Soil Lead Contamination from the Exide Battery Smelter: The Role of Spatial Scale in Cleanup Efforts

by

Monica A. Finnstrom

A Thesis Presented to the
Faculty of the USC Graduate School
University of Southern California
In Partial Fulfillment of the
Requirements for the Degree
Master of Science
(Geographic Information Science and Technology)

August 2019

To my parents, Richard “Rick” A. Finnstrom and Alice M. Finnstrom, for always supporting me.

&

In Loving Memory of my two Grandpas:

Victor P. Avila and Richard “Dick” Finnstrom

Table of Contents

List of Figures

List of Tables

Acknowledgments

I am so grateful for my advisor, Dr. Karen Kemp, for providing me with valuable direction and assistance. I appreciate all of her time and effort in helping me produce this thesis and am thankful for all of those extended Blue Jeans meetings. I would also like to express my gratitude to my committee members, Dr. An-Min Wu and Dr. Su Jin Lee, for providing me with valuable insight and feedback. Without the inspiration for the topic and early guidance from Dr. Wu, this thesis would not have come to fruition. I am also fortunate in having had the opportunity to work with Dr. Lee as an Undergraduate Researcher for two years, helping me cultivate my interest in GIS. I consider Dr. Kemp, Dr. Wu, and Dr. Lee as mentors and am grateful for all the opportunities SSI has afforded me.

I would also like to acknowledge Dr. Jill Johnston for providing her expertise and valuable insight into the subject matter. I would like to recognize the California Department of Toxic Substances Control for making the data publicly available for use, which provided the backbone for this thesis.

Most importantly, I want to thank my parents for being so supportive, encouraging, and patient throughout this entire process. They are my true inspiration and their love and support has motivated me to constantly be and do my best.

List of Abbreviations

AIN	Assessor Identification Number
APN	Assessor's Parcel Number
ASE	Average Standard Error
bgs	Below Ground Surface
BLL	Blood Lead Level
CDC	Centers for Disease Control
DTSC	Department of Toxic Substances Control
Exide	Exide Technologies
EBK	Empirical Bayesian Kriging
GIS	Geographic Information System
HQ	Hazard Quotient
IQ	Intelligence Quotient
Mg/kg	Milligram per kilogram
Pb	Lead
PIA	Preliminary Investigation Area
Ppm	Parts per million
RMS	Root Mean Square
SSI	Spatial Sciences Institute
UCL	Upper Confidence Limit
USC	University of Southern California
USEPA	United States Environmental Protection Agency
XRF	X-Ray Fluorescence

Abstract

Lead is a significant health threat to people, especially for children where elevated absorption of lead into the bloodstream can cause permanent damage. One site for concern of lead exposure is the surrounding communities of the retired Exide Technologies lead-acid battery smelter in Vernon, California. The California Department of Toxic Substances Control (DTSC) is leading an extensive cleanup effort to remove lead-contaminated soil from affected residences and eliminate the negative health risks posed by the contamination. Soil sampling conducted for approximately 8,500 parcels serves as the primary dataset for this research. While DTSC is currently undertaking the cleanup process on a parcel-by-parcel basis, this thesis works toward understanding the effect of geographic scale in the estimation of levels of lead contamination. It also offers alternatives for identifying priority areas for cleanup by using various aggregation methods and examining how the resulting values may be affected by scale. This research used Empirical Bayesian Kriging to produce interpolated surfaces of lead concentration values. Various aggregation methods were then utilized to aggregate the surfaces into easily defined geographical units of different scales, including block groups, blocks, and parcels. The resulting aggregation values include the mean, percent area, and a Hazard Quotient, an index value for determining health risk. The results demonstrate that the larger areas of the block groups moderate high lead concentration values and thus have lower overall aggregation values for the block groups. In contrast, blocks have a greater tendency to include these high lead concentrations in the aggregations resulting in higher overall values and wider ranges of values for the blocks. This research provides alternative approaches for prioritizing the cleanup of contaminated sites that could be more effective to address the health risks associated with contamination and can be applied to other areas faced with the same problem in the future.

Chapter 1 Introduction

Lead (Pb) poses a significant environmental health risk to people. Children are especially vulnerable because of their ability to absorb more ingested lead, which can lead to permanent neurologic and developmental disorders (Wu et al. 2010). The recently closed Exide Technologies lead-acid battery smelter in southeast Los Angeles has demonstrated a significant threat of lead exposure to the community, with many possible health risks affecting the community.

This 15-acre facility was one of only two west of the Rocky Mountains that could melt lead from old car batteries for use in the production of new ones and had operated since 1922, with Exide taking over the facility in 2000 (Barboza 2015). Smelter operations in violation of numerous environmental regulations resulted in the release of lead, arsenic, and other heavy metals into the local environment through aerial release. Violations included lead and acid leaks, large cracks in the floors, hazardous levels of lead in the soil outside, and an overflowing pond of toxic sludge (Barboza 2015).

Despite local, state, and federal officials citing the plant of these violations several times over the years, the plant was allowed to remain open. In addition, the state allowed this smelter to operate for 33 years without a full permit. The California Department of Toxic Substances Control (DTSC) had known of these violations as well but failed to correct them. It was only when the company was threatened with a serious federal investigation in 2014 that the company was forced to sign an agreement to permanently close in 2015 (Barboza 2015).

Currently, the smelter site has become one of the most extensive cleanups of its kind by targeting the removal of lead-contaminated soil from thousands of homes within an approximate 1.7-mile radius of the facility. The California Department of Toxic Substances Control is now

acting as the primary agency overseeing the cleanup efforts. The agency has established a Preliminary Investigation Area (PIA) where soil sampling was conducted at approximately 8,500 properties within the community for heavy metal analysis, with an emphasis on the lead concentration results (Figure 1). Figure 2 displays the PIA boundary and its regional location. Both figures are directly sourced from DTSC's Final Environmental Impact Report.

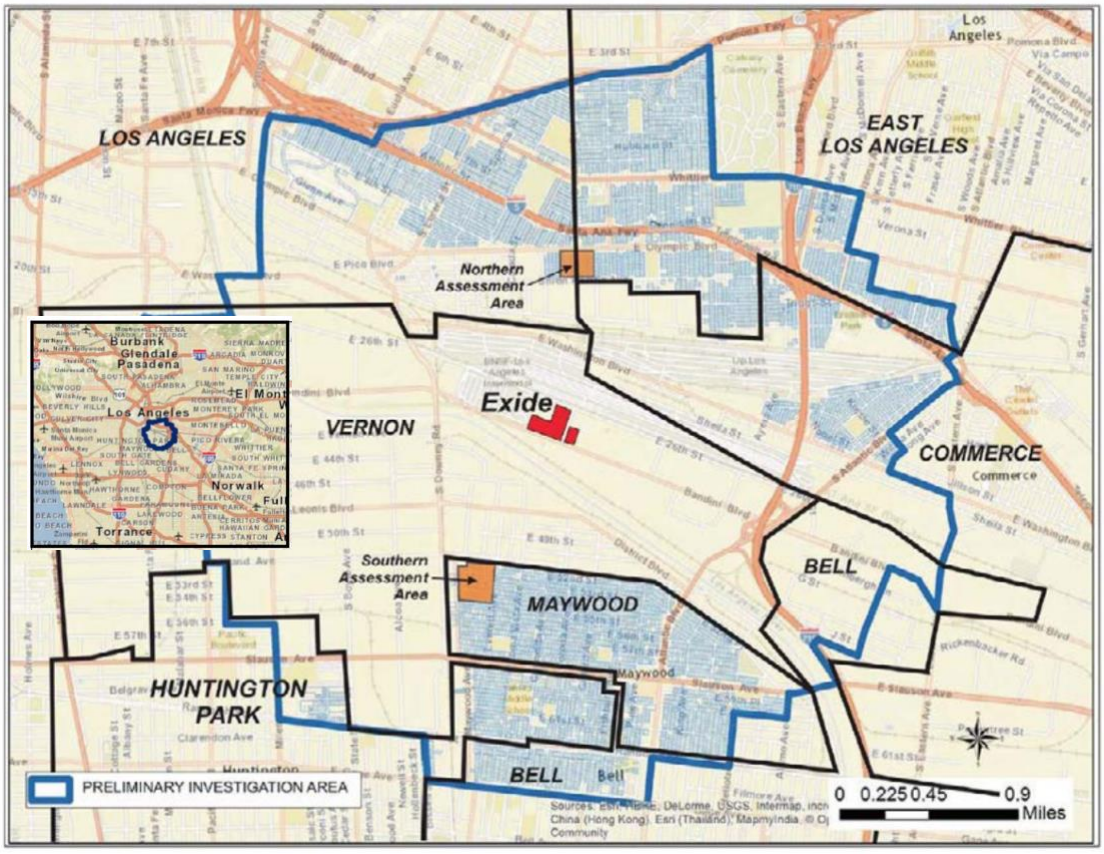


Figure 1 Preliminary Investigation Area (PIA) determined by DTSC for the Cleanup Plan of soil lead contamination (Source: directly from the Final Environmental Impact Report DTSC 2017, S1-5)

Although a majority of the properties are determined to have lead concentrations that exceed California's standard for residential soil (80 ppm), current funding limits the cleanup efforts to only 2,500 of the impacted properties. The most recent Cleanup Plan prioritizes the following properties for cleanup (DTSC 2017, ES-4):

- "Residential properties with a representative soil lead concentration [95% upper confidence limit] of 400 ppm or higher; and
- Residential properties with a representative soil lead concentration of less than 400 ppm, but where any soil sampling result of 1,000 ppm or higher is detected; and
- Daycare and child care centers with a representative soil lead concentration of 80 ppm or higher that have not yet been cleaned up."

While these values used for prioritization provide sufficient measures of risk for lead exposure, the parcel-by-parcel approach taken is inefficient. The Exide facility has been closed for three years now. However, only 496 total parcels have been cleaned as of October 2018, with 166 of these properties being in the recent July 2017 Cleanup Plan. Meanwhile, residents have grown frustrated with the slow pace of cleanup and have limited their children's playtime outside, as they are fearful for their children's exposure to lead (Barboza 2018). Community groups and county health officials have become critical of the Cleanup Plan and believe the continuation of the cleanup on a parcel-by-parcel basis is insufficient, leaving contaminated properties still intermixed with cleaned properties (Barboza 2018). This thesis attempts to understand the effect of geographic scale in estimating the level of lead contamination and offers other approaches for identifying priority areas for cleanup that could be more effective than a parcel-by-parcel approach.

The following sections discuss important background information on lead's impact on human health, soil as a source of lead exposure, further context on Exide, and concludes with the research goals and organization of this thesis.

1.1. Lead's Impact on Human Health

With sufficient exposure, lead can wreak havoc on human health, exerting severe and chronic health effects. Blood-lead levels (BLLs) typically reported as micrograms of lead per deciliter of blood (μg/dL), is considered to be one of the primary biomarkers for lead exposure

(Juberg, Kleiman, and Kwon 1997). While adults absorb roughly 5 to 15% of ingested lead, only retaining less than 5% of these lead values, children can absorb approximately 30 to 40% of ingested lead because of metabolic and physiological differences. After absorption in the blood, lead is "distributed primarily among three compartments – blood, soft tissue (kidney, bone marrow, liver, and brain), and mineral tissues (bone and teeth)" (Juberg, Kleiman and Kwon 1997, 167). Lead initially absorbed by the bloodstream has the potential to be stored in bone for years, which then becomes a long-term source of Pb back into the bloodstream (Laidlaw et al. 2017). Initial research has determined an association between elevated Pb in children and lower Intelligence Quotient (IQ), behavioral problems, and learning disorders. The impacts are also age dependent, with Pb presence in blood interfering with proper neuron formation in young children. Recent research provides a greater understanding of chronic health effects with lead exposure. There is evidence now that specifies strong associations between lead and many diseases including "motor neuron disease, autism, preeclampsia developmental delays in children, heart disease, ADHD, dementia, mental illness, and brain cancer" (Laidlaw et al. 2017, 16). With such negative health risks present from lead exposure, it is vital that DTSC address the health risks posed by the Exide contamination in a timely and effective manner.

1.2. Soil as a Source of Lead Exposure

Although lead occurs naturally in the soil at concentrations ranging from 10 to 50 mg/kg, human influence has caused lead to become more prominent in soils, especially in urban areas (Stehouwer 2010). The widespread use of lead-based paint before the mid-1970s and the use of lead additives in gasoline before the mid-1980s have contributed as major sources of lead in soil. The peak of lead-based paint usage happened in the 1920s, while the peak of leaded gasoline usage occurred in the early 1970s (Mielke and Reagan 1998). It is now estimated that leaded

gasoline left a residue of roughly 4 to 5 million metric tons of Pb in the environment (Mielke and Reagan 1998).

Banned in 1978, lead-based paint is estimated to have been used on approximately 75% of houses built before then (Stehouwer 2010). Although less widespread, airborne lead from industrial sources such as smelters has significant potential to contaminate nearby residential soils. Due to these sources, urban soils tend to have higher soil lead concentrations than normal background levels or soils in rural areas (Markus and McBratney 2001; McClintock 2012; Stehouwer 2010; Mielke and Reagan 1998). These urban lead concentrations can range from 150 mg/kg to even 10,000 mg/kg, if at the base of a home with lead-based paint (Stehouwer 2010).

1.2.1. How Lead Behaves in Soil

According to Chaney, Mielke, and Sterrett (1989, 2), research has "repeatedly shown that small Pb-rich particles reaching the surface of a soil profile largely remain on or near the surface for a prolonged period." This is because organic matter particles and very fine clay hold onto soil lead very tightly, in which the lead typically accumulates in the upper 1-2 inches of soil (Stehouwer 2010). However, the mixing of soil by humans for the health of plants and the lawn and the movement of soil by creatures such as earthworms allow lead to penetrate deeper into the soil strata, which can make it more difficult to clean up. In addition, lead becomes most concentrated in very fine soil particles. These particles tend to form airborne soil dust and can stick to clothing and even skin.

Furthermore, chemical factors play an important role in the bioavailability of lead, as not all of the lead is readily absorbed by plants or human bodies. The availability of lead in soil largely depends on two factors: solubility (how much the lead dissolves in water) and how tightly it is held by soil particles. Lead is more soluble and held less tightly in acidic conditions

(pH < 5), while less soluble and held more strongly in neutral (pH 5-6.5) to basic conditions (pH > 6.5) (Stehouwer 2010). Since lead is held tightly by soil organic matter, lead availability decreases as organic matter increases. The bioavailability of lead also has important implications within the human body. Adsorbed Pb becomes soluble within the human stomach, due to the acidic environment. When it enters the small intestine, the pH rises, in which soil then adsorbs the Pb and reduces its solubility. Due to this adsorption equilibrium process, "higher soil Pb concentration should strongly increase the bioavailability of soil Pb" (Chaney, Mielke and Sterrett 1989, 1).

1.2.2. Exposure to Lead from Soil

In the past decades, researchers disagreed on soil being an important pathway for lead exposure to humans, as some researchers argued that lead-based paint was the most important source of lead exposure, as described by Mielke and Reagan (1998). However, further research now proves soil to be a relatively significant source for lead exposure and must be considered in order to have effective primary lead prevention (Mielke and Reagan 1998).

Pathways of exposure commonly emphasized for humans, especially children, are that of ingestion and inhalation, with ingestion as the primary pathway (Laidlaw et al. 2017). Pb exposure from soil can occur through various forms including "track-in via soil particles attached to shoes, pets tracking Pb indoors on their fur, and direct contact with soils when the weather is favorable and children play outdoors" (Laidlaw et al. 2017, 15). Younger children are also known to ingest more dirt than adults relative to their body mass. Since children can absorb greater amounts of lead than adults, Pb exposure from soil is deemed particularly dangerous for this vulnerable population.

In studies done after the use of leaded gasoline was prohibited, it was found that blood Pb levels declined concurrently with the decrease of air lead from the gasoline sources (Laidlaw et al. 2017). Similarly, child blood Pb levels declined sharply after the Bunker Hill Pb smelter in Idaho was closed and Pb aerosols therefore ceased. These studies indicate the importance of inhalation as a pathway for exposure and not to be overlooked.

A small, yet growing number of studies have examined the spatial relationships between soil lead and BLLs. One such study in New Orleans utilized 5,467 soil Pb samples spanning 286 census tracts and geo-referenced BLL data for 55,551 children in New Orleans (Laidlaw et al. 2017). The results of the study indicated that the BLLs of the children are spatially associated with the soil Pb levels, in which 67% of the variation in children's BLL could be explained by soil Pb sample location variables.

1.2.3. Soil Pb Standards and BLL Guidelines

Having established soil as an important source for lead exposure, what soil Pb standards and BLL guidelines are in existence? Current soil standards are inadequately developed, ranging from 20 mg/kg to over 1,000 mg/kg (Laidlaw et al. 2017). However, the most widely cited soil Pb standard that a majority of guidelines are aligned with is the U.S. EPA 400 mg/kg value, which is not a health-based standard. As mentioned by Laidlaw (2017), several studies and reviews suggest health-based soil Pb guidelines, correlating to target levels of BLL. A prominent review proposed a standard of <100 mg/kg centered on evidenced-based data from studies and the assumption that 10 μg/dL BLL is safe (this is now considered to be still not safe according to clinical studies) (Laidlaw et al. 2017). A New Orleans study in 1999 determined a soil Pb guideline of around 80 mg/kg, with the goal of preventing Pb exposure ≥ 10 μg/dL for children. A repeat of this study ten years after Hurricane Katrina revised this soil Pb standard value to ≤ 40

mg/kg, with the goal of preventing Pb exposure ≥ 5 μg/dL for children (a version of a revised CDC BLL reference value).

According to the California Code of Regulations, lead contaminated soil is defined as "… bare soil that contains an amount of lead equal to, or in excess of, four hundred parts per million (400 ppm) in children's play areas and one thousand parts per million (1,000 ppm) in all other areas" (DTSC 2017, S2-24). DTSC's screening level for lead in residential properties is 80ppm, based off a revised toxicity evaluation for lead that now determines the threshold for children's blood level concern as 1 μg/dL, as opposed to the previous 5 μg/dL (DTSC 2017). Because children are the most sensitive group to lead exposure and are the group used for determining these standards, the places where they spend most of their time present the greatest risk for exposure, which can be considered as sensitive land uses. A majority of the parcels within the Preliminary Investigation Area are residential, parks, schools, child care facilities, and day care centers, all of which are sensitive land uses and should adhere to the guidelines.

1.3. Exide Facility Site Characterization and Background Information

To better understand the nature of the site and critical background information, this section provides a description of the conceptual site model as laid out by DTSC in their Final Removal Action Plan (Cleanup Plan) and describes a preliminary background lead levels study performed by Exide. The conceptual site model aids in planning efforts by determining exposure pathways from the facility. The preliminary background lead levels study, carried out by Exide and enforced by DTSC, ensures that the contribution of background lead concentrations have been considered and indicates Exide as a primary source for the lead concentrations.

1.3.1. Conceptual Site Model

Conceptual site models are important planning tools utilized to support the decision-making process of impacted properties, such as that of the lead-contaminated properties in the PIA (DTSC 2017). A conceptual site model provides insight into contaminant sources and release mechanisms by describing the potential movement of chemicals throughout a particular area and its exposure pathways to potential affected populations. In the case of the Exide site, lead-acid battery activities from the years of operation, inadequate waste management practices, and insufficient air pollution control are the likely causes for the lead contamination on affected properties (DTSC 2017). The primary sources for this contamination include aerial releases from the smokestacks and facility in general, as well as trucks transporting the materials (Figure 2).

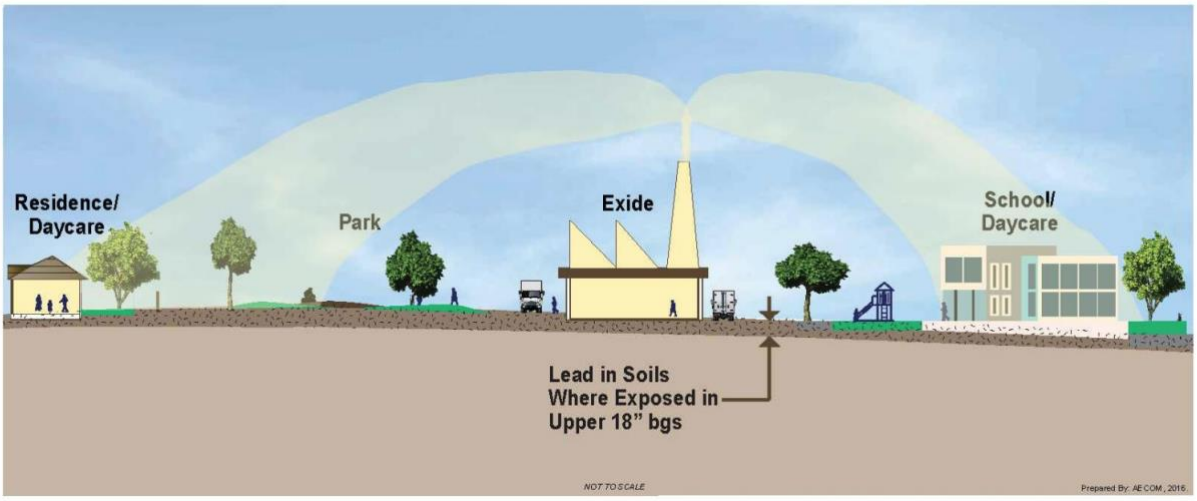


Figure 2 Pictorial Conceptual Site Model for Exide (Source: DTSC 2017, S2-19). Note: "bgs" stands for below ground surface.

Within the conceptual site model, it is acknowledged that other lead sources are also possible, including lead-based paint, leaded gasoline, and historical industrial operations from other facilities operating in the same industrial area. The key pathways to exposure identified include particulate inhalation, dermal contact, and accidental ingestion. The main population identified for lead exposure is the residents within the PIA, including sensitive individuals such

as children and pregnant women. The most likely source of ongoing exposure involving lead from the Exide facility during years of operation is surface soil and soil or dust around plants.

1.3.2. Preliminary Background Lead Levels Study

DTSC had Exide do a background lead study in 2014 to ensure that the selected soil screening level of 80 ppm for cleanup did not fall below background, or typical, levels of lead in soil within the area. For this background study, Exide chose a background area located roughly 14 miles south of the Exide facility in the City of Long Beach to serve as a control for the soil samples. The area was "residential and considered similar to the PIA, but without potential lead contamination from the former Exide facility, because of its proximity to major freeways, a historically industrial area, a sizable rail yard with intermodal facility and switching yard, and housing of similar size and density" (DTSC 2017, 25-26). As a part of the study, 19 residential properties were sampled. According to the results, surface lead concentrations ranged from 29 to 195 ppm, with a median value of 54.8 ppm and a representative soil lead concentration of 76.6 ppm. The representative soil lead concentration is a property-wide lead concentration calculated from the soil sample results and is more health protective than an average of the soil sample concentrations. This term is further explained in Chapter 3. These concentrations are significantly lower than the lead concentrations sampled from soil in the Initial Assessment Areas and are below the DTSC screening level for lead. These conclusions mean that the representative soil lead concentrations within the PIA are unlikely attributable to background concentrations; rather, they indicate Exide as a primary source for the lead concentrations.

1.4. Research Goals

This thesis works toward understanding the effect of geographic scale in estimating levels of lead contamination to help determine the effectiveness of using a scaled-up approach for directing the decision-making process for cleanup efforts. Currently, DTSC is undertaking the cleanup process on a parcel-by-parcel basis, with only parcels where samples exceed defined soil lead concentration limits are chosen for cleanup. Since the distribution of lead concentrations in the soil is a continuously varying surface, this research approaches the problem spatially. It considers alternative ways to delineate areas for cleanup based on estimating single summary values from an estimated surface of varying lead concentrations in the soil for various sized spatial units. If larger areas can be identified as having sufficiently higher lead concentrations than other parts of the study area, it could help decision-makers determine more efficient ways to identify priority areas for cleanup.

Thus, the research aims to answer these questions:

1. Based on the available sampled data, what is the continuous spatial distribution of lead concentrations in the soil within the Exide Preliminary Investigation Area?
2. How does scale (i.e. size of the spatial zones) affect the values allocated to various sized zones that can be used to identify priority areas for cleanup of lead contaminated soils?

While DTSC is using a parcel level representative soil lead concentration to determine priority areas for cleanup, this thesis aims to provide alternatives for identifying priority areas through the use of numeric aggregation methods. These aggregation methods provide other estimates of lead concentration values on which to base the decision for priority areas of cleanup, including the statistical mean of values for an area, the percentage of an area where lead

concentration values exceed the national standard of lead in soil concentration of 400 ppm, and a Hazard Quotient, an index value that indicates the level of risk to human health. In addition, this thesis aims to look at how scale affects these values when assigning them to different geographical units of differing scales. These geographical units include block groups, blocks, and parcels, those typically utilized in policy implementation. When looking into how scale affects these values, the results provide insight into which scale may be best for managing cleanup of lead contaminated soils. Although DTSC may have already decided how they are approaching the cleanup for Exide, this research provides other approaches for prioritizing cleaning up of contaminated sites that could be applied to other areas faced with the same problem in the future.

1.5. Study Organization

This study includes four additional chapters. Chapter Two provides a literature review regarding approaches to understanding soil contamination and highlights studies that use geostatistical approaches. This chapter also presents an overview into how scale affects spatial analysis and examines various studies that inform the methodologies for this thesis. Chapter Three provides a detailed description of the soil sample dataset used in this study and presents the methodology, which includes the use of Empirical Bayesian Kriging to create predictive surfaces of lead concentrations followed by various aggregation methods used to aggregate the surface results into the different scales of analysis. Chapter Four provides the results of the analysis. Chapter Five concludes with further discussion of the results and their implications, limitations of the study, and suggestions for further research.

Chapter 2 Related Work

Studies utilizing GIS and spatial analysis to examine soil contamination of heavy metals, including lead, have increased through the years. This section starts by discussing literature that has taken various approaches to understanding soil contamination, with a focus on geostatistical methods. The next part of this section examines scale by looking at literature concerning how scale affects spatial analysis. Different aggregation methods for the results of geostatistical processes into different units of analysis are explored as well. This determination requires examining literature from different fields of study, those other than contamination studies, to look for existing frameworks that can be applied to soil contamination.

2.1. Geostatistical Approaches to Understanding Soil Contamination

The use of GIS and geostatistical approaches are increasingly being used to study soil contamination of heavy metals, including lead. Mapping contaminant distribution, such as lead, makes detecting patterns within these distributions easier and allows for the identification of areas that contain potentially hazardous concentrations by showing how the contaminant changes with space (Markus and McBratney 2001). Knowledge of the spatial distribution of a contaminant is an essential factor in site assessment and for any succeeding risk assessment. Due to expensive costs of large sampling efforts and chemical analyses that follow, a technique known as interpolation has been used to estimate concentrations at locations between sampling sites. In particular, Kriging, a geostatistical method of interpolation, has become popular within studies to map the distribution of lead and other heavy metal concentrations in soil (Markus and McBratney 2001). In addition to providing a predicted surface of concentration values, Kriging also provides other useful information, including a measure of uncertainty associated with the

predicted values. This makes Kriging a desirable method for interpolation in soil contamination studies.

The following studies provide examples of how Kriging was used in understanding soil contamination. One study examined soil metal contamination of Cd, Cr, Cu, Ni, Pb, and Zn in the highly urbanized Kowloon area of Hong Kong, generating geochemical maps via a Kriging method that shows hot-spots of heavy metal contamination in soils (Li et al. 2004). The researchers used Principle Component Analysis and Cluster Analysis to determine significant spatial relationships for the metals. Further overlay analysis, which included comparing the proximity of hot-spot results to other features, concluded that road junctions, major roads, and industrial buildings were possible sources for the metals in urban soils.

Another study that utilizes geostatistics focused on the soil heavy metal concentrations in the rice paddy fields in the Hangzhou-Jiaxing-Huzhou (HJH) Plain (Liu, Wu and Xu 2006). The study employed ordinary Kriging and lognormal Kriging to map the spatial patterns of heavy metals including Cu, Zn, Pb, Cr, and Cd. Using 450 soil samples spread through the study area, contour maps of the heavy metals were produced. The authors use these maps of the spatial distributions as a way to quantify risk assessment. They argue that for a variety of practical problems in environmental management concerning heavy metals in soil and their relative threshold values, information is needed at unsampled sites (Liu, Wu and Xu 2006). The resulting maps produced using the geostatistical methods serve as a solution and inform risk assessment of environmental pollution and associated decision-making.

Kriging has also been used as the geostatistical mapping method in studies that center on smelter-impacted soils. For example, multivariate and geostatistical analysis was used to investigate the spatial variation of heavy metals in the soils of a mining-smelting area in the

Hunan Province of China (Wei, Wang and Yang 2009). Ordinary Kriging for As, Cd, Pb, and Zn and Inverse Distance Weighting for Cr and Cu were performed, demonstrating hot spots and similar dispersion patterns that indicated the smelter and mining source areas.

In another study concerning a former coal-mining area in France, ordinary Kriging was also used in its investigation into the spatial variability of the pseudototal concentrations of Cd, Pb, and Zn (Pelfrêne, Détriché and Douay 2014). The researchers used exploratory spatial data analysis to characterize data distribution to aid in the Kriging process. Unlike the previous studies, this one combined its geostatistical analyses with the incorporation of oral bioaccessibility to improve the assessment of the population's exposure to metals in the smelter-impacted soils. As exemplified in these studies, the use of Kriging has become a primary approach in understanding the distribution of soil contamination, with ordinary Kriging proving to be the most commonly used type of Kriging.

In regard to this thesis, Kriging serves as the main method for analyzing the spatial distribution of lead in the soil of the PIA. Although Kriging is widely used in soil contamination studies, the ease of implementing Kriging within a GIS environment, without the full understanding of its various factors, can often produce unreliable and possibly misleading results (Oliver and Webster 2014). Users must therefore take caution when carrying out the Kriging process, making sure they understand all of the important data considerations. Sound Kriging requires a plausible function for the variogram. A paper by Oliver and Webster provides guidelines for choosing suitable functions, among other important considerations with respect to Kriging. This paper, as well as other documentation, are key in guiding this project's appropriate use of Kriging.

In addition to Kriging, a variety of other methods are used to characterize the spatial distribution of lead in soils and were informative for this thesis. One study investigated the process of identifying pollution hotspots using Pb concentrations in the urban soils of Galway, Ireland and determining which factors influence hot spot identification (Zhang et al. 2008). The researchers in this study used local Moran's I-test to identify pollution hotspots of Pb contamination, classifying them into spatial clusters and outliers. Factors influencing the determination of results included definition of weight function, data transformation, and the existence of extreme values.

In another study, researchers analyzed the bioavailable soil lead concentrations in Los Angeles. They sought to understand the contribution of lead to soil from residential lead-based paint and from cars using leaded-gasoline based on traffic variables (Wu et al. 2010). Utilizing several variables including house age, traffic index, proximity to freeways and highways, and land use patterns, the researchers used multi-variable regression models to explain soil lead concentrations and mapped the spatial distribution of these concentrations. This study provides a good example of understanding lead contamination in soils in Los Angeles. Although the lead in the soil near the Exide facility could possibly be explained by a few of these variables, Exide is considered the primary source of soil lead in the surrounding area. Regression models could be applied to this study of lead within the PIA to determine the extent of each variable's contribution to lead contamination. However, this addition is outside the scope of this thesis since the focus is on exploring the effect of the scale of spatial units for which values are aggregated from the soil contamination surface.

2.2. The Effect of Spatial Scale on Analysis

Much research in the social sciences focuses on how structural characteristics affect various outcomes, with one form studying if structural characteristics of neighborhoods affect various aggregate outcomes (Hipp 2007). According to Hipp, despite "the variety of research paradigms focusing on the importance of neighborhoods, a commonality of many studies is that less attention is paid to the appropriate level of aggregation for such neighborhood effects" (Hipp 2007, 659). Although studies attempt to understand the effect of neighborhoods on outcomes, the definition of neighborhood itself tends to get lost in the methodological details. Studies have consistently used various geographic units including blocks, block groups, tracts, and zip codes to test the effects of structural characteristics.

However, this strategy hardly considers if the geographic unit is appropriate for the outcome of interest (Hipp 2007). The significance and challenge of determining appropriate aggregation levels is inherent in the modifiable areal unit problem (MAUP). The issue of MAUP persists in any study that analyzes a difference between scales and affects the decision of aggregation level. Whereas Hipp attempts to determine the appropriate geographic level of aggregation using crime and disorder as an example, this thesis seeks to determine the best geographic level of aggregation for cleanup of soil lead contamination, using Exide as a case study.

Similar to Hipp's study (2007), Root (2012) also points out the weakness a lack of definition of neighborhood can have on health studies. Root (2012) states that it results in little consideration of the spatial scale at which socioeconomic factors influence a certain health outcome. While neighborhood studies have been popular in health literature, earlier studies hardly attempted to understand the limitations of using a single geographic unit for analysis. In

the case of Exide, this limitation is present in that DTSC has only considered the parcel scale as the single geographic unit for determining the prioritization of cleanup. Through her study in examining the role of spatial scale in neighborhood effects on health, Root explores how geographic and statistical methods can aid in defining neighborhoods and selecting the appropriate scale. She accomplished this through a case study that examined the relation of socioeconomic status (SES) to orofacial clefts at different spatial scales: 4,000m Buffer, census tract, and census block group (Root 2012). Similar geographic and statistical methods can be used to explore different spatial scales that have not been considered by DTSC, allowing for the recommendation of possible better approaches to prioritizing cleanup.

While Hipp (2007) and Root's (2012) studies emphasize the problem of limiting studies to a single geographic unit for analysis, other studies have demonstrated how the use of multiple scales can actually change the analytical results. For example, Su and Ang examine the effects of spatial aggregation on energy-related CO_2 emissions embodied in trade using the input-output analysis, with China as the study area (Su and Ang 2010). They used three different spatial levels – China as a whole, 3 regional groups, and 8 regions – for analysis and concluded that the embodied emissions depend greatly on the spatial aggregation scheme.

Another study explored the modifiable areal unit problem in the relationship between exposure to NO_2 and respiratory health (Parenteau and Sawada 2011). According to the researchers, previous Canadian population health studies have succumbed to the limitation of using a single geographic unit for analysis by using census tracts as proxies for neighborhoods. The researchers attempt to remedy this by providing a study on the relationship between NO_2 and respiratory health using three different spatial structures to demonstrate the effects of spatial units on analytical results. They used Moran's I and regression analysis for each of the units of

analysis; however, found no significant effect of NO_2 exposure on respiratory health (Parenteau and Sawada 2011). Both of these studies demonstrate the importance of using multiple scales for analysis to account for the differences in results. This thesis attempts to take that into consideration by using different analytical scales for determining an approach to prioritizing cleanup.

In a study that assesses soil lead contamination in Oakland, California, multiple scales were used to determine the extent of contamination being an obstacle to the expansion of urban agriculture within the city (McClintock 2012). Through mapping soil samples via GIS and reconstructed land histories, McClintock performed spatial analysis at city, neighborhood, and site-specific scales. Clusters of Pb contamination at the city and neighborhood levels were found to be related to land use history, while contamination at the site-scale revealed high variability (McClintock 2012). Although McClintock's study on soil contamination employs multiple scales, it is different from the others in that sampling methods were used to differentiate between scales, rather than data being aggregated into units.

2.3. Aggregation Methods

Census geographic units of varying sizes were explored in this study to investigate their use for efficient targeting of cleanup operations. However, summarizing the values across a continuously varying surface, estimated through interpolation as a raster layer, into a single value for each unit examined requires determining appropriate methods for this aggregation.

2.3.1. Using the Statistical Mean as an Aggregation Method

Researchers undertaking a quantitative estimation of health risk to residents from contaminated groundwater and soils in the Slovak Republic utilize municipality, district, and regional boundaries to map potential risk areas (Fajčíková et al. 2014). In their analysis, they

calculated the arithmetic means of each studied compound for each administrative unit of analysis based on interpolated gridded data. This example serves as one potential method for aggregating the geostatistical results of soil lead contamination. The researchers also convey that the use of basic administrative units to map potential risk areas provides for easy discussion with policy- and decision-makers (Fajčíková et al. 2014). This provides support for the use of census geographic units for this project.

In a thematically different study using similar approaches, Liang and Weng (2008) utilized GIS data and remote sensing to perform a multi-scale analysis of the urban heat island (UHI) in Indianapolis using the census-based units of block, block group, and tract. Selected variables for the urban landscape and land surface temperatures (LST) were aggregated at the various census levels, in which correlation analysis and linear regression modeling were then performed at each level. This aggregation into the various census levels allowed for the examination of the scale effect, or the sensitivity of the relationship to aggregation. The study also worked to "establish a method for evaluating urban thermal environment within geographical (census) units" (Liang and Weng 2008, 129). It was successful by allowing for the calculation of census-based variations of land surface temperatures. The researchers aggregated numerous vector and raster variables into the three census levels. For land surface temperatures provided as raster data, the statistical mean values of LST were calculated for each census zone. This approach is similar to the previous study and supports using the statistical mean value as potential method for aggregation.

2.3.2. Using Percentage as an Aggregation Method

In a similar fashion to the area proportions used for grouping vector and raster features in Liang and Weng's study, the use of percentages could be used as a potential aggregation scheme.

A study carried out by Hanna-Attisha and her colleagues (2016) analyzed the differences in pediatric blood lead level incidence pre- and post- the Flint drinking water crisis. Percentage of elevated blood lead levels were assessed from both time periods and identified geographical locations through analysis (Hanna-Attisha et al. 2016). As part of the methods, ordinary Kriging was used to create a predicted surface of child blood lead levels throughout Flint. In their map of the Kriging results, the researchers overlaid the results with Flint City Wards, to determine the percentage of the ward areas where water samples exceeded 15ppb. This seemed to be used as a validation method for their predicted surface. However, this approach could be applied to the aggregation of soil lead distribution by indicating the percent of the area in each geographical unit where the lead concentrations exceed a certain threshold, such as 80ppm or 400ppm, the values determined to be significant for lead exposure, based on California and EPA guidelines, respectively.

2.3.3. Using Risk Assessment to Inform an Aggregation Method

A subset of literature examining contaminated sites uses a combination of spatial statistics and health risk assessment to aid in remediation decisions. This subset offers insights into different approaches when tackling contamination and remediation issues. In their study, Gay and Korre propose a methodology that combines spatial statistical methods with quantitative probabilistic human health risk assessment of produce an assessment of risk to human health from contaminated land (Gay and Korre 2006). A population is mapped across the contaminated area and an adaptation of the Contaminated Land Exposure Assessment model is used to probabilistically calculate the intake of soil contaminants by individuals. Essentially, exposure maps are combined with population data to provide risk evaluation and enable efficient risk reduction measures.

The use of geostatistics for helping to characterize risk assessment has also grown and can be seen in the following studies. Guastaldi and Del Frate utilize geostatistical simulations of pollutants measured in terrain samples to assess the uncertainty for risk analysis of a contaminated industrial site in northern Italy (Guasaldi and Del Frate 2011). These geostatistical simulations quantify the risk through simulation of possible realities and how many exceed contamination thresholds; they provide a means for visualizing risk. These models are then used in aiding the decision-making processes of deciding areas targeted for remediation.

A prior study suggests that sample-specific risk can be used in a geostatistical procedure to produce maps of block-specific risk at the site to aid in decision-making processes with remediation strategies (Ginevan and Splitstone 1997). The researchers employ probability Kriging of sample-specific risks, with five indicator levels of risk for the Kriging procedure. This resulted in the estimation of conditional cumulative distribution functions for each of 3,911 conceptual remediation blocks, which defined the surface soil of the site. Median and 95% upper percentile risks of the blocks were then mapped according to the indicator levels.

A commonality among many of the risk assessment studies include the use of a hazard quotient (HQ) and/or a hazard index (HI) to quantify the means of risk. This suggests a third method for aggregating the geostatistical results of soil lead contamination. Zhao et al. (2012) examined heavy metal contamination near the Dabaoshan Mine in Southern China. The researchers produced spatial distribution maps of the various metals and used land uses and uptake models to create a dose-response model for human health risks (Zhao et al. 2012). The hazard quotients for the various metals were then mapped on continuous surfaces, in which the spatial patterns indicated Cd as the primary pollutant contributing to health risk for humans. Other studies previously discussed also engage the use of hazard quotients and/or hazard indexes

in components of their analysis (Fajčíková et al. 2014); (Pelfrêne, Détriché and Douay 2014). The simple equation for hazard quotient is

$$\text{Hazard Quotient} = \text{Exposure Concentration} / \text{Reference Concentration} \quad (1)$$

The reference concentration is typically determined from credible sources, such as the EPA. If HQ < 1, then there is no risk to human health. If HQ > 1, then some degree of risk exists. Hazard quotients can be calculated for each parcel and combined with the percentage method as a combined approach. This would result in values showing the percentage of area with hazard quotients over 1. However, it is important to note that, like all the other methods, risk calculations will change as geographic boundaries change (Fajčíková et al. 2014). Using the municipality, district, and regional boundaries in their study, Fajčíková and colleagues pointed out that these boundaries bear no relationship to the geochemistry of contamination, in which the resulting spatial associations of the studied chemicals could differ from actuality. This lends urgency to the need to consider spatial scale in the assessment of risk.

Chapter 3 Methods

The purpose of this study is to provide valuable alternatives for identifying priority areas for cleanup by using various aggregation methods and examining how the resulting values may be affected by the scale of spatial units considered. Here, block groups, blocks, and parcels were used in this analysis to provide insight into what scales are best for planning and managing cleanup of lead contaminated soils. This chapter begins with a description of the soil sample data used for this study and details the process of preparing these data for analysis, following with a section on data exploration that delves into data quality issues and the importance of outliers on this analysis. Next, the study area boundaries and the scales for analysis are established. The study then employs Empirical Bayesian Kriging for the development of a suitable Kriging model to create interpolated surfaces of lead concentration values, using cross-validation as a guide. These surfaces are then utilized in various aggregation methods, resulting in the mean, percent area, and Hazard Quotient (HQ) values for spatial units at each scale. Figure 3 displays the overall workflow for this thesis.

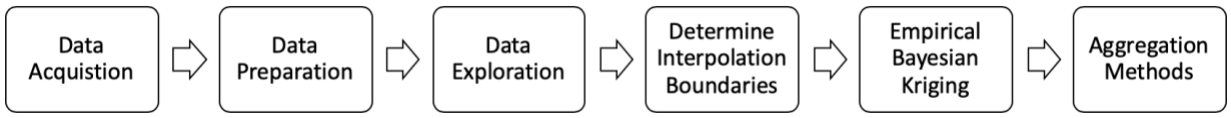


Figure 3 Summary of Workflow

3.1. Soil Sample Data

This research utilizes soil sample data collected as part of the Exide Preliminary Investigation and provided by DTSC. The Preliminary Investigation Area (PIA) covers an area surrounding the smelter site within an approximately 1.7-mile radius. The dataset was provided by DTSC in the form of a table in Excel format. In this Excel spreadsheet provided by DTSC, there are four sheets/tabs: Data Table (contains all of the soil sample data and attributes), Data Legend (provides additional clarifying information about the data and its attributes), List of Acronyms (acronyms used in the spreadsheet), and Multiple APN Properties (listing a primary APN to represent properties that contain multiple parcels).

Soil was sampled on approximately 8,500 properties within the PIA, with a total of 328, 069 soil samples in the Excel table. Roughly 10 to 15 soil samples were taken from each property at several locations in the front, back, and side yards. The sampling matrixes included soil and paint, with soil being the focus for this analysis. The samples were either sent to a laboratory for chemical analysis or analyzed in the field using X-Ray Fluorescence (XRF). Approximately 20% of the samples from each property were analyzed in a laboratory.

According to the Final Sampling Workplan for DTSC, there were several steps taken to select the soil sampling areas. This included choosing sampling locations that targeted "bare exposed soils that have not been recently disturbed and open grassy areas away from structures or thick trees" and areas "in which maximum deposition and exposure potential are likely" (DTSC, Final Workplan 2015, S3, 1). It was noted that approximately 15 sample locations would be taken at each property, marked by pin flags. The sampling distribution criteria is as follows: "five locations in the front yard; five locations in the back yard; five locations distributed in drip zones, near downspouts, side yards, and other open bare soils areas; and two additional

contingency sample locations if a play area is present" (DTSC, Final Workplan 2015, S3, 2). Samples for all sample locations were collected at the 0- to 3- inch depth interval, with additional depth intervals taken (up to 18 inches) for the two locations detected with the highest lead concentrations. Information regarding how coordinates were captured is provided in the Final Sampling Workplan as follows: "The location of each sample will be measured from a reference point at the property and marked on a field sketch. In addition, coordinates of each soil sampling location will be recorded using a global positioning system (GPS) unit. GPS coordinates of each sampling location will also be recorded in the field notes" (DTSC, Final Workplan 2015, S3, 3).

Furthermore, the Final Sampling Workplan for DTSC indicates that sampling reports were also provided for each property. These sampling reports included the following information: "a description of the property, a map showing the sampling locations, coordinates of the sampling locations, sampling results in tabular form and electronic format (MS Excel), screening of the results against criteria established in the Workplan to determine if further action is required at the property, photographs of the sampling locations, laboratory analysis reports, an evaluation of the quality of the data, [and] an explanation from any deviation from the Workplan" (DTSC, Final Workplan 2015, S4, 1).

In addition to providing the locational coordinate information, several attributes are associated with each of the sample results in the Excel table. A unique property ID assigned by the sampling contractor and address locations are given for each of the sampled sites, along with the facility type, such as if it is residential, a child care center, commercial, a park, school, or other type. The primary Assessor Parcel Number identified for the properties is also listed for each soil sample. Detailed information about the samples is also provided: Sample ID, dates samples were collected, their location, depth taken (in inches), sampling matrix, and location

codes (specifying location on properties). In addition to these attributes, characteristics of the analysis are also included: analysis type (in the field or lab), sample type, test type, metal analyzed, the analysis results and their units (ppm, mg/kg, and mg/cm2) and detection limit. A list of the main attributes and their descriptors can be found in Table 1, with information taken directly from the 'Data Legend' within the source Excel document.

Using all samples from each property, a Representative Soil Lead Concentration was also determined by DTSC for each property, summarizing all the lead values into one value assigned to each property. This value was used by the DTSC in prioritizing cleanup by parcel. The value was calculated using the USEPA's ProUCL software and is the 95 percent upper confidence level limit of the lead concentrations, which is typically greater than the average concentration of all samples from the property, but below the maximum concentration. It is important to note that the 95% UCL values were only calculated when there are more than 8 samples per property. When there are less than 8 samples per property, the maximum sample concentration value in the 0-3-inch depth interval is used as the Representative Soil Lead Concentration. This Representative Soil Lead Concentration is also listed as an attribute for each soil sample result and more details concerning this value can be found in Table 1. Although the dataset includes samples taken at various depths within the PIA and information about multiple metals, this research used only data on lead in samples collected from a depth of 0-3 inches below ground surface.

Table 1 Main Attributes for Soil Samples and Comments (most of the table is sourced directly from the 'Data Legend' tab in the original source Excel spreadsheet)

Category	Field Name	Comments
Property Information	Property ID	This number is a unique identifier assigned by the sampling contractor
	Address	The address of the property (separated into different columns for numeric address, street name, city, state, and zip code)
	Primary APN	Primary Assessor Parcel Number (APN) associated with a Property ID (Some properties have multiple APNs; for example, some schools have many APNs but for sampling and cleanup purposes they are considered as a single property)
Sample Location Information	Sample ID	Each Sample collected has a unique identification number (XXXXX)-(XX)-(XX). Typically, this is the Property ID, followed by the location number at the property, followed by the depth the sample was collected (refer to "Depth" column for depth definitions). The location number is a sequentially assigned number up to the total number of samples collected at a property.
	Sample Location	This is the numeric identification number for where samples were collected at a property (XXXXX)-(XX). The first part is the Property ID and the second is the location number at the property. The location number is a sequentially assigned number up to the total number of samples collected at a property
	Location Code	The type of location on the property where a sample was collected. BY = Back Yard, Chip = Paint sample, DL = Drip Line, FY = Front Yard, G = Garden, GCComp = Composite Sample, O = Other, P = Property Boundary, PA = Play Area, SY = Side Yard
Analysis Information	Analysis Type	Samples were either analyzed in the field (XRF) or sent to a laboratory. FI = Field Sample (XRF) or Lb = Laboratory Sample
	Sample Type	Sample type indicates whether the sample was a normal sample or a duplicate sample.
	Metal Analyzed	Metal that either the soil or paint was analyzed for: Lead, Antimony, Arsenic, Cadmium, Copper, Zinc
	Result	Result of the XRF or laboratory analysis.
	Result Units	The unit the results are presented in: ppm = parts per million (soil analyzed by XRF), mg/kg = milligram per kilogram (soil or paint analyzed in a laboratory), mg/cm2 = milligram per square centimeter (paint analyzed by XRF)

Category	Field Name	Comments
Representative Soil Lead Concentration Information	Representative Soil Lead Concentration	The representative property-wide lead concentration is based on the 95 percent upper confidence limit (UCL) statistical analysis, which is typically greater than the average concentration of lead on the property, but below the maximum concentration. The 95 percent UCL concentration represents the concentration that would be greater than the true mean value of the samples on the property 95 percent of the time. The 95 percent UCL and maximum lead concentrations are used to determine the level of exposure to lead at each property. 95 percent UCL value calculated by ProUCL using the lead soil results for the 0-3" depth. Properties with fewer than 8 samples from 0-3" do not have 95 percent UCL values calculated. For these properties the maximum concentration in the 0-3" depth interval was used as the representative lead concentration. Some properties have fewer than 8 samples because there are often size limitations of the surface area of unpaved surfaces at a property. In instances where properties have small surface areas from which to collect samples, professional judgment is used to determine an appropriate number of samples to collect to represent the conditions at a property.
Coordinate Information	Longitude and Latitude	On a map, x,y coordinates are used to represent features at the location they are found on the earth's spherical surface. Georeferencing in progress = Using Geographic Information System (GIS) software to determine the X (longitude) and Y (latitude) coordinates of sampling locations from the maps provided in the sampling reports

3.2. Data Preparation

This section discusses how the soil sample data from the Excel spreadsheet were prepared for use in a GIS and the steps involved in the preparation for their use in the study's analysis. Data preparation included the geocoding of some sample points and a reduction in the dataset to only the attributes that were needed for the analysis. It also included the integration of duplicate soil samples through a process of averaging. Most of the data preparation steps were carried out using ArcGIS Pro.

3.2.1. Initial Data Preparation

Since the soil sample data were supplied in table format, some initial preparation was needed so that the data could be used in a GIS. This included converting the Longitude and Latitude columns from the Excel data table from text to numeric data type and saving the file in CSV format to be imported into ArcGIS Pro so that it could be converted into vector-based discrete points.

Out of the 328,069 soil samples, 322,761 samples had coordinates and could be directly uploaded as points in ArcGIS. However, 5,308 samples of all metal types and depths were missing coordinates, listed as 'Georeferencing in progress' in the original Excel data table, and some had coordinates that did not fall within the PIA boundary, ranging from some just missing the boundary to some points appearing in the ocean. The phrase 'Georeferencing in progress' was defined in the 'Data Legend' as "using Geographic Information System (GIS) software to determine the X (longitude) and Y (latitude) coordinates of sampling locations from the maps provided in the sampling reports." It is unclear what is meant by this term in relation to the location coordinates. The points with missing coordinates and inaccurate coordinates were extracted from the main dataset in Excel and geocoding was carried out in ArcGIS for all samples that could not be directly or correctly mapped, using the addresses associated with each of the samples. In addition to the geocoding, the rest of the data preparation took place in ArcGIS Pro.

After being re-geocoded, the original incorrectly located points did fall into the PIA boundary, demonstrating potential mistakes in the original data table. These mistakes could have arisen from the transferring of the field data into the Excel table. Aside from these mistakes, the rest of the coordinates appear to be relatively accurate, given that they were determined by GPS

units. The possible uncertainty that arises from these mistakes does not present a concern for this analysis, as the soil samples were used to create an approximate interpolated surface.

The total number of geocoded points for samples used in the analysis, those originally with no coordinates and those redone, was 399 samples. While the Esri World Geocoding Service provided the approximate locations of the sample points, the geocoded sample points were then manually placed within their properties, respective to the additional location codes provided for the samples, such as front yard and backyard, etc., and confirming they were associated with the correct property ID. The geocoded points were then merged with the rest of the dataset. The complete set of points is shown in Figure 4. The red points are the samples that were geocoded, while the blue points are the samples that had existing coordinates within the study boundary. Note that the solid blue areas indicate the density of individual sample locations. The zoomed in map illustrates the distribution of samples within a small area of the PIA.

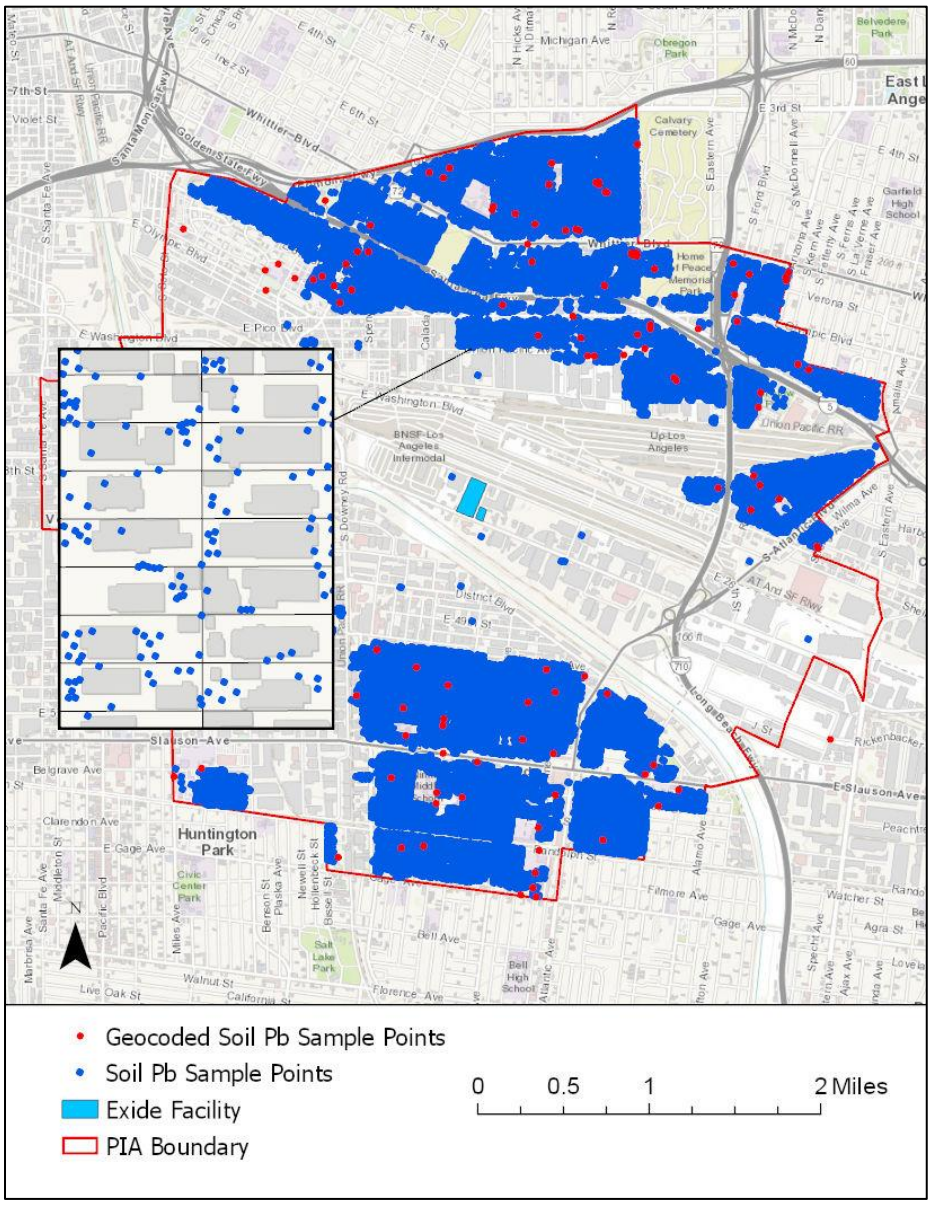


Figure 4 Exide soil Pb sample locations.

Since the full version of the dataset includes analysis for multiple metals within the soil and differing matrixes of soil and paint, the dataset needed to be filtered to include the proper parameters for analysis. In particular, the Selection by Attribute tool was used to reduce the dataset and create a new feature class with samples only having these necessary attributes: Metal Analyzed – Lead, Matrix – Soil, and Depth – 3 inches (depth interval from 0-3 inches below ground surface, considered the surface layer). The surface values are important for this analysis,

as lead typically stays towards the surface of soil and does not penetrate too deeply into the soil. Although other depths were measured, the values are not necessary for this analysis.

Though a Representative Soil Lead Concentration was determined for each property and could have been used for this analysis to eliminate uncertainty of the sample data, it was determined that the raw data from the soil samples would be more valuable for this analysis. Dr. Jill Johnston, Assistant Professor of Preventative Medicine and Director of USC's Environmental Health Centers Outreach Program, has become an expert in the situation concerning Exide and provided this recommendation.

Since this study used the raw data from the soil samples, the dataset was further filtered to include only samples that were analyzed in the field with XRF, while lab samples were removed from the dataset for this study's analysis. This was done to address a concern regarding the difference in measuring methods and sample selections between field and lab samples. In addition, the field sample concentration units are parts per million (ppm) and the lab concentration units are milligrams per kilogram (mg/kg). While these units are ultimately the same scale, a noticeable difference in the range of lab values versus fields values suggested important measuring differences. This decision to remove the lab samples was also supported by Dr. Jill Johnston. The total number of field samples was determined to be 52,947 out of the combined field and lab sample total of 115,584 samples.

To further prepare the data for use in analysis, the soil sample point data was projected into the NAD 83 California State Plane coordinate system, Zone V. This coordinate system was chosen due to its high accuracy in each zone and its high utilization by state and local governments. Since the soil sampling data is highly localized, the State Plane coordinate system is appropriate for use in this study.

3.2.2. Integrating Duplicates by Averaging

In the initial Excel spreadsheet of the soil sample data, it was noted in the 'Data Legend' tab that there was marked duplicate soil samples. This distinction was made known in a column titled Sample Type that established whether a sample was a normal sample or a duplicate sample. Duplicate samples arose when two or more samples were taken for a specific location and depth. This procedure was done to account for variability within the soil and is a standard procedure for soil sampling in the field. Although duplicate samples are important when analyzing the soil from a scientific perspective, these redundant points can pose issues when creating interpolated surfaces through geostatistical means. Hence, each set of original and corresponding duplicate samples was averaged to produce a single value at each sample location. The averaging of such duplicates, a common practice with such data, ensures that a good estimate for each sample location is considered when making the interpolated surfaces.

Upon further examination into the soil sample data, it was noted that there were additional coincident sample points, beyond the samples marked as duplicates. This was determined by a one-to-many relationship between sample coordinates and Sample Location, the unique numeric identifier assigned to each sample location. With this in mind, the Summary Statistics tool in ArcGIS Pro was used to average the coincident point values with the longitude and latitude as the case fields, rather than using Sample Location as a case field. The longitude and latitude case fields meant statistics were calculated for each unique coordinate pairing, considering the coincident points and those that were marked as duplicates. This resulted in a table with a count of 50,952 records.

The attributes included in the Summary Statistics table were the following: Property ID (unique ID for each property), the primary APN (parcel number associated with the property), Sample Location (unique sample location ID), the mean of the results, the result units, and the

representative soil lead concentration. Since the output of the Summary Statistics tool is a table, the longitude and latitude values in the table were used to create a new feature class of the newly averaged soil sample data and was projected into the appropriate State Plane coordinate system. The resulting dataset is the set of soil samples, with any coincident point values averaged, that was used for all of this study's subsequent analyses.

The initial data preparation as well as the steps of integrating duplicate samples by averaging are summarized in the following workflow diagram (Figure 5).

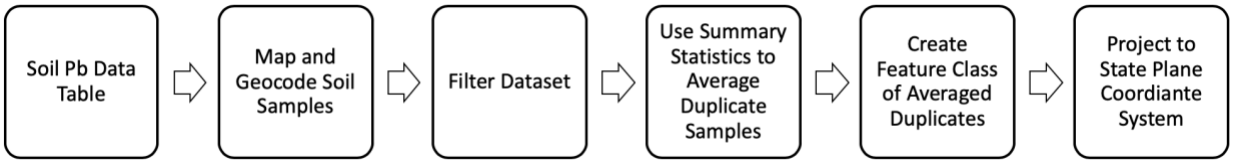


Figure 5 Data Preparation Process

3.3. Data Exploration

After preparing the soil sample data for use in the analysis, further exploration of the dataset investigated potential data quality issues and provided an assessment of outliers within the data.

3.3.1. Data Quality Issues Regarding Recorded Locations

One data quality concern is related to the uncertainty associated with the X and Y coordinates for the sample locations in the Excel file. Although the coordinates were recorded using a GPS unit and were sampled by property, some sample points appear to be on top of buildings when visualized in a GIS, indicating the possible use of parcel centroids or inaccuracies resulting from the GPS unit. To investigate these inaccuracies, building outline data was acquired from the LA County GIS Data Portal and spatially joined to the soil sampling point

data to determine how many points are on buildings. Out of the 50,952 data points, only 2,999 points are on buildings, with the highest number of points on a building being 13 points. This indicates minor horizontal errors in the coordinates and the chances of the samples being on the incorrect parcel are slim. These inaccuracies can be attributed to human error in the process of transferring the locational information from the sampling reports to the Excel file, GPS positioning accuracy from the devices used, possible errors in the building outline data, or even errors in the ArcGIS base map data. As mentioned earlier, the use of the phrase 'Georeferencing in progress' could also present confusion in the process and suggests possible error in the preparation of the Excel table. Despite these uncertainties within regards to the coordinates associated with the samples, the error resulting from these uncertainties does not significantly affect this study's analysis, as the samples were utilized in the creation of an interpolated surface.

Another indicator of potential data quality concern was whether the Assessor ID (AIN) in the parcel boundaries polygon feature class matched the primary Assessor Parcel Numbers (APN) associated with the soil sample data on each property. These numbers are formatted identically so they can be directly compared. The AIN of the parcels and the APN of the properties should match each other and provides an indication of the accuracy of the soil sample data. To investigate this concern, the parcel data was spatially joined to the soil sampling data and a query was performed to indicate the number of instances where the APN from the soil data did not match the AIN from the parcel data. The resulting record count was 3,185. However, critical examination into these records revealed that the APN and AIN of these records were only off due to adjacent parcels being listed as the APN for the soil data.

An explanation for this data quality concern stems from the source Excel spreadsheet, where the tab 'Multiple APN Properties' mentions that a primary APN was selected to represent

properties that contain multiple parcels and thus multiple APNs. This explains why the AIN of some parcels do not match with the APN of the corresponding properties, as the AIN could correspond to any of the APNs associated with the property instead of the primary APN. Although the AIN of the parcels do not match exactly with the APN of the properties in the soil sample data, the explanation for it validates the accuracy of the soil sample point locations and does not present itself as a primary concern for this analysis.

3.3.2. Outlier Assessment

Although outliers within data could be a cause for concern in most analyses, outliers serve an important role in this analysis. Typically, outliers should be excluded from the data before analysis takes place. However, for this study, outliers with higher concentration values indicate areas with higher levels of lead contamination, signifying areas that should be a health concern. Even though outliers are important to this analysis, the maximum value in the dataset is 21,854 ppm, which can be considered unreasonably high and likely be due to human error in the compilation of data. Therefore, the sample with this particular outlier value was excluded to make for better models in the geostatistical analysis. Excluding the high outlier, the minimum value of the dataset is 8.82 ppm, the maximum value is 7,348.04, and the mean is 268.53.

Figure 6 displays the remaining outliers in the soil sample data at both ends of the data's standard deviation spectrum. The soil sample data was symbolized using standard deviation classification on a logarithmic scale. The symbology was set to only display outlier values less than -2.5 standard deviations and greater than 2.5 standard deviations. The distribution of outliers appears to be evenly scattered within the study areas.

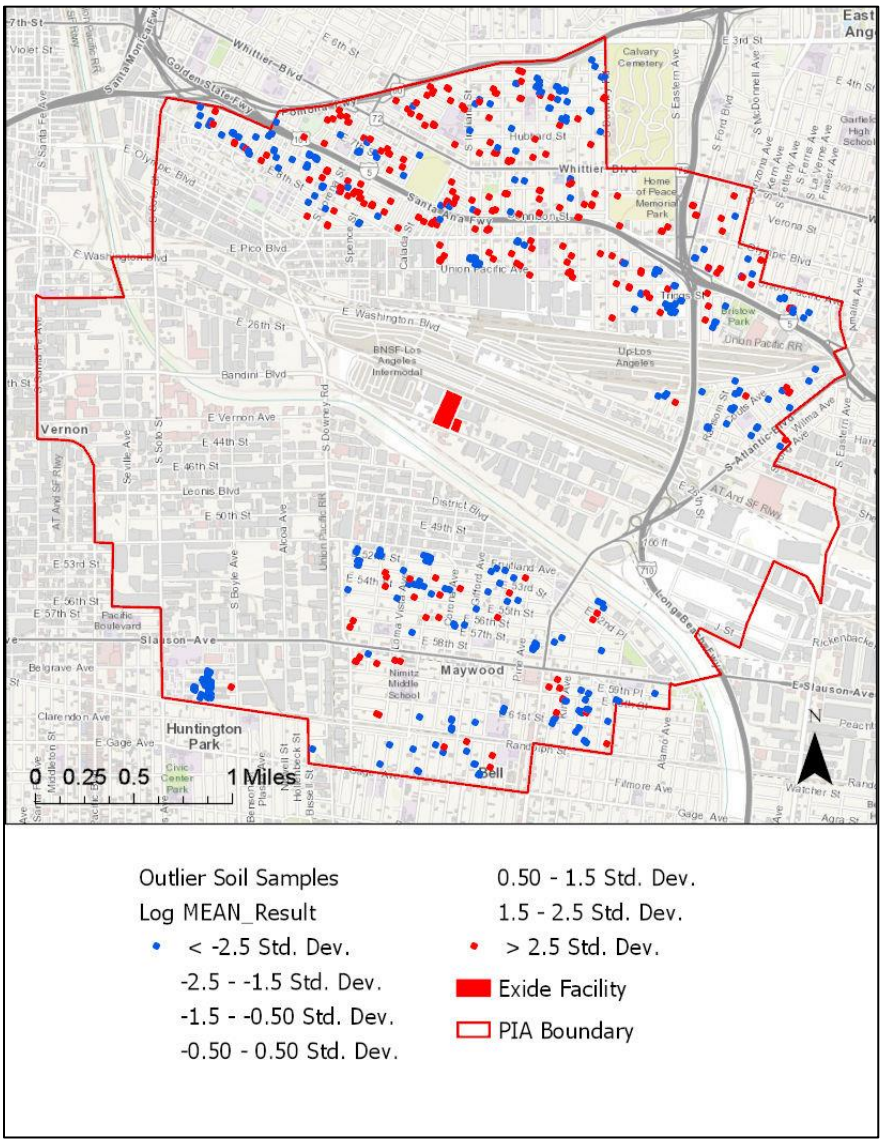


Figure 6 Outlier Soil Samples

3.4. Study Areas and Scales of Analysis

This section discusses how three separate study areas were determined within the overall PIA boundary, using the sample points as a guideline, and the decisions made regarding the different scales of analysis.

3.4.1. Study Areas

The Exide Preliminary Investigation Area, which encompasses an approximately 1.7-mile radius around the smelter site, aids in determining the study areas for this analysis. The Preliminary Investigation Area includes parts of the following cities: Los Angeles, East Los Angeles, Maywood, Boyle Heights, Huntington Park, Commerce, Bell, and Vernon. To capture the PIA, the PIA boundary was created as a new feature, using a map made by DTSC as a guide. However, since the sampling points are densely concentrated in particular sections within the PIA boundary, three separate study area boundaries were created, as shown in Figure 7, which ultimately was needed for the geostatistical analysis.

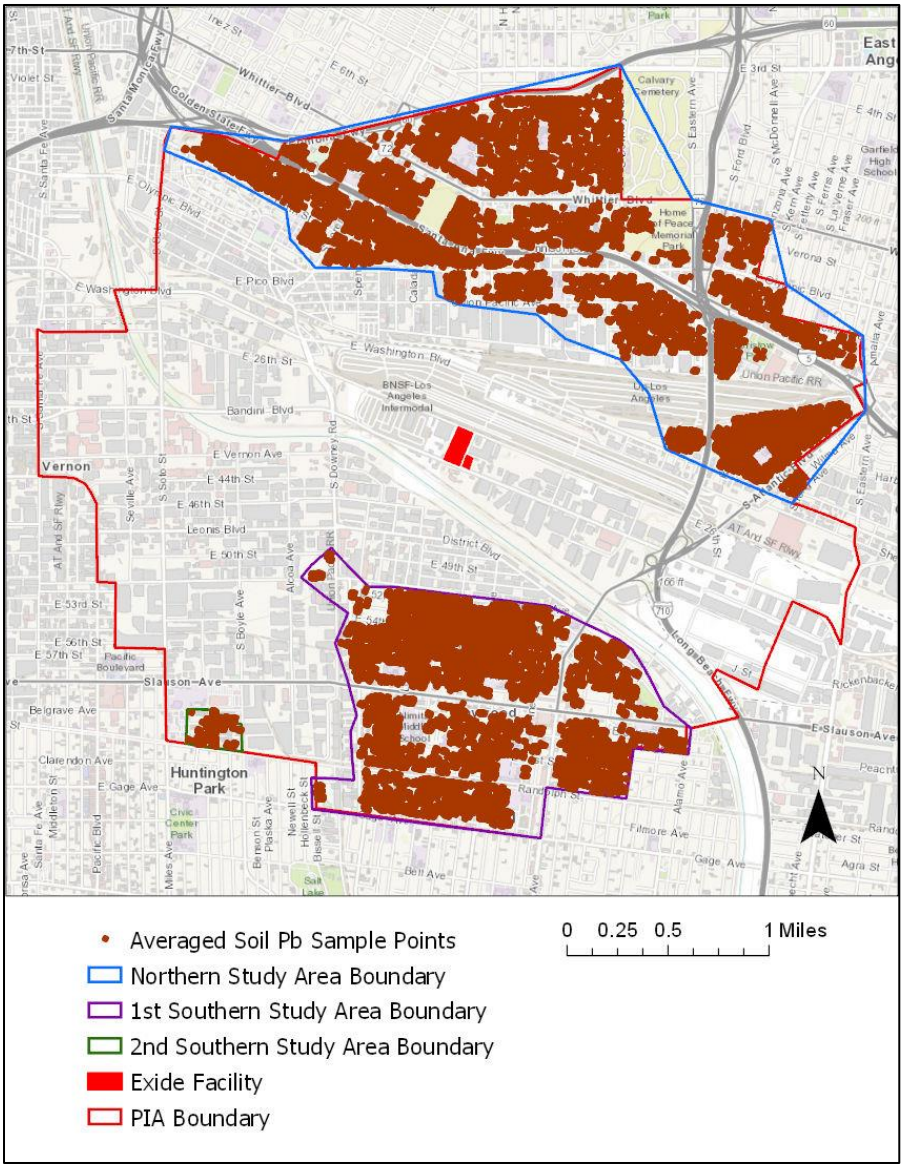


Figure 7 Study Area Boundaries

To produce these three separate study areas, the Aggregate Points tool, which creates polygons around clustered point features, was used. The polygons generated from this tool and the PIA boundary were used to determine and smooth the newly created separate study area boundaries. One of the three concentrated areas of the sample points is located north of Washington Boulevard in the northern part of the PIA boundary, while the other two concentrated areas of the sample points are located in the southern part of the PIA boundary,

with a smaller cluster to the left. Corresponding to their respective concentrated areas of points, the three separate study area boundaries are defined as follows: Northern Study Area Boundary, 1st Southern Study Area Boundary (the larger cluster of points in the southern area of the PIA), and 2nd Southern Study Area Boundary (the smaller cluster of points in the southern area of the PIA). The soil sample points were then clipped to their respective study area boundaries and the feature classes analyzed separately.

3.4.2. Scales of Analysis

In determining the scales of analysis for this study, spatial units typically used for policy implementation and that are easily defined geographically were of interest. Hence, it was decided that census units would form the basis for the scales of analysis for this study, along with parcel boundaries, which is the current aggregation unit being used to prioritize cleanup of the contaminated soils. The aggregation units used for this study include block groups, blocks, and parcels. These zones of differing scales were used to aggregate the results of the interpolated surfaces created from Kriging through various aggregation methods. These diverse aggregation units provide for demonstration of how the values assigned to these units vary based on scale and offer other approaches for delineating priority areas to clean up the lead contaminated soils, in addition to the parcel-based approach. Table 2 lists the sources of these feature layers.

Table 2 Polygon Feature Layers Used for Scales of Analysis

Dataset	File Type	Data Type	Details	Source
Census Block Groups (2018)	Shapefile (.shp)	Polygon Feature Class	Boundaries of block groups	U.S. Census Bureau TIGER Products
Census Blocks (2018)	Shapefile (.shp)	Polygon Feature Class	Boundaries of blocks	U.S. Census Bureau TIGER Products
LA County Assessor Parcels (2016)	Geodatabase (.gdb)	Polygon Feature Class	Parcel boundaries for LA County	Los Angeles County GIS Data Portal

3.5. Creating Surfaces through Geostatistical Analysis

Section 3.5 describes in detail the process for creating the interpolated surfaces using the Empirical Bayesian Kriging geostatistical method. It first defines and explains what Empirical Bayesian Kriging entails and then discusses how an appropriate Kriging model was determined, using cross-validation as a guide. The section then lays out the logic for determining the cell size of the output surfaces and what was done post-Kriging to prepare the interpolated surfaces to be used in the various aggregations.

3.5.1. Empirical Bayesian Kriging

The geostatistical method of Empirical Bayesian Kriging (EBK) was utilized in this study to obtain interpolated surfaces of the soil samples. Unlike other Kriging methods in Esri's Geostatistical Analyst, Empirical Bayesian Kriging automates some of the more challenging aspects of creating a valid Kriging model by using a large number of subsets of the data to create and compare various simulated models rather than requiring the user to manually adjust the parameters to find a suitable Kriging model. This latter approach can result in a model based on arbitrary decisions (Esri 2018). EBK also accounts for a possible error when estimating

semivariograms by creating a collection of semivariograms for the simulated models and averaging them to create a suitable Kriging model. This contrasts to other Kriging methods which are dependent on a single semivariogram, calculated from all of the known sample points, to estimate the values at unknown locations. Since these other methods rely on the single semivariogram, they can underestimate the true standard errors of the predictions (Esri 2018).

3.5.2. Determining the Appropriate Model

Even though EBK significantly aids the process of building a valid Kriging model, there are still decisions that need to be made when considering some of EBK's parameters. Esri's Geostatistical Wizard helps guide this decision-making process of constructing and evaluating each model's performance through a set of interactive pages (Figure 8).

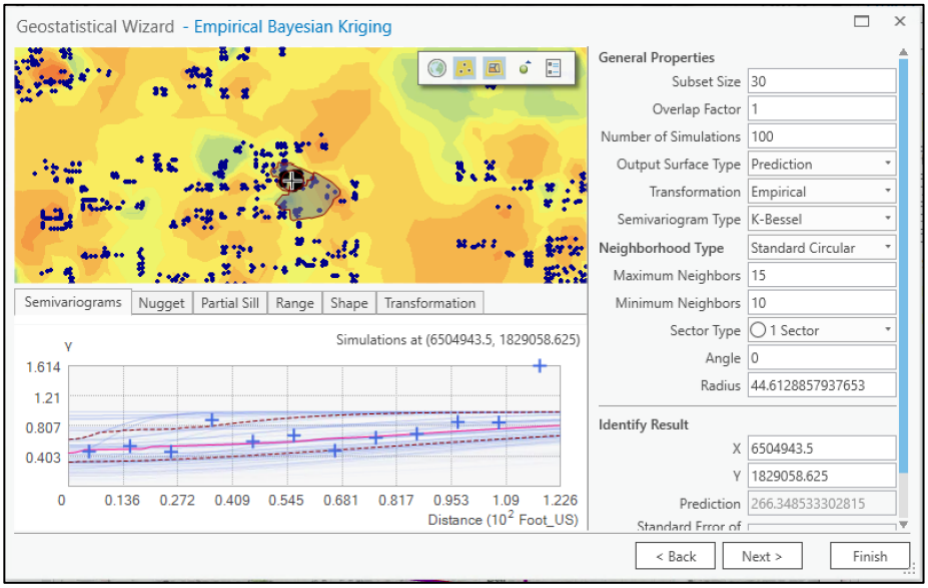


Figure 8 Empirical Bayesian Kriging in Geostatistical Wizard

One of the main considerations in building a successful Kriging model is determining the type of semivariogram that would be best for the model, depending on if a transformation is applied. The Empirical Bayesian Kriging method offers several different types of semivariograms along with multiplicative skewing normal score transformation, with Empirical

and Log Empirical as its two base distributions. The normal score transformation transforms the dataset to closely resemble that of a normal distribution. If the Transformation parameter is set to None, then only the Power, Linear, and Thin Plate Spline options are available. If the Transformation parameter is set to Empirical or Log Empirical, then Exponential, Whittle, K-Bessel and their detrended counterparts are available options. Each of these semivariogram types have their advantages and disadvantages.

Other considerations in building a successful Kriging model using EBK include deciding the number of simulations for determining the semivariograms and the subset size for the local models. The default number of simulations per subset is 100 and was the number used for this analysis, as the greater number of simulations produce greater precision in the predictions. The subset size is the maximum number of points in each local model. Depending on the size and dispersion of the dataset, the subset size can play a crucial role in determining valid Kriging models.

3.5.2.1. Using Cross-Validation as a Guide

To determine the appropriate subset size and semivariogram model for use in this study, a technique known as cross-validation was used to evaluate each of the models. Cross-validation aids in making informed decisions by evaluating which of the models provide the best predictions. The process of cross-validation involves the removal of one sample point from the model and then using the rest of the sample points to predict the value at the location of the omitted point, repeating the process for all of the points in the dataset. The predicted value is compared to the measured value. The process results in useful information in the form of statistical diagnostics, which enables the comparison of models.

When comparing models, certain guidelines based on the statistical diagnostics are considered to determine the best model. Four of the most important statistics that pertain to the guidelines are the standardized mean, the standardized Root-Mean-Square, the Root-Mean-Square (RMS), and the Average Standard Error (ASE) (Esri 2019). The standardized mean, or mean of the cross-validation errors, assesses bias in the model, such as if the model predicts too low or too high. This value should be the closest to zero, indicating a model with the least bias. The Root-Mean-Square value directly measures the accuracy of the predictions. Generally, a smaller value signifies greater accuracy of the cross-validation predictions to the measured values (Esri 2019). The standardized Root-Mean-Squared prediction error should be the closest to 1. Lastly, the Average Standard Error should be nearest to the Root-Mean-Squared prediction error (Esri 2019).

These guidelines influenced the workflow for determining the most appropriate Kriging model to use for all the study areas. After initial testing using the different types of transformations (None, Empirical, and Log Empirical), the Empirical transformation consistently proved to have the best results after cross-validation and was the transformation used for building the Kriging models. It is also important to note that the Log Empirical transformation was purposely not used due to its sensitivity with outliers, in which resulting predictions could be orders of magnitudes smaller or larger than the actual values (Esri 2018). Since this study relies on the importance of outliers, this transformation was not suitable for use.

Next, the ideal range of subset sizes was chosen. Using the Exponential semivariogram for the Northern Study Area as a control, the subset sizes of 20, 25, 30, 50, and 100 (the default) were tested, with the range of 20 to 30 performing best in cross-validation. To determine the best semivariogram model and the ideal subset size, the semivariogram types of Exponential, Whittle,

and K-Bessel with the subset sizes of 20, 25, and 30 were tested for each of the study areas (North, 1st South, and 2nd South). The full model results can be found in Appendix A.

All of the models performed reasonably well, emphasizing the significance of the guidelines for the statistical diagnostics resulting from cross-validation. The models that had the standardized means closest to 0 and the standardized Root-Mean-Squares closest to 1 had a higher Root-Mean-Square and greater differences between this value and the Average Standard Error. The models with the lower RMSs and smaller differences between this value and the ASE had higher standardized means and standardized RMSs not the closest to 1. The top three models were chosen for each of the study areas by balancing the two outcomes based on meeting the objectives of the statistical guidelines.

After taking all of the objectives into consideration, the Kriging model determined to be the best suited for all of the study areas is the K-Bessel semivariogram using a subset size of 30 and it was the chosen model for making the interpolated surfaces. This model produced standardized means closest to 0, standardized Root-Mean-Squares closest to 1, and relatively small differences between the RMS and ASE values (Table 3). Although additional factors and greater scrutiny of the cross-validation results could be considered for determining the best Kriging model for use, it is important to note that this study only requires the development of suitable interpolated surfaces of each of the study areas for use in exploration of the impacts of aggregation of different scales. The interpolated surfaces are a means to the end, rather than the end itself.

Table 3 EBK Cross Validation Results for 30 Subset Size K-Bessel Model

Study Area	Standardized Mean	Standardized Root Mean Squared	Root Mean Square (RMS)	Average Standard Error	RMS & Avg. Standard Error Approx. Difference
North	0.00390	0.99888	201.127	184.207	17
1st South	-0.00199	1.00246	132.653	119.816	13
2nd South	0.00519	0.92088	84.120	83.981	1

3.5.3. Determining the Cell Size

For the purpose of the aggregation comparison, the gridded surfaces created through EBK need to have a cell size that made it possible to capture the variation of the surface at all scales. Given the wide range of parcel sizes, it was determined there should be at least 2 cells per parcel. To determine this cell size, the mean, minimum, and maximum parcel area of all the parcels within the study area boundaries were determined. The mean parcel area was 10,717 sq. ft. (996 sq. meters), the minimum parcel area was 38 sq. ft. (4 sq. meters), and the maximum parcel area was 2,117,852 sq. ft (196,755 sq. meters). Thus, a cell size of 50 feet (15.24 meters) was chosen so that it would create a 2,500 square foot cell (232 sq. meters), fitting all but the few smallest parcel sizes. It is important to note that the coordinate system used for analysis is in U.S. feet and was chosen for this study because it is the common coordinate system used by Los Angeles County.

3.5.4. Post-Kriging

After producing the geostatistical layers for each of the study areas using EBK, the interpolation output surfaces were extracted to rasters, with the appropriate cell size of 50 feet (15.24 meters) clipped to the study areas using the mask function.

3.6. Aggregation Methods

This section discusses the different aggregation methods used to summarize the interpolated surface results into each geographical unit of block groups and blocks, which include taking the mean of surface result values, the percent area, and determining a Hazard Quotient. It also describes the use of a different method for the parcel unit.

3.6.1. Aggregation Methods for Block Groups and Blocks

To aggregate the interpolated surface results into each analysis unit (block groups and blocks), the tool Zonal Statistics as a Table was utilized. This tool works by summarizing the values of a raster into zones defined by polygons that are overlaid on the raster. Results are stored in the form of a table with a row for each of the overlaid polygons and columns providing summary statistics within each polygon. Figure 9 displays a workflow for the aggregation methods used for blocks and block groups.

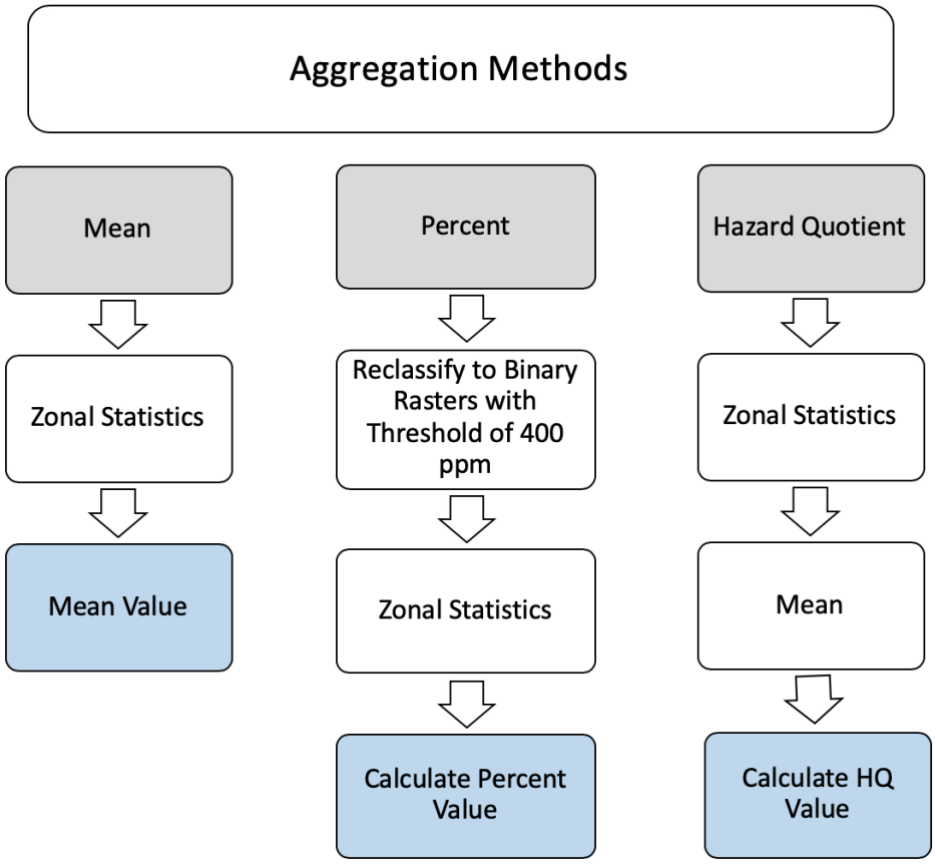


Figure 9 Aggregation Methods Workflow for Block Groups and Blocks

The summary values assigned to each aggregation unit are the statistical mean, the percent area, and a Hazard Quotient (HQ). All of these values are different ways to identify priority areas for cleanup of the lead contaminated soils.

The easiest summary value to calculate was the statistical mean as this is determined directly as one of the statistics produced from zonal statistics. It is the statistical mean of all lead concentrations found in the set of cells encompassed by each polygon. Using a polygon overlay algorithm, zonal statistics involves zone features (the geographical units in this case) overlapping with the cell centers of the value raster (the interpolated surfaces). Thus, zonal statistics works by providing statistics for the various zones by using the cell values from the input raster where

their centers fall within the boundaries of the zones. If zone features do not overlap with any cell centers of the input value raster, then these zone features will not be represented in the output, as determined through the algorithm.

To calculate the percent of the area of a polygon where lead concentration values exceed 400 ppm (USEPA standard for lead exposure), the interpolated surfaces were first reclassified into binary rasters, where any value below 400 became 0 and any value greater than or equal to 400 became a 1. After creating zonal statistics tables from these reclassified rasters, the sum and count statistics were utilized to calculate the areal percentages. The sum statistic represents the sum of all the cells in a polygon containing the value of 1, while the count statistic represents the count of all the cells per polygon. A new percentage field was then calculated using the sum divided by the count multiplied by 100. This percentage represents the percent of the area of a polygon (block group or block) where lead concentration values exceed 400 ppm.

The third summary value assigned to each aggregation unit is a Hazard Quotient. The simple equation for determining a Hazard Quotient is Exposure Concentration / Reference Concentration. If HQ < 1, then there is no risk to human health. If HQ >1, then some degree of risk exists. The reference concentration utilized for this study was 400 ppm, the USA EPA standard for lead exposure. The means of the cells per polygon calculated from the zonal statistics were used as the exposure concentrations. A new field for Hazard Quotients was calculated by dividing the means by the reference concentration of 400ppm. This creates an index value that standardizes the mean values and offers proportions that can be used to easily identify which areas may present some degree of risk to human health from the lead contamination.

The tables of the resulting summary values – the mean, percentage, and the Hazard Quotients – were then joined to their respective spatial aggregation units for each of the study areas. These aggregation units included block groups and blocks.

3.6.2. Aggregation Method for Parcels

For the parcel scale, recognizing that in many cases only small portions of a small number of grid cells on the interpolated surface would fall within individual parcels, it was determined that the polygon overlay algorithm in zonal statistics would not provide a valid result. So, it was decided that it would be useful to compare a single value extracted from the interpolated surface with the Representative Soil Lead Concentration values currently used. To do this, centroids were calculated for each parcel polygon, then the Extract Multi Values to Points tool was utilized to extract the cell values from the interpolated surfaces at each centroid. These values were then subtracted from the original Representative Soil Lead Concentration values determined by DTSC to calculate the magnitude of differences.

Chapter 4 Results

Chapter 4 presents the results of the analysis described in the previous chapter. The first section discusses the interpolated surfaces produced from Empirical Bayesian Kriging which show the spatial distribution of lead concentrations in the soil within the various study areas. The second section describes the results from the various aggregation methods used to summarize the interpolated surfaces into the different scales of block groups and blocks. The results indicate how scale affects the values allocated into these various zones which might be used to identify priority areas for cleanup. The values are the mean, percent area, and Hazard Quotient (HQ) for each geographical unit. The results also compare which areas would be chosen for cleanup, depending on the aggregation value and scale used to base the decision. The last section discusses the results for the parcels aggregation method and compares the findings to the Representative Soil Lead Concentration DTSC is using for the current cleanup plan by mapping the differences in values.

4.1. Empirical Bayesian Kriging

The minimum and maximum interpolated values for each study area, as well as their comparative minimum and maximum sampled values, can be found in Table 4. The interpolated surfaces produced by Empirical Bayesian Kriging are displayed in Figures 10 to 12. The surfaces demonstrate the wide range of lead concentration values distributed throughout the study areas, with noticeably higher concentrations in the Northern Study Area. The Northern Study Area maximum value of 7,299.61 determined the upper limit of the classification for the maps, 7,300 ppm, while the significant value of 400 ppm, as the US EPA standard for lead exposure, helped determine the remaining classification numbers.

Table 4 Minimum and Maximum – Sampled vs. Interpolated Values for Each Study Area

Minimum and Maximum – Sampled vs. Interpolated Values			
	Northern Study Area	**1st Southern Study Area**	**2nd Southern Study Area**
Minimum Sampled Value (ppm)	8.82	10	19
Minimum Interpolated Value (ppm)	13.33	6.64	25.93
Maximum Sampled Value (ppm)	7,348.04	4,902.00	1,699.50
Maximum Interpolated Value (ppm)	7,299.61	1,973.16	584.59

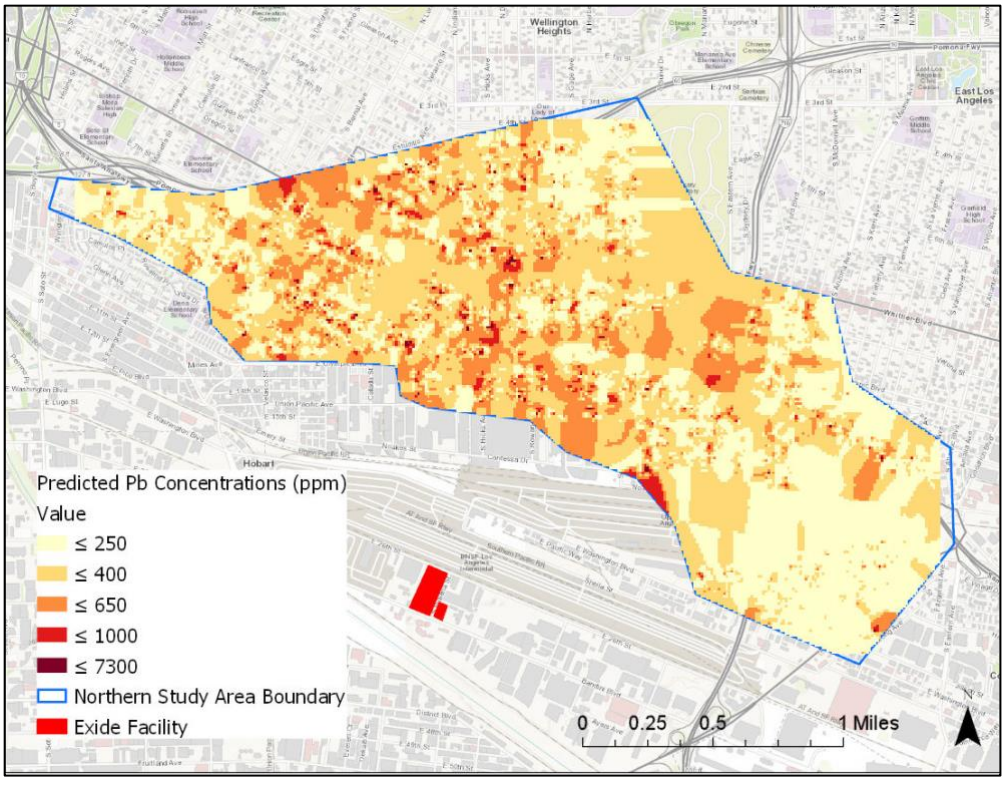


Figure 10 Northern Study Area Interpolated Surface

A majority of the predicted lead concentration values in the Northern Study Area appear to be in the 0-250 ppm and the 250-400 ppm ranges, while the 0-250 ppm range is the majority

in both of the Southern Study Areas. Both the Northern and the 1st Southern Study Areas have 1,000-7,300 ppm as their highest concentration range. The ranges with higher concentrations are dispersed throughout the two study areas as well, indicating the wide range of area in which higher lead contamination values may exist. These study areas are also the closest to the Exide facility, while the 2nd Southern Study Area is the furthest. There is a cluster of high lead concentration values on the west side of the 1st Southern Study Area, while the 2nd Southern Study Area exhibits a cluster of values in the 250-400 ppm and 450-600 ppm ranges on the east side.

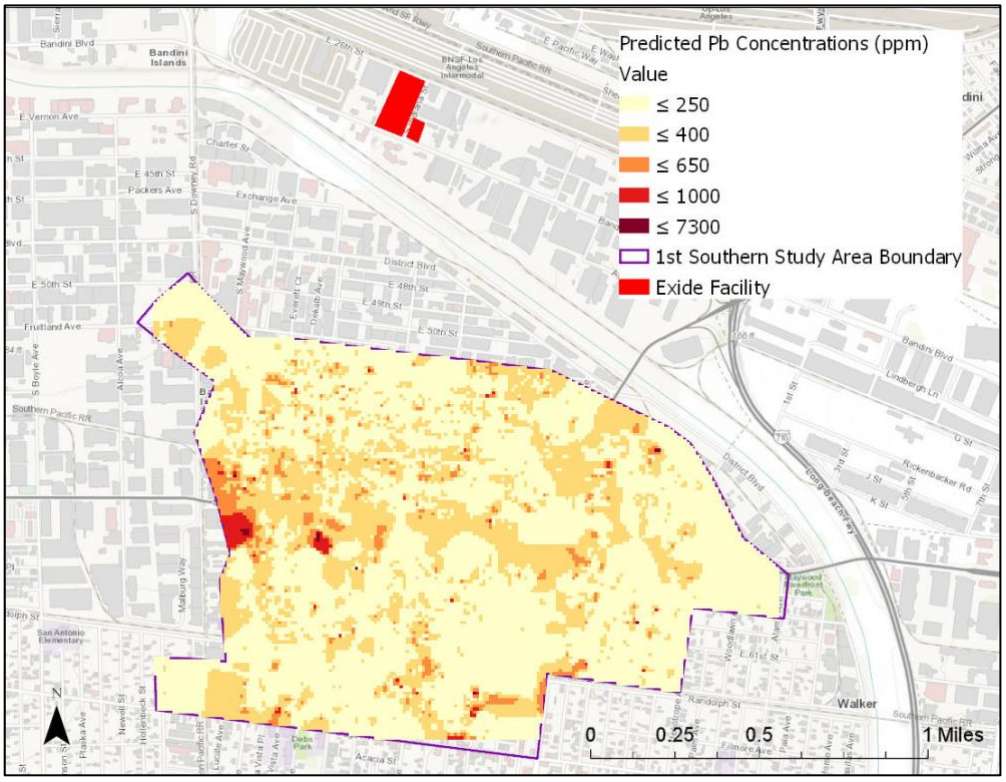


Figure 11 1st Southern Study Area Interpolated Surface

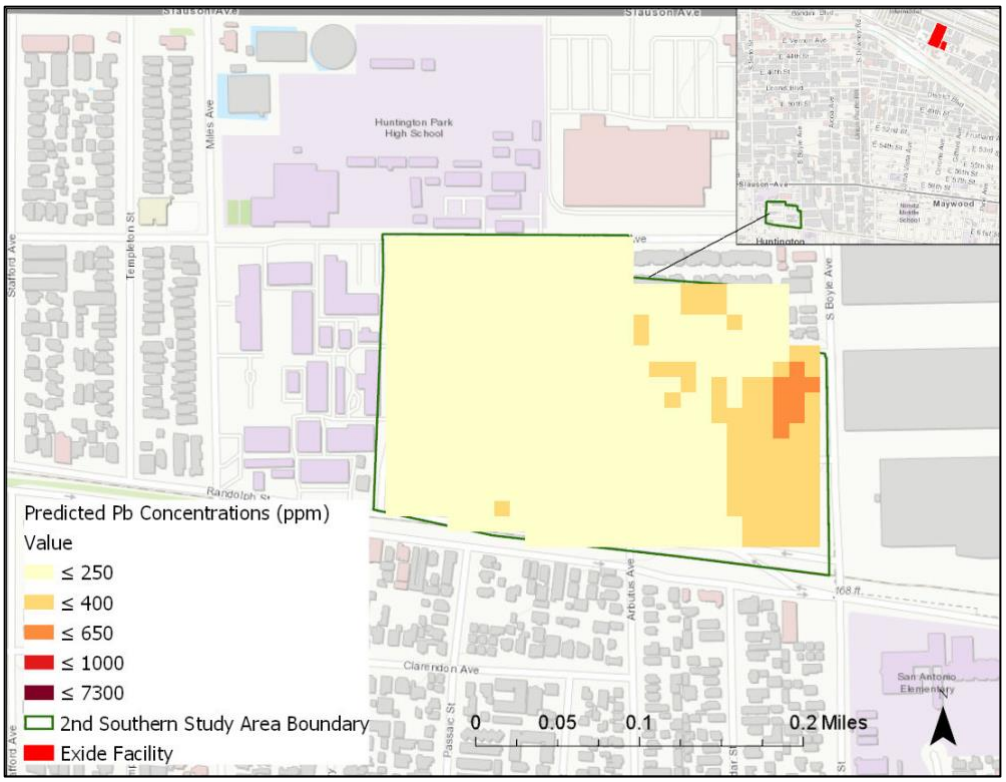


Figure 12 2nd Southern Study Area Interpolated Surface

4.2. Aggregation Results for Block Groups and Blocks

As expected, the results of the analysis demonstrate how scale affects the values allocated to various zones that may be used for determining priority areas for cleanup. Generally, the values of mean, percent area, and Hazard Quotients assigned to block groups are smaller than when assigned to blocks. To give a big picture of the difference in scale between block groups and blocks and how these geographical units may affect cleanup decisions, Table 5 shows the total number of block groups and blocks that would be selected for cleanup where the statistical mean value exceeds 400ppm, along with the number of parcels that would be selected for cleanup within those units. The number of parcels is used as a reference to compare block groups and blocks, indicating that this would be the number of parcels that would need to be cleaned if the whole block or block group were slated for cleanup.

Table 5 Comparison between the number of block groups and blocks within each study area and the number of units selected for potential cleanup

	Northern Study Area		**1st Southern Study Area**		**2nd Southern Study Area**	
	Blocks	**Block Groups**	**Blocks**	**Block Groups**	**Blocks**	**Block Groups**
Total # of Units	492	47	268	30	18	3
# of Selected Units for Cleanup	59	2	11	0	0	0
# of Parcels Selected for Cleanup	795	301	91	0	0	0
Total # of Parcels in the Study Area	6,619 parcels		4,467 parcels		353 parcels	

In the Northern Study Area, 2 out of 47 block groups and 59 out of 492 blocks had mean values greater than or equal to 400 ppm and are thus selected for potential cleanup. This is equivalent to 301 and 795 parcels out of 6,619 parcels, respectively. The effect of scale on deciding priority areas for cleanup is clearly evident in these results, in addition to the fact that only 11 blocks and 0 block groups were selected for potential cleanup in the 1st Southern Study Area and 0 blocks or block groups were selected in the 2nd Southern Study Area. The following subsections detail the results according to study area.

4.2.1. Northern Study Area

The results for the Northern Study Area are shown in Table 6, as well as in the following maps. Table 6 displays summary statistics of the result tables for blocks and block groups in the Northern Study Area, including the mean, minimum, maximum, range, and standard deviation for each value: mean, percent area, and Hazard Quotient.

Table 6 Summary Statistics for the Northern Study Area

Northern Study Area				
		Analysis Approaches		
Scale	**Summary Statistic**	**Mean**	**Percent Area Over 400 ppm**	**HQ**
Blocks				
	Mean	307.03	18%	0.77
	Minimum	111.96	0%	0.28
	Maximum	759.58	100%	1.90
	Range	647.62	100%	1.62
	Standard Deviation	90.27	23.6%	0.23
Block Groups				
	Mean	311.98	19.3%	0.78
	Minimum	183.79	0%	0.46
	Maximum	526.00	98.1%	1.31
	Range	342.21	98.1%	0.86
	Standard Deviation	68.48	19.2%	0.17

While all of the statistics provide insight into how scale affects the allocation of values, the mean statistics is one of the more indicative statistics. The average for all of the mean values in the blocks is 307.03 ppm, while the average for all of the mean values in the block groups is 311.98 ppm. These results are contrary to the tendency for values to decrease when the area increases. This is due to an edge effect, where boundaries on the edges of the study area get assigned values from only a small part of the raster surface that intersects with them (Figure 13). In addition, some block groups on the edges contain extra blocks that were not used in the analysis for the block scale (Figure 13). The consequences of the edge effect are discussed in Chapter 5.

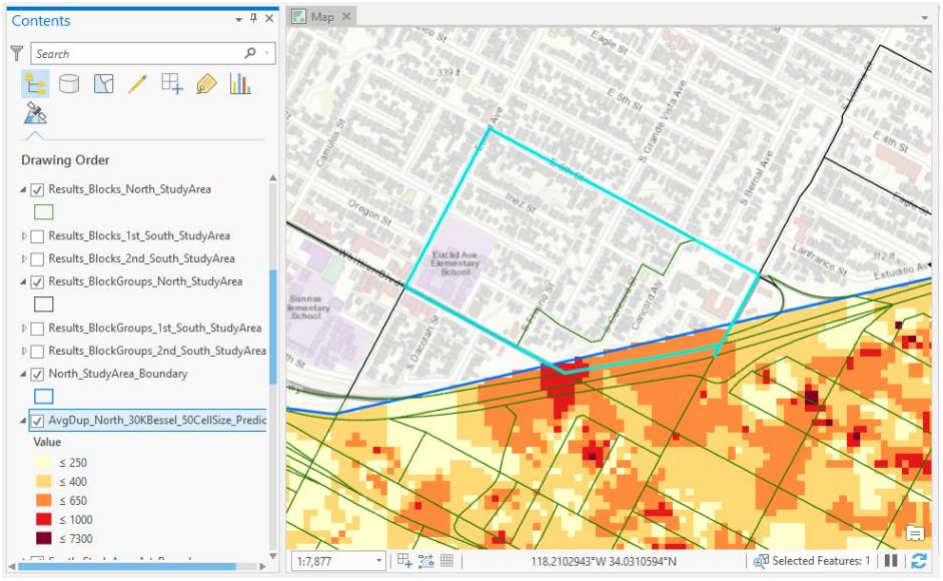


Figure 13 Edge Effect of Boundaries

The Hazard Quotient values correspond with the mean values, since these values are directly calculated using the mean values. The average HQ for blocks is 0.77, while the average HQ for block groups is 0.78. The average percent area value for blocks is 18%, while the average percent area value for block groups is 19.3%. This indicates blocks having 18% of their area where lead concentration values exceed the nationally recommended exposure value of 400 ppm. The values for percent area on the edge boundaries have also been affected by the edge effect and are examined more in Chapter 5.

Figure 14 displays the range of mean and HQ values for block groups and blocks in the Northern Study Area. The Quantile classification method was used to classify the values for all of the result maps for each study area. Although this classification method is considered to have limitations, the Quantile classification works well in displaying the results for this analysis. The breaks of ranges between block groups and blocks using the Quantile classification are so similar that any differences can be considered negligible. As expected, the block results have a wider range of values than the block groups. This can be seen in the maps of block groups versus

blocks. The block groups that have higher values do however correspond to blocks also with high values. The white hatched fill demonstrates areas that would be designated for cleanup on the basis that their mean value is greater than or equal to 400 ppm.

Figure 15 displays the range of percent area values for block groups and blocks in the Northern Study Area. While the block groups and blocks generally correspond and have a similar range in values, the maps show the difference a larger or smaller area has when determining how much of the area is over the 400 ppm lead concentration. The percent area values are also affected by the edge effect of the boundaries.

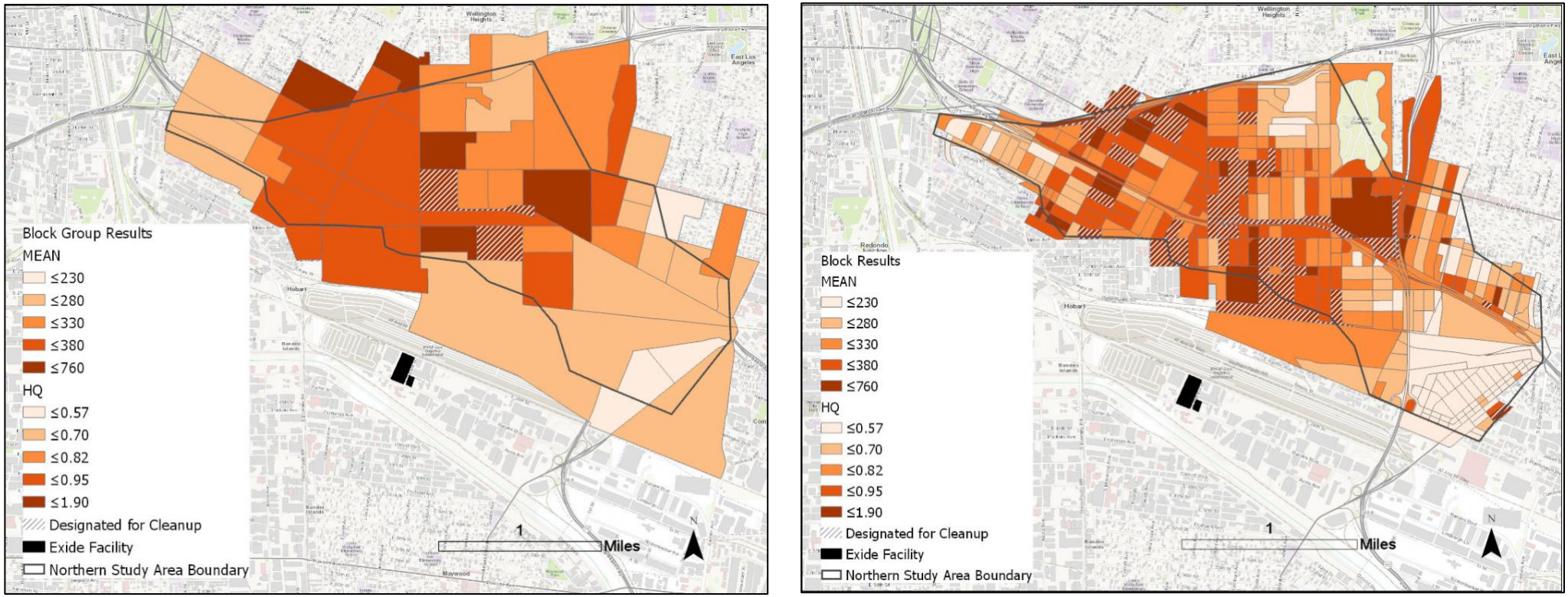


Figure 14 Block Group and Block Results for the Northern Study Area – Mean and HQ

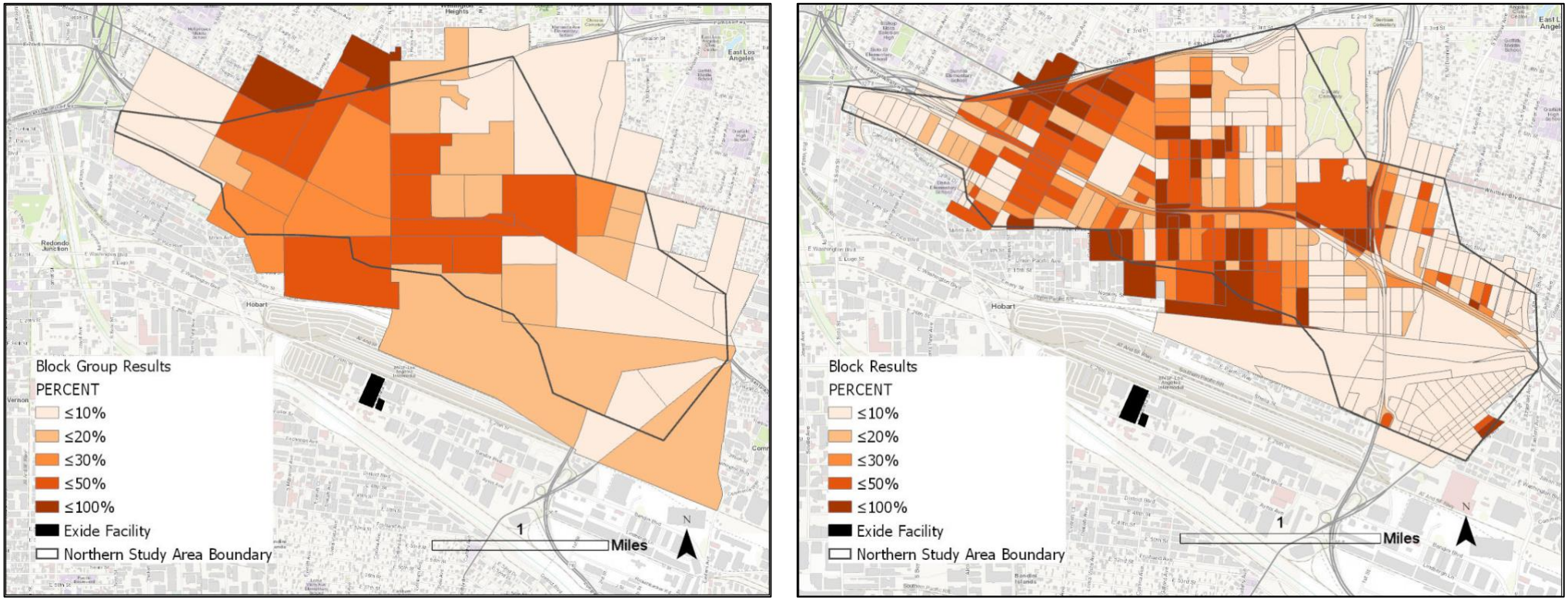


Figure 15 Block Group and Block Results for the Northern Study Area – Percent Area

4.2.2. 1st Southern Study Area

The results for the 1st Southern Study Area are shown in Table 7, in addition to the maps depicted in the following figures. Table 7 contains summary statistics of the result tables for blocks and block groups in the 1st Southern Study Area, including the mean, minimum, maximum, range, and standard deviation for each value: mean, percent area, and Hazard Quotient. The average for all of the mean values in the blocks is 237.65 ppm, while the average for all of the mean values in the block groups is 225.12 ppm. The average HQ for blocks is 0.59, while the average HQ for block groups is 0.56. These results are consistent with the tendency for values to decrease when there is an increase in area. The average percent area value for blocks is 6.2%, while the average percent area value for block groups is 4%.

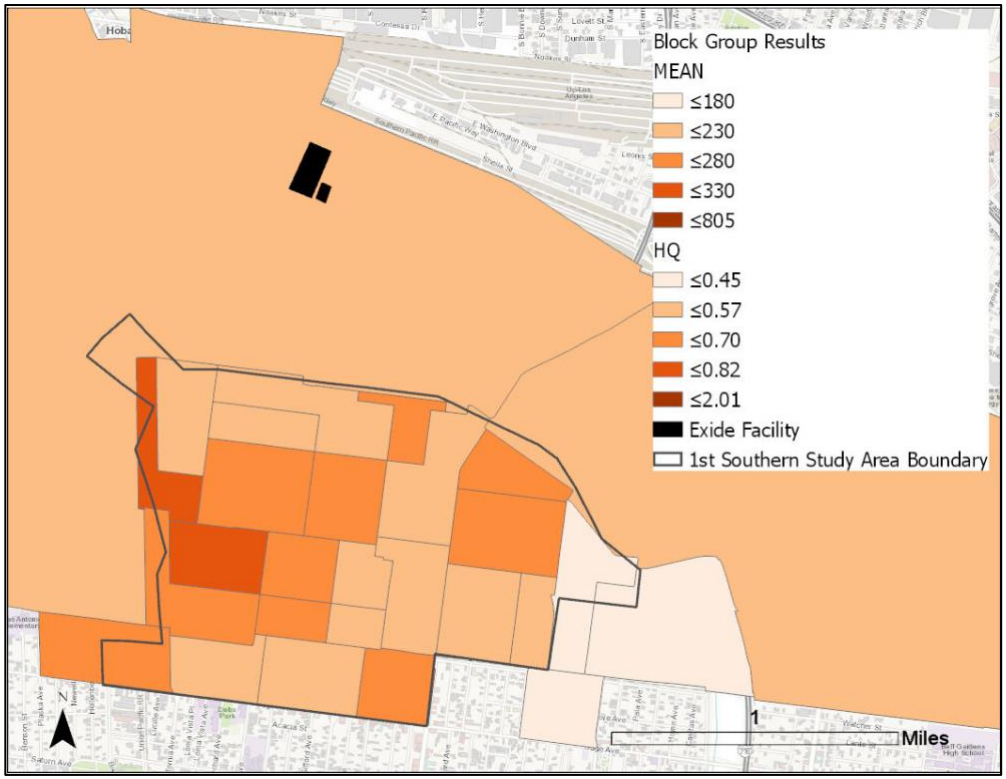
Table 7 Summary Statistics for the 1st Southern Study Area

1st Southern Study Area				
		Analysis Approaches		
Scale	**Summary Statistic**	**Mean**	**Percent Area Over 400 ppm**	**HQ**
Blocks				
	Mean	237.65	6.2%	0.59
	Minimum	115.06	0%	0.29
	Maximum	804.34	100%	2.01
	Range	689.28	100%	1.72
	Standard Deviation	83.83	15.4%	0.21
Block Groups				
	Mean	225.12	4%	0.56
	Minimum	129.46	0%	0.32
	Maximum	313.41	22.4%	0.78
	Range	183.95	22.4%	0.46
	Standard Deviation	40.18	5.1%	0.10

Figure 16 displays the range of mean and HQ values for block groups and blocks in the 1st Southern Study Area. Similar to the Northern Study Area, the block results have a wider range of values compared to the block group results. Both the block group and block maps show

the highly concentrated area of lead on the western side of the study area. The results for the 1st Southern Study Area carry the same observations noted in the previous subsection for the Northern Study Area. It is also important to note that a few of the block groups on the edges of the study area only have a portion of their zones intersecting with the study area boundary. This is noticeably observed in the two very large block groups that extend past the 1st Southern Study Area boundary, covering a large portion of the map (Figures 17 and 18). This also demonstrates the consequences of the edge effect.

Figure 17 displays the range of percent area values for block groups and blocks in the 1st Southern Study Area. The maps for this study area clearly demonstrate the notion that larger areas (block groups) will have a greater number of concentration values used towards the calculation of the aggregation values to be applied to their areas, while smaller areas (blocks) will have a smaller amount of concentration values used towards the calculation of the aggregation values for the zones. This is demonstrated on the eastern side of the study area, where the larger areas of the block groups likely had a greater number of high concentration values in determining the percentage of the block groups where lead concentrations are 400 ppm or higher. In contrast, since the blocks generally have smaller areas compared to the block groups, certain blocks will not have the high concentration values picked up by the larger area of the block group that encompasses it. This can be seen in the same area on the eastern side where there is a greater number of blocks colored the lightest shade, representing 0 percent of the area having concentrations greater than or equal to 400 ppm. The percent area value also emphasizes the high lead concentrations on the western side of the study area.



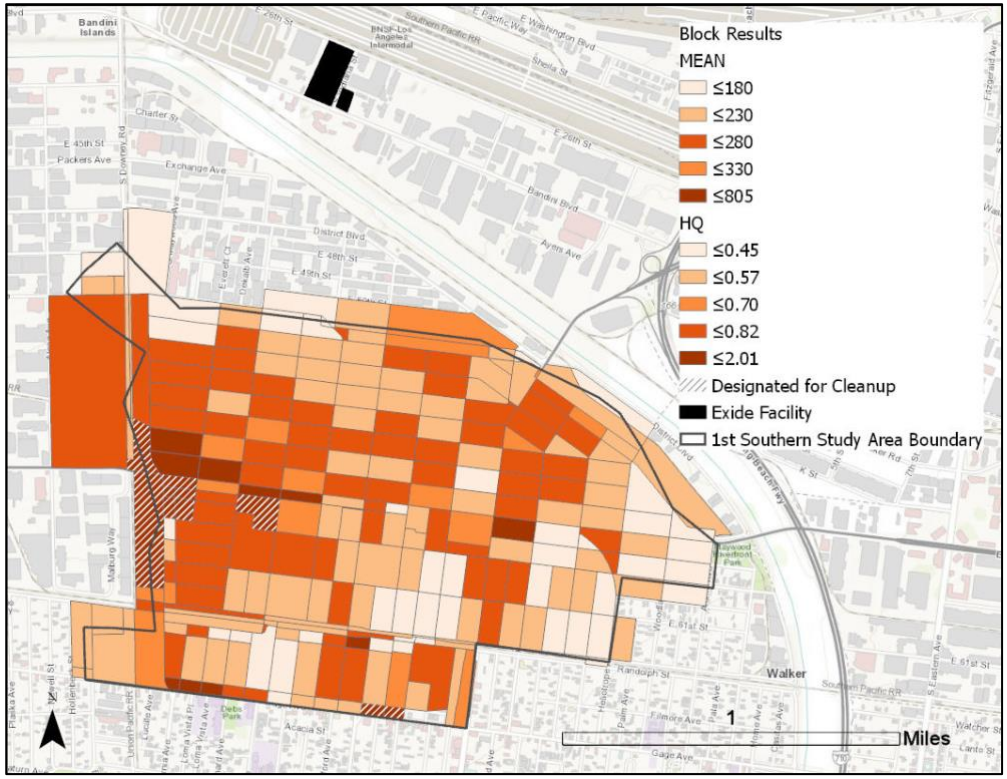


Figure 16 Block Group and Block Results for the 1st Southern Study Area – Mean and HQ

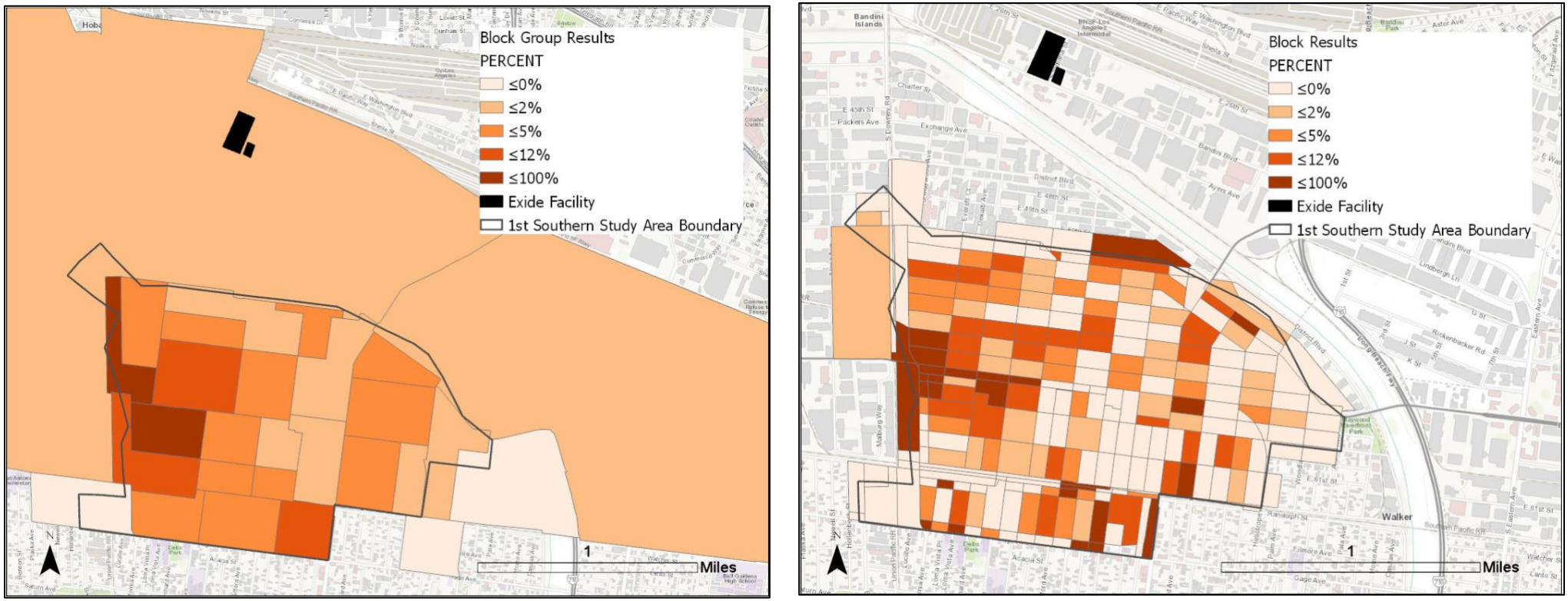


Figure 17 Block Group and Block Results for the 1st Southern Study Area – Percent

4.2.3. 2nd Southern Study Area

The results for the 2nd Southern Study Area are shown in Table 8, in addition to the maps represented in the following figures. Table 8 shows summary statistics of the result tables for blocks and block groups in the 2nd Southern Study Area, including the mean, minimum, maximum, range, and standard deviation for each value: mean, percent area, and Hazard Quotient. The average for all of the mean values in the blocks is 134.78 ppm, while the average for all of the mean values in the block groups is 124.84 ppm. The average HQ for blocks is 0.34, while the average HQ for block groups is 0.31. These results are also in agreement with the tendency for values to decrease when there is an increase in area. The average percent area value for blocks is 0.8%, while the average percent area value for block groups is 0.9%.

Table 8 Summary Statistics for the 2nd Southern Study Area

2nd Southern Study Area				
		Analysis Approaches		
Scale	**Summary Statistic**	**Mean**	**Percent Area Over 400 ppm**	**HQ**
Blocks				
	Mean	134.78	0.8%	0.34
	Minimum	65.77	0%	0.16
	Maximum	294.04	10.1%	0.74
	Range	228.27	10.1%	0.57
	Standard Deviation	71.91	2.9%	0.18
Block Groups				
	Mean	124.84	0.9%	0.31
	Minimum	108.22	0%	0.27
	Maximum	141.45	1.8%	0.35
	Range	33.23	1.8%	0.08
	Standard Deviation	23.5	1.3%	0.06

Figure 18 displays the range of mean and HQ values for block groups and blocks in the 2nd Southern Study Area. Due to the overall small size of the 2nd Southern Study Area, the maps clearly demonstrate how values get partitioned into the different scales of geographical units and

can change which class/range the units get assigned to, based on how many individual values are being considered when the overall value is calculated for that unit. It can be easily seen that the one block group that covers the entire study area gets assigned to a lower class/range than its block counterparts. The blocks have a more diverse set of ranges, with the eastern side of this study area containing higher concentration values, thus putting some of the blocks in a higher class/range. With such a small study area, it is clearly shown how the larger area of a block group can moderate the high concentration values that would be noticed in the smaller areas of blocks. It is also important to note that the edge effect is seen within this study area as well, where only a small portion of some blocks and the one block group intersect with the study area boundary.

Figure 19 displays the range of percent area values for block groups and blocks in the 2nd Southern Study Area. The small size of the study area also provides insight into the previous observations on the size of areas affecting percent area values. As shown in the map, the one high value in the one block is accounted for in the whole block group that covers the entire study area. This is somewhat contrary to the notion previously presented that larger areas moderate high concentration values. While although that may be true for the mean and HQ values, this contrast shows a difference in outcomes between the mean and HQ values and the percent area values. For percent area, a larger area means that all the values in the blocks get accounted for in the block group, which can affect the outcome of the percent area where lead concentrations exceed the 400 ppm value.

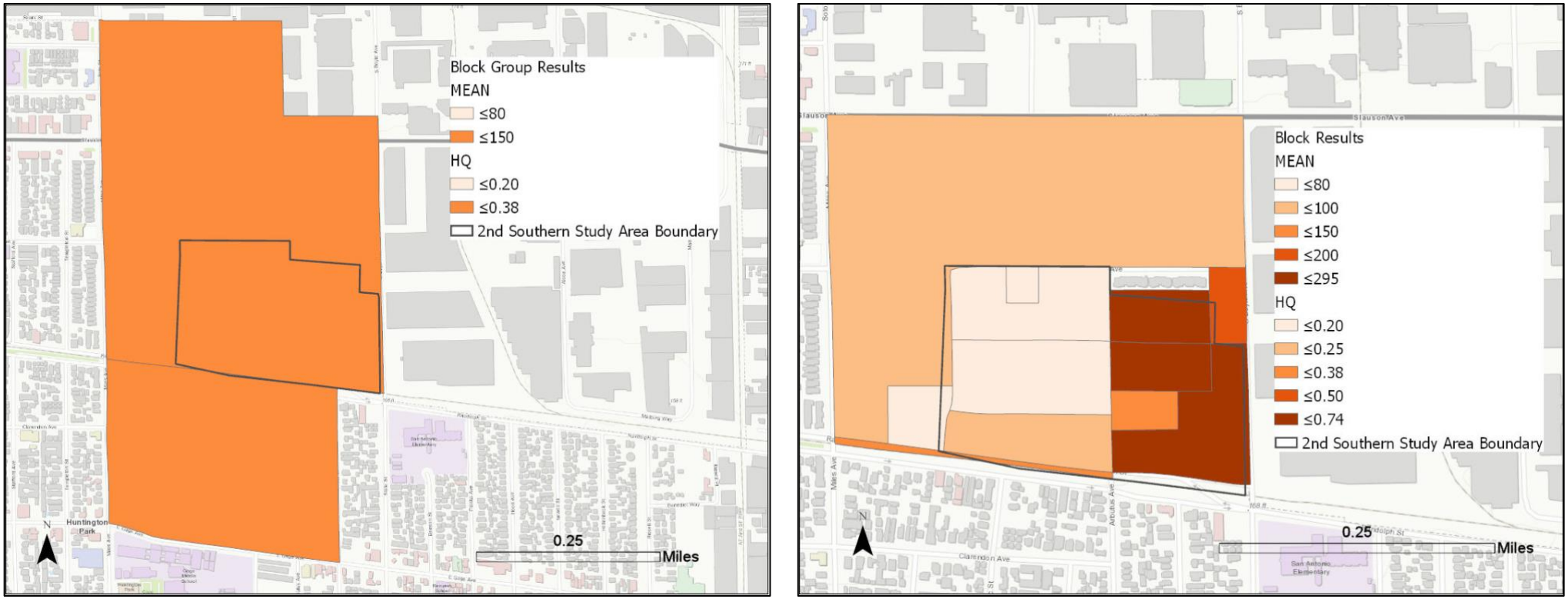
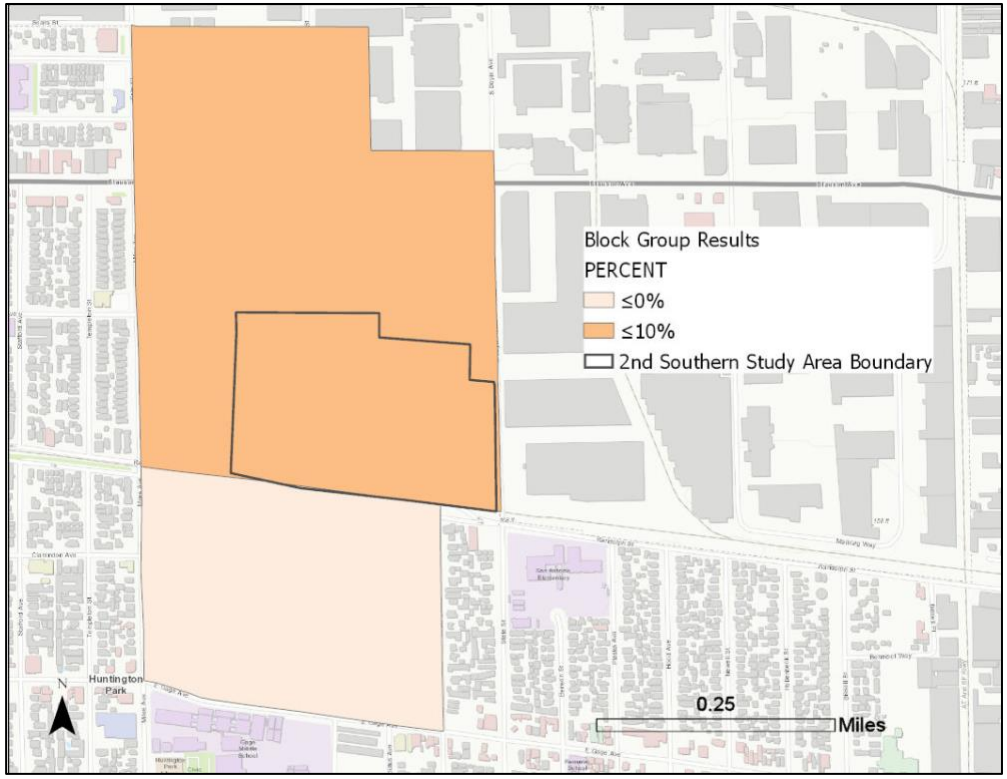


Figure 18 Block Group and Block Results for the 2nd Southern Study Area – Mean and HQ



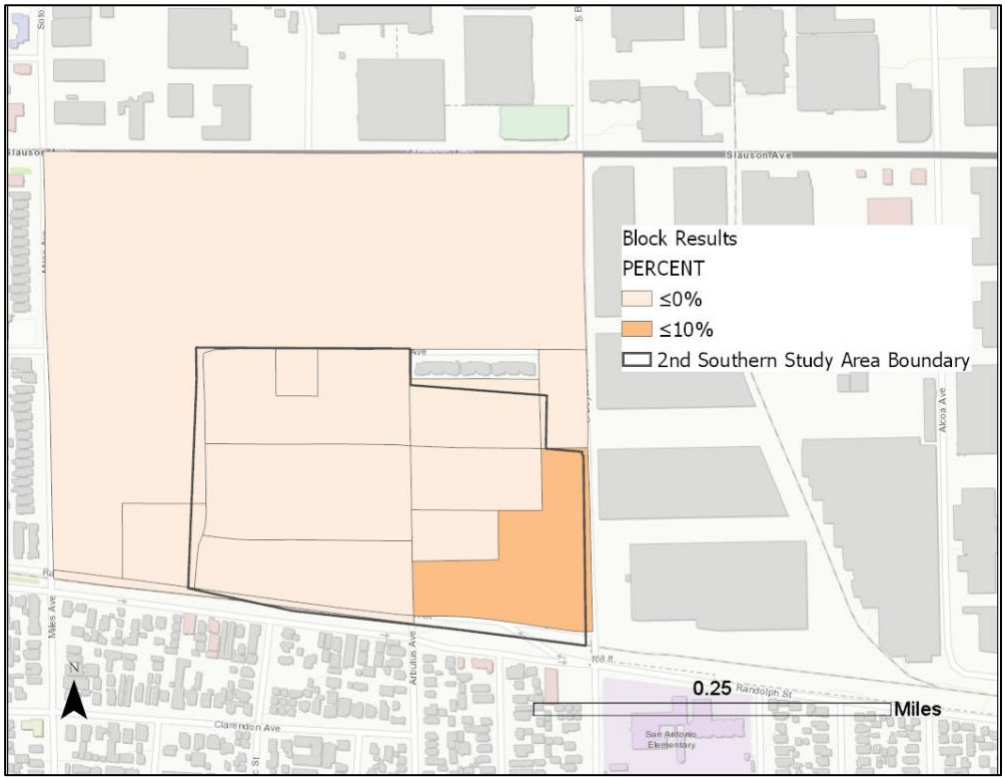


Figure 19 Block Group and Block Results for the 2nd Southern Study Area – Percent

4.3. Aggregation Results for Parcels

For the parcel aggregation, the method utilized a tool that extracted the cell values from the rasters at the centroid of each parcel. The results for all the study areas are presented as summary statistics in Table 9. The resulting values are compared to the original Representative Soil Lead Concentration values determined by DTSC through differences. The difference calculation is as follows: Representative Soil Lead Concentrations minus resulting values from the cells. The positive differences indicate that the Representative Soil Lead Concentration is greater than the cell value, while the negative differences indicate that the cell value is greater than the Representative Soil Lead Concentration. These differences were then mapped, as shown in Figures 20-22. Overall, there are more positive differences than negative differences, meaning in general, the interpolation surface method produced generally lower concentration values.

Table 9 Summary Statistics for Cell Values Extracted from Interpolated Surfaces at Parcel Scale

Summary Statistics for Parcels			
Summary Statistic	**Northern Study Area**	**1st Southern Study Area**	**2nd Southern Study Area**
Mean	328.58	241.51	122.48
Minimum	21.30	30.39	25.93
Maximum	7299.61	1973.16	576.03
Range	7278.31	1942.77	550.10
Standard Deviation	182.41	106.5	88.35

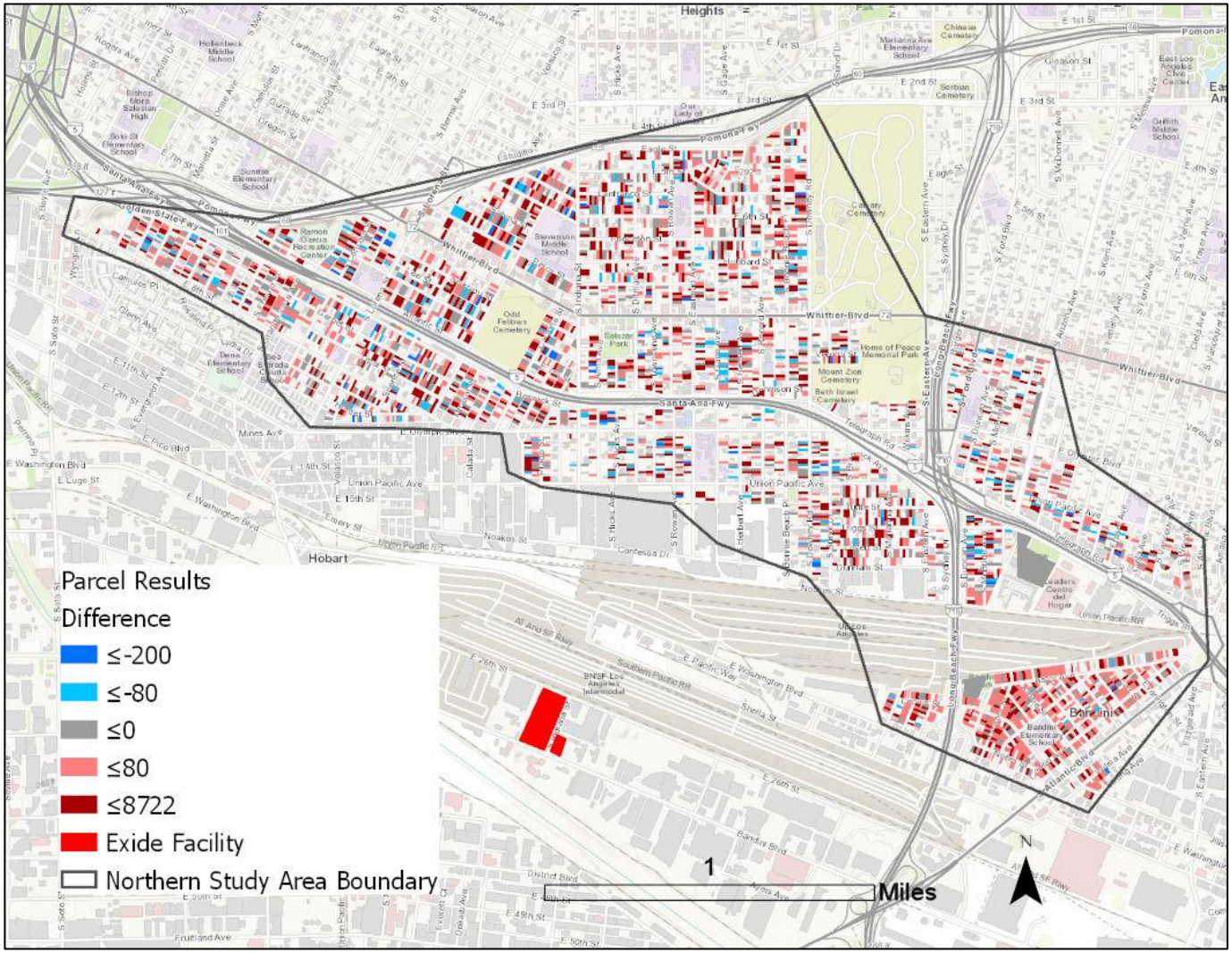


Figure 20 Difference in Parcel Values for the Northern Study Area

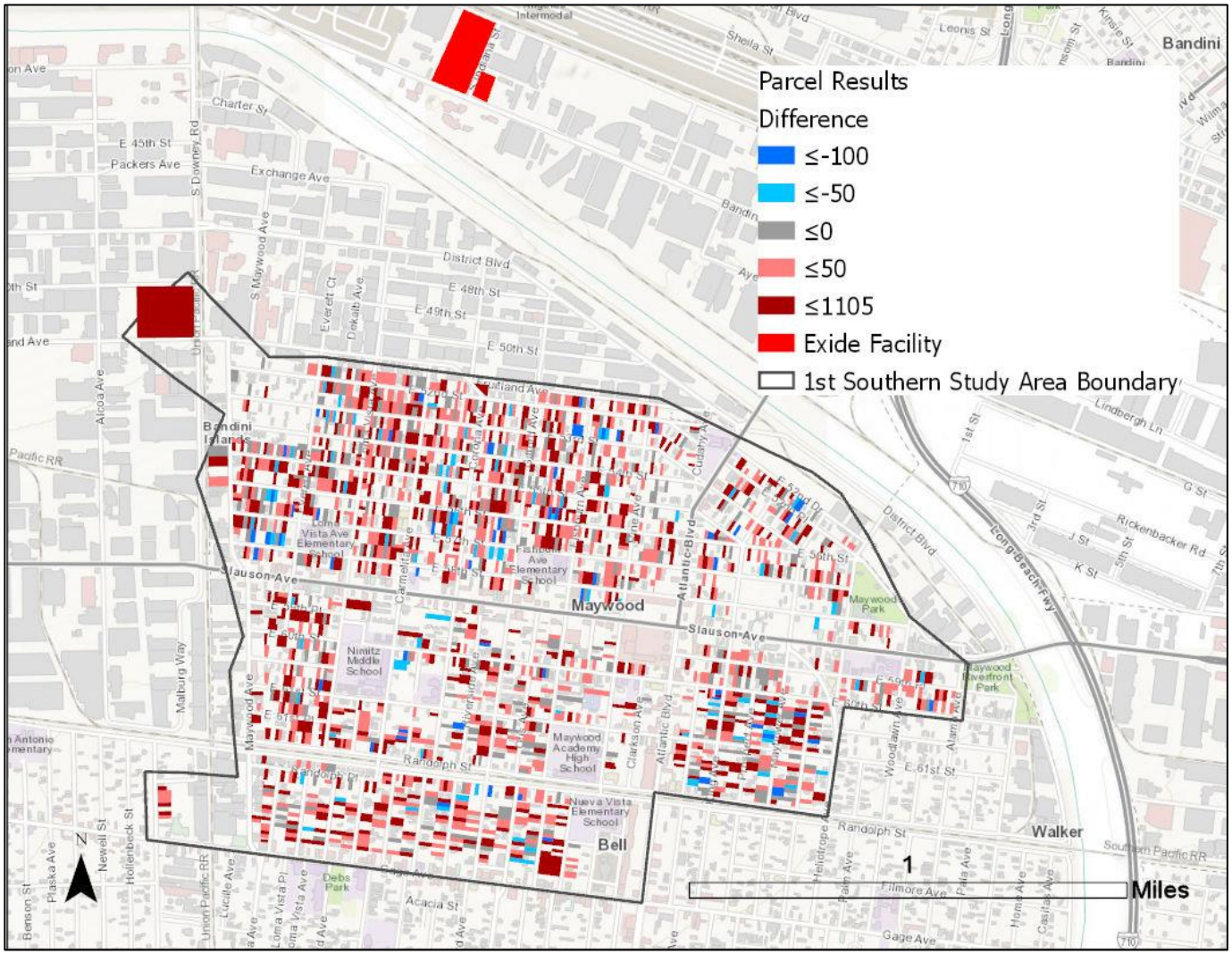


Figure 21 Difference in Parcel Values for the 1st Southern Study Area

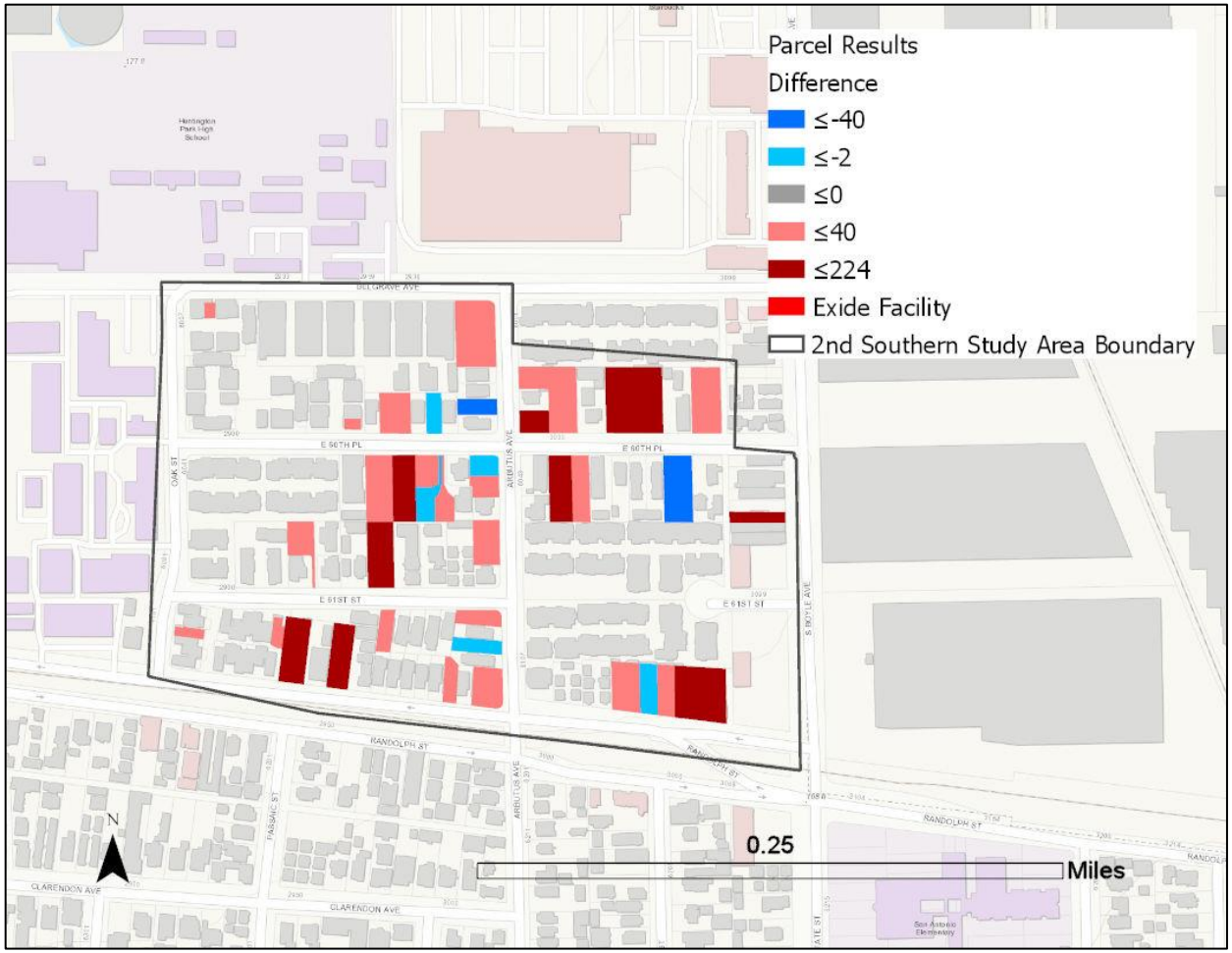


Figure 22 Difference in Parcel Values for the 2nd Southern Study Area

Chapter 5 Discussion and Conclusions

This study examined the effect of scale on aggregated values into various zones and the role spatial scale has in cleanup efforts. This chapter concludes this thesis and provides a deeper discussion of the results, with explanations for the results and the effects of using these boundaries. The chapter concludes with limitations of the study, overall implications, and suggestions for further research.

5.1. Discussion

This section provides a deeper discussion of the results from the analysis, offering explanations for both the results and the effects of using block groups, blocks, and parcels as aggregation boundaries. It also points out general conclusions from the results, both from the block groups and blocks aggregation as well as the parcels aggregation.

5.1.1. Block Groups and Blocks Aggregation Discussion

As evidenced by the results, this study clearly demonstrates how scale affects the values allocated to different zones used for determining priority areas for cleanup. For all of the study areas, the mean and HQ values for the blocks have wider ranges than those for the block groups. This can be attributed to the high outlier values being more influential in the block aggregations than the block group aggregations. Since the block groups are a larger unit and have a greater area, the same few high values present in the blocks get moderated through the averaging with more lower values present in the larger area of the block groups. This notion explains why the mean and HQ values for both Southern Study Areas are higher for blocks and lower for block groups. This finding is also present in the inherent nature of the dataset. As visualized by the

interpolated surfaces, the dataset contains a lot of low to moderate lead concentration values, with a few high lead concentration values dispersed throughout the study areas.

In addition to scale affecting the mean and HQ values, different scales affect the percent area values in much of the same way. Although the ranges for the percent area values are similar for both block groups and blocks among all the study areas, the larger area of the block groups contains more values in the calculation for how these values spread across the block group areas, while the blocks have less values in the calculation. Since the values for the percent area were calculated using binary rasters, where any cell containing a concentration value greater than or equal to 400 ppm became a 1, the larger areas of the block groups have a greater chance of containing more cells reclassified as 1's than blocks. The result, a percentage representing the percent area of a polygon where lead concentration values exceed 400 ppm, is then reflected in the percentage values for the differing scales of block groups and blocks.

Although the results are largely indicative of how scale affects the mean, percent area, and HQ values partitioned into the geographical units, there are some inaccuracies present due to edge effects that occur with the boundaries. For both the block group and block results, there are certain block groups and blocks on the edges of the study areas that receive assigned values from only a small part of the interpolated surface that intersects with them. Essentially, the small part of the surface that does intersect with these boundaries determines the values for the whole boundaries.

In addition, some of the same block groups on the edges contain extra blocks that were not used in the analysis for the block scale, as those blocks were not present when selecting the block boundaries based on the study areas. These edge effects impact the summary statistics tables and show in the maps for each of the study areas, particularly in the Northern Study Area.

One of the noticeable consequences of the edge effects is in the mean and HQ values in the summary statistics table for the blocks and block groups in the Northern Study Area. The edge effects explain why the average values for the block groups are higher than the blocks, contrary to how it should be. It also causes some blocks and block groups to be selected as potential cleanup areas, even though no soil samples were taken in those areas.

Despite the consequences of the edge effects, it is a natural outcome from constructing the study area boundaries around the soil sample points. It is a compounding limit of boundaries in a sense. The study area boundaries were determined from the soil sample points, then the block group and block boundaries that intersected with the study area boundaries were selected using Select by Location. Thus, it is the polygons themselves that impacted the results, as opposed to decisions made.

5.1.2. Parcels Aggregation Discussion

For the parcel aggregation, the larger number of positive values compared to negative values suggest that the Representative Soil Lead Concentrations are higher than the polygon centroid cell values pulled from the raster surfaces. This signifies that the values pulled from the surfaces tend to be lower than the determined Representative value for that parcel. This could pose problems if this method was to be used for potential cleanup, as it will miss the higher outlier values present in the dataset, which are important in determining health effects. This aggregation method for the parcels proves to be more useful for comparing the surface values to the Representative Soil Lead Concentrations and can be used for determining the accuracy of the surfaces, as opposed to an actual method used for decision-making on cleanup.

5.2. Limitations

Although this study provides a successful look into the effect of geographic scale in lead contamination, it was not without limitations. One of the main limitations for this study was issues in data quality. While most of the soil sample data had coordinates associated with the soil sample points, there were a handful of samples that needed to be geocoded and another handful of samples that had been inaccurately geocoded and therefore needed to be geocoded again. While these problems were fairly simple fixes, they could have produced inaccuracies in the mapping of the soil samples in terms of where their actual locations may have been in the field. Data quality could also be improved in the initial phase of data acquisition through choice of sample sites and methods. More sample sites, or differently chosen sample sites, could lead to greater accuracy in the interpolated surfaces.

In addition, there is an issue with coordinate accuracy from the original soil sample data. Some of the coordinates associated with the samples have less than six decimal places, which has a precision of 7-10 meters. This could cause these samples to be placed in roughly 1-2 parcels away from the actual sampled parcel, if not in the same parcel, depending on the size of the parcel. To solve this issue, the coordinates with less than six decimal points would need to have their addresses geocoded again as well. This issue of coordinate accuracy could explain the data quality issues seen in the earlier data exploration, where some of the points fell on top of buildings and some sample's Primary Assessor Parcel Numbers did not match with the Assessor IDs for the parcels in the parcel data table.

Another major limitation for this thesis was not being able to use the same aggregation methods and values used for the block group and block scales, for the parcel scale. Initially, the aggregation methods and values of mean, percent area, and HQ were intended to be used for all

the scales of analysis, including block groups, blocks, and parcels. This approach would have made the effects of scale on lead contamination and the determination of priority areas for cleanup even more clear.

However, when attempting to use these methods for the parcel scale, it was discovered that the Zonal Statistics as a Table tool only produces results that use the center of the cells to determine which polygon the cell values get partitioned to, as opposed to splitting the cell so that part of the cell's value goes to one polygon and the other part goes to another. This would need to happen if a single cell overlaps two boundaries. At the parcel level, some parcels do not contain any centroids of the interpolated cells, which created "holes" in the aggregation results, as the results do not include parcels that do not contain any centroids of cells. Numerous tools and solutions, such as turning the rasters into polygons or decreasing the cell size used, were attempted, but the same problem still existed. This indicates that the parcel scale may be too small to be used for these aggregations, unless the method is found to appropriately apportion the cell values into the different parcels.

As demonstrated by the edge effects, a third limitation of this study includes the boundary issues that arise from these edge effects. The limitation being that the areas at the edge boundaries may have some inaccuracies in values. A solution would be to instead use boundaries that are only completely within the study areas. However, Figure 23 depicts that if this solution were to be used, it would severely trim down the number of blocks and block groups used for the analysis and the boundaries would not cover the whole study areas, which means not using some of the soil samples provided within the study areas. This would lead to an even greater limitation in the results. Therefore, it must be accepted that the edge effects will occur when using these specific geographic units for analysis.

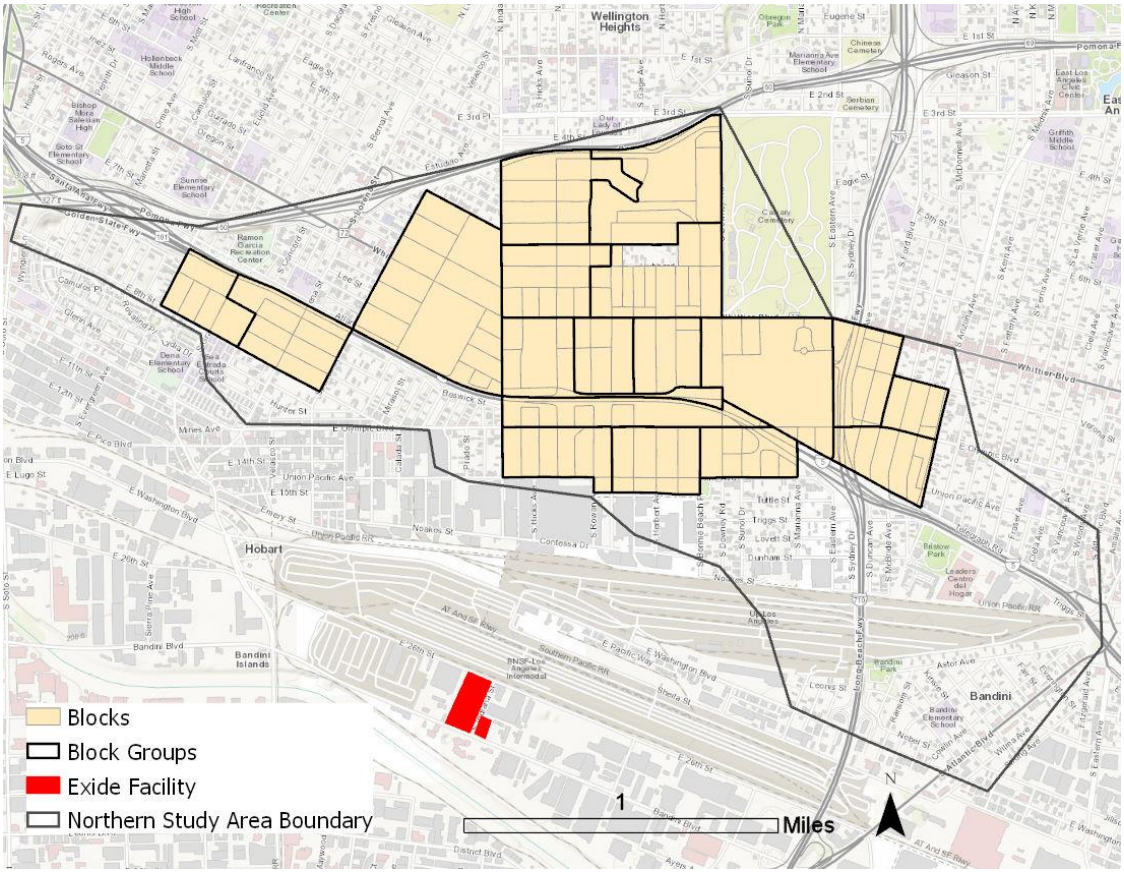


Figure 23 Boundaries Completely Within Northern Study Area

Considering the edge effects that occur with the choice of these geographical units for analysis, it would be better to set the boundary of the study areas to the geographic unit boundaries, rather than the soil samples. Because blocks and block groups were chosen as the units for analysis, framing the study area boundaries around these units would provide for less complications in the analysis, while still including the important locations of the points.

5.3. Future Research and Implications

Although this thesis successfully met its objective in demonstrating how scale can affect the allocation of values into the various zones that may be utilized to prioritize potential areas for cleanup, there are several improvements that could be incorporated into this research and ideas for future studies. One improvement that could be made is to make the interpolated surfaces

more accurate by considering the error that is associated with them. Since the aim of this thesis was to see how scale affects the aggregation of values into different areas, it was only necessary to develop suitable surfaces for the aggregation, as the surfaces were a means to an end and not the final result. However, one of the benefits of Empirical Bayesian Kriging is the option to create error surfaces associated with the predictions, which could be utilized in future studies to evaluate the interpolated surfaces. The use of different cell sizes for the creation of the surfaces could provide another insight into how scale may affect the aggregation of values as well.

In addition to making the surfaces more accurate, the use of the Representative Soil Lead Concentration values determined by DTSC for the creation of the interpolated surfaces rather than the raw values may help smooth the surfaces. Currently, the surfaces are more rigid and "bumpy," a reflection of the raw values and the parcel sizes. The use of the Representative Concentrations would smooth the surface of the raster, as only one value is associated with each parcel, as opposed to 10-15 sample values. The addition of the lab samples into the raw values used for analysis could also improve accuracy in the interpolated surfaces and may provide differing results.

To improve the accuracy of the aggregations, building outline data could be used to clip the rasters, thereby eliminating some of the impervious surfaces that get assigned values due to the nature of raster creation. Since soil lead concentrations are really only prominent in permeable surfaces, taking out the areas where buildings are in the raster surfaces could better demonstrate where the distributions of the concentrations are. It would be interesting to see if high outlier values exist where buildings actually exist. Another future study could employ the use of different boundaries other than block groups, blocks, and parcels for the aggregations of

the values. This would change the results according to the boundaries used. It is a possibility that there could be better determined boundaries to use for policy implementation.

Furthermore, the soil sample dataset is quite extensive and contains samples and data not utilized in this analysis. Future research can take these additional data into account. There are samples taken from paint chips within the parcels. These samples could possibly be utilized in a correlation analysis with the soil samples to determine how much the concentrations are affected by the paint from the age of the homes. Samples of different depths were also taken, which could be utilized in a depth variation analysis. In addition, other heavy metals, such as arsenic, were sampled in the soil sampling process. The distribution of these heavy metal samples could be determined and examine how scale may affect the allocation of these values into zones for cleanup. With this knowledge, it could be determined if the highly concentrated areas of the other heavy metals overlap with the highly concentrated areas of lead.

While there is much future research that could be accomplished in relation to Exide and the soil samples gathered, this thesis provides important implications for the cleanup of lead contaminated soils. This thesis suggests alternatives for identifying priority areas for cleanup through different aggregation methods. The values resulting from these aggregations – the statistical mean, the percentage of an area where lead concentration values exceed the nationally recommended exposure value of 400 ppm, and a Hazard Quotient, an index value that determines the risk to human health – provide options for deciding which areas should be cleaned. The examination of scale, using block groups, blocks, and parcels, also provides insight into which scale would be best for managing the cleanup of areas. There is much that still needs to be considered in the decision-making process for cleanup, such as specified criteria, budget, timeline, permissions, etc. However, this thesis may help demonstrate to decision makers that

using a spatially informed methodology can help them in the decision-making process. Although these options may not be considered for the Exide site specifically, the conclusions from this study can be applied to other areas faced with the same problem in the future.

References

Barboza, Tony. 2015. “How a Battery Recycler Contaminated L.A.-Area Homes for Decades.” *Los Angeles Times*, 2015. http://www.latimes.com/local/lanow/la-me-exide-cleanup-story-so-far-20151121-story.html

Barboza, Tony, and Ben Poston. 2018. “The Exide plant in Vernon closed 3 years ago. The vast majority of lead-contaminated properties remain uncleaned.” *Los Angeles Times*, 2018. http://www.latimes.com/local/lanow/la-me-exide-cleanup-20180426-story.html

Chaney, Rufus, Howard Mielke, and Susan Sterrett. January 1989. "Speciation, Mobility, and Bioavailability of Soil Lead." [Proc. Intern. Conf. Lead in Soils: Issues and Guidelines. B.E. Davies & B.G. Wixon (eds.)]. *Environmental Geochemistry and Health* 11 (Supplement): 105-129).

Department of Toxic Substances Control. 2015. "Final Workplan Sampling and Analysis of Properties in the Vicinity of the Exide Facility (Vernon, California)." Pasadena: Parsons. https://www.dtsc.ca.gov/HazardousWaste/Projects/upload/Exide_Adv-Grp_Sampling-workplan.pdf

Department of Toxic Substances Control. 2017. "Final Removal Action Plan (Cleanup Plan) Offsite Properties Within the Exide Preliminary Investigation Area." Sacramento: State Clearinghouse. http://www.dtsc.ca.gov/HazardousWaste/Projects/Residential-Cleanup.cfm

Department of Toxic Substances Control. 2017. "Final Environmental Impact Report: Volume 1 Chapter 1 Through Chapter 2, Section 2.3.4." Pasadena: State Clearinghouse. https://www.envirostor.dtsc.ca.gov/public/community_involvement/9913273453/Final_EIR_Vol_1_Chapter_1_Executive_Summary.pdf

Department of Toxic Substances Control. 2018. *Soil Sampling Data for the Exide Preliminary Investigation Area*. July 12, 2018 version. Distributed by California Department of Toxic Substances Control. https://www.dtsc.ca.gov/HazardousWaste/Projects/pia-sampling-data.cfm

Esri. 2018. “What is Empirical Bayesian Kriging? — ArcGIS Pro | ArcGIS Desktop.” *Pro.Arcgis.Com.* https://pro.arcgis.com/en/pro-app/help/analysis/geostatistical-analyst/what-is-empirical-bayesian-Kriging-.htm.

Esri. 2019. "Comparing Models — Help | ArcGIS Desktop.” *Desktop.Arcgis.Com*. http://desktop.arcgis.com/en/arcmap/latest/extensions/geostatistical-analyst/comparing-models.htm.

Fajčíková, K., V. Cvečková, A. Stewart, and S. Rapant. 2014. "Health Risk Estimates for Groundwater and Soil Contamination in the Slovak Republic: A Convenient Tool for Identification and Mapping of Risk Areas." *Environmental Geochemistry and Health* 36 (5): 973-986. doi:10.1007/s10653-014-9612-9.

Gay, J. Rebecca, and Anna Korre. 2006. "A Spatially-Evaluated Methodology for Assessing Risk to a Population from Contaminated Land." *Environmental Pollution* 142 (2): 227-234. doi:10.1016/j.envpol.2005.10.035.

Ginevan, Michael, and Douglas Splitstone. 1997. "Improving Remediation Decisions at Hazardous Waste Sites with Risk-Based Geostatistical Analysis." *Environmental Science and Technology* 31 (2). doi:10.1021/es972129n.

Guastaldi, Enrico, and Andrea Alessandro Del Frate. 2011. "Risk Analysis for Remediation of Contaminated Sites: The Geostatistical Approach." *Environmental Earth Sciences* 65 (3): 897-916. doi:10.1007/s12665-011-1133-6.

Hanna-Attisha, Mona, Jenny LaChance, Richard Casey Sadler, and Allison Champney Schnepp. 2016. "Elevated Blood Lead Levels in Children Associated with the Flint Drinking Water Crisis: A Spatial Analysis of Risk and Public Health Response." *American Journal of Public Health* 106 (2): 283-290. doi:10.2105/ajph.2015.303003.

Hipp, John R. 2007. "Block, Tract, And Levels of Aggregation: Neighborhood Structure and Crime and Disorder as a Case in Point." *American Sociological Review* 72 (5): 659-680. doi:10.1177/000312240707200501.

Juberg, Daland R., Cindy F. Kleiman, and Simona C. Kwon. 1997. "Position Paper of the American Council on Science and Health: Lead and Human Health." *Ecotoxicology and Environmental Safety* 38 (3): 162-180. doi:10.1006/eesa.1997.1591.

"Kriging Interpolation - The Prediction is Strong in this One - GIS Geography." 2018. *GIS Geography*. https://gisgeography.com/Kriging-interpolation-prediction/.

Laidlaw, Mark A.S., Gabriel M. Filippelli, Sally Brown, Jorge Paz-Ferreiro, Suzie M. Reichman, Pacian Netherway, Adam Truskewycz, Andrew S. Ball, and Howard W. Mielke. 2017. "Case Studies and Evidence-Based Approaches to Addressing Urban Soil Lead Contamination." *Applied Geochemistry* 83: 14-30. doi:10.1016/j.apgeochem.2017.02.015.

Li, Xiangdong, Siu-lan Lee, Sze-chung Wong, Wenzhong Shi, and Iain Thornton. 2004. "The Study of Metal Contamination in Urban Soils of Hong Kong Using a GIS-Based Approach." *Environmental Pollution* 129 (1): 113-124. doi:10.1016/j.envpol.2003.09.030.

Liang, Bingqing, and Qihao Weng. 2008. "Multiscale Analysis of Census-Based Land Surface Temperature Variations and Determinants in Indianapolis, United States." *Journal of Urban Planning And Development* 134 (3): 129-139. doi:10.1061/(asce)0733-9488(2008)134:3(129).

Liu, Xingmei, Jianjun Wu, and Jianming Xu. 2006. "Characterizing the Risk Assessment of Heavy Metals and Sampling Uncertainty Analysis in Paddy Field By Geostatistics and GIS." *Environmental Pollution* 141 (2): 257-264. doi:10.1016/j.envpol.2005.08.048.

Markus, Julie, and Alex B. McBratney. 2001. "A Review of The Contamination of Soil with Lead: II. Spatial Distribution and Risk Assessment of Soil Lead." *Environment International* 27 (5): 399-411. doi:10.1016/s0160-4120(01)00049-6.

McClintock, Nathan. 2012. "Assessing Soil Lead Contamination at Multiple Scales in Oakland, California: Implications for Urban Agriculture and Environmental Justice." *Applied Geography* 35 (1-2): 460-473. doi:10.1016/j.apgeog.2012.10.001.

Mielke, Howard W., and Patrick L. Reagan. 1998. "Soil Is an Important Pathway of Human Lead Exposure." *Environmental Health Perspectives* 106 (1): 217-229. doi:10.2307/3433922.

Oliver, M.A., and R. Webster. 2014. "A Tutorial Guide to Geostatistics: Computing and Modelling Variograms and Kriging." *CATENA* 113: 56-69. doi:10.1016/j.catena.2013.09.006.

Parenteau, Marie-Pierre, and Michael C Sawada. 2011. "The Modifiable Areal Unit Problem (MAUP) in the Relationship Between Exposure to NO2 and Respiratory Health." *International Journal of Health Geographics* 10 (1): 58. doi:10.1186/1476-072x-10-58.

Pelfrêne, Aurélie, Sébastien Détriché, and Francis Douay. 2014. "Combining Spatial Distribution with Oral Bioaccessibility of Metals in Smelter-Impacted Soils: Implications for Human Health Risk Assessment." *Environmental Geochemistry and Health* 37 (1): 49-62. doi:10.1007/s10653-014-9629-0.

Root, Elisabeth Dowling. 2012. "Moving Neighborhoods and Health Research Forward: Using Geographic Methods to Examine the Role of Spatial Scale in Neighborhood Effects on Health." *Annals of the Association of American Geographers* 102 (5): 986-995. doi:10.1080/00045608.2012.659621.

Stehouwer, Richard. 2010. "Lead in Residential Soils: Sources, Testing, and Reducing Exposure." *Penn State Extension*. https://extension.psu.edu/lead-in-residential-soils-sources-testing-and-reducing-exposure.

Su, Bin, and B.W. Ang. 2010. "Input–Output Analysis of CO2 Emissions Embodied in Trade: the Effects of Spatial Aggregation." *Ecological Economics* 70 (1): 10-18. doi:10.1016/j.ecolecon.2010.08.016.

Wei, Chaoyang, Cheng Wang, and Linsheng Yang. 2009. "Characterizing Spatial Distribution and Sources of Heavy Metals in the Soils from Mining-Smelting Activities in Shuikoushan, Hunan Province, China." *Journal of Environmental Sciences* 21 (9): 1230-1236. doi:10.1016/s1001-0742(08)62409-2.

Wu, Jun, Rufus Edwards, Xueqin Elaine He, Zhen Liu, and Michael Kleinman. 2010. "Spatial analysis of bioavailable soil lead concentrations in Los Angeles, California." *Environmental Research* 110 (4): 309-317. https://doi.org/10.1016/j.envres.2010.02.004.

Zhang, Chaosheng, Lin Luo, Weilin Xu, and Valerie Ledwith. 2008. "Use of Local Moran's I and GIS to Identify Pollution Hotspots of Pb In Urban Soils of Galway, Ireland." *Science of the Total Environment* 398 (1-3): 212-221. doi:10.1016/j.scitotenv.2008.03.011.

Zhao, Huarong, Beicheng Xia, Chen Fan, Peng Zhao, and Shili Shen. 2012. "Human Health Risk from Soil Heavy Metal Contamination Under Different Land Uses Near Dabaoshan Mine, Southern China." *Science of The Total Environment* 417-418: 45-54. doi:10.1016/j.scitotenv.2011.12.047.

Appendix A: EBK Model Results

Table 10 Comparisons between EBK model results for the Northern Study Area (The red colored model results indicate 1st best choice, while those colored in blue indicate 2nd and 3rd best choices.)

Northern Study Area						
Semivariogram Type	**Subset Size**	**Standardized Mean**	**Standardized Root Mean Squared**	**Root Mean Square**	**Average Standard Error**	**RMS & ASE Approx. Difference**
Objective:		Nearest to 0	Nearest to 1	Lowest value		Lowest difference
Exponential	20	0.01394	0.95593	198.592	184.016	14
Exponential	25	0.01426	0.94336	197.772	185.177	12
Exponential	30	0.00699	0.99130	200.489	179.566	21
Exponential	50	0.00186	1.02557	200.505	176.432	24
Exponential	100	-0.00721	1.09192	201.016	171.059	30
Whittle	20	0.01014	0.96454	197.523	181.316	16
Whittle	25	0.01192	0.94652	196.099	183.357	13
Whittle	**30**	**0.00579**	**0.99109**	**200.556**	**181.266**	**19**
K-Bessel	**20**	**0.00825**	**0.96884**	**196.519**	**182.353**	**14**
K-Bessel	25	0.01094	0.94974	197.573	186.959	11
K-Bessel	**30**	**0.00390**	**0.99888**	**201.127**	**184.207**	**17**

Table 11 Comparisons between EBK model results for the 1st Southern Study Area (The red colored model results indicate 1st best choice, while those colored in blue indicate 2nd and 3rd best choices.)

1st Southern Study Area						
Semivariogram Type	**Subset Size**	**Standardized Mean**	**Standardized Root Mean Squared**	**Root Mean Square**	**Average Standard Error**	**RMS & ASE Approx. Difference**
Objective:		Nearest to 0	Nearest to 1	Lowest value		Lowest difference
Exponential	20	0.00332	0.97220	133.279	122.782	10
Exponential	25	0.00432	0.96033	133.495	122.005	11
Exponential	30	-0.00003	0.99747	134.175	119.453	15
Whittle	**20**	**0.00144**	**0.97616**	**132.465**	**122.493**	**10**
Whittle	25	0.00372	0.95872	132.138	121.983	10
Whittle	30	-0.00123	1.00310	133.258	118.816	14
K-Bessel	**20**	**0.00083**	**0.97739**	**131.627**	**123.761**	**8**
K-Bessel	25	0.00256	0.96184	130.917	121.109	10
K-Bessel	**30**	**-0.00199**	**1.00246**	**132.653**	**119.816**	**13**

Table 12 Comparisons between EBK model results for the 2nd Southern Study Area (The red colored model results indicate 1st best choice, while those colored in blue indicate 2nd and 3rd best choices.)

2nd Southern Study Area						
Semivariogram Type	**Subset Size**	**Standardized Mean**	**Standardized Root Mean Squared**	**Root Mean Square**	**Average Standard Error**	**RMS & ASE Approx. Difference**
Objective:		Nearest to 0	Nearest to 1	Lowest value		Lowest difference
Exponential	20	0.01995	0.85843	83.840	83.432	1
Exponential	25	0.01732	0.89699	83.069	82.977	1
Exponential	30	0.00391	0.91733	84.799	83.121	2
Whittle	20	0.01788	0.86920	82.646	80.403	2
Whittle	**25**	**0.01084**	**0.91876**	**82.068**	**81.163**	**1**
Whittle	**30**	**0.00465**	**0.90616**	**83.822**	**84.533**	**1**
K-Bessel	20	0.02061	0.86413	82.807	79.532	3
K-Bessel	25	0.01879	0.90930	81.724	81.599	1
K-Bessel	**30**	**0.00519**	**0.92088**	**84.120**	**83.981**	**1**