the information using the Identity tool in ArcGIS.

Dependent and Explanatory Variables

Many of the explanatory variables were used in both models while others could not be used in both models or were the same variables but were attributed at different spatial resolutions. While this is a long list of variables, many were eliminated due to their insignificance in trial runs conducted to find the best-fitting models. Non-rainfall summary statistics for the rare events logistic regression dataset are in Table 1 and the spatial econometric model dataset summary statistics are in Table 2. The rainfall summary statistics for the rare events logistic regression are in Figure 2 in Chapter 2 and on Figure 4 later in this chapter.

Variable	Description	Unit	Mean	Std. Dev.	Min	Max	Type
Claim_Y	Claim Variable (1 = claim)	Binary	0.040	0.197	0	1	Binary
grid_pct	Tree density	Percentage	0.285	0.245	0	1	Percentage
GRIDCODE	Restaurant density	Density Level	41.165	35.714	0	199	Categorical
ln07TOTVAL	Log 2007 parcel value	Dollars	13.640	1.478	6.215	20.729	Continuous
PPHH	People per household	-	2.060	0.407	1.082	3.179	Continuous
lnPPSQMI	Log population per sq. mi.	-	9.182	0.611	6.178	10.668	Continuous
lnHHSQMI	Log households per sq. mi.	-	8.480	0.711	5.279	10.414	Continuous
AGE	Pipe age	Years	70.725	26.672	0	118	Categorical
DEPTH	Pipe depth	Feet	11.594	5.344	-41	77.55	Categorical
ELEV	Pipe elevation	Feet	186.117	120.788	-19.055	494.95	Categorical
MNL_LENGTH	Pipe length	Inches	242.814	207.035	2.32	8740.96	Categorical
MNL_WIDTH_	Pipe width	Inches	12.681	14.129	0	150	Categorical
MNL_HEIGHT	Pipe height	Inches	12.710	13.947	0	144	Categorical
MNL_SLOPE_	Pipe slope	-	3.610	5.741	0	80.8	Continuous
CLAYD	Clay pipe dummy variable	0 or 1	0.395	0.489	0	1	Binary
CONCRETED	Concrete pipe dummy variable	0 or 1	0.532	0.499	0	1	Binary
PROBSAND	Sanitary flow dummy variable	0 or 1	0.374	0.455	0	1	Binary
PROBCOMD	Combined flow dummy variable	0 or 1	0.626	0.500	0	1	Binary

Table 1: Summary statistics for the non-rainfall variables used in the rare events logistic regression models (Claim_Y statistics are in reference to the December 14th, 2006 storm)

Variable	Description	Unit	Mean	Std. Dev.	Min.	Max.	Type
Amount_Pai	Claim amount	dollars	$16,519	$24,914	$50	$195,000	Continuous
LN_AMT	Log claim amt.	log dollars	8.70	1.57	3.89	12.18	Continuous
PROBSAND	Sanitary dummy variable	0 or 1	0.23	0.42	0	1	Binary
PROBCOMD	Combined dummy variable	0 or 1	0.77	0.42	0	1	Binary
CLAYD	Clay dummy variable	0 or 1	0.54	0.50	0	1	Binary
CONCRETED	Concrete dummy variable	0 or 1	0.41	0.49	0	1	Binary
MNL_FCNLEN	Pipe length	feet	302.91	98.59	5.58	845.20	Continuous
WIDTH	Pipe width	inches	11.54	8.22	6.00	54.00	Discrete
SLOPE	Pipe slope	unitless	2.40	3.42	0	47.90	Continuous
DEPTH	Pipe depth	feet	11.62	3.41	0	25.81	Continuous
ELEV	Pipe elevation	feet	184.04	112.45	3.91	466.95	Continuous
CCTVYN	CCTV dummy variable	0 or 1	0.05	0.21	0	1	Binary
AGE	Pipe age	years	76.54	23.82	11.00	115.00	Discrete
RINT	Rainfall intensity on date of loss (DOL)	inches/hour	0.10	0.07	0.00	0.23	Continuous
RDOL	Rainfall amount on DOL	inches	1.34	1.30	0.00	5.33	Continuous
RDUR	Amount of hours where rainfall was recorded on DOL	hours	10.56	5.96	0	24	Continuous
L3	Prev. 3 day rainfall	inches	1.22	0.81	0	2.64	Continuous
N15	Prev. 4 to 18 day rainfall	inches	1.45	1.24	0	7.19	Continuous
SSAT	Soil saturation	percentage	0.44	0.24	0	1.37	Continuous
GREASE_COD	Restaurant density	density units	73.67	43.03	5	194	Discrete
grid_pct	Tree density	percentage	0.27	0.20	0	0.94	Discrete
lnHH	Log household density	log HH/sqmi.	8.50	0.53	5.90	10.10	Continuous
lnPOP	Log population density	log pop/sqmi.	9.22	0.44	6.55	10.43	Continuous
sqftlot	parcel area	square feet	19484	54341	873	344323	Continuous
lnYOLVAL	Log property value in year of loss	log dollars	13.29	1.22	11.29	17.94	Continuous

Table 2: Spatial econometric model variable summary statistics

Dependent Variables

The damage model uses log claim amounts because the log distribution is more normal distributed than the actual dollar values. The dependent variable for the probability model is a binary variable and zeros were attached to the sample parcels and ones were attached to the claim parcels for the specific storm events. The sample parcels will be discussed later in the chapter.

The three highest claim-producing storms were the May 27th, 2006 storm, the December 14th, 2006 storm and the December 3rd, 2007 storm. These storms produced 32, 147 and 48 positive sewer backup claims respectively, which is nearly half of all sewer backup claims in the dataset. These claims are the ones in the binary dependent variable and the zeros are a sample of parcels where claims did not occur and the latter will be discussed later in this chapter.

Meteorological Variables

For the damage model, the sewer backup claims were due to dozens of storms that occurred in a period of seven years, so creating raster files for every storm was not feasible. The only information available for the timing of the storm in the dataset was the date of loss. As a result, rainfall intensity was calculated as the amount of rain on the date of loss divided by the number hours of positive recorded rainfall. It is postulated that higher rainfall intensity should correlate with higher damage amounts.

Soil saturation was another postulated explanatory variable for why backups occur and no direct measure was found. However, the USGS (2012) has a measurement that portrays the level of susceptibility to landslide events called the Cumulative Precipitation Threshold (Figure 3). One

of their study areas is Seattle, so the application for this proxy variable should be adequate for measuring soil saturation. The equation for this threshold is shown in Equation 1:

(1) $$\text{(3-day precip)} = 3.5 - 0.67*\text{(previous 15-day precip)}$$

The precipitation values in Equation 1 are the previous rainfall recorded in the last three days and the fifteen days before the previous three days of a given examined time respectively.

As one can see from the graph, this threshold is used to predict when conditions are such that landslides are likely to occur. This threshold allows for previous four to eighteen day rainfall amounts and previous three day rainfall amounts to influence landslide susceptibility at different levels. More recent rainfall (previous three day rainfall) is weighted higher than rainfall that occurred in the more distant past (previous four to eighteen day rainfall). Though this study is not concerned with landslides, the threshold can be used as an indicator of soil moisture conditions. The threshold levels can be measured as an explanatory variable for malfunctions in the sewer system.

The measurement would ideally be the cumulative density based on where the precipitation values are on the x and y axes and the resulting values. However, some of the dataset contains claims where no rainfall was recorded in either the three previous days before the date of loss or the previous fifteen days before the three day period. The probability model uses the cumulative density since the storm events examined had rainfall in both periods and this density (the triangular area created by positive previous three day and previous four to eighteen day rainfall amounts divided by the area of the "landslides unlikely" triangle, 9.1875) is measured using Equation 2:

(2) $$[0.5 * (15\ day\ cumulative\ precip) * (3\ day\ cumulative\ precip)]/9.1875$$

For the damage model, the higher number between the two threshold percentages (the three-day rainfall period divided by 3.5 inches or the fifteen-day rainfall divided by 5.25 inches) will be the soil saturation measure. This does allow for a percentage above the threshold in cases where such rainfall previous to the sewer backup claim occurred. In both calculations, higher soil saturation levels should translate to higher damage amounts as, the closer the level gets to landslide danger, the less room there is for water in the soil.

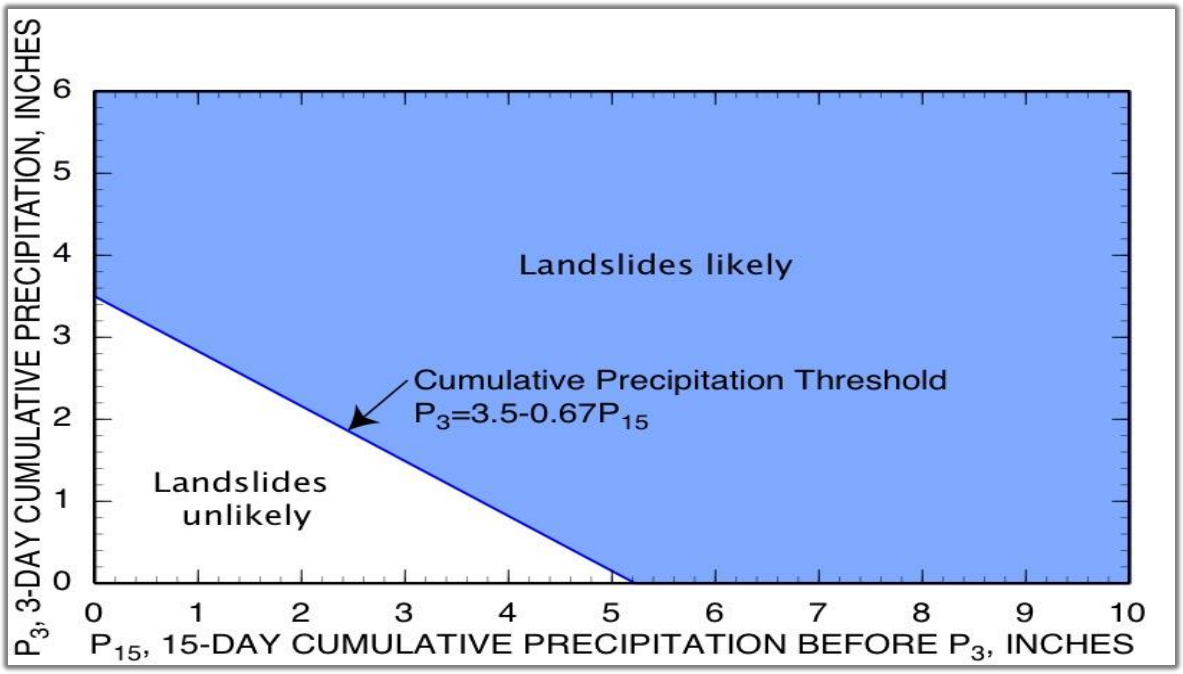


Figure 3: Cumulative Precipitation Threshold graph (Courtesy of USGS)

It may also be possible that soil saturation has no effect on the damage amounts at all. The sewer system is not completely sealed off from groundwater leakage and soils saturated at the depth level of the pipes may add to the flow in high levels of rainfall. Soil saturation is a variable worth measuring given the subterranean location of the system and the tendency for soils to be saturated during the period of the year when heavy rain storms occur. The soil saturation

densities for the three major storms are displayed in Figure 4. They will apply to the analysis and the results in Chapter 4.

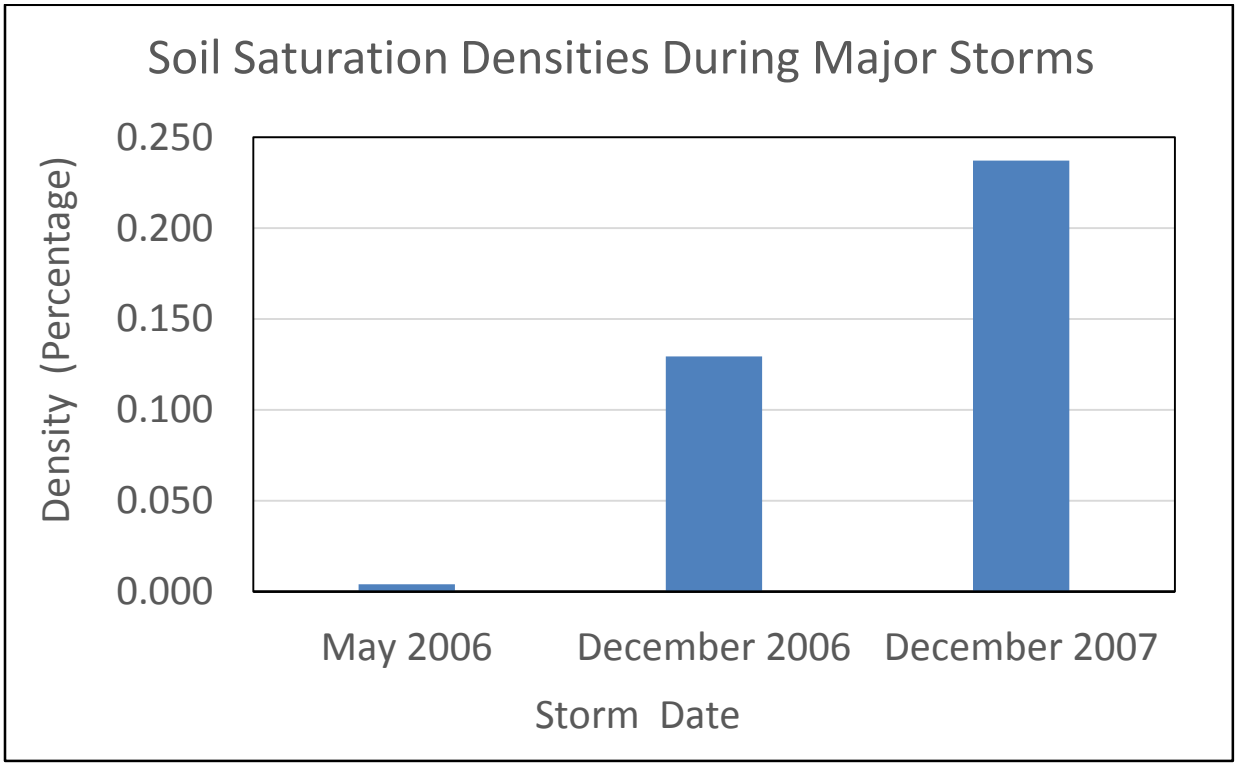


Figure 4: Soil saturation densities for the three major storms

Rainfall variables are specific to a specific rainfall event so the production of rainfall rasters is justified. For the probability model, rasters were created for the following variables using the Inverse Distance Weighting tool in ArcGIS: peak five-minute rainfall, peak ten-minute rainfall, peak hour rainfall, peak three hour rainfall, rainfall intensity in the period when the storm occurred, soil saturation (same specification of the damage model), previous 1, 2 and 3 day rainfall as well as the previous 15-day rainfall period before three day interval previous to the storm occurred (the x-axis component on the Cumulative Precipitation Threshold graph). The rainfall intensity period was established as the beginning to the end of the recorded rainfall

period where the peaks were recorded and was calculated by dividing the amount of rain recorded during the period by the number of hours in the period.

The rainfall variables in the damage model were, in addition to soil saturation and rainfall intensity, the previous three days of rainfall (as a percentage of the threshold), the next fifteen days previous to the three-day period (also as a percentage of the threshold) and rainfall duration on the date of loss and total rainfall on date of loss.

Tree Density Variable

The densities were measures of tree cover between 0 and 1, with 1 being all trees and 0 being no trees. Vegetation can have much influence on the system, directly and indirectly. Directly, trees can sprout roots that will work their way into the system through cracks or by biological processes. Once inside the pipe, the roots expand and take up space in system, lowering the capacity of flow and collecting material such as grease, floating solids and trash in the system.

Indirectly, trees can delay stormwater from entering the system, which can help the system conduct rain water more steadily and not create the pressure and capacity issues that cause backups. Trees also retain water that falls through their leaves, which can reduce the amount of rain water in the system.

Naturally, the direct and indirect effects counteract one another. Roots in the system may cause more backups than if the roots were not present. Vegetation may reduce the amount of backups by reducing peak flows. The sign of the coefficient for this variable should indicate which effect has a stronger influence on the system.

Structural Variables

Since the structural data, also known as sewer pipe data, were attached to the claims manually, every pipe contained descriptive information in the shapefile including its age (in years), length (feet), width (diameter, in inches), material, slope, elevation (feet), depth (feet) and probable flow in the system. Probable flow is defined by what is supposed to be flowing in the system; wastewater flow (sewage only) is called sanitary flow and sewage and stormwater which is called combined flow.

The types of pipe materials were vitrified clay, reinforced concrete, concrete, brick, asbestos cement, PVC and ductile iron, though brick and both concrete materials made up the vast majority (~ 93%) of the pipes, so those materials were the only ones examined. A dummy variable was created for clay and for concrete (concrete and reinforced concrete pipe were coupled for the concrete variable).

Average elevation was measured by adding the upstream elevation and the downstream elevation and dividing the sum by two. Average depth was computed in the same way as elevation. Probable flow was a dummy variable that can either be sanitary or combined (sanitary and drainage water). Age was calculated for the year that the backup occurred.

The sign of the age coefficient is difficult to project. Age can be a measure of deterioration and can assumed to have a positive correlation with probability and damage from backups but the condition of the pipe may not be correlated with backup occurrence and damage. The sign of pipe elevation should be negative since, as elevation decreases, higher volumes of flow will be entering the system. As elevation increases, there is less land area that has rainfall on it, draining into the system.

Pipe length should have a positive coefficient in both models due to the nature of surveillance costs. The longer the pipe, the more costly it is to inspect a pipe since many CCTV (closed-circuit television recordings via remote cameras) surveillance contracts are charged by the foot of pipe inspected. Not all pipes in the system are inspected so length may have no effect on claims attached to pipes that would not be inspected no matter what the length. Depth is nearly the same case as length. The further the pipe is below the ground, the more expensive it is to replace or repair.

The slope coefficient should be negative since higher slopes should equate to higher velocities of water running through the pipe and a lower propensity for the water in the pipe to stagnate and be a part of backup flow. The width coefficient should be negative as well. The greater the pipe width, the more capacity there is in the pipe, though a better measure might be a measure that compares the width to the volume of flow to the relative volume of flow the pipe is expected to conduct but that data is not available.

The material variables, clay and concrete, do not have a clear postulated sign. Clay pipes are more resilient to deterioration than concrete pipes (Martin, 2012) but that refers to their propensity to deteriorate, not whether or not they will contribute to backup cause or claim amounts. The probable flow variables should have different signs. The combined flow dummy variable, with 1 being combined and 0 being sanitary, should be positive since combined pipes carry waste water and storm water instead of just waste water. The sanitary flow dummy variable should then be negative since it is the opposite of the combined variable. These variables could be important if they prove to be significant as combined pipes would then have more of a priority for maintenance and surveillance.

Demographic Variables

Restaurant density was transferred to all the variables in ArcGIS to be used as a proxy for grease in the system. The density units ranged from 0 (lowest density) to 199 (highest density). The density data were collected in 2011, so the further the claims are behind this date, the less accurate of a portrayal of restaurant density the variable is. Without the raster, there really is not a good measure of how much grease is in the system. The author requested CCTV reports for pipes that were attached to sewer backup claims and were identified as having heavy amounts of grease after being surveyed but only nine pipes fit those criteria.

The household and population densities are logged estimates of households per square mile and population per square mile, respectively. They were converted to log form to better represent normally distributed variables. The parcel values are the log total values of the parcel, which consists of the addition of appraised land value and appraised improvement value for a parcel. Those values appear in the probability model. The damage model utilizes the value per square feet, which is the log value of the parcel divided by the log square foot of the parcel, and the log square footage of the parcel.

The sign of the restaurant density coefficient should be positive as higher restaurant densities should mean more contributions of grease into the system. Though laws and programs are in place to prevent grease from going into the system, the density can act as some kind of proxy for the past state when grease into the system was not regulated. In addition, a large amount of restaurants in an area will invariably translate into some level of grease into the system. The significance level should be a test as to the strength of the proxy.

The signs of population and household densities should be positive since higher levels of population and households should translate into larger contributions of waste water flow into the system. For the parcels, higher parcel values should correlate into higher amounts of possession and home values. Hence, for the probability and the damage model, the parcel value coefficient should be positive.

The sign of value per square feet has no intuitive sign to be postulated. Since the distribution of parcel values and parcel claims for the entire city resembled exponential distributions, the log 2007 parcel value was divided by the log square footage of the parcel. In these conversions, different scales of parcel sizes (single-family, apartment buildings, commercial buildings and museums among others) and parcels should be rendered equal and the measure itself a unitless ratio.

Lastly, for the damage model, log square footage should be a positive coefficient since the amount of basement area and bathrooms should correlate with the amount of square footage the parcel takes up. This presumption makes more sense in an urban area where residential and commercial buildings take up much of the parcel area.

Spatial Econometric Model Methodology

The spatial damage model analysis was run using R 2.12.0 and the 'spdep', 'rgdal' and 'maptools' packages. Since the data are hypothesized to be spatially autocorrelated and heteroskedastic, a spatial weights matrix was created to test for the presence of spatial autocorrelation and heteroskedasticity.

The spatial weights matrix was row standardized (the sum of every row in the matrix equals 1) with the neighbors defined as the k amount of neighbors closest to a certain point with $k = 15$.

Spatial autocorrelation was tested for using Moran's I test (Moran, 1950) and heteroskedasticity was tested for using the Breusch-Pagan test (Breusch and Pagan, 1979).

The Moran's I statistic is defined by ESRI (2012) as:

(3) $$I = \frac{n}{S_0}\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,j} z_i z_j}{\sum_{i=1}^{n} z_i^2}$$

where z_i is the deviation of an attribute for feature *i* from its mean $(x_i - X)$, $w_{i,j}$ is the spatial weight between feature i and j, n is equal to the number of features and S_0 is an aggregation of spatial weights:

(4) $$S_0 = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,j}$$

The expectation and variance of Moran's I (ESRI, 2012) are:

(5) $$E[I] = -1/(n-1)$$

(6) $$\mathrm{V}[I] = \mathrm{E}[I^2] - E[I]^2$$

For the Breusch-Pagan test, there exists a linear model represented as such:

(7) $$y_i = x_i'\beta + u_i,$$

where β is a $k\ x\ 1$ vector of coefficient parameters, x_i is a vector of explanatory variables and the disturbances u_i are normally and independently distributed with a mean of zero and a variance:

(8) $$\sigma_i^2 = h(z_i'\alpha),$$

where h $(\cdot)$ is not indexed by t and is assumed to possess first and second derivatives, α is a $p\ x\ 1$ vector of unrestricted parameters functionally unrelated to the β coefficients, and the first element in z_i (the same as x_i) is unity (Breusch and Pagan, 1979). The null (homoscedastic) hypothesis is $H_0 = \alpha_2 = \cdots = \alpha_p$ and the Lagrange Multiplier statistic for testing this hypothesis is one-half of the explained sum of squares in a regression of $g_i = \hat{\sigma}^{-2}\hat{u}_i^2$ on z_i (Breusch and Pagan, 1979). This test statistic is asymptotically distributed as χ^2 with $p - 1$ degrees of freedom under the null hypothesis (Breusch and Pagan, 1979).

First, the claim damage amounts were regressed on all of the explanatory variables in an OLS regression model. Insignificant variables were eliminated and re-added until the best fitting model remained given that only fairly significant variables remained in the model. The residuals of the remaining variables were tested for spatial dependence using Moran's I. When spatial dependence was detected, a Lagrange Multiplier test was conducted to determine which model to use. In this case, the spatial lag model was the obvious choice as can be seen by the Lagrange Multiplier results in the Chapter 4 (Table 5).

The spatial lag model adjusts for spatial autocorrelation and is expressed in matrix form in Equation 9 as seen in LeSage and Pace (2009):

(9) $$Y = X\beta + \rho Wy + \varepsilon,\ \varepsilon \sim N(0, \sigma^2 I_n)$$

W is an $n\ x\ n$ spatial weights matrix and ρ is a scalar parameter that ranges between (0,1) and represents the level of spatial dependence. Y and ε are $n\ x\ 1$ vectors and the latter is a normally distributed error term with a mean of 0 and a variance of σ^2 times an $n\ x\ n$ identity matrix. X is a $n\ x\ k$ matrix where n is the amount of claims and k k is the amount of significant variables

remaining from the OLS trials. Also, β is a $k\ x\ 1$ vector. Equation 9 can be inverted to read as such:

(10) $$Y = (I_n - \rho W)^{-1} X\beta + (I_n - \rho W)^{-1}\varepsilon,$$

where $(I_n - \rho W)^{-1}$ is a geometrically-deteriorating lag function. This model allows for the influence of the nearest neighbors of an asset to be weighted the highest and more distant assets further away to be weighted less the further they are from the asset in question.

The spatial lag model was run with the same variables as the best-fitting OLS model. The spatial error model was also run and, as anticipated from the Lagrange Multiplier results, did not perform as well as the spatial lag model.

Rare Events Logistic Regression Model Methodology

The data used in the rare events logistic regression model are variables associated with claims from the major storms mentioned earlier and a sample of parcels where claims were not filed (and presumably backups did not occur). The model will compare the variables where claims happened and did not happen. The differences will aid in the explanation of why backups occurred where they did in the city.

The usual strategy [for sample data collection] is either random sampling, where all observations (X, Y) are selected at random, or exogenous stratified sampling, which allows Y to be randomly selected within categories defined by X (King and Zeng, 2001). This study uses exogenous stratified sampling where the parcels were randomly selected within each neighborhood using ArcGIS. This will allow for the creation of a citywide sewer backup risk profile by using the sample to forecast probabilities, potential damage and expected sewer backup damage. The

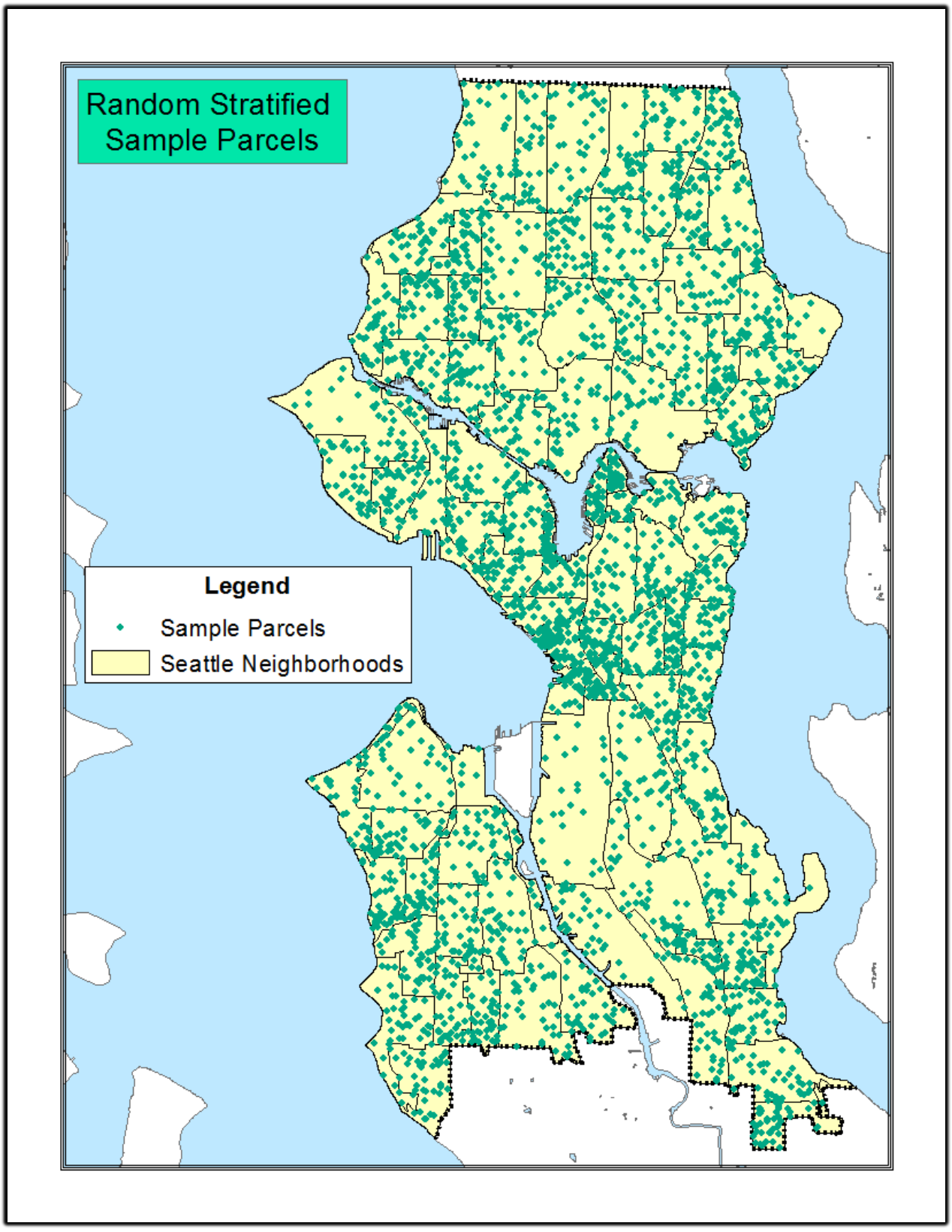


Figure 5: Random stratified parcel sample displayed as points

methods for the calculations are discussed later in this chapter. This random stratified sample of parcels is shown in Figure 5.

There are eighty-nine neighborhoods in Seattle and the first sample collected produced 40 points in each neighborhood using the Create Random Points tool in ArcGIS. The points were then joined to the Seattle parcel layer using the Spatial Join tool in ArcGIS. Some points were joined to the same parcel. When the joined results were exported to Excel, the duplicates were removed. A smaller sample of parcels was collected in the same fashion as the first to supplement the loss of parcels in the first sample. The parcels from the second parcel were added after being sorted by their neighborhood ID so their neighborhood could be identified and sorted by another ID so the selection process would be unbiased.

Once the supplemental parcels were joined to the first sample, a few neighborhoods remained under-populated. This was mainly due to the amount of parcels available in the neighborhood. The final amount of parcels was 3494, though the goal was 3560 (which would have been 40 parcels for each neighborhood). The variables attached to the claim parcels were also attached to the sample parcels in ArcGIS. The claim parcels and the sample parcels were joined.

The binary dependent variable consists of zeros (the sampled non-claim parcels) and ones (the claims related to a storm). The dependent variable was regressed on the available variables in a logit, or logistic regression, model in order to remove insignificant variables. King and Zeng (2001) defines logistic regression where a single variable Y_i ($i = 1,...,n$) follows a Bernoulli probability function (Equation 11) that takes on the value 1 with probability π_i (Equation 12) and 0 with probability of $1 - \pi_i$. Then π_i varies over the observations as an inverse logistic

function of a vector x_i , which includes a constant and $k-1$ explanatory variables (King and Zeng, 2001).

(11) $$Y_i \sim Bernoulli\ (Y_i|\pi_i)$$

(12) $$\pi_i = \frac{1}{1+ e^{-x_i\beta}}$$

Y_i is the binary dependent variable and β is an estimated vector of coefficients. The probability function of the Bernoulli is $L(\beta|y) = {\pi_i}^{Y_i}(1 - \pi_i)^{1-Y_i}$and the unknown parameter to be solved for ($\beta = [\beta_0, {\beta_1}']'$) is $k\ x\ 1$ vector, where β_0 is a scalar constant term and $\beta_1$is a vector with elements corresponding to the explanatory variables (King and Zeng, 2001).

The parameters are estimated by maximum likelihood, with the likelihood function formed by assuming independence over the observations $\prod_{i=1}^{n} {\pi_i}^{Y_i}(1-\pi_i)^{1-Y_i}$ (King and Zeng, 2001). Greene (2012) notes that it is simpler to work with the log of the likelihood function. Thus, the full and simplified log likelihood function is shown in Equation 13:

(13) $$lnL(\beta|y) = \sum_{Y_i=1} \ln(\pi_i) + \sum_{Y_i=0} \ln(1-\pi_i) = -\sum_{i=1}^{n} \ln(1 + e^{(1-2Y_i)x_i\beta})$$

The goal is to find the maximum value of the function ($\hat{\beta}$) and compute the standard errors with the variance matrix, $V(\beta)$. The process of running the logit model, removing insignificant variables and rerunning the model, was repeated until the best-fitting model remained (based on the pseudo R squared value) given that insignificant variables were absent. When that process was complete, the rare events logistic regression specification was applied.

For the logit model and the Rare Events logistic regression model, Stata 12.1 was used along with the 'logit', 'relogit', 'setx' and 'relogitq' packages. The rare events logistic regression model functions much like the logistic regression model but the rare events model uses statistical correction to properly weight the estimates (King and Zeng, 2001). Logistic regression without the rare events correction assumes that the number of zeros and ones are proportional to the population's distribution of zeros and ones.

Rare events logistic regression was used because it reduces the amount of data collection necessary by adjusting the estimates based on the true population of ones and zeros. Researchers can collect all (or all available) ones and a small random sample of zeros and not lose consistency or even much efficiency relative to the full sample, drastically changing the optimal trade-off between more observations and better variables, enabling scholars to focus data collection efforts where they matter most (King and Zeng, 2001).

The two types of corrections are prior correction and weighted correction. Prior correction involves computing the usual logistic regression MLE and correcting the estimates based on prior information about the fraction of ones in the population, τ (or tau), and the observed fraction of ones in the sample (or sampling probability), $\bar{y}$ (King and Zeng, 2001). Since the amount of parcels in Seattle is known, the τ is known and is the number of claims produced by the storm divided by 177,154 (the number of parcels in Seattle minus Harbor Island and parcels outside the bounds of the neighborhoods, which was removed from the analysis due to its industrial composition). The respective τ values for the May 2006, December 2006 and December 2007 storms are 0.000181, 0.00083 and 0.000271. The correction of the intercept $\widehat{\beta_0}$

via prior correction is specified in Equation 14 and this correction is consistent for β_0 (King and Zeng, 2001):

(14) $$\widehat{\beta_0} - ln\left[\left(\frac{1-\tau}{\tau}\right)\left(\frac{\bar{y}}{1-\bar{y}}\right)\right]$$

Prior correction can only be used when τ is known or can be estimated. Prior correction is easy to use but, if the model is misspecified, estimates of both $\widehat{\beta_0}$ and $\widehat{\beta_1}$ are slightly less robust than weighting (King and Zeng, 2001; from Xie and Manski 1988).

Weighted correction involves weighting the data to compensate for differences in the sample ($\bar{y}$) and population (τ) fractions of ones induced by choice-based sampling (King and Zeng, 2001). The weighted log-likelihood is then maximized using Equation 15 (from King and Zeng, 2001):

(15) $$\ln L_w(\beta|y) = \; w_1 \sum_{(Y_1=1)} \ln(\pi_i) + \; w_0 \sum_{(Y_i=0)} \ln(1 - \; \pi_i)$$

$$= \; - \sum_{i=1}^{n} w_i \; \ln(1 + \; e^{(1-2y_i)x_i\beta}),$$

where the weights are $w_1 = \; \tau/\bar{y}$ and $w_0 = (1 - \; \tau)/(1 - \; \bar{y})$, and where

(16) $$w_i = \; w_1 Y_i \; + \; w_0(1 - Y_i)$$

Weighting can outperform prior correction when both a large sample is available and the functional form is misspecified (King and Zeng, 2001; from Xie and Manski, 1988).

The standard errors were computed with the asymptotic variance matrix (King and Zeng, 2001):

(17) $$V(\beta) = \; [\sum_{i=1}^{n} \pi_i(1 - \pi_i)x_i' x_i]^{-1}$$

In rare events data, ones are more informative than zeros, since the part of the matrix affected by rare events is $\pi_i(1-\pi_i)$ and this term will be larger for ones than zeros, thus making the contribution to the variance smaller for ones than zeros (King and Zeng, 2001).

Spatial effects were tested for using Moran's I test in ArcGIS. It is postulated that the random stratified sample would displace the clustered sample of claims from the three major storms per Robertson et al. (2009).

As stated before, the model results were be used to determine a probability of positive sewer backup claim damage occurring given the rainfall variables of a specific storm. In essence, this means that the probabilities are a prediction of probabilities for a past event. While some of the results will explain what caused the backups, it may be more of a portrayal of where the backups occurred. Given the importance of rainfall in the occurrences of sewer backups, inconsistencies in the postulated signs of the coefficients and the results may be explained by where the rainfall occurred and the non-rainfall variables may not directly explain why the backups occurred.

Expected Sewer Backup Damage Calculations and Maps

The results of the rare events logistic regression model produce probabilities that a backup would occur for all of the sample (or non-claim parcels). The probability was calculated in Equation 18 (and is the same equation as Equation 12) where the probability of a backup *i* is a function of the variables in parcel *i* that were used in the rare events logistic regression (RELR) results that included the rainfall characteristics from storm *s*:

$$(18) \qquad (Probablity_i | \ storm\ occurrence) = \frac{1}{1+e^{-x_i\beta^S_{RELR}}} = \frac{e^{x_i\beta^S_{RELR}}}{1+e^{x_i\beta^S_{RELR}}}$$

The spatial econometric model (damage model) provided the framework to project sewer backup damage potential for out-of-sample parcels. The results produce a damage amount for parcel *i* and are a function of the variables in parcel *i* used in the spatial lag model (SLM) and the coefficients from the results of the SLM. Since those results are in natural log form, Euler's number (e) must be used to transform the damage into actual dollars in Equation 19:

$$Damage_i = e^{x_i \beta_{SEM}} \quad (19)$$

The probabilities and damage amounts were multiplied together in Equation 20 to get an expected sewer backup claim damage (ESBD) amounts for each parcel *i*:

$$ESBD_i = E[probablity_i * Damage_i] \quad (20)$$

The probabilities, damage potential amounts and ESBD results from the December 14th, 2006 storm were joined with the parcel layer in ArcGIS. The parcels were converted to points and the IDW tool was used to create citywide sewer backup probability, damage potential and ESBD maps. The ESBD results were transposed onto the Seattle Parcel layer to get a calculation of citywide ESBD from the storm.

First Differences and Risk Ratios

First difference and rate ratio calculations can be used to see changes in probability and changes in magnitudes of probability occur respectively where different variable values can be compared with the results. In order to do this, the data must be averaged or aggregated in some other manner using 'setx' in Stata (rare events software) to make these calculations. Then, two different values of x can be compared to calculate these quantities of interest using 'relogitq.'

First difference values (FD) are defined by Imai et al. (2007) in Equation 21:

(21) $$FD = Prob(Y = 1|x_1, \tau) - Prob(Y = 1|x, \tau)$$

Risk ratios (RR) are defined by Imai et al. (2007) in Equation 22:

(22) $$RR = Prob(Y = 1|x_1, \tau)/Prob(Y = 1|x, \tau)$$

In these cases, x is the true value of x and x_1 is some transformation away from x. The tau value, τ, is the same proportion value of ones to the total population of ones and zeroes (in the binary dependent variable context).

These measures will be applied to significant variables that can be conceivably altered in real-life or variables that can appear in different variations, the latter being the rainfall variables which are specific to a given storm. First difference values show the change in probability of a backup occurring as a result of a change in a variable value with all the other variables being constant (in their mean-transformed state). Risk ratios show how much more likely a backup is to occur given the change in variables with everything else remaining constant.

Chapter 4: Results and Discussion

Spatial Econometric Model Results

The log claim amounts were regressed on the variables in the sewer backup dataset using ordinary least squares (OLS) and the insignificant variables were removed. The results of the best-fitting OLS model (given that all significant variables were retained) are in Table 3.

Variables	Description	Coefficient	Std. Error	z-value	Pr(>|z|)
Intercept	Constant	9.100035	1.353446	6.724	0.000
RINT	Rainfall intensity	4.6776412	1.517188	3.083	0.002
SSAT	Soil saturation	4.149707	0.797756	5.202	0.000
L3PCT	Prev. 3 day rainfall	-2.6119691	0.683532	-3.821	0.000
N15PCT	Prev. 15 day rainfall before L3	-4.1040017	0.619467	-6.625	0.000
grid_pct	Tree density	-0.745461	0.356909	-2.089	0.037
lnHH	Log household density	-0.2926563	0.123201	-2.375	0.018
CCTVYN	Pipe surveyed before date of loss	-0.6837433	0.319952	-2.137	0.033
valsqft	Log parcel value/log parcel sq. ft.	-0.0013507	0.000601	-2.249	0.025
lnsqftlot	Log square feet of parcel	0.2689738	0.073261	3.671	0.000
Residual standard error: 1.41 on 449 degrees of freedom					
Multiple R-squared: 0.213, Adjusted R-squared: 0.1972					
F-statistic: 13.5 on 9 and 449 DF, p-value: < 2.2e-16					

Table 3: OLS results of the best-fitting model with insignificant variables excluded

The best-fitting OLS model had an R-squared value of 0.213 and an adjusted R-squared value of 0.1972. It appears that there are many factors that explain amounts of sewer backup claims which are not accounted for in the model. Accounting for the assumed spatial autocorrelation

would add to the fit of the model if spatial autocorrelation is present as explained in the literature review (Griffith and Layne, 1999).

Moran's I tests and the Bresuch-Pagan test were run to test for spatial autocorrelation in the dependent variable and the residuals of the variables and for heteroskedasticity in the residuals, respectively. Table 4 has the results of the Moran tests and the results of the Breusch-Pagan tests. Spatial autocorrelation is present in the dependent variable and the residuals, though significantly more so in the dependent variable. Slight heteroskedasticity is present in the residuals but, with a p-value of 0.08769, it was not deemed significant enough to reduce the level of heteroskedasticity through the omission of variables or other means.

Data Tested	p-value	Moran I statistic
Dependent Variable	0.00000	0.21927
OLS Residuals	0.02815	0.03175

Data Tested	p-value	Breusch-Pagan statistic
OLS Residuals	0.08769	15.1202

Table 4: Moran's I and Breusch-Pagan test results

The application of a spatial econometric model that adjusts for spatial autocorrelation was necessary. The Lagrange Multiplier test results in Table 5 suggests that the spatial lag model should be utilized as the standard and robust results for the lag model were far more significant than the error model results. This seems intuitive when the level of spatial autocorrelation results

Model	p-value
LMerror	0.0443
LMlag	0.000304
RobustLMerr	0.05978
RobustLMlag	0.000398
Degrees of Freedom: 1	

Table 5: Lagrange Multiplier results

for the dependent variable and the residuals are compared.

Figure 6 shows the Moran scatterplot of the Moran's I test on the dependent variable. The labels on the outliers refer to the neighborhoods where the individual claims resulted from. The Harrison/Denny-Blaine neighborhood is where the largest cluster in the dataset is located so it stands to reason that the outliers would originate from that neighborhood.

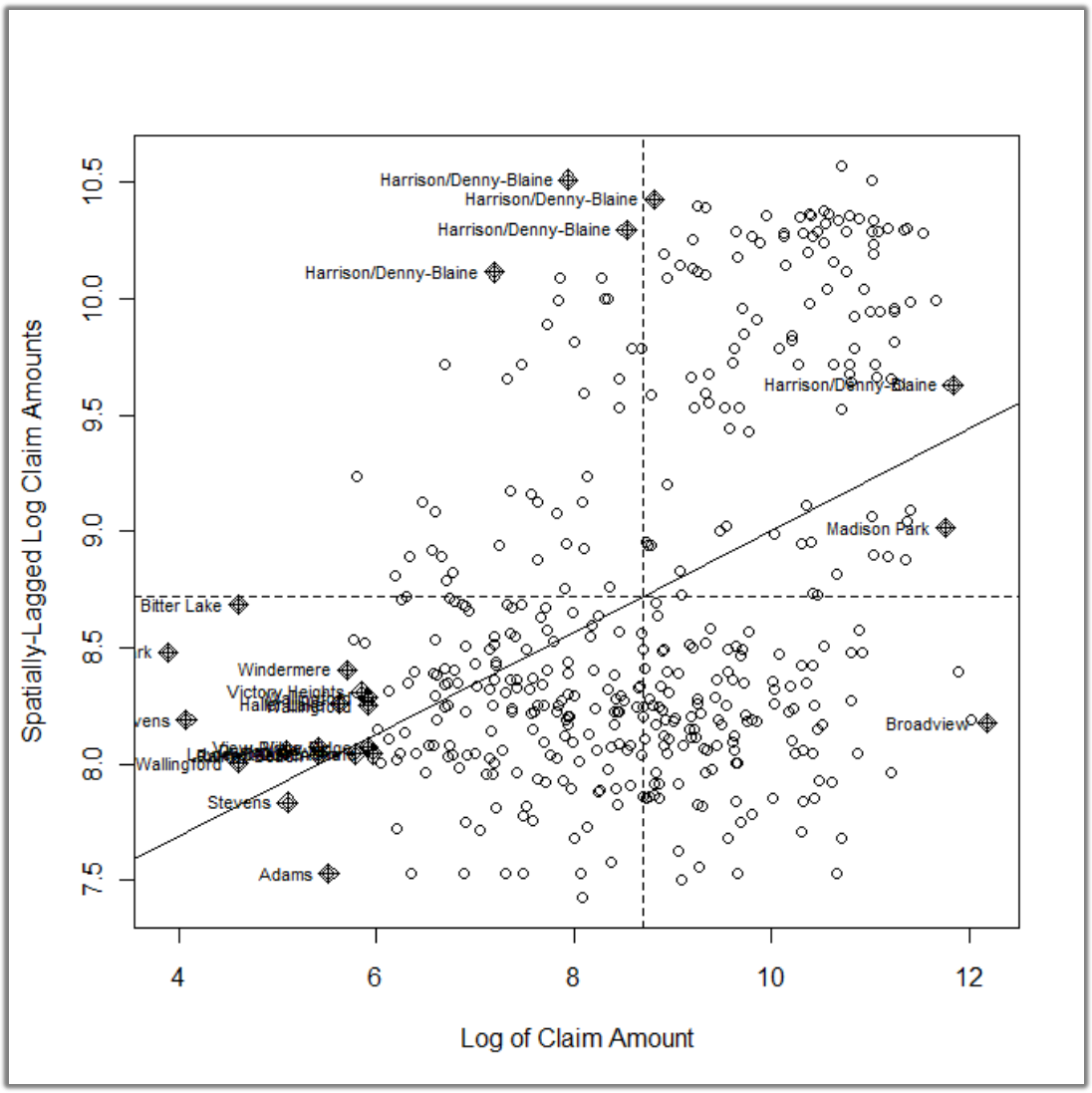


Figure 6: Claim and spatially-lagged claim scatterplot

The spatial lag model specifications were used to regress the log claims amount on the variables from the best-fitting OLS model to start and variables were removed if they were no longer significant in the spatial lag model. The best-fitting model results are the results that were used to calculate potential damage. The best-fitting spatial lag model results are shown in Table 6 and the correlation matrix with these variables is shown in Table 7.

Both the RINT (rainfall intensity) and the SSAT (soil saturation) variable from the OLS model and in the spatial lag model are positive while the coefficients of the two variables that are used

Variables	Description	Coefficient	Std. Error	z-value	Pr(>|z|)
Intercept	Constant	6.50696	1.73749	3.745	0.000
RINT	Rainfall intensity	4.27851	1.4963	2.8594	0.004
SSAT	Soil saturation	3.39523	0.81893	4.1459	0.000
L3PCT	Prev. 3 day rainfall	-2.24538	0.68608	-3.2728	0.001
N15PCT	Prev. 15 day rainfall before L3	-3.39467	0.65208	-5.2059	0.000
grid_pct	Tree density	-0.65398	0.3492	-1.8728	0.061
lnHH	Log household density	-0.19554	0.12784	-1.5296	0.126
CCTVYN	Pipe CCTV'd previous to date of loss	-0.62321	0.31339	-1.9886	0.047
valsqft	Log parcel value/log parcel sqft	-0.00119	0.00059	-2.0222	0.043
lnsqftlot	Log sqft of parcel	0.23012	0.07276	3.1628	0.002

Rho: 0.23604 Likelihood Ratio test value: 5.714 p-value: 0.016830

Asymptotic standard error: 0.098193, z-value:2.4039, p-value: 0.016223

Wald statistic: 5.7785, p-value: 0.016223

Log likelihood: -801.2281 for lag model

Number of observations: 459

AIC: 1626.5, (AIC for lm: 1630.2)

LM test for residual autocorrelation

test value: 0.10472, p-value: 0.74624

Table 6: Spatial lag model results

to compute SSAT are negative. Similar results were seen in the rare events logistic regression results. On Table 7, the respective correlations between L3PCT and SSAT as well as N15PCT and SSAT are 0.6051 and 0.718. This indicates a strong correlative relationship between the L3PCT, N15PCT and SSAT, which was anticipated since L3PCT and N15PCT are used to compute SSAT. The countering signs on the coefficients indicate that all these variables are significant but the variables' marginal contributions counteract one another.

Variable	RINT	L3PCT	N15PCT	SSAT	grid_pct	lnsqftlot	lnYOLTOT	valsqft	lnHH	CCTVYN
RINT	1									
L3PCT	0.763	1								
N15PCT	0.047	0.066	1							
SSAT	0.4	0.605	0.718	1						
grid_pct	0.04	0.028	-0.067	-0.01	1					
lnsqftlot	-0.12	-0.06	0.001	-0.1	-0.26	1				
lnYOLTOT	-0.19	-0.25	-0.077	-0.25	-0.25	0.738	1			
valsqft	-0.1	-0.25	-0.113	-0.19	0.05	-0.52	0.192	1		
lnHH	-0.33	-0.34	0.06	-0.32	-0.05	0.147	0.243	0.111	1	
CCTVYN	-0.11	-0.04	0.151	0.08	-0.18	0.014	0.0002	-0.015	0.07	1

Table 7: Correlation matrix for the spatial lag model results

The coefficient for tree density (grid_pct) is negative, meaning that higher tree densities correspond with lower damage amounts. It appears that the indirect effect of trees and foliage has more influence on decreasing the amounts of damage claims than the direct effect of tree roots burrowing into sewer pipes and decreasing their capacity.

The coefficient for the log household density is negative as well, though the coefficient was postulated to have been positive. According to the results, more households in an area mean less damage. This may be due to problems inherent in the sewer system where much of the claims are clustered. These areas may not be as dense as other areas where fewer problems are observed.

Also, areas with higher population densities may be tapped into a system with more capacity and more of a maintenance priority, which may suggest some endogeneity with CCTVYN.

The coefficient for the CCTVYN variable is negative. Since this variable is a binary variable, the results suggest that when a pipe was viewed via CCTV before the date of loss, the damage is less than if it were not viewed beforehand. Obviously, the act of viewing the pipe does not prevent the damage but the results imply that subsequent maintenance actions were taken in places where pipes were viewed before the date of loss. Given that the data are backup claims, the backups still occurred after these measures were conducted but the damage was lessened as a result of the presumed maintenance actions.

The coefficient for the log square feet (lnsqftlot) of the parcel was positive while the value per parcel square feet (valsqft) coefficient was negative. The log parcel value was not a significant variable on its own. These results suggest that larger parcels received larger amounts of claim damage but parcel value had a very slight negative effect on claim amounts. Damages could be mitigated by prioritizing maintenance for sewer pipes that are connected to large parcels and presumably large buildings, especially in areas with lower household densities. The same counteractive effect can be seen with the lnsqftlot and valsqft variables as their correlation is -0.52.

The spatial lag model appears to have reduced the spatial autocorrelation in the dependent variable according to the p-value of the likelihood ratio test value (0.01683). While this value still suggests that slight spatial autocorrelation is present in the dependent variable, it has been significantly reduced from the uncorrected (OLS) state. The p-value of the result of the Lagrange

Multiplier test on the residual autocorrelation (0.74624) suggests that spatial autocorrelation of the residuals was removed with the spatial lag specification.

Rare Events Logistic Regression Results - The December 2006 Storm[6]

The three storms mentioned in Chapter 2 were analyzed in a logit model but the logit results of the December 14, 2006 storm will serve as the example for the purposes of brevity. However, it is worthy of note that the logit model in the December 3rd, 2007 storm had a pseudo-R squared value of 0.9351. While the results are not adjusted for the true population of sewer backups, this result (as well as the other logit results) was encouraging given the goodness-of-fit of the best-fitting OLS model that is displayed in this section.

All of the variables were placed in a logit model to eliminate insignificant variables so the remaining significant variables can be used in the rare events logistic regression model. The results of the first run of the logit model are displayed in Table 8. The peak five minute rainfall variable did not allow for an output in the December 14th, 2006 storm due to a lack of variance. The previous three days of rainfall was equal to the previous two days of rainfall since no rain was recorded on December 11th, 2006, so the variable for the previous three days of rainfall was omitted by Stata due to multicollinearity between the previous two and three day rainfall variables.

[6] 88 sample parcel observations were omitted by Stata due to no values in the age variable. The other two storms had no observations removed as the age variable was not significant in the results of either model and the other variables were not missing any values.

The coefficients for 10 minute rainfall and soil saturation are negative while the rest of the rainfall variables are positive. This is not meant to be interpreted as 10-minute rainfall and soil saturation having a negative effect on whether or not a backup occurs. Instead, it means that there exists non-linearity within the various rainfall durations. This can be explained because the RL2 and the N15 variables are correlated with the SSAT variable since the combination of the two former variables are used to derive the latter

Variables	Description	Robust Coefficient	Std. Err.	z	P>z	[95% Conf.	Interval]
RPK10MIN_121406	Peak 10 minute rainfall	-123.6878	58.87154	-2.100	0.036	-239.0739	-8.3016
RPEAKHR_121406	Peak hour of rainfall	50.26696	17.79439	2.820	0.005	15.39059	85.1433
RPEAK3HR_121406	Peak 3 hour rainfall	36.05604	3.576258	10.08	0.000	29.0467	43.0653
RL2_121406	Previous 2 days rainfall	18.97426	9.976632	1.900	0.057	-0.579577	38.5281
N15_121406	Rainfall 4 to 18 days before event	57.38016	13.94567	4.110	0.000	30.04715	84.7131
SSAT_121406	Soil saturation	-526.9178	158.5381	-3.320	0.001	-837.6467	-216.18
GRIDCODE	Restaurant density	0.0114485	0.004497	2.550	0.011	0.0026354	0.02026
ln07TOTVAL	Log 2007 parcel value	-0.826698	0.141362	-5.850	0.000	-1.103762	-0.5496
lnHHSQMI	Log households per sq. mile	-1.473217	0.298763	-4.930	0.000	-2.058781	-0.8876
AGE	Pipe age	-0.035070	0.005441	-6.450	0.000	-0.045734	-0.0244
ELEV	Pipe elevation	-0.014588	0.002884	-5.060	0.000	-0.020241	-0.0089
PROBCOMD	Combined flow dummy variable	3.774504	0.560863	6.730	0.000	2.675233	4.87377
Constant	Intercept	-35.03142	6.032621	-5.810	0.000	-46.85514	-23.207
Number of obs = 3553		Wald chi2(12) = 201.14			Prob > chi2 = 0.0000		
Log pseudolikelihood = -184.02315					Pseudo R2 = 0.6994		

Table 8: Best-fitting logit model for the December 14th, 2006 storm

In the same way, the 10 minute peak rainfall is correlated with the one and three hour rainfall since the 10 minute peak rainfall occurred within the peak one hour rainfall and the one hour peak rainfall occurred within the three hour peak rainfall. When the coefficients are multiplied by the variables for the probability calculation, the rainfall has a net positive effect on the probability of a sewer backup occurring. The use of many different rainfall variables is warranted since they are all significant and hence contribute to the explanation of why backups occur.

The remaining variables in the last logit model run were placed into a rare events logistic regression model using both the prior and weighted correction specifications. For the May 2006 and December 2007 storm, the τ was too low for the weighted correction model, so a higher value was placed in the model until results were achieved. However, due to the imposed inflation of the tau value for the May 2006 and December 2007 storms, only the prior correction model results will be examined.

The final run of the logit model for the December 2006 storm (Table 8) had a pseudo R-squared value of 0.6994 and 12 statistically significant variables remained. The variables were placed in a *relogit* model with the prior correction specification and insignificant variables were removed after each output until no insignificant variables remained (variables with a p-value >= 0.1). The same process was completed in a rare events logistic regression model with the weighted correction specification. The final results of the model using the weighted correction specification are displayed in Appendix A, Table A. The model results with the prior correction specification applied are displayed in Table 9.

Variables	Description	Robust Coefficient	Std. Err.	z	P>z	[95% Conf.	Interval]
RPEAKHR_121406	Peak 1 hour rainfall	11.32741	3.1177	3.63	0.000	5.21668	17.4381
RPEAK3HR_121406	Peak 3 hour rainfall	34.54804	3.7414	9.23	0.000	27.2150	41.8811
N15_121406	Rainfall 4 to 18 days before event	36.83012	4.5139	8.16	0.000	27.983	45.6772
ln07TOTVAL	Log 2007 parcel value	-0.86731	0.1469	-5.9	0.000	-1.1552	-0.5793
lnHHSQMI	Log households per square mile	-1.34926	0.2530	-5.33	0.000	-1.8451	-0.8533
SSAT_1214	Soil saturation	-349.625	43.128	-8.11	0.000	-434.15	-265.09
AGE	Pipe age	-0.03431	0.0052	-6.58	0.000	-0.0445	-0.0241
ELEV	Pipe elevation	-0.01381	0.0032	-4.29	0.000	-0.0201	-0.0075
PROBCOMD	Combined flow dummy variable	3.593501	0.7132	5.04	0.000	2.19564	4.99136
Constant	Intercept	-27.158	4.9085	-5.53	0.000	-36.778	-17.537
Number of observations: 3553							

Table 9: Results from rare events logistic regression using December 14th, 2006 storm data and the prior correction specification

It was decided that the prior correction will be the preferred specification for this project since the prior correction is the only valid specification for models for other two storms and the true τ is known. When the researcher is confident of the functional form and explanatory variables, prior correction is called for; otherwise, our corrected version of weighting with rare event corrections would seem preferable (King and Zeng, 2001). Since the logit models fit well even before being adjusted for the τ level, it can be said that confidence can be placed in the functional form and the explanatory variables used in this thesis. Also, weighting is asymptotically less

efficient than prior correction, an effect that can be seen in small samples, though the differences are not large (King and Zeng, 2001).

Variable	PEAKHR	PEAK3HR	N15	ln07 TOTVAL	lnHH	SSAT	AGE	ELEV	PROB COMD
PEAKHR	1.000								
PEAK3HR	0.963	1.000							
N15	-0.581	-0.633	1.000						
ln07TOTVAL	0.139	0.142	-0.082	1.000					
lnHH	0.057	0.053	-0.03	0.123	1.000				
SSAT	-0.286	-0.356	0.905	-0.102	-0.152	1.000			
AGE	0.269	0.282	-0.21	0.065	0.301	-0.253	1.000		
ELEV	-0.162	-0.18	0.213	-0.254	0.139	0.195	0.021	1.000	
PROBCOMD	0.227	0.247	-0.165	0.099	0.132	-0.168	0.386	-0.125	1.000

Table 10: Correlation matrix for December 2006 storm results

The claims and the peak three hour rainfall amounts for this storm are displayed in Figure 7. Most of the claims seem to lie within the areas where the peak three hour rainfall amounts were above 0.8 inches. This subset of claims has two large clusters and two smaller ones; the rest are dispersed in the areas elsewhere within the heavy bands of rainfall.

There were four rainfall variables that were significant in the prior correction results: peak one hour, peak three hour, previous 4 to 18 days of rainfall and soil saturation. The coefficients for the first three were positive while the soil saturation's coefficient was negative. As seen in the spatial econometric model and the logit model results, the variable that is part of the soil saturation equation (Equation 2) and the soil saturation variable itself appear to be counteracting one another.

Table 10 shows high correlation (0.905) between the soil saturation variable (SSAT) and the previous four to eighteen day rainfall (N15). This suggests that the other variable (L3) that

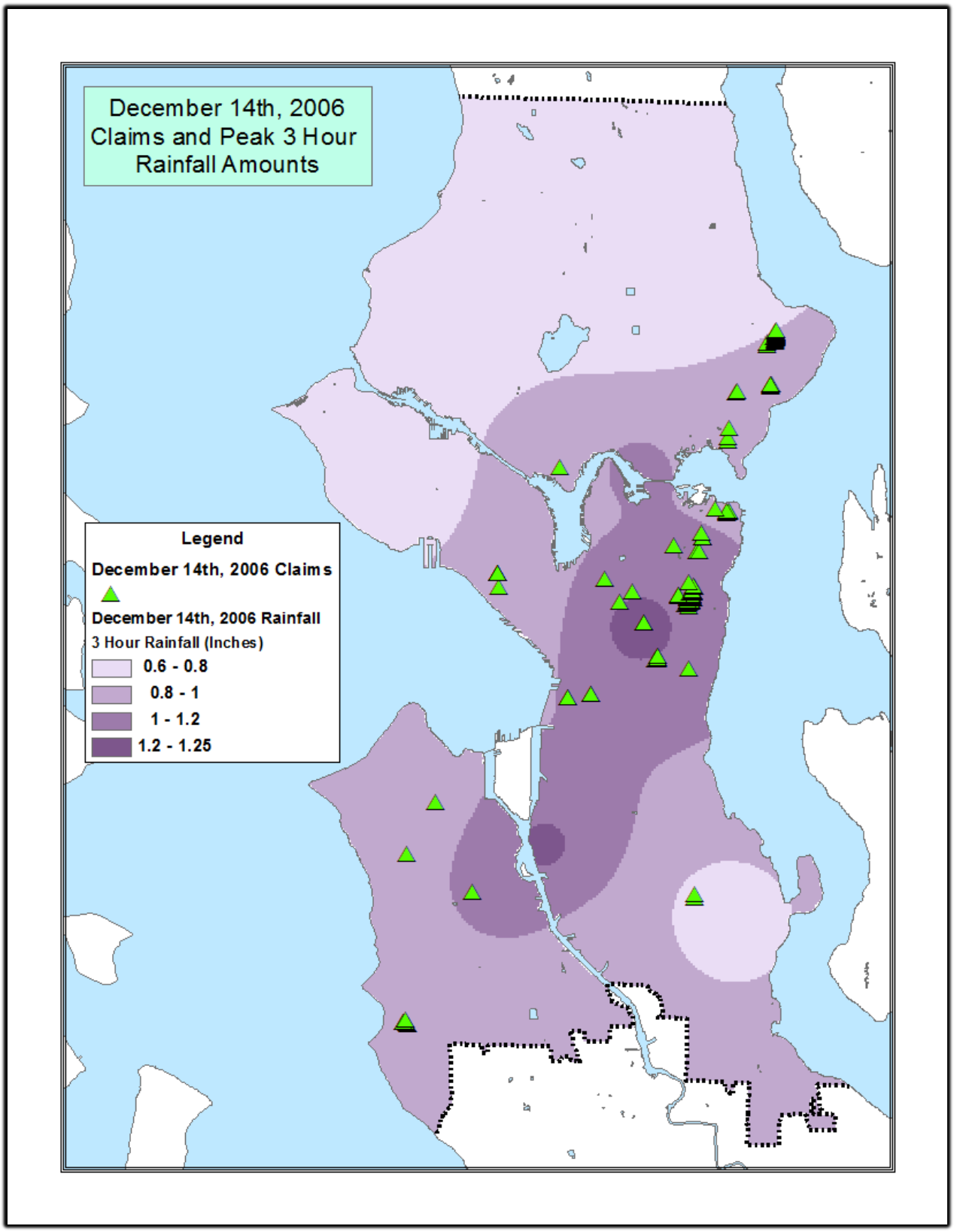


Figure 7: Claims and peak three hour rainfall amounts from the December 14th, 2006 Storm

makes up the soil saturation variable has influence on soil saturation. The 10 minute rainfall variable was no longer significant when the rare events specification was applied.

The log parcel value and log household density coefficients were negative and such results were opposite of was postulated. An explanation for the negative sign of the parcel value is the presence of endogeneity in the model where the parcel value and the error term are correlated. As for the sign of the household density, endogeneity could apply to this situation as well. In this case, endogeneity manifests itself by suggesting that lower levels of population density predict higher probabilities of sewer backups when these areas were more likely to have been disproportionately affected by the rain storm.

The coefficients for pipe age and elevation were negative and were also very low numbers. As for age, it may be the case that backups occurred in areas where the infrastructure is younger than many of the sample pipes. Backups can occur in pipes of many different ages and the slight negative coefficient can be a function of the uneven amounts of rainfall that fell in the city. The sign of the pipe elevation coefficient was expected, though the effect was not large. The combined flow dummy variable coefficient was positive and this result was expected. The significance conveys that the heavy rainfall amounts added to the sanitary flow in the combined flow sewer pipes increases the probability of backups.

Spatial effects in the December 2006 claims and the sample parcels were tested for using Moran's I in ArcGIS. The results are in Figure C in Appendix D. The Moran's I test reveals no significant spatial clustering with a test value of -0.000986 and a p-value of 0.14049. Since the December 2006 storm contains the most amount of claims and visibly has the highest amount of clustering in the claims, the other results from the other two storms can be considered absolved

of significant levels of spatial autocorrelation due to the use of the same random stratified sample of parcels.

December 2006 Model Interpretation: First Differences and Rate Ratios

First difference values were calculated for the significant variables from the December 2006 storm that would conceivably vary with an event or could be reasonably altered for the purposes of sewer backup damage claim mitigation. The only variables that applied to these criteria were the peak one hour, peak three hour, previous four to eighteen day rainfall and soil saturation. The first difference values are changes in the probability of a backup occurring given the application of the percentage increases or decreases away from the mean and that all the other significant variables from the rare events logistic regression results remain constant. These variables are recorded in Table 11.

Changes in the one hour rainfall level in either direction never registered above 0.000005 in absolute change. Increases in the three hour rainfall amounts produced higher amounts of probability changes than the decreases applied. There were significant increases in sewer backup probability with the applied increases in previous four to eighteen day rainfall. A 30% increase in the mean previous four to eighteen day rainfall increases the probability of a backup by 66.2%! Soil saturation was the only variable with a negative coefficient and positive changes were ineffective in influencing backup probabilities while negative changes increased the probability of a backup occurring, especially at the 30% decrease.

The total probability changes were dominated by the probability change of the most influential variables. All variable changes suggested that there was a higher probability of a backup occurring. The 30% increases and decreases brought about significant positive changes in sewer

backup probability. The risk ratios, shown in Table 12, demonstrate how significant those changes in probability are. Since these changes are applied to all of the variables at once, each risk ratio is the effect of the variable's change and the total effects are the combinations of those ratios.

Variable	Mean Value	10% increase	20% increase	30% increase	10% decrease	20% decrease	30% decrease
Peak 1 hour rainfall	0.54183	0.00000	0.00000	0.00001	-0.00000	-0.00000	-0.00000
Peak 3 hour rainfall	0.86922	0.00002	0.00042	0.00821	-0.00000	-0.00000	-0.00000
Prev. 4-18 day rainfall	1.30552	0.00013	0.01558	0.66168	-0.00000	-0.00000	-0.00000
Soil saturation	0.12781	-0.00000	-0.00000	-0.00000	0.00009	0.00761	0.42532

Values with all zeros are less than +/- 0.000005 depending upon the values' sign

Table 11: First differences of rainfall variables' change in probability depending upon the percentage changes away from mean values

The probability increases in the percentage changes for the peak rainfall variables are small but the rate ratios imply that the a 30% increase in the mean value of the peak one hour rainfall makes the occurrence for a backup over six times more likely than a peak one hour rainfall amount and a 30% increase in the mean value of the peak three hour value makes the probability of a backup 8000 times more likely on its own. This implies that a uniform storm (similar to December 2007) but with higher peaks may produce more backups than the city is accustomed to.

Variable	10% increase	20% increase	30% increase	10% decrease	20% decrease	30% decrease
Peak 1 hour rainfall	1.85996	3.41220	6.18657	0.53862	0.29837	0.15744
Peak 3 hour rainfall	20.39637	410.00	8000.00	0.05044	0.00251	0.00012
Prev. 4-18 day rainfall	130.00	14000.00	580000.00	0.00806	0.00007	0.00000
Soil saturation	0.01124	0.00014	0.00001	88.59306	7200.00	380000.00
Values with all zeros are less than +/- 0.000005						

Table 12: Risk ratios of rainfall variables (change in variable percentage to mean variable value)

Given the results of the first difference and risk ratios for the two peak rainfall variables, large risk ratios can be expected for the two soil saturation variables due to their much higher first difference values. A 30% increase in the previous four to eighteen day rainfall amount makes a backup 580,000 times more likely than the mean amount. A 30% decrease in soil saturation makes a backup 380,000 times more likely. Given the differences in first difference values and risk ratios, it is clear that the two soil saturation variables analyzed here counteract one another. It would be even clearer if the previous three day rainfall were significant.

As with the ESBD calculations, looking at relative probabilities is more helpful in determining where the risks are. Risk ratios allow for the analysis of the sensitivity of the coefficients. The coefficients on their own convey little more than their effect on probability, whether that be positive or negative. Examining how the probabilities change as a result of marginal changes in the variables better informs how the rainfall variables affected the occurrence of backups within a given storm. It is clear that, regarding the December 2006 storm, average levels of rainfall did not produce the amount of backups that the higher levels of rainfall did. This reinforces what was seen in the model results and in Figure 7.

Model Validation - Predicted Backups for December 2006 Storm

The values of the probability raster (displayed in Figure 8) were transferred to every parcel in the Seattle Parcel layer and those probabilities were summed to predict the amount of backups conditional on the December 2006 storm occurrence. The predicted amount of backups is 75, barely more than half of the actual backups (147). The four rainfall variables (one hour peak rainfall, three hour peak rainfall, previous four to eighteen day rainfall and soil saturation) were increased and decreased by 5% and 10% and the probabilities given those changes (and holding all other significant variables constant.

A 5% and 10% increase in the rainfall variables summed to 592 and 3002 backups, respectively. A 5% and 10% decrease in the rainfall variables summed to 8 backups and 1 backup, respectively. These results suggest that there is a fine line between a few backups and thousands of backups. Considering the fact that the totals estimated from the increases in rainfall variables have not been seen while the totals from the decreases are common, it is reasonable to be skeptical of the former though the results of Salathe et al. (2010) and Zhu (2012) suggest rainfall intensity increases from the December 2006 rainfall variables are possible.

Other model validation measures were attempted and they include shifting the peak rainfall and soil saturation amounts to different parts of the city and using the December 2007 rainfall variable results but the results of these model validations were not viable. The results suggest that the Stata results for one storm cannot be transferred to another storm but the results of each storm produce conditional probabilities for that storm and the accumulation of storm analysis should be combined via model averaging. As models from claim-producing storms increase, model averaging should reveal what parts of the city experience sewer backups and what parts of the city file claims for sewer backup damage.

Expected Damage Results and Maps

As stated in the Methods section, the coefficients of the rare events logistic regression model (in Table 9) and the spatial lag model (in Table 6) were multiplied by the relevant variables associated with the non-claim parcels in the rare events results from the 2006 storm to derive a probability and damage amount, respectively. The probabilities and the damage amounts for each parcel were multiplied together to get an expected sewer backup damage (ESBD) amount.

The citywide probability, damage and ESBD amount IDW rasters are displayed on maps in Figures 8, 9 and 10, respectively. While the significant variables can be used as reference for newer storms that can be analyzed, the rainfall variables are characteristic of the storm that happened and the many of the non-rainfall variables are static or cannot be altered for the purposes of lowering backup probability.

Figure 8 shows the predicted probabilities that were calculated using the variables attached to the parcels and the coefficients in Table 9 using Equation 19. The two areas with the highest probabilities correspond with the clusters of backups shown in Figure 7. Probabilities of backups at the lower end of the spectrum never reached zero though they had as many as eleven zeroes to the right of the decimal point.

The significant variable coefficients from the results of the spatial econometric model with the spatial lag specification (Table 6) and the variables from the sample parcels were used to calculate potential damage using Equation 19. Figure 9 shows the potential damage extrapolated over the entire city of Seattle. The amounts are called potential damage because the amount reflects the amount of predicted damage that would occur given the occurrence of a backup and corresponding paid claims. The ESBD raster was used to compute the values on every single

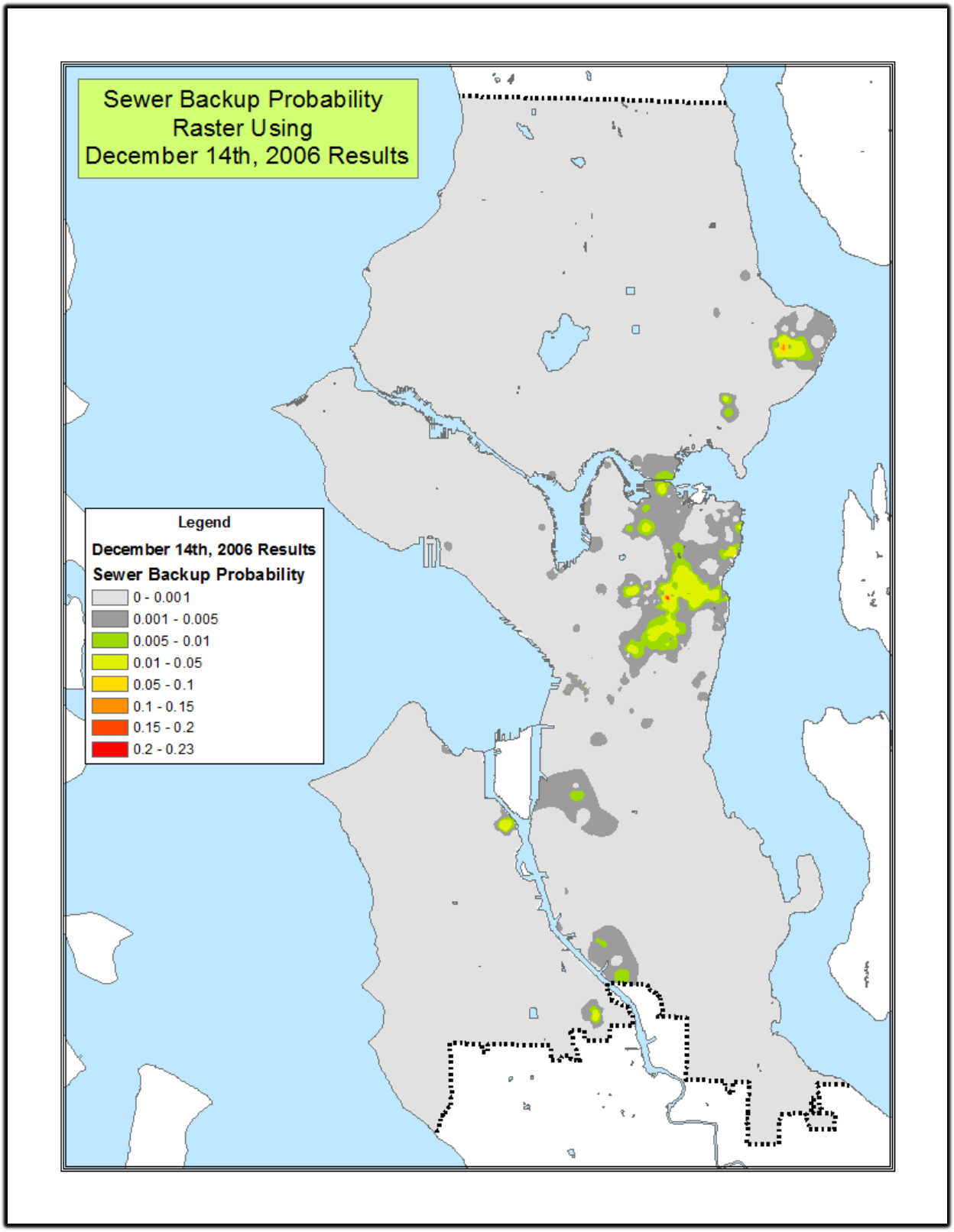


Figure 8: Sewer backup probability raster using the December 2006 rare events logistic regression model results

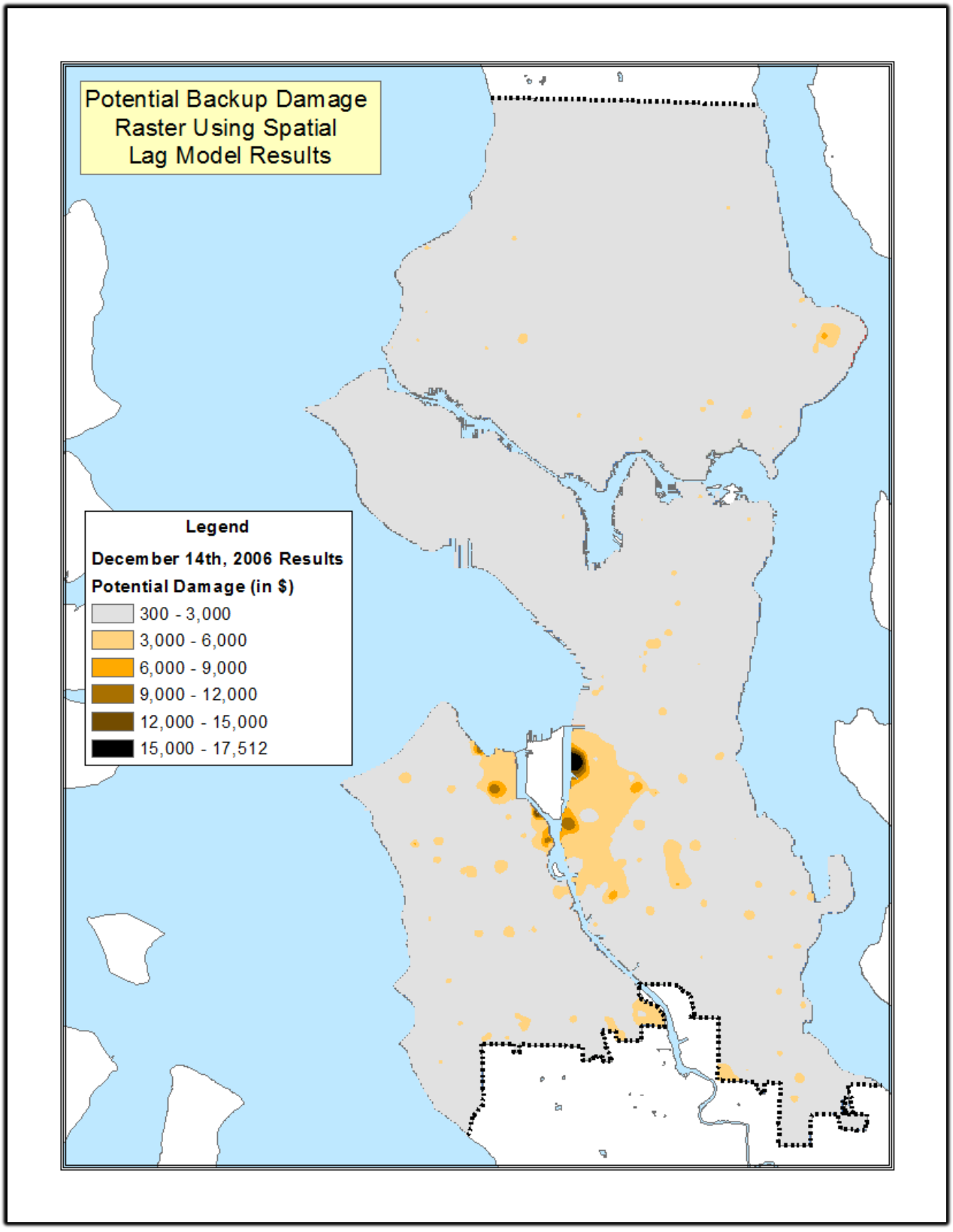


Figure 9: Potential backup damage raster using the spatial lag model results

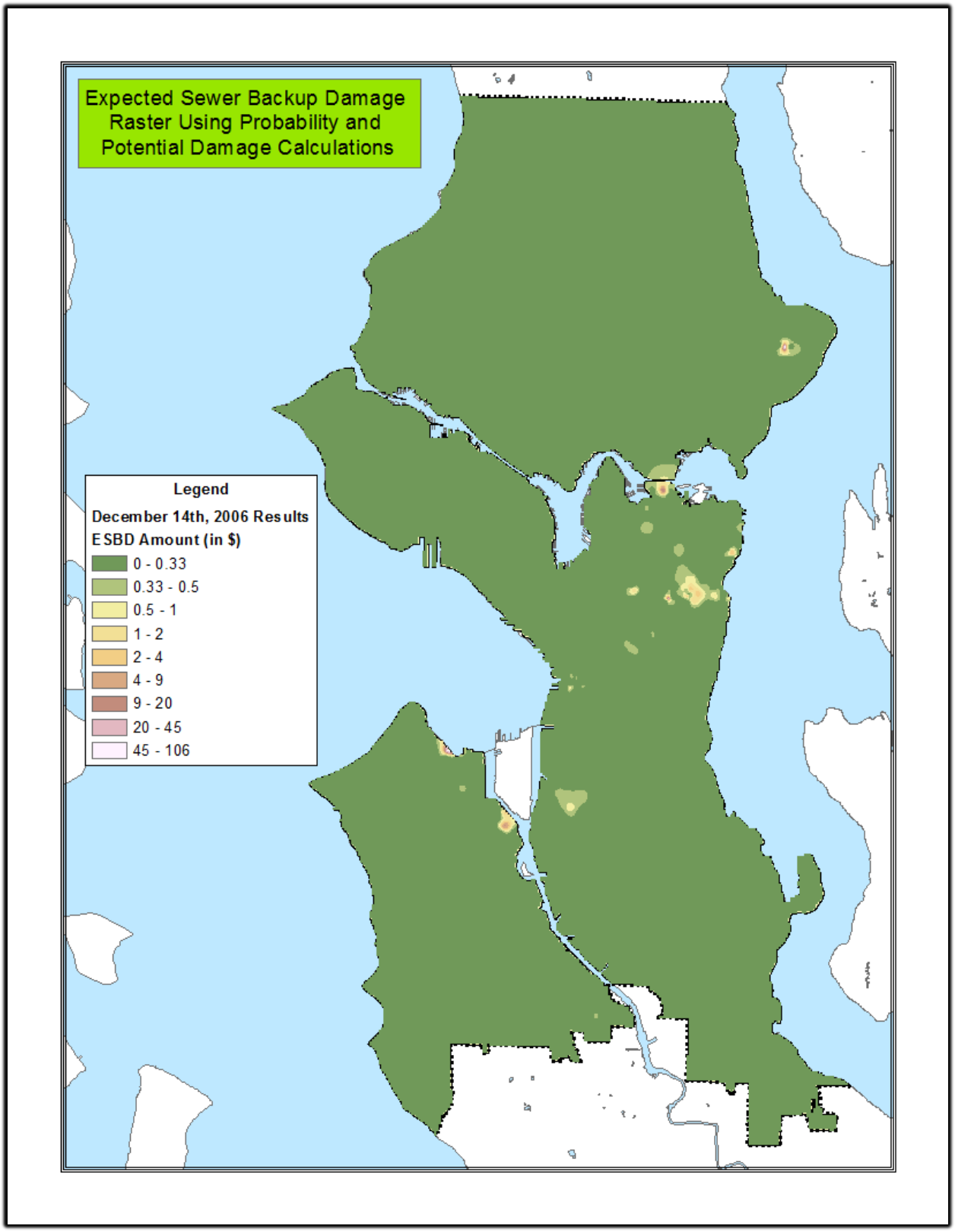


Figure 10: Expected Sewer Backup Damage raster map using results from the December 2006 rare events logistic regression model and the spatial lag model

parcel in the Seattle parcels layer. The total calculated amount of expected damage using this method was $6,378. In Table 2, the mean claim amount is $16,519. Given the highest potential damage amount in Figure 9, $17,512, along with the fact that the average paid claim amount for the December 2006 storm was $26,660, the damage results seem to underestimate high amounts of damage. This may mean that high levels of damage are difficult to predict. It may also mean that large storms produce a lot of damage and the amounts of damage from many of the individual claims resulting from the large storm are above the mean.

Another explanation of the low amounts of predicted damage may have to do with the resolution of the rainfall data used in the spatial econometric model. From a goodness-of-fit perspective, the rare events logistic regression model better explains the probability of a sewer backup occurring on a parcel than the spatial econometric model explains the variance of damage claim amounts resulting from sewer backups. Recalling what was explained in the methods section about the level of data collection, the results reinforce that the level of detail in peak rainfall data is the difference between the two models.

Chapter 5: Implications and Conclusion

Future Storms and Storm Damage Expectations

Another element that the asset managers of Seattle's sewer system may face is planning for the probability of a claim producing storm or multiple storms in a given year. Granted, claims are produced in every year of the sewer backup claim dataset but the storms that produce large amounts of claims may not occur every year yet they pose the largest costs when they occur. A study on the changing frequency of claim-producing storms given the assumed effects of global climate change may be necessary to obtain the best assessment of risk for planning and preventative maintenance.

The probability and damage models provide expected sewer backup damage provided that the storm occurs. Examining the historical distribution of heavy rainfall will reveal areas that are geographically-prone to conditions where sewer backups occur. The landscape of Seattle clearly is clearly dynamic with the lowest areas lying at sea level and the highest areas lie 500 feet above sea level. The highest backup occurred due to a pipe at an elevation of 466.95 feet.

Another important element of measuring expected sewer backup damage is the knowledge of storm frequencies. Finding the likelihood of a storm in a given year may be just as important as determining where the risks of potential sewer backup claims. Part of this risk may include accounting for the risk of multiple claim-producing storms of the same magnitude or storms of different magnitude.

The equation for estimating the expected sewer backup claim damage in a year t is conditional on the amount and severity of claim damage. This equation is derived from the law of total

expectations (David, 2008) for a set of various sized storms $S = \sum_{j=0}^{n} S_j$ (where *j* is a bin that represents an interval of a certain storm magnitude or storm return period i.e. 100 year storms, 50 year storms, etc.) that can occur in a given year and for expected sewer backup claim damage in a given year is $ESBD_t$ ($\sum_{i=1}^{n} ESBD_i = \sum_{i=1}^{n} p_i D_i$) resulting from the claims *i* paid produced by all the storms and is written in Equation 23:

(23) $$E(ESBD_t) = \sum_j E\left(ESBD\middle|S_j\right)P(S_j)$$

Given the recommendations of Salathé Jr. et al. (2010) and Rosenberg et al. (2010), the probabilities of various-sized claim-producing storms occurring need to be recalculated. A logical next step for accounting for the true risk of backup claims in a given year is to determine a probability density function for storm return periods or rainfall intensity. This will allow for a better estimate of the set of storms that make up the set of possible storms ($\sum_j S_j$). When the set of storms is defined in discrete form, as opposed to the continuous form (which is the true definition of the probabilistic nature for how storms actually occur), it does diminish the probability for multiple storms of the same size occurring.

An accurate estimate of the preventative mitigation needed for the expected damage from a storm of a magnitude level *j* can cover the expected damage from an infinite amount of storms (though the maximum number of mutually exclusive storms that can occur in a given year is finite) at the same magnitude level of *j* by using the Taylor series expansion (Shynk, 2012):

(24) $$\sum_j \sum_{n=m}^{\infty} s_j^n = \sum_j \frac{s_j^m}{1-s_j}$$

where m is the number of occurrences of a storm and is equal or greater than 1 since no sewer backup damage is expected if zero storms occur. Since one storm is the lower bound of storm amounts to consider, the true total expectations equation is given in Equation 25:

(25) $$E(ESBD_t) = \sum_j E\left(ESBD\middle|S_j\right)P(\sum_j \frac{s_j}{1-s_j})$$

This would allow potential storms to be placed into bins of various magnitudes and allow for possibility that more than one of the storms within those bins will occur. Having a well-estimated probability density function would help risk managers accurately assess the probabilities of storm return periods and rainfall intensities at various levels and combine with the estimates of $ESBD$ to get the best estimate of $E(ESBD_t)$ given the level of information available.

Regional Models and Further Storm Analysis

The probability model is limited in extrapolating the results of one storm to another storm. Though future storms may fit the archetype of one of the three storms used in this project, claims resulting from the storm must be in the dataset along with parcels that did not submit a backup claim. The December 2010 storm was not analyzed using the probability model because the storm only produced 23 claims. That storm has similar rainfall patterns to the December 3rd, 2007 storm with 24 hour rainfall amounts ranging between 3 and 4 inches in 2010 compared to 4.2 to 5.2 inches in a 24 hour period in 2007. This storm was the fourth largest storm from a claim-producing standpoint and could be added to the expected sewer backup damage analysis provided that the τ level is high enough to be used in the rare events logistic regression model specification.

The models used in this thesis are models that take into account backups that occurred within Seattle. Given the amount of claims that the three most severe storms produced and the levels of clustering where these backups occurred, the models could single out backups and events that occurred in a particular space and/or a particular time. A regional damage model could analyze claims that occurred within a given neighborhood or other type of sub-region of the city. That model would analyze specific aspects of the system and would allow for examination of the causes in places where claims seem to occur often or a lot of claims occurred all at once. There may be some issues with multicollinearity due to the nature of the homogeneity of neighborhood sections of the sewer system and populations but not all aspects are homogenous. There may be potential for certain sewer system variables to aid in the explaination of sewer backup occurrences and damage.

A damage model could also be applied using the claims that occurred within a given storm. The December 2006 storm produced nearly a third of the claims in the claim dataset. Given the level of under-prediction of high claim amounts from the citywide damage model, this storm could allow for the accurate estimation of damages from future storms of this magnitude. A subset of all of the claim producing storms within the sewer backup dataset, including the December 2010 storm, may be an even more accurate predictor of potential damage.

Flood Claims

While this study only applies these methods to sewer backup claims, there were over 200 flood claims in the same original database where the sewer backup claims were exported from that could be analyzed in the same manner. The flood claims were not analyzed in this project because they could not be linked with the sewer system. Given the results of this thesis, the

characteristics of the sewer system are largely inconsequential in explaining sewer backup damage. This knowledge would aid a future project in terms of measuring flood damage by ignoring the fact that some floods are produced by stream overflow and others by system overflow.

New variables may include distance to creek, surface elevations (as opposed to pipe elevations), and stream flow data among others. While causes of flood damage are not quite as mysterious as sewer backup damage causes, using these same methods to measure flood damage and probability would provide some insights on the pattern of damages and what variables may portend flood damage claims.

Adding to the Sewer Backup Occurrence Database

This project is concerned with measuring probabilities and damages of sewer backups resulting in paid claim damage amounts and this damage are ultimately incurred by the city after the claims are paid out. Claims that were not awarded were presumably valid but were not awarded compensation due to a lack of liability on the part of the city. Beyond those claims, sewer backups may have caused damage in Seattle parcels but the parties eligible to make a claim may have elected to pay for the damages themselves or they may have made an insurance claim.

If another study were conducted with the intent to estimate the total recorded damage of backups that occurred in the city, water damage cleaning company and insurance records would need to be sought to add to the damage model. When paid sewer backup damage claims are the only damage accounts in the damage model (as is the case for this study), the results provide a lower bound for ESBD estimates as it is known that these backups occurred at the very least and that

there were most likely more backups that occurred than just those in the sewer backup claim dataset.

Another method of ascertaining the level that backups affect the city is by conducting a random stratified survey of the entire city about a certain storm soon after it occurs. The rare events logistic regression model could be used to compare those that reveal they had backup damage and those that did not. The true τ level for the number of backups would not be known but the rare events logistic regression specification does allow for a range of tau levels which would have to be projected based off of the result. This survey could also seek backup data for storms beyond a specific storm and that data could be devoted to strengthening the claim dataset.

Model Limitations and Interpretations

When this project was started, the rare events logistic regression model was not a part of the research design. The spatial econometric model results were intended to be used with a rainfall event simulation model. The data was collected and processed for the spatial econometric model first and the idea to integrate peak levels of rainfall at different time periods was not conceived until the rare events logistic regression model design was explored. Knowing what was discovered through the application of different levels of peak rainfall, such rainfall data would have been collected for the spatial econometric model as well.

What was known during the data collection period for the spatial econometric model was that rainfall intensity influences whether or not backups occur and that other factors may be involved that could be measured in order to add to the explanation. This statement is still true but, in hindsight, dedicating more data collection and processing time to rainfall intensity levels would have certainly added to the goodness-of-fit for the model.

Before the rare events logistic regression model results were examined, it was assumed that, while rainfall intensity explains sewer backup damages, it was not known where in the mass of the storm was the catalyst for the sewer backup damage and the practice was to treat the storm as a whole and measure the rainfall as a mass rather than measure peak rainfall.

It can be said that if faced with a problem that can be explained by complex mathematical methods and data needs to be collected for such an assignment, investing time appropriate to the level of influence one believes has anecdotal knowledge or apparent knowledge that a variable has on the cause of the problem would be good practice. In this case, it is apparent that, without heavy rainfall, sewer backups are far less likely to occur. While it was apparent that rainfall intensity was important, it was not apparent on what level rainfall intensity was important in explaining sewer backup claim damage amounts.

While having more information about what time of day the backup occurred or when it was discovered may have led to the conclusion for higher resolution rainfall intensity, that information was not available. Given the results from the rare events logistic regression model, this information may aid in explaining what causes these backups. While it is known that rainfall intensities can explain sewer backup occurrences, having this better resolution of data could be used to study sewer backups by using time series data to better explain sewer backup occurrences and sewer backup claims.

Spekkers et al. (2012) had similar goodness-of-fit problems and these problems add to the discussion of how difficult it can be to predict damage resulting from storms. They had the rainfall variables but did not have a lot of the other types of variables used in this project. Having

the combination of high-resolution rainfall density variables and other types of variables may better explain the damage that occurs.

That being said, the damage model may have some useful applications on its own. The relative amounts of potential damage projected on the map can be used to prioritize maintenance for a given storm. In the same manner, the map displaying ESBD (Figure 10) may also have some informative aspects from a relative sense. The sewer pipes in areas with highest amounts of potential damage and ESBD would be prioritized for CCTV surveillance and subsequent grease and root treatments if deemed necessary given an exogenously determined budget constraint.

Present Externalities and Mitigation

The presence of externalities was mentioned in the introduction briefly and it seems that there are two present in the occurrence of sewer backups and sewer backup claims. There is a dual externality that is both positive and negative, trees, and a strictly negative externality, restaurant and residential grease in the system. There were direct and indirect reasons that were postulated to counteract one another. The direct effect (the negative portion) is the damage caused by tree roots that burrow into the sewer system and decrease the capacity in the system. The indirect effect (the positive portion) is the ability of the trees and the greenspaces to delay a portion of the high amounts of rain from entering the system.

Trees have the ability to add to the cause of backups and prevent them during a large rain event. The results say that the positives of tree cover outweigh the negatives given the sign of the tree density coefficient in the spatial econometric model. Mitigating the negative externality of tree cover remains a task of finding the areas at the highest risk of a backup and removing the tree roots when found. Tree roots can be removed by a chemical process with an earth-friendly

application or by a physical process with a remote saw-like power tool that (temporarily) removes blockages in the system (Martin, 2012). These applications are relatively low cost when compared to an alternative of complete asset removal. Such mitigation options allow for them to be applied in many pipes.

The other externality present in the sewer backup problems is the strictly negative externality of restaurant and residential grease. Though the restaurant density was only significant in the May 2006 storm results (Appendix A), grease is still present in the system in places and decreases system capacity. Seattle Public Utilities has implemented a program called the Fats, Oils and Grease (F.O.G.) Disposal program (FOG, 2013). This program inspects restaurants for grease disposal methods and promotes awareness of destructive properties of fats, oils and greases.

Conclusion

The end goal of risk mitigation for a utility should be to examine and prevent damage from the entire suite of risks that are present within a line of business. Sewer backup damage is just one risk that affects a utility and its customers. This project's purpose was to assess the risk of sewer backups and claim damage resulting from sewer backups using models that analyze many different types of factors that may explain why and how sewer backups occur in Seattle and to do so by analyzing the spatial patterns of backup occurrences. Econometric methods used in this project have not been applied before to these types of claims.

This thesis has established the design for efficiently assessing sewer backup claim risk and that assessment can be expanded beyond claims with further data collection. As with any project, further data will always be more useful and, in this case, more storms will further aid in the estimation of sewer backup risk. In terms of future Pacific Northwest rainfall regimes, it is quite

possible that higher amounts of excess rainfall may be imposed on Seattle and the Puget Sound region more often than they have been in the past and present.

Mitigating the increased risk of sewer backup damage from claim-producing storms will require that, with each additional storm, the level of information added to the assessment through future claims and the characteristics of future claim-producing storms will allow for the assessment of risk to keep up with the future rainfall regime, whatever that may be.

Sources:

Bell, K. P., and N. E. Bockstael. (2000). Applying the generalized-moments estimation approach to spatial problems involving microlevel data. Review of Economics and Statistics 82: 72-82.

Braitman, L. E., & Rosenbaum, P. R. (2002). Rare outcomes, common treatments: analytic strategies using propensity scores. *Annals of Internal Medicine*, *137*(8), 693-695.

Breusch, T. S., & Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica: Journal of the Econometric Society*, 1287-1294.

David, N. (2008). Law of Total Probability. *Available at SSRN 1310502*.

Dulière, V., Zhang, Y., & Salathé Jr, E. P. (2011). Extreme Precipitation and Temperature over the US Pacific Northwest: A Comparison between Observations, Reanalysis Data, and Regional Models*. *Journal of Climate*, *24*(7), 1950-1964.

ESRI: ArcGIS Desktop. (2012, March 7). *Help for Previous Versions | ArcGIS Resources*. Retrieved June 2, 2013, from http://help.arcgis.com/en/arcgisdesktop/10.0/help/index.html#/How_Spatial_Autocorrelation_Global_Moran_s_I_works/005p0000000t000000/

"Fats Oils and Grease ." *seattle.gov*. N.p., n.d. Web. 2 June 2013. <www.seattle.gov/util/MyServices/DrainageSewer/FatsOilsGrease/index.htm

Greene, William H. "Econometrics."*Econometric analysis*. 7. ed. Harlow: Pearson Addison Wesley, 2012. 1-2.

Griffith, Daniel A., Larry J. Layne, J. K. Ord, and Akio Sone. *A casebook for spatial statistical data analysis: a compilation of analyses of different thematic data sets*. New York: Oxford University Press, 1999.

Guns, M., & Vanacker, V. (2012). Logistic regression applied to natural hazards: rare event logistic regression with replications. *Nat. Hazards Earth Syst. Sci*, *12*, 1937-1947.

Hall, J., Dawson, R., Speight, L., Djordjevic, S., Savic, D., & Leandro, J. (2007). Sensitivity based attribution of flood risk. *NOVATECH 2007*.

Hydstra. Kisters Pty Ltd. June 12, 2012

Imai, K., King, G., & Lau, O. (2007). Relogit: rare events logistic regression for dichotomous dependent variables. *URL: http://gking.harvard.edu/zelig*

King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political analysis*, *9*(2), 137-163.

King, G., & Zeng, L. (2001a). Explaining rare events in international relations.*International Organization*, *55*(3), 693-715.

King, G., & Zeng, L. (2001b). Improving forecasts of state failure. *World Politics*,*53*(4), 623-658.

King, G., & Zeng, L. (2002). Estimating risk and rate levels, ratios and differences in case-control studies. *Statistics in medicine*, *21*(10), 1409-1427

King, Gary. "Rare Events." *Gary King*, 1 May 2013. http://gking.harvard.edu/category/research-interests/methods/rare-events

King County GIS Data Center. "KCGIS Center." *King County*, 1 May 2013 http://www.kingcounty.gov/operations/GIS/GISData/GISDataDistribution.aspx

Lee, S. and Chung, S. (2003). "Infrastructure Asset Management - Methodologies for Infrastructure Asset Management System in U.S", Proceedings of 2003 KICEM, Korea, pp. 66-72.

LeSage, James P., and Robert Kelley Pace. "Introduction ." *Introduction to spatial econometrics*. Boca Raton, FL: CRC Press, 2009. 5-6. Print.

Lubini, Alain and Musandji Fuamba (2011). "Modeling of the deterioration timeline of sewer systems." *Canadian Journal of Civil Engineering* 38: 1381-1390.

Martin, Terry. Personal interviews. March and April 2012.

Moran, P. A. (1950). Notes on continuous stochastic phenomena. *Biometrika*,*37*(1/2), 17-23.

Pattanayak, Subhrendu and David Butry (2002). "Complementarity of Forests and Farms: A Spatial Econometric Approach to Economic Valuation in Indonesia." *RTI international* Working Paper 02_01: 1-27. *RTI International*. Web. 26 Apr. 2012.

Phelps (1978). "Public Works in Seattle: A Narrative History of the Engineering Department 1875-1975." *Seattle: Seattle Engineering Department.*

Robertson, R. D., Nelson, G. C., & De Pinto, A. (2009). Investigating the predictive capabilities of discrete choice models in the presence of spatial effects. *Papers in Regional Science*, *88*(2), 367–388.

Rosenberg, E. A., Keys, P. W., Booth, D. B., Hartley, D., Burkey, J., Steinemann, A. C., & Lettenmaier, D. P. (2010). Precipitation extremes and the impacts of climate change on stormwater infrastructure in Washington State. *Climatic Change*, *102*(1-2), 319–349.

Salathe Jr, E. P., Leung, L. R., Qian, Y., & Zhang, Y. (2010). Regional climate model projections for the State of Washington. *Climatic Change*, *102*(1-2), 51-75.

Sharpley, Jason. Personal Interview, March 2012.

Shynk, J. J. (2012). *Probability, Random Variables, and Random Processes: Theory and Signal Processing Applications*. Wiley-Interscience.

Spekkers, M. H., Kok, M., Clemens, F. H. L. R., & Ten Veldhuis, J. A. E. (2012). A statistical analysis of insurance damage claims related to rainfall extremes. *Hydrology and Earth System Sciences Discussions*, *9*(10), 11615-11640.

Tomz, Michael, Gary King, and Langche Zeng. 1999. RELOGIT: Rare Events Logistic Regression, Version 1.1 Cambridge, MA: Harvard University, October 1, http://gking.harvard.edu/

"USGS: Seattle Area, Washington." *Landslide Hazards Program*. N.p., n.d. Web. 8 Apr. 2012. http://landslides.usgs.gov/monitoring

Van Tassell, L. W., Yang, B., & Phillips, C. (2000). Depredation claim behavior and tolerance of wildlife in Wyoming. *Journal of Agricultural and Applied Economics*, *32*(1), 175-188.

Vospernik, S. (2006). Probability of bark stripping damage by red deer (Cervus elaphus) in Austria. *Silva Fennica*, *40*(4), 589.

Yoder, J. (2002). Estimation of wildlife-inflicted property damage and abatement based on compensation program claims data. *Land Economics*, *78*(1), 45-59.

Zhou, Qianqian, et al. (2012). "Framework for economic pluvial flood risk assessment considering climate change effects and adaptation benefits." *Journal of Hydrology* 414(1): 539-549.

Zhu, J. (2012). Impact of Climate Change on Extreme Rainfall Across the United States. *Journal of Hydrologic Engineering*, 121018101344009. doi:10.1061/(ASCE)HE.1943-5584.0000725

Appendix A: Rare Events Logistic Regression Results - The May 2006 Storm

The results shown in Table A are the rare events logistic regression model output using the prior correction specification and the May 27th, 2006 rainfall variables. The claims and the rainfall amounts for the peak five minute rainfall variable are displayed on Figure A. There were three rainfall variables, all peak amounts of rainfall at different periods of return. The sign of the coefficient for the five minute and three hour rainfall variables were positive while the one hour rainfall variable coefficient was negative, indicating non-linearity within the variables. Given the brevity of the storm, the peak five minute rainfall amount is contained within the peak hour amount which is contained within the three hour amount, it is clear why one variable is offsetting the other two.

The log parcel value and the log household density coefficients were negative while the log population density was positive. It was presumed from the results of the December 2006 storm that the backups disproportionally hit areas with lower parcel values and lower household densities. Since there is a positive sign on the population density coefficient but a negative sign on the household density coefficient, it appears that this storm affected areas where there are high population densities within lower household densities. These areas could be characterized as residential areas that contain single-family homes. The sign of the parcel value coefficient reinforces that hypothesis since single-family homes are less valuable than multi-family properties and apartment buildings.

The restaurant density coefficient was slightly positive indicating that higher restaurant densities translate into higher backup probabilities. This could be a product of where the rainfall fell on

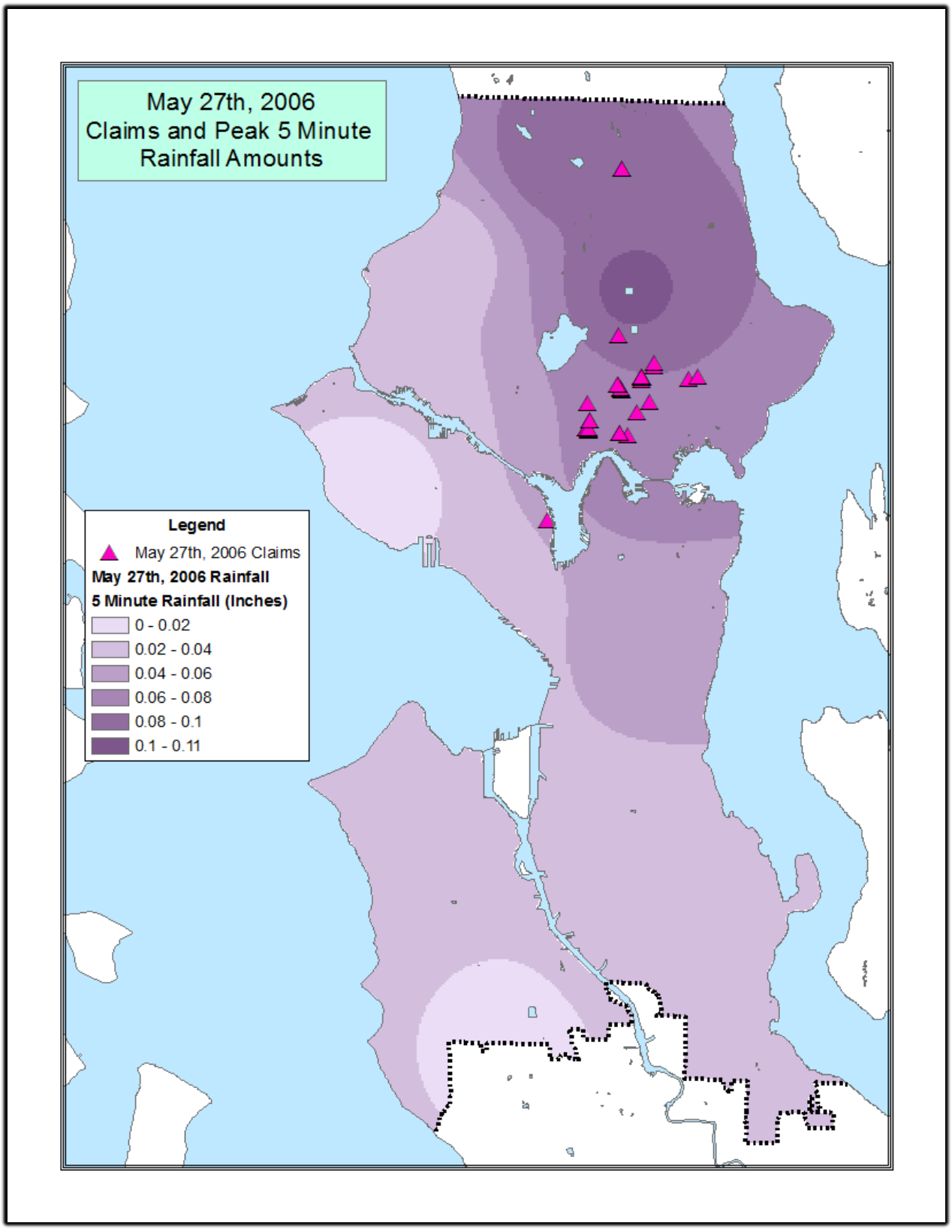


Figure A: Claims and peak five minute rainfall amounts from the May 27th, 2006 storm

Variable	Description	Robust Coef.	Std. Err.	z	P>z	[95% Conf.	Interval]
RPK5MIN_052706	Peak 5 minute rainfall	86.2289	17.0625	5.05	0.000	52.7869	119.6708
RPEAKHR_052706	Peak 1 hour rainfall	-78.9521	19.7374	-4.00	0.000	-117.6367	-40.2675
RPEAK3HR_052706	Peak 3 hour rainfall	44.7962	10.9443	4.09	0.000	23.3457	66.2467
ln07TOTVAL	2007 log parcel value	-0.5022	0.1500	-3.35	0.001	-0.7962	-0.2081
lnPPSQMI	Log population/sq. mi.	4.8931	1.4902	3.28	0.001	1.9724	7.8138
lnHHSQMI	Log household/sq. mi.	-3.4359	1.0702	-3.21	0.001	-5.5334	-1.3384
GRIDCODE	Restaurant density	0.0320	0.0061	5.28	0.000	0.0201	0.0439
MNL_LENGTH	Pipe length	0.0013	0.0003	4.20	0.000	0.0007	0.0019
PROBCOMD	Combined dummy flow variable	1.3778	0.5989	2.30	0.021	0.2041	2.5516
_cons	Intercept	-23.8209	6.3863	-3.73	0.000	-36.3377	-11.3040
Number of observations: 3553							

Table A: Results from rare events logistic regression model using May 27th, 2006 rainfall variables and the prior correction specification

the city since parts of the city did not receive any rainfall. Given that the coefficient is close to zero, the effect cannot be considered a significant factor in why sewer backups occurred.

The pipe length coefficient is slightly above zero. A similar conclusion to the restaurant density could be drawn for this variable's coefficient. The combined flow dummy variable coefficient was positive though not as large or as significant as the variable results for the December 2006 storm. This may be a function of sample size since there are five times more claims used in the December 2006 model as there were in this model.

Appendix B: Rare Events Logistic Regression Results - The December 2007 Storm

The results shown in Table B are the rare events logistic regression model output using the prior correction specification and the December 3rd, 2007 rainfall variables. There are three rainfall variables in the results of this model with two of the coefficients, the peak one hour and peak three hour rainfall amounts, being positive and the other, the previous two day rainfall amount, being negative. This establishes an apparent inevitability of non-linearity inherent in variables where at least one variable is counteracting the positive influence of rainfall on backup claim probability. The claims resulting from this storm along with the peak one hour rainfall amounts are displayed in Figure B.

Variable	Description	Robust Coefficient	Std. Err.	z	P>z	[95% Conf.	Interval]
RPEAKHR_1203	Peak 1 hour rainfall	121.6240	57.38	2.1	0.034	9.1613	234.087
RPEAK3HR_1203	Peak 3 hour rainfall	-47.6812	19.89	-2.4	0.017	-86.67	-8.6933
RL2_120307	Previous 2 day rainfall	-23.1593	2.818	-8.22	0.000	-28.69	-17.635
MNL_LENGTH	Pipe length	0.0029	0.001	19	0.000	0.0026	0.0032
PROBSAND	Sanitary flow dummy variable	3.3551	0.976	3.44	0.001	1.4411	5.2691
ELEV	Pipe elevation	0.0045	0.002	2.56	0.011	0.0011	0.0080
_cons	Intercept	32.4356	5.351	6.06	0.000	21.949	42.9226
Number of observations: 3542							

Table B: Results from rare events logistic regression model using December 3rd, 2007 rainfall variables and the prior correction specification

The pipe length and pipe elevation coefficients were barely positive. These results were very similar to the results for the same variables in the May 2006 model results. However, instead of

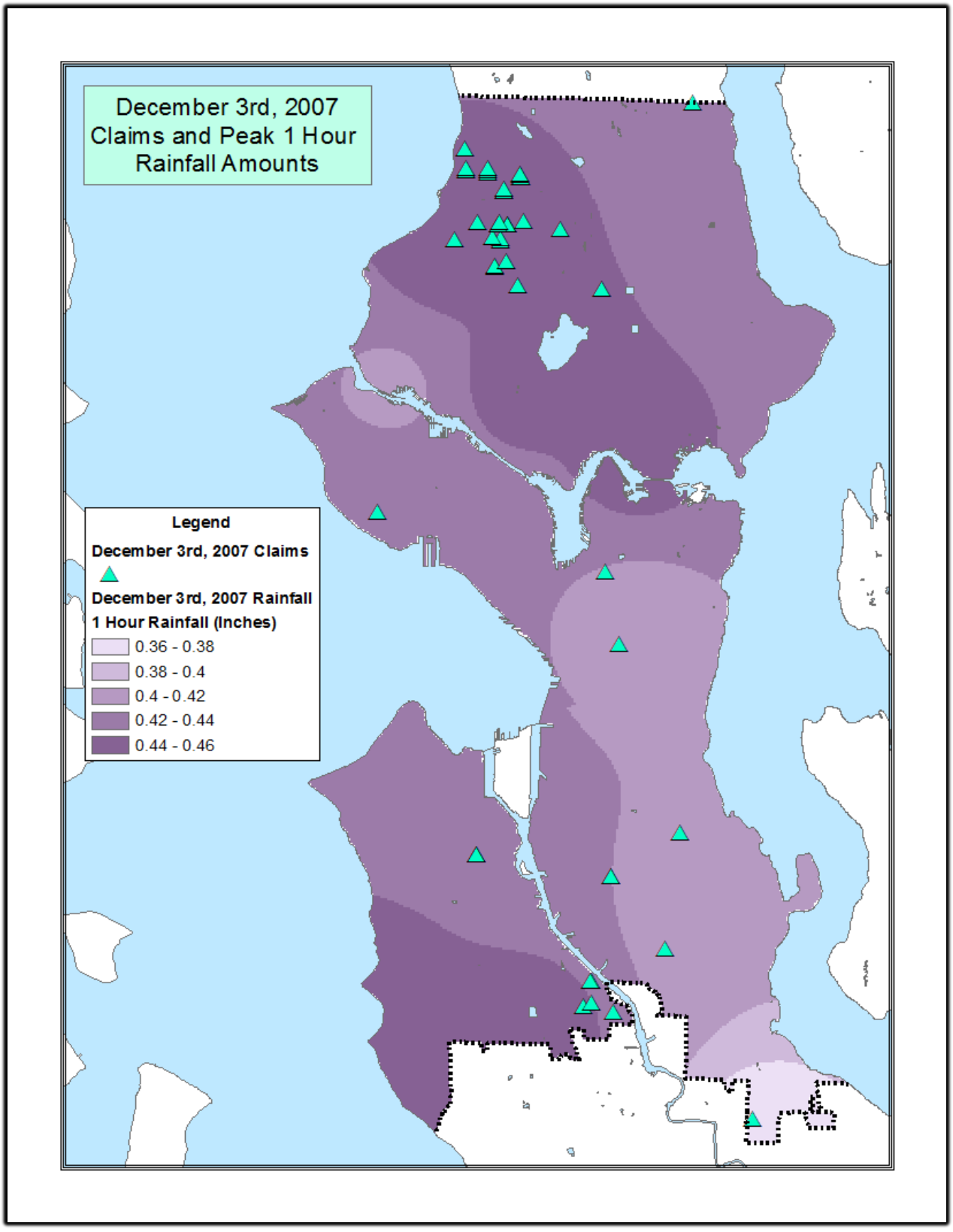


Figure B: Claims and peak five minute rainfall amounts from the December 3rd, 2007 Storm

the combined flow dummy variable being significant and positive, the sanitary flow dummy variable is significant and positive.

Examining Figure 1, where the various sewer system types are displayed, and Figure B, which displays where the claims occurred, it appears that most of the backups occurred in a sanitary system. These results show that some of the variables that are significant in the results of the three models are more descriptive of the geographic distribution of the storm and less of a description of where backups are likely to occur.

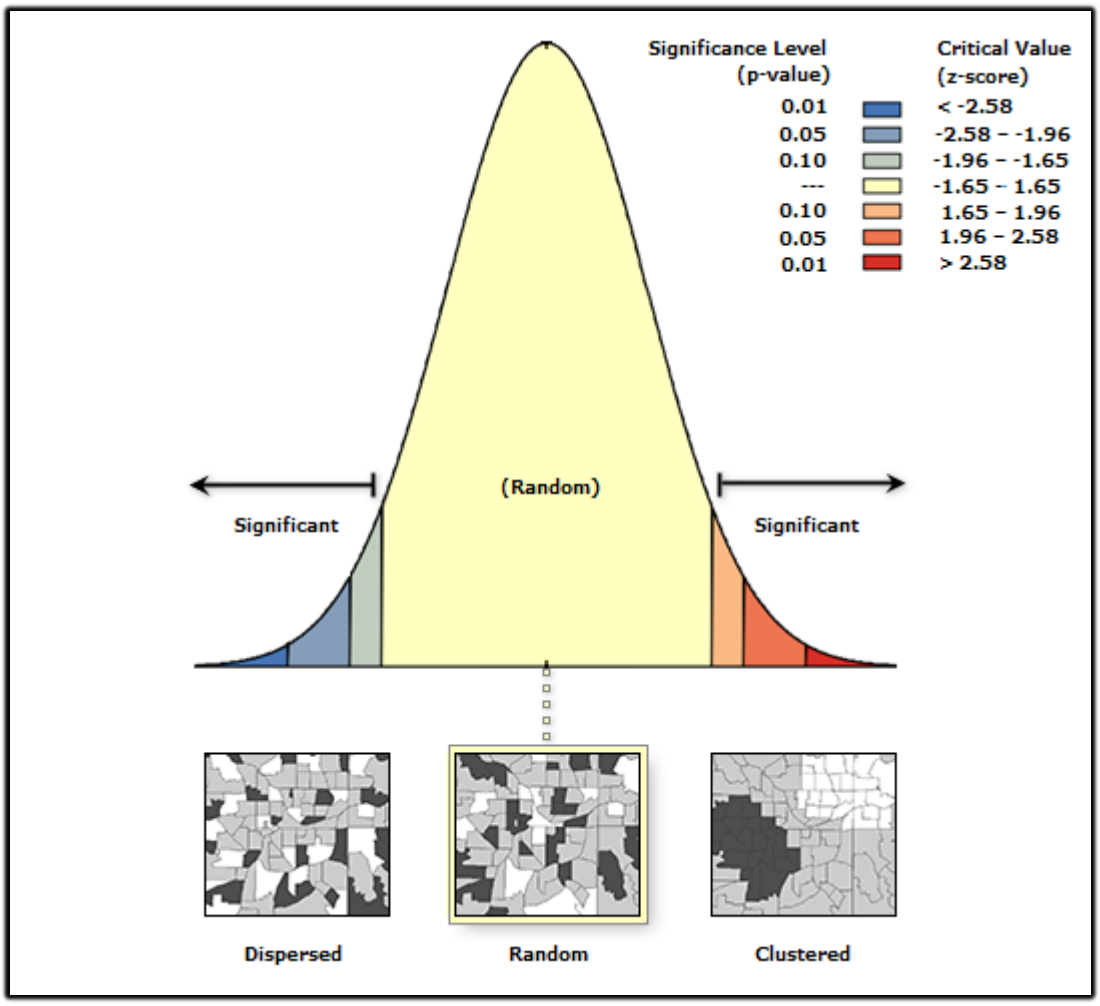
Appendix C: Weighted Correction Results

Variables	Description	Robust Coefficient	Std. Err.	z	P>z	[95% Conf.	Interval]
RPEAKHR_121406	Peak hour of rainfall	16.0553	4.3921	3.66	0.000	7.4469	24.66369
RPEAK3HR_121406	Peak 3 hour rainfall	30.3374	4.66825	6.5	0.000	21.187	39.48702
N15_121406	Rainfall 4 to 18 days before event	34.5365	4.36361	7.91	0.000	25.984	43.08906
ln07TOTVAL	Log 2007 parcel value	-0.6844	0.10581	-6.47	0.000	-0.8917	-0.477
lnHHSQMI	Log households per square mile	-1.2787	0.23835	-5.36	0.000	-1.7458	-0.81155
SSAT_1214	Soil saturation	-339.76	42.5202	-7.99	0.000	-423.10	-256.426
AGE	Pipe age	-0.035	0.00574	-6.11	0.000	-0.0462	-0.0238
ELEV	Pipe elevation	-0.0167	0.00323	-5.17	0.000	-0.0229	-0.01035
PROBCOMD	Combined flow dummy variable	3.43629	0.67355	5.1	0.000	2.1161	4.756425
MNL_LENGTH	Pipe length	0.00073	0.00031	2.38	0.017	0.0001	0.001341
Constant	Intercept	-27.286	4.31422	-6.32	0.000	-35.747	-18.8298
Number of observations: 3553							

Table C: Results from rare events logistic regression using December 14th, 2006 storm data with the weighted correction specification

Looking at the difference in the results in Table C, the pipe length variable was retained in the weighted correction model but not in the prior correction model. The coefficients are slightly different but the common variable coefficients have the same signs and levels of significance. The weighted correction model estimates higher probabilities for the sample parcels and the differences in the coefficients allow for different estimates in expected sewer backup claim damage (ESBD). However, since the other two storms could not be estimated via weighted correction and would not be valid if the models were averaged, these results are not displayed.

Appendix D: Moran's I Results For The Sample Parcels



Global Moran's I Summary

Moran's Index:	-0.000968
Expected Index:	-0.000319
Variance:	0.000000
z-score:	-1.473970
p-value:	0.140490

Dataset Information

Input Feature Class:	RELR_Parcels_Points
Input Field:	CLAIM_Y

Conceptualization:	INVERSE_DISTANCE
Distance Method:	EUCLIDEAN
Row Standardization:	False
Distance Threshold.:	2000.0
Weights Matrix File:	None

Figure C: Moran's I Results for the December 14th, 2006 Paid Backup Claims and the Sample Parcels

Appendix E: Spatial Lag Model Code (in R)

```
#Start code

library(RODBC)

channel <- odbcConnectExcel("H:\\SEPOS_012413")

SEdata <- sqlFetch(channel, "SEPOS_012413")

odbcClose(channel)

SEdata$lnsqftlot <- log(SEdata$sqftlot)

SEdata$valsqft <- SEdata$lnYOLTOT/SEdata$lnsqftlot

install.packages(“ctv”)

library(“ctv”)

install.views(“Spatial”)

library(maptools)

library(rgdal)

library(spdep)

loc.sp = SpatialPoints(cbind(SEdata$X,SEdata$Y))

SE_nbq <- knn2nb(knearneigh(loc.sp,k=15,RANN=T))

SEdata_nbq_w<- nb2listw(SE_nbq)

SEdata_nbq_w

names(SEdata)

moran.test(SEdata$OBJECTID_1, listw=SEdata_nbq_w,alternative="two.sided")

moran.test(SEdata$LN_AMT, listw=SEdata_nbq_w,alternative="two.sided")

moran.plot(SEdata$LN_AMT, SEdata_nbq_w, labels=as.character(SEdata$LN_AMT),
xlab=”Log of Claim Amount”, ylab=”Spatially-Lagged Log Claim Amounts”)

names(SEdata)

Alm <-lm(LN_AMT~RINT + SSAT + L3PCT + N15PCT + grid_pct + lnHH1 + CCTVYN +
valsqft + lnsqftlot, data=SEdata)

Alm

summary(Alm)
```

```
SEdata$lmresid<-residuals(Alm)

moran.test(SEdata$lmresid, listw=SEdata_nbq_w,alternative="two.sided")

lm.LMtests(Alm,SEdata_nbq_w, test="all") #SARMA has lowest p-value

install.packages(“lmtest”)

library(lmtest)

bptest(Alm)

SElag<-lagsarlm(LN_AMT~RINT + SSAT + L3PCT + N15PCT + grid_pct + lnHH1 +
CCTVYN + valsqft + lnsqftlot, data=SEdata,SEdata_nbq_w)

SElag

summary(SElag)

bptest.sarlm(SElag)

SEerr<-errorsarlm(LN_AMT~RINT + SSAT + L3PCT + N15PCT + grid_pct + lnHH1 +
CCTVYN + valsqft + lnsqftlot, data=SEdata, SEdata_nbq_w)

SEerr

summary(SEerr)

bptest.sarlm(SEerr)

#end code
```

Appendix F: Rare Events Logistic Regression code (in Stata)

```
import excel "H:\DATASET_121406.xlsx", sheet("Sheet1") firstrow

logit Claim_Y RPK5MIN_121406 RPK10MIN_121406 RPEAKHR_121406
RPEAK3HR_121406 RL1_121406 RL2_121406 RL3_121406 N15_121406 RINT_12140
RSSAT_1214 grid_pct GRIDCODE ln12TOTVAL PPHH lnPPSQMI lnHHSQMI AGE DEPTH
ELEV MNL_LENGTH MNL_WIDTH_ MNL_HEIGHT MNL_SLOPE_ CLAYD
CONCRETED PROBSAND PROBCOMD, robust

outcome = RPK5MIN_121406 <= .0499563 predicts data perfectly

r(2000);

logit Claim_Y RPK10MIN_121406 RPEAKHR_121406 RPEAK3HR_121406 RL1_121406
RL2_121406 RL3_121406 N15_121406 RINT_12140 RSSAT_1214 grid_pct GRIDCODE
ln12TOTVAL PPHH lnPPSQMI lnHHSQMI AGE DEPTH ELEV MNL_LENGTH
MNL_WIDTH_ MNL_HEIGHT MNL_SLOPE_ CLAYD CONCRETED PROBSAND
PROBCOMD, robust /* removing 5 minute variable gave readout */

logit Claim_Y RPK10MIN_121406 RPEAKHR_121406 RPEAK3HR_121406 RL1_121406
RL2_121406 RL3_121406 N15_121406 RSSAT_1214 grid_pct GRIDCODE ln12TOTVAL
lnPPSQMI lnHHSQMI AGE ELEV MNL_LENGTH MNL_HEIGHT PROBCOMD, robust

logit Claim_Y RPK10MIN_121406 RPEAKHR_121406 RPEAK3HR_121406 RL1_121406
RL2_121406 N15_121406 grid_pct GRIDCODE ln12TOTVAL lnHHSQMI AGE ELEV
MNL_LENGTH MNL_HEIGHT PROBCOMD, robust

logit Claim_Y RPEAKHR_121406 N15_121406 ln12TOTVAL lnHHSQMI AGE ELEV
PROBCOMD, robust /* var with low p-values */

logit Claim_Y RPK10MIN_121406 RPEAKHR_121406 RPEAK3HR_121406 RL1_121406
RL2_121406 N15_121406 GRIDCODE N15_121406 ln12TOTVAL lnHHSQMI AGE ELEV
MNL_HEIGHT MNL_LENGTH PROBCOMD, robust

logit Claim_Y RPK10MIN_121406 RPEAKHR_121406 RPEAK3HR_121406 RL1_121406
RL2_121406 N15_121406 GRIDCODE ln12TOTVAL lnHHSQMI AGE ELEV
MNL_WIDTH_ MNL_LENGTH PROBCOMD, robust

logit Claim_Y RPK10MIN_121406 RPEAKHR_121406 RPEAK3HR_121406 RL1_121406
RL2_121406 N15_121406 GRIDCODE ln12TOTVAL lnHHSQMI AGE ELEV PROBCOMD
CLAYD CONCRETED , robust
```

logit Claim_Y RPK10MIN_121406 RPEAKHR_121406 RPEAK3HR_121406 RL1_121406 RL2_121406 N15_121406 GRIDCODE ln12TOTVAL lnHHSQMI AGE ELEV PROBCOMD CLAYD CONCRETED RPK5MIN_121406 , robust

logit Claim_Y RPK10MIN_121406 RPEAKHR_121406 RPEAK3HR_121406 RL2_121406 N15_121406 SSAT_121406 GRIDCODE ln12TOTVAL lnHHSQMI AGE ELEV PROBCOMD, robust

relogit Claim_Y RPK10MIN_121406 RPEAKHR_121406 RPEAK3HR_121406 RL1_121406 RL2_121406 N15_121406 GRIDCODE ln12TOTVAL lnHHSQMI AGE ELEV PROBCOMD grid_pct MNL_WIDTH_ MNL_LENGTH, wc(0.00083)

relogit Claim_Y RPEAKHR_121406 RPEAK3HR_121406 RL1_121406 RL2_121406 N15_121406 GRIDCODE ln12TOTVAL lnHHSQMI AGE ELEV PROBCOMD grid_pct MNL_WIDTH_ MNL_LENGTH, wc(0.00083)

relogit Claim_Y RPEAKHR_121406 RPEAK3HR_121406 N15_121406 GRIDCODE ln12TOTVAL lnHHSQMI AGE ELEV PROBCOMD grid_pct MNL_WIDTH_ MNL_LENGTH, wc(0.00083)

relogit Claim_Y RPK10MIN_121406 RPEAKHR_121406 RPEAK3HR_121406 N15_121406 GRIDCODE ln12TOTVAL lnHHSQMI AGE ELEV PROBCOMD grid_pct MNL_WIDTH_ MNL_LENGTH, wc(0.00083)

relogit Claim_Y RPK10MIN_121406 RPEAKHR_121406 RPEAK3HR_121406 N15_121406 GRIDCODE ln12TOTVAL lnHHSQMI AGE ELEV PROBCOMD grid_pct MNL_WIDTH_ MNL_LENGTH, wc(0.00083)

relogit Claim_Y PCT10MIN PCT1HR RPK10MIN_121406 RPEAKHR_121406 RPEAK3HR_121406 N15_121406 GRIDCODE ln12TOTVAL lnHHSQMI AGE ELEV PROBCOMD grid_pct MNL_WIDTH_ MNL_LENGTH, wc(0.00083)

relogit Claim_Y PCT10MIN PCT1HR RPEAK3HR_121406 N15_121406 GRIDCODE ln12TOTVAL lnHHSQMI AGE ELEV PROBCOMD grid_pct MNL_WIDTH_ MNL_LENGTH, wc(0.00083)

relogit Claim_Y RPEAKHR_121406 RPEAK3HR_121406 RSSAT_1214 RL3_121406 N15_121406 GRIDCODE ln12TOTVAL lnHHSQMI AGE ELEV PROBCOMD grid_pct MNL_WIDTH_ MNL_LENGTH, wc(0.00083)

relogit Claim_Y RPEAKHR_121406 RPEAK3HR_121406 RSSAT_1214 N15_121406 GRIDCODE ln12TOTVAL lnHHSQMI AGE ELEV PROBCOMD grid_pct MNL_WIDTH_ MNL_LENGTH, wc(0.00083)

relogit Claim_Y RPEAKHR_121406 RPEAK3HR_121406 RSSAT_1214 N15_121406 GRIDCODE ln12TOTVAL lnHHSQMI AGE ELEV PROBCOMD MNL_WIDTH_ MNL_LENGTH, wc(0.00083)

relogit Claim_Y RPEAKHR_121406 RPEAK3HR_121406 RL3_121406 N15_121406 GRIDCODE ln12TOTVAL lnHHSQMI AGE ELEV PROBCOMD MNL_WIDTH_ MNL_LENGTH, wc(0.00083)

relogit Claim_Y RPEAKHR_121406 RPEAK3HR_121406 RL3_121406 N15_121406 ln12TOTVAL lnHHSQMI AGE ELEV PROBCOMD MNL_WIDTH_ MNL_LENGTH, wc(0.00083)

relogit Claim_Y RPK10MIN_121406 RPEAKHR_121406 RPEAK3HR_121406 N15_121406 GRIDCODE ln12TOTVAL lnHHSQMI AGE ELEV PROBCOMD MNL_WIDTH_ MNL_LENGTH, wc(0.00083) /* preferred relogit model for 121406 */

relogit Claim_Y RPK10MIN_121406 RPEAKHR_121406 RPEAK3HR_121406 N15_121406 GRIDCODE ln08TOTVAL lnHHSQMI AGE ELEV PROBCOMD MNL_WIDTH_ MNL_LENGTH, wc(0.00083) /* preferred relogit model for 121406 */

relogit Claim_Y RPEAKHR_121406 RPEAK3HR_121406 N15_121406 ln08TOTVAL lnHHSQMI SSAT_1214 AGE ELEV PROBCOMD MNL_LENGTH, wc(0.00083)

relogit Claim_Y RPEAKHR_121406 RPEAK3HR_121406 N15_121406 ln07TOTVAL lnHHSQMI SSAT_1214 AGE ELEV PROBCOMD MNL_LENGTH, wc(0.00083)

relogit Claim_Y RPEAKHR_121406 RPEAK3HR_121406 N15_121406 ln07TOTVAL lnHHSQMI SSAT_1214 AGE ELEV PROBCOMD MNL_LENGTH, pc(0.00083)

relogit Claim_Y RPEAKHR_121406 RPEAK3HR_121406 N15_121406 ln07TOTVAL lnHHSQMI SSAT_1214 AGE ELEV PROBCOMD, pc(0.00083)

setx (RPEAKHR_121406 RPEAK3HR_121406 N15_121406 ln07TOTVAL lnHHSQMI SSAT_1214 AGE ELEV PROBCOMD) mean

setx

relogitq, fd(pr) changex(RPEAKHR_121406 .5418274 .59601014 & RPEAK3HR_121406 .8692247 .95614717 & N15_121406 1.305519 1.43607 & SSAT_1214 .1278081 .140589)

/* 10% increase in all rainfall variables */

relogitq, fd(pr) changex(RPEAKHR_121406 .5418274 .65019 & RPEAK3HR_121406 .8692247 1.04307 & N15_121406 1.305519 1.56662 & SSAT_1214 .1278081 .15337)

/* 20% increase in all rainfall variables */

relogitq, fd(pr) changex(RPEAKHR_121406 .5418274 .70438 & RPEAK3HR_121406 .8692247 1.12999 & N15_121406 1.305519 1.69717 & SSAT_1214 .1278081 .16615)

/* 30% increase in all rainfall variables */

relogitq, fd(pr) changex(RPEAKHR_121406 .5418274 .48764 & RPEAK3HR_121406 .8692247 .7823 & N15_121406 1.305519 1.17497 & SSAT_1214 .1278081 .11503)

/* 10% reduction in all rainfall variables */

relogitq, fd(pr) changex(RPEAKHR_121406 .5418274 .43346 & RPEAK3HR_121406 .8692247 .69538 & N15_121406 1.305519 1.04442 & SSAT_1214 .1278081 .10225)

/* 20% reduction in all rainfall variables */

relogitq, fd(pr) changex(RPEAKHR_121406 .5418274 .37928 & RPEAK3HR_121406 .8692247 .60846 & N15_121406 1.305519 0.91386 & SSAT_1214 .1278081 .08947)

/* 30% reduction in all rainfall variables */

relogitq, rr(RPEAKHR_121406 .5418274 .59601014 & RPEAK3HR_121406 .8692247 .95614717 & N15_121406 1.305519 1.43607 & SSAT_1214 .1278081 .140589)

/* 10% increase in all rainfall variables */

relogitq, rr(RPEAKHR_121406 .5418274 .65019 & RPEAK3HR_121406 .8692247 1.04307 & N15_121406 1.305519 1.56662 & SSAT_1214 .1278081 .15337)

/* 20% increase in all rainfall variables */

relogitq, rr(RPEAKHR_121406 .5418274 .70438 & RPEAK3HR_121406 .8692247 1.12999 & N15_121406 1.305519 1.69717 & SSAT_1214 .1278081 .16615)

/* 30% increase in all rainfall variables */

relogitq, rr(RPEAKHR_121406 .5418274 .48764 & RPEAK3HR_121406 .8692247 .7823 & N15_121406 1.305519 1.17497 & SSAT_1214 .1278081 .11503)

/* 10% reduction in all rainfall variables */

relogitq, rr(RPEAKHR_121406 .5418274 .43346 & RPEAK3HR_121406 .8692247 .69538 & N15_121406 1.305519 1.04442 & SSAT_1214 .1278081 .10225)

/* 20% reduction in all rainfall variables */

relogitq, rr(RPEAKHR_121406 .5418274 .37928 & RPEAK3HR_121406 .8692247 .60846 & N15_121406 1.305519 0.91386 & SSAT_1214 .1278081 .08947)

/* 30% reduction in all rainfall variables */