OCTOBER 3, 2007
EPA 260-R-08-003

FINAL REPORT FOR THE PILOT STUDY OF TARGETING ELEVATED BLOOD-LEAD LEVELS IN CHILDREN

Prepared by

BATTELLE

Prepared for:

Margaret Conomos, Work Assignment Manager
Barry Nussbaum, Technical Adviser
Analytical Products Branch
Environmental Analysis Division

Sineta Wooten, Project Officer
Program Assessment and Outreach Branch
National Program Chemical Division
Office of Pollution Prevention and Toxics

U.S. Environmental Protection Agency
1200 Pennsylvania Avenue NW (7404T)
Washington D.C. 20460

BATTELLE DISCLAIMER

This report is a work prepared for the United States government by Battelle. In no event shall either the United States government or Battelle have any responsibility or liability for any consequences of any use, misuse, inability to use, or reliance upon the information contained herein, nor does either warrant or otherwise represent in any way the accuracy, adequacy, efficacy, or applicability of the contents hereof.

ACKNOWELDGEMENTS

The EPA and the authors thank the organizations whose contributions made this report possible including the Lead Poisoning Prevention Branch at the Centers for Disease Control and Prevention, the Childhood Lead Poisoning Prevention Program at the Massachusetts Department of Public Health, and the Office of Healthy Homes and Lead Hazard Control at the U.S. Department of Housing and Urban Development.

This report was based on work conducted by Battelle, with significant contributions from Warren Strauss, Tim Pivetz, Elizabeth Slone, Jyothi Nagaraja, Nicole Iroz-Elardo, Rona Boehm, Michael Schlatt, Darlene Wells, Jennifer Zewatsky, Michele Morara, and Bruce Buxton.

TABLE OF CONTENTS

Page

Page

LIST OF TABLES

LIST OF FIGURES

EXECUTIVE SUMMARY

This pilot study seeks to develop statistical models to predict risk of childhood lead poisoning within specified geographic areas based on a combination of demographic, environmental, and programmatic information sources. Exposure factors associated with childhood lead poisoning were investigated within census tracts for a community-focused set of models in Massachusetts, as well as within counties across the United States in a series of national models. Aggregated summary measures of the proportion of children screened at or above 5, 10, 15 and 25 μg/dL within defined geographic areas (census tracts and counties) were used as the response variable in the statistical models. These summary measures were constructed at 3-month (quarterly) intervals from 1995 through 2005, in counties across the nation using data from CDC's National Surveillance Database, and from 2000-2006 in census tracts within the Commonwealth of Massachusetts based on data provided by the Massachusetts Department of Public Health.

The results of this study suggest that longitudinal predictive models can be developed at the county level across the nation, based on the use of quarterly summary information from CDC's National Surveillance Database, and at the census-tract level within states that have a long history of universal screening and reporting, such as Massachusetts. These models can be used to describe how risk of childhood lead poisoning changes over time within different regions of the country, as well as within small geographic areas within states (e.g., counties) and even smaller geographic areas within counties (e.g., census tracts). They can be used to predict the risk of childhood lead poisoning in counties (or census tracts) with little or no surveillance data, and also can be used to identify those counties (or census tracts) that are at highest risk at the end of the period of observation.

The statistical model chosen (a random effects model with separate intercepts and slopes estimated within each county or census tract) also allows ranking of geographic areas based on the rate of decline over time after accounting for the fixed-effects variables of the model (although only among those areas that provided adequate surveillance data). These random-effects models were fit to the exceedance proportions within the context of a logistic regression model. Within the context of the Broad-Based National Model, these random effects allow EPA to identify those counties that are experiencing a more rapid reduction in risk of childhood lead poisoning over time (to identify best practices) and those counties that are experiencing a significantly less rapid decline over time (to identify areas in need of additional attention and resources for combating lead poisoning), after already accounting for the demographic, programmatic, and environmental factors included in the multivariate model.

Within the series of national models at the county level of geographic specificity, the data suggest that there are significant differences in the distribution of childhood blood-lead concentrations among the different regions of the country, and that the manner in which these distributions change over time and are impacted by seasonality also is regionally specific. The risk of childhood lead poisoning had a statistically significant downward trend over time in all areas of the country.

After accounting for these regional differences, a number of demographic, environmental, and programmatic variables were found to be highly predictive of childhood blood-lead

concentrations among the different response variables modeled within this project. The specific variables that were found to be predictive within the multivariate models varied based on the response variable; however, there were certainly some variables that were found to be predictive in multiple models. In addition to various census demographic variables that were identified in previous risk modeling efforts (e.g., age of housing, percent single parent families, race/ethnicity), air modeling data, variables constructed from EPA's Safe Drinking Water Information System, and programmatic funding information from HUD and CDC were found to be highly predictive in the multivariate models.

Within the context of the high-resolution model developed using data from the Commonwealth of Massachusetts, a highly significant downward trend in the risk of childhood lead poisoning also was identified among the five models developed. Due to a very small number of children observed at or above 25 μg/dL within Massachusetts over the 2000-2006 period of observation, we were unable to fit this sixth model. After accounting for the long-term reduction over time and seasonality using similar methods that were employed in the Broad-Based National Model, only the demographic and programmatic variables were included in the multivariate models for risk of childhood lead poisoning at the census-tract level. Of particular interest were the variables that described the proportion of housing units within each census tract that were found to be in compliance and out of compliance with the Massachusetts Standard of Care. In all five of the multivariate models, the risk of childhood lead poisoning was significantly reduced as the proportion of housing units in compliance increased within a census tract. In addition, for the last two models (which predicted proportion of children at or above 10 and 15 μg/dL), the risk of childhood lead poisoning increased significantly as the proportion of housing units out of compliance increased within a census tract.

The observed and predicted values from the multivariate models (including predicted values where there were no observed surveillance data) were used to generate static maps using Arc-View software, and were loaded into a customized dynamic visualization tool that allows users to interact with the modeling results to assess how risk of childhood lead poisoning changes over time within specific regions of the country. This tool will help EPA and others identify areas that remain at risk for childhood lead poisoning as we approach the 2010 goal of elimination of this preventable adverse health outcome.

1.0 INTRODUCTION

1.1 Background and Purpose of Study

Over the past 15 years, various childhood lead poisoning prevention programs (CLPPPs) throughout the United States have conducted analyses of their screening data to develop "risk indices," or mathematical models for predicting the prevalence of childhood lead poisoning in different geographic areas within their regions of concern. These modeling efforts generally are intended to characterize the extent of the prevalence of childhood lead poisoning within their geographic areas and to support the development of targeted screening and outreach plans in order to reach the 2010 goal of eliminating childhood lead poisoning throughout the United States.

To date, the majority of modeling efforts have focused on combining blood-lead testing information and demographic data available from the U.S. Census. Previous studies have combined childhood surveillance data (aggregated at the zip-code or census-tract level) with demographic predictor variables from the Census Bureau for the purposes of targeting geographic areas at higher risk of childhood lead poisoning (Miranda, Dolinoy, and Overstreet 2002; Miranda et al. 2005; Strauss et al. 2001a). These studies have led to recommendations for using age of housing and percent of population below the poverty line for targeting neighborhoods that may be of increased risk for childhood lead poisoning (CDC 1997). Numerous studies also have been used to document the relationship between children's blood-lead concentrations and measures of lead in residential environmental media (dust, soil, air, water, and food) (HUD 1995; Lanphear et al. 1998; Strauss et al. 2001b). These studies have contributed to EPA and HUD regulations and policies for identifying and reducing residential childhood lead exposures (24 CFR Part 35; 40 CFR Part 745; 40 CFR Part 745; U.S. Department of Housing and Urban Development September 15, 1999). Other studies have combined blood-lead surveillance data with programmatic information on housing units treated to determine the positive impact of housing-based intervention programs (Strauss et al. 2006).

The goal of this study is to explore models based on a hierarchical combination of demographic, environmental, and programmatic information sources in order to predict the number of children at risk of elevated blood-lead levels for a given geographic area. While the models are highly dependent on available data, this study provides a statistical methodology that combines each data source in an appropriate manner, adjusting for global and local trends over time. In doing so, the models build upon concepts of hierarchical modeling and longitudinal data analysis.

As EPA, CDC, and other federal and state agencies prepare to meet the 2010 goal of eliminating childhood lead poisoning, this pilot study of integrating several different types of data sources hopefully improves the predictive power of models that rely on a single information source. This allows for more efficient targeting of those geographic areas that need the most help in eliminating childhood lead poisoning.

1.2 Study Objectives

1.2.1 Objective 1 – Combine and Manage Multiple Data Sources

The first objective of the study was to combine multiple sources of information in order to assess the impacts of various factors on children's blood-lead levels. The study had to obtain and manage data relating to blood-lead levels, environmental exposure, demographic characteristics, and programmatic support to state and local childhood lead-poisoning prevention efforts. Missing, incomplete, or error-prone data were identified for each data source and steps were taken to resolve data problems. Databases were developed to store and later combine each data source in a manner that supported the development of predictive models. Master databases that integrated multiple data sources were developed to enable efficient access to data required for statistical analyses. A data dictionary was prepared to document the various study databases.

1.2.2 Objective 2 – Conduct Analyses to Identify Predictive Variables and Model Children's Blood-Lead Levels

The second study objective was to conduct statistical analyses in order to develop models that are predictive of risk of childhood lead poisoning within defined geographic areas as a function of various different environmental, programmatic, and demographic factors. As part of this objective, a National model was developed for predicting risk at the county level based on surveillance data from the U.S. Centers for Disease Control and Prevention (CDC), and a local model was developed at the census-tract level using blood-lead surveillance data from within the Commonwealth of Massachusetts. As part of the model building process at both the national and local levels, the various data sources underwent exploratory analyses to investigate data distributions, identify relationships between variables, and determine appropriate variables to include in subsequent statistical models. Part of the exploratory analyses included an effort to identify which environmental, programmatic, and demographic factors were most predictive of risk of childhood lead poisoning. Multivariate statistical models then were developed using appropriate statistical software to combine the various data sources within a single model that accounted for trends in risk of childhood lead poisoning over time within defined geographic areas. Model diagnostics were reviewed, and models with the best fit were identified.

1.2.3 Objective 3 – Develop Visualization Tool to Graphically Model Predicted Blood-Lead Levels

The third study objective was to develop an appropriate visualization tool that allows users to interact with the results of the statistical model predicting children's blood-lead levels across the United States. This tool provides the user with the flexibility to visually compare the predicted blood-lead levels across areas of the country and also to drill down into individual counties or census tracts to assess the input data that generated the predicted value.

2.0 STUDY METHODOLOGY

2.1 General Approach

This pilot study sought to develop models to predict the number of children at risk of elevated blood-lead levels for a given geographic area based on a hierarchical combination of demographic, environmental, and programmatic information sources. Doing so required looking at both the mechanisms of childhood lead risk assessment and control activities at the local level as well as at broad trends across the United States. The two main analysis goals correspond to developing predictive models at two different levels of geographic specificity, and appear as follows:

1. Broad Coverage (Low-Resolution) Model: This type of model is intended to be able to characterize broad trends over time in the prevalence of childhood lead poisoning at the county level across the entire United States. This model was based on quarterly county-level aggregated surveillance data from the CDC and augmented with environmental data from a variety of sources, demographic data from the U.S. Census, and programmatic (level of federal funding) information.

2. High-Resolution Model: This type of model represents the effort to assess the relative contribution of various exposure sources associated with elevated blood-lead concentrations within select communities. This type of model certainly reflects the idea that exposures that contribute to childhood lead poisoning are likely to be community specific. This analysis goal was met through modeling census-tract level surveillance data within Massachusetts as well as housing unit lead assessment and/or control activities. These data sources were augmented with all of the environmental, demographic, and programmatic information used in the national model with the addition of state programmatic funding levels.

The primary objective of this pilot study was to utilize combined information from different sources at various levels of geographic and temporal specificity to more accurately target geographic areas at high risk for not meeting the 2010 goal of eliminating childhood lead-poisoning. As such, the study required careful integration of a variety of data sources with various characteristics and documentation. Data to support this study were gathered from multiple sources, including federal, state and local lead poisoning prevention programs, as well as publicly available data that were downloaded from the internet (e.g., census data, EPA's Toxics Release Inventory).

2.2 Data Management

When each data source was received, the data and supporting documentation were reviewed to gain knowledge on the structure, relationship, and quality of the data. Database managers worked with the project team to determine the final format for each database, the desired uses of the databases; as well as the requirements for maintaining the databases. Based on this information, master databases were constructed in SQL server for both the national low-resolution model and for the high-resolution model based on Massachusetts data that integrated the various environmental, demographic, and programmatic variables, and facilitated statistical

analyses of the combined data. These datasets were translated directly to SAS datasets for statistical analysis, and also were transferred to Microsoft Access for delivery to EPA. The Microsoft Access database includes a compact version of each database utilized in the statistical analysis, with any extraneous variables removed. In addition, the Microsoft Access database includes a copy of the integrated longitudinal dataset used to support the final multivariate models developed within this project.

Throughout the development process, checks for completeness were conducted on all study databases, and the project team worked with data-sharing collaborators and EPA to attempt to complete missing data as necessary to complete the proposed statistical analyses. Any changes to the databases (corrections, additions, deletions, etc.) were documented in appropriate meta-data files, and reported to EPA within the data dictionary attached to this report as Appendix H. As part of constructing and maintaining these databases, the project team will develop appropriate documentation of the combined master databases.

Standard Operating Procedures (SOPs) were followed to ensure the proper storage, backup, and retrieval of datasets created and analyzed for this study. Additional details of these SOPs can be found in the Quality Management Plan prepared for this project (Battelle 2007).

2.3 Descriptive Data Analyses

The analysis began with an assessment of the study sample, i.e., the proportion of counties and census tracts in the sample with complete data for both the response variable and the explanatory variables. Prior to the fitting of any descriptive statistics to assess the predictive ability of any of the explanatory variables, the blood-lead response variables needed to be constructed based on the CDC and Massachusetts blood-lead surveillance data. These data sources contain information on individual blood-lead testing results on children, and were aggregated into quarterly summary statistics (number of children observed, arithmetic and/or geometric mean[1], and number of children observed at or above 5, 10, 15, and 25 μg/dL) at the county level (for the CDC data) and the census-tract level (for the Massachusetts data). An executable was developed to extract these quarterly summary statistics from each county from CDC's SQL server database for children aged 6-36 months, and a similar executable was deployed to create parallel summary statistics at the census-tract level for the Massachusetts surveillance data. Because of confidentiality restrictions, county/quarter (or census tract/quarter) combinations with fewer than 5 observations were automatically eliminated from the dataset. Data reported prior to 1995 also were eliminated from the analysis database prior to statistical analysis.

Once the aggregated summary datasets were constructed, they were reviewed for possible problems associated with childhood lead poisoning prevention programs not following universal reporting protocols (for some localities, data were only transmitted to the CDC National Surveillance Database for children with elevated blood-lead concentrations over certain periods of time). A screening algorithm was developed to remove these suspect data from the analysis dataset – resulting in the elimination of less than 3 percent of the aggregate summary records from the National database. The screening algorithm also was applied to the Massachusetts data – however no records were eliminated, as Massachusetts was following universal screening and

[1] The CDC reported only the arithmetic mean, while Massachusetts reported both arithmetic and geometric means.

reporting guidelines over the 2000-2006 time period for which they provided data. Additional detail on the manner in which the blood-lead response variables were constructed can be found in Section 3.1.

In preparation for developing longitudinal statistical models, univariate summaries of each variable as a function of time were generated and comparisons were made of these distributions using side-by-side box-plots for continuous data or bar-charts for categorical data. This helped verify that the data were clean and ready for analysis and identified cells with sparse data. Such descriptive analyses were conducted on each database, to characterize the distributions of all observed variables using frequency distributions for categorical variables, and simple summary statistics (mean, median, mode, minimum, maximum, and select percentiles) for continuous variables. Distributional assumptions also were explored for certain variables, as appropriate, in preparation for more sophisticated models. For example, some environmental concentration data may depart from normality, and follow a log-normal distribution. In these cases, we additionally reported the geometric mean and geometric standard deviation as part of the simple descriptive summary.

The univariate descriptions then were followed by fitting a series of cross-sectional bivariate relationships between the blood-lead response variable(s) and each candidate explanatory variable. These cross-sectional relationships were explored as a function of time to better understand the stability of these relationships, and whether they change over time, so that they can be modeled appropriately in the more sophisticated longitudinal analyses. These analyses also will help identify which explanatory variables are most predictive of the blood-lead response variable.

In preparation for more sophisticated statistical analyses, such as the Generalized Linear Mixed Logistical Regression Model outlined below, relevant stratified analyses were performed to investigate interactions discussed in the data analysis plan. For example, the population density variable was investigated in this manner, as density may serve as a surrogate to differentiate between rural and urban geographic areas in the analyses – and exposure variables may be different in these types of areas. Similarly, EPA regions were investigated as a potential stratification variable. If variation in the measure of effect is not observed (e.g., odds ratios) across the levels of a third variable; however, the third variable can likely be treated as a potential confounder in the multivariate model, rather than as an effect modifier. If the odds ratios differ markedly—e.g., the effect appears to be protective in one subgroup and hazardous in another subgroup—the third variable must be considered as an effect modifier.

Specific variables within each type were explored using four general approaches – (1) histograms or side-by-side box-plots of the candidate explanatory variable, (2) simple regression line plots exploring the relationship between predicted risk of lead poisoning and the explanatory variable for each of the four specified time periods, (3) distributional summaries of the explanatory variable across the three time periods, and (4) statistical modeling of the relationship between the explanatory variable and various blood-lead response variables after adjusting for the effects of time and seasonality within different regions of the country for the National (Low-Resolution) model and for the effects of time in the Massachusetts (High-Resolution) Model.

Histograms or Side-by-Side Box-Plots of Potential Explanatory Variables

Using one record for each quarterly county- or census-tract -level data point, a histogram illustrating the distribution of the explanatory variable is presented. A fitted line assists with assessing the distribution of each potential explanatory variable (e.g., whether the data are approximately normally distributed). Histograms were plotted for potential predictor variables that were time invariant. For predictor variables that varied over time within the analysis dataset, side-by-side box-plots were used to characterize the distribution over the time periods, using an average of the predictor variable across the quarters in which blood-lead concentrations were observed within each time-period and area.

Logit Probability Plots for each Explanatory Variable

The county-level quarterly proportion of screened children exceeding 10 μg/dL reported by the CDC were modeled as a function of each candidate explanatory variable, with separate logit curves used to represent each of the time periods. This analysis allows comparison of the relationship between the explanatory variable and predicted blood-lead level trends across time periods. If the relationship is stable across time, roughly parallel curves are evident. If the effect of the variable on blood-lead varies over time, non-parallel (and perhaps intersecting) curves are observed. In this case, the longitudinal analyses may need to be adjusted to allow for the effect of the covariate to change over time.

Plots of Predicted GM Blood-Lead Levels and Explanatory Variables

The census-tract-level quarterly blood-lead data available from Massachusetts were fit to each explanatory variable to generate predicted GM blood-lead levels across the range of the explanatory variable for each of the time periods. A simple linear regression line plot summarizes this analysis with one line for each time period. This analysis allows comparison of the relationship between the explanatory variable and predicted blood-lead level trends across time periods. If the relationship is stable across time, roughly parallel lines are evident. If the effect of the variable on blood lead varies over time, non-parallel (and perhaps intersecting) lines are observed. In this case, separate slopes may need to be fit for these variables over different periods of time in the more sophisticated longitudinal analyses.

Distributional Summaries

The first table presented for each explanatory variable contains a series of summary statistics for each of the time periods including sample size, number missing, mean, and standard error. The sample size is relative to the number of quarters represented in the analysis dataset; therefore, these distributions correspond to the analysis dataset (and not necessarily to the distribution of the variable across the nation or state). The distribution of the data for each time period also is presented (minimum, median, and maximum and 10^{th}, 25^{th}, 75^{th}, and 90^{th} percentiles). Comparing the summary data across time allows assessment of changes in the explanatory variable over time for the groups of tracts included in the analysis for each time period. Generally, the mix of counties and Massachusetts census tracts included in each of the time periods is similar, so that the distribution of the data from each period also is similar.

Statistical Modeling of Relationship between Explanatory Variables and Exceedance of Blood-Lead Thresholds for the National (Low-Resolution) Model:

For each candidate predictor variable being considered for the National (Low-Resolution) Model, the following generalized linear mixed models approach was used to model the proportion of children exceeding certain thresholds as a function of the predictor variable after adjusting for Region-specific intercepts, slopes over time and effects of seasonality:

$$\log it(E[Y_{ij} / n_{ij}]) = \mathrm{Re}\,gion_{ik} \cdot (\beta_{0k} + \beta_{1k} \cdot t_{ij} + \vec{\beta}_{2k} \cdot Season_{ij}) + \beta_{3\cdot} X_{ij} + \delta_{0i} + \delta_{1i} \cdot t_{ij}$$

Where (i indexes county, j indexes time, and k indexes the region of the country), Y_{ij} represents the number of children observed above the blood-lead threshold in the i^{th} county at time j, n_{ij} represents the number of children tested in the i^{th} county at time j, t_{ij} and $Season_{ij}$ are fixed effects variables corresponding to a time-trend (in years) and seasonality, X_{ij} is the candidate predictor variable being investigated, the beta parameters (β) represent a vector of fixed effects, and the delta parameters (δ) represent random effects that allow each county to have its own trend over time. The X_{ij} predictor variable is mean centered in this series of models, allowing the intercept term to be relatively stable across the multiple predictor variables being investigated. In this model, it can be assumed that δ_{0i} and δ_{1i} jointly follow a multivariate normal distribution with mean zero and covariance matrix $\Sigma = \begin{bmatrix} \sigma_{00}^2 & \sigma_{01}^2 \\ \sigma_{10}^2 & \sigma_{11}^2 \end{bmatrix}$.

- Model 1 follows the above approach – where Y_{ij} represents the number of children observed with blood-lead concentrations at or above 5 μg/dL, and n_{ij} represents the total number of children screened within each record (county/quarter).
- Model 2 follows the above approach – where Y_{ij} represents the number of children observed with blood-lead concentrations at or above 10 μg/dL, and n_{ij} represents the total number of children screened within each record (county/quarter).
- Model 3 follows the above approach – where Y_{ij} represents the number of children observed with blood-lead concentrations at or above 15 μg/dL, and n_{ij} represents the total number of children screened within each record (county/quarter).
- Model 4 follows the above approach – where Y_{ij} represents the number of children observed with blood-lead concentrations at or above 25 μg/dL, and n_{ij} represents the total number of children screened within each record (county/quarter).

In addition to the above models, the project team explored whether the effect of each candidate predictor variable on the exceedence proportions varied over time. This was done by exploring the interaction between each candidate predictor variable and (1) a linear effect of time, (2) a quadratic effect of time, and (3) a 4-level categorical effect of time.

Statistical Modeling of Relationship between Explanatory Variables and Exceedance of Blood-Lead Thresholds for the Regional (High-Resolution) Model:

The Regional (High-Resolution) Models developed for the Massachusetts data at the census-tract level of geographic specificity included models for both continuous data (geometric mean) and binomial data (exceedence proportions). Therefore, each explanatory variable being considered for these models were explored using models for both continuous and binomial data as described below:

Continuous Data: The following mixed models analysis of variance (i.e., a random-effects model for continuous data) was used to model the geometric mean (GM) blood-lead concentration as a function of a candidate predictor variable:

$$GM_{ij} = \beta_0 + \beta_1 \cdot t_{ij} + \beta_2 X_{ij} + \delta_{0i} + \delta_{1i} \cdot t_{ij} + \varepsilon_{ij}$$

Where (i indexes census tract, j indexes time), GM_{ij} represents the geometric mean blood-lead concentration in the i^{th} census tract at time j, t_{ij} is a fixed-effects variable corresponding to a time-trend (in years), X_{ij} is the candidate predictor variable being investigated, the beta parameters (β) represent a vector of fixed effects, and the delta parameters (δ) represent random-effects that allow each county or census tract to have their own trend over time. The X_{ij} variable typically is mean centered in this series of models, allowing the intercept term to be relatively stable across the multiple predictor variables being investigated. In this model, it can be assumed that δ_{0i} and δ_{1i} jointly follow a multivariate normal distribution with mean zero and covariance matrix $\Sigma = \begin{bmatrix} \sigma_{00}^2 & \sigma_{01}^2 \\ \sigma_{10}^2 & \sigma_{11}^2 \end{bmatrix}$, and the residual error also is assumed to follow a normal distribution with mean zero and variance $\sigma^2_{\varepsilon_{ij}}$.

- Model 1 follows the above approach – where the responses are weighted equally.
- Model 2 follows the above approach – where the responses (GM) are weighted by the number of children observed (screened) within each record (census tract/quarter).

Binomial Data: The following generalized linear mixed model (i.e., a random-effects model for binomial data) was used to model the proportion of children exceeding certain thresholds as a function of a candidate predictor variable:

$$\log it(E[Y_{ij} / n_{ij}]) = \beta_0 + \beta_1 \cdot t_{ij} + \bar{\beta}_2 \cdot Season_{ij} + \beta_3 X_{ij} + \delta_{0i} + \delta_{1i} \cdot t_{ij}$$

Where (i indexes census tract and j indexes time), Y_{ij} represents the number of children observed above the blood-lead threshold in the i^{th} census tract at time j, n_{ij} represents the number of children tested in the i^{th} census tract at time j, t_{ij} is a fixed effects variable corresponding to a time-trend (in years), X_{ij} is the candidate predictor variable being investigated, the beta parameters (β) represent a vector of fixed effects, and the delta parameters (δ) represent random effects that allow each census tract to have its own trend over time. The X_{ij} variable also is mean centered in this series of

models, allowing the intercept term to be relatively stable across the multiple predictor variables being investigated. In this model, it can be assumed that δ_{0i} and δ_{1i} jointly follow a multivariate normal distribution with mean zero and covariance matrix $\Sigma = \begin{bmatrix} \sigma_{00}^2 & \sigma_{01}^2 \\ \sigma_{10}^2 & \sigma_{11}^2 \end{bmatrix}$.

- Model 3 follows the above approach – where Y_{ij} represents the number of children observed with blood-lead concentrations at or above 5 µg/dL, and n_{ij} represents the total number of children screened within each record (census tract/quarter).
- Model 4 follows the above approach – where Y_{ij} represents the number of children observed with blood-lead concentrations at or above 10 µg/dL, and n_{ij} represents the total number of children screened within each record (census tract /quarter).
- Model 5 follows the above approach – where Y_{ij} represents the number of children observed with blood-lead concentrations at or above 15 µg/dL, and n_{ij} represents the total number of children screened within each record (census tract /quarter).
- Model 6 follows the above approach – where Y_{ij} represents the number of children observed with blood-lead concentrations at or above 25 µg/dL, and n_{ij} represents the total number of children screened within each record (census tract /quarter).

To allow comparison of the different variables explored within each variable type, tables are included in Section 4 that present the log-likelihood statistic from each model run and presented in Appendices A and B. Within each variable category, the variable that provided the best fit for each of the six models is highlighted in yellow. To ensure compatibility in the likelihood-based statistics being used to make comparisons among the different candidate predictor variables, missing values for predictor variables were imputed using the mean of the distribution. The number of imputed values that were necessary is provided by the nmiss column in the table of distributional summaries described above. The project team choose whether to adjust for changes in the slope for a candidate predictor variable over time based on a comparison of the likelihood statistics after adjusting for the number of degrees of freedom used in the model for the effects of the explanatory variable (over time) on the response. Those variables highlighted in yellow have the largest likelihood statistic after adjusting for differences in the degrees of freedom, and were considered as strong candidate predictors for the multivariate statistical models.

Due to the iterative nature and complexity of the Mixed Models Analysis of Variance and Generalized Linear Mixed Modeling Approaches, these models did not always converge. Models that failed to converge for a particular predictor variable are discussed in the results sections within Appendices D and E, and also are indicated in Tables 4-1 and 4-2, as well as in the summary pages of Appendices A and B by blank cells. Cases in which model convergence is not attained likely will translate to exclusion of that particular variable when building the multivariate model. Note that because of the sparseness of data for children with blood-lead levels at or above 25 µg/dL within the Massachusetts data, Model 6 failed to converge across all variables. Thus, Model 6 results are not presented or discussed for the Massachusetts models.

2.4 Development of Multivariate Statistical Models

2.4.1 Statistical Models for the Broad Coverage – Low-Resolution Model

This model is being used to characterize broad trends over time in the prevalence of childhood lead poisoning across the entire United States. The various surveillance, environmental sources, demographic characteristics, and programmatic support data sources were aggregated to the county level for all localities with universal screening and reporting. Quarterly estimates of each candidate predictor variable were created for each county within the United States, including those county/quarter combinations that did not include observed blood-lead response variable information (allowing for the extrapolation of the model predictions to geographic areas and time-points that were not represented within CDC's National Surveillance Database).

In addition to investigating the predictive ability of each potential environmental, programmatic and demographic variable as described earlier, various different stratification variables (region of the country, population density) and covariates (time trend and seasonality) were investigated. As a result, all four multivariate statistical models adjust for a categorical variable that differentiates among the risk of childhood lead poisoning within the 10 EPA regions. Within each EPA region, a separate intercept, trend over time, and seasonality term (based on fitting intercepts for each quarter of time) was included in the multivariate statistical model. For the purposes of discussion, it was assumed that the modeling approach will focus on a logistic regression model for the proportion of children that have elevated blood-lead concentrations ($\geq$10 μg/dL). The temporal nature of declining childhood lead poisoning will be addressed via classic concepts of longitudinal data modeling of the low resolution data. Let

Y_{ij} represent the number of children that were detected with blood-lead concentration above 10 μg/dL from the i^{th} county and j^{th} point in time (quarter),

n_{ij} represent the number of children that had their blood-lead concentration tested from within the i^{th} county and j^{th} point in time (quarter),

Please note that we expect that $n_{ij}<N_{ij}$, where N_{ij} represents the total population of children in the i^{th} county and j^{th} point in time.

t_{ij} represent time (in years) corresponding to the Y_{ij} response variable,

$Region_{ik}$ represents the region of the country that the ith county is located within (where k==1,…,10 and is representative of the 10 established EPA regions),

$Season_{ij}$ represents a series of 4 indicator variables that differentiate between the 4 different quarters captured by the j-index, and

X_{ij} represent a series of predictor variables associated with the Y_{ij} response variable. These predictor variables may represent air monitoring data, drinking water data, census demographic data, programmatic data on federal financial support for lead poisoning prevention, and other related information as detailed above that can potentially help predict the prevalence of lead poisoning at the county level.

We introduce the following as a potential baseline model:

$$\log it(E[Y_{ij} / n_{ij}]) = \mathrm{Re}\, gion_{ik} \cdot (\beta_{0k} + \beta_{1k} \cdot t_{ij} + \vec{\beta}_{2k} \cdot Season_{ij}) + \vec{\beta}_{3.} X_{ij} + \delta_{0i} + \delta_{1i} \cdot t_{ij}$$

Where the beta parameters (β) represent a vector of fixed effects, and the delta parameters (δ) represent random effects that allow each county to have their own trend over time. In this model, it can be assumed that δ_{0i} and δ_{1i} jointly follow a multivariate normal distribution with mean zero and covariance matrix $\Sigma = \begin{bmatrix} \sigma_{00}^2 & \sigma_{01}^2 \\ \sigma_{10}^2 & \sigma_{11}^2 \end{bmatrix}$.

Counties with larger δ_1 parameters estimates represent areas where lead-poisoning has not significantly decreased over time. Similarly, the parameter estimates can be used to identify those counties with the highest predicted prevalence of childhood lead poisoning at various time points in the future.

In building the multivariate statistical model for the Broad-Based Modeling Objective, the project team first evaluated the predictive ability of each candidate predictor variable that was considered within the exploratory analyses. For the environmental predictor variables, in particular (information from EPA's 1999 National Air Toxics Assessment, Safe Drinking Water Information System, and Toxics Release Inventory), the data were largely concentrated at zero. Therefore, a series of zero/one indicator variables that represent county/quarter combinations at or above the 95th and 99th percentile of observed values of these environmental predictor variables within the analysis dataset also were investigated.

Once the predictive ability of each candidate variable was established within the exploratory analyses described earlier, candidate predictor variables were classified into groups (e.g., housing age, income, education, air modeling, programmatic financial support) and then the single best predictor variable within each group was selected for possible inclusion within each of the six multivariate statistical models being developed. If the selected variable demonstrated a relationship with risk of lead poisoning that changes over time (as evidenced by intersecting lines in the plots generated in the exploratory analyses), then this interaction was taken into consideration within the evaluation of the predictive ability of the candidate variable(s).

The approach to determining which environmental, programmatic, and demographic variables were included in the model followed a backward elimination process – in which each variable group's best predictor variable identified earlier was included in the first model – with variables being eliminated from the model when they were not deemed to be highly significant. This model building process also was aided by investigation of the selected environmental, programmatic, and demographic variables for issues of potential colinearity via investigation of correlation matrices and principal components analysis. The resulting multivariate statistical models were parsimonious – and in most cases only included variables that were highly statistically significant. In a few cases, a variable was left in the model without being highly significant – because its elimination caused a large drop in the model log-likelihood (suggesting

that the model is significantly improved with the addition of a variable whose slope is not significantly different from zero).

After the multivariate statistical models were developed, model fit diagnostics were evaluated and documented.

The parameter estimates for the four National Multivariate Statistical Models are provided in the results section. The results of these models also were explored in multiple ways. Maps were generated to demonstrate observed and predicted proportion of children at or above 10 μg/dL within each EPA region for the Years 2000 and 2005 (data were appropriately averaged across the four quarters in each of these years prior to mapping). Lists also were generated to identify the 150 highest risk counties across the United States at the end of the observation period (2006) as predicted by each of the six models, as well as the 10 highest risk counties within each state.

Finally, the predicted values from these multivariate statistical models (extrapolated to county/quarter combinations not represented in the CDC surveillance database) were integrated in a unique data visualization tool. The product of this effort is a time-series of maps (or a movie) that spatially interpolates risk of childhood lead poisoning as a function of various appropriate predictor variables. The visualization tool allows users to interact with the modeling results at different levels of temporal and geographic specificity. The tool allows the user to select an appropriate response variable (e.g., proportion of children with blood-lead concentrations above 5 μg/dL) and play a movie that displays a time-series of maps that displays how the predicted (or observed) risk changes over time across the various counties within a selected state. The user can zoom in on a rectangular area, to see these results with a higher degree of geographic specificity. The user also can stop the movie (or rewind, or fast-forward) to isolate specific points in time. By using the mouse, the user also can select a specific county and the tool will display the observed and predicted data for that particular county in a separate window. The visualization tool was written in C++, and was built in a manner that will allow EPA to modify the model and for the project team to quickly import the resulting data from a modification into the tool.

2.4.2 Statistical Models for the High-Resolution Model within Massachusetts

High-resolution models will be utilized to identify the relative contribution of various types of exposure sources in elevated risk for childhood lead poisoning within select communities within the Commonwealth of Massachusetts. These types of sources include housing factors, broader environmental exposure, demographic composition, and programmatic resources. While this type of model reflects the idea that exposures contributing to childhood lead poisoning likely are community-specific, analysis of the high-resolution models may have certain limitations including selection bias and generalizability to other geographic areas.

The Massachusetts Department of Public Health (MDPH) entered into a limited use data sharing agreement with the project team, allowing them to provide blood-lead testing results on individual children (aged 6-36 months) and housing inspection data in a format that preserves linkages through a housing unit identification variable. These data will be utilized in two different modeling approaches. The first modeling approach will seek to develop census tract quarterly summary measures similar to the National Model for blood-lead (e.g., exceedance

proportions and geometric means), as well as summary measures for the proportion of housing units in each census tract that are known to be in (or out of) compliance with the Massachusetts standard of care (for use as a potential explanatory variable). MDPH also has provided the project team with summary information regarding HUD and state funding of residential housing interventions (lead hazard control and abatement) – which will be used to develop a longitudinal summary of current and cumulative per-capita spending on residential intervention within each census tract (using various assumptions on the allocation of such dollars). Other explanatory variables, such as the U.S. Census, EPA Toxics Release Inventory, 1999 National Air Toxics Assessment, and water quality data will be available for use in these models.

These census-tract level summary data (both response variable and explanatory variables) were modeled using a similar approach to what is being proposed for the National (Low-Resolution) Model – only the unit of clustering was census tract rather than county.

3.0 DATA SOURCES AND DATABASE DEVELOPMENT

The main goal of the statistical analysis were to develop a series of predictive models that help provide a better understanding of (1) the relative importance of various exposure sources in addition to leaded paint in housing and (2) the geographic areas across the United States that remain at increased risk for childhood lead poisoning. To do so, blood-lead data were combined with various environmental, demographic, and programmatic datasets at different levels of geographic specificity and coverage. A description of each of these data sources, as well as discussion of how they were combined, is included in this section.

3.1 Children's Blood-Lead Measurements

The statistical models are based upon blood-lead levels of children corresponding to the various geographic areas studied. To enable national analyses, CDC's Lead Poisoning Prevention Branch provided quarterly summary data from their national surveillance database for children aged 6-36 months within each county that had submitted data. These summary measures included the number of children screened, percentage of children who exceeded certain blood-lead thresholds, and arithmetic mean blood-lead concentration for state/local grantees with a history of universal reporting.

The intention was to have the models reflect the annual prevalence of childhood lead poisoning over time. Thus, the data were summarized so that each child could only be reported once a year. An algorithm was developed to select representative screening test(s) for children with multiple results with an objective of having children represented in the analysis dataset maximally once a year. For a given patient with multiple testing results, the algorithm preferentially selected tests confirming elevated blood-lead levels and then selected follow-up tests taken beyond nine months of the previously selected test. Screening tests were selected when no confirming record was available.

The response variable consists of quarterly summary statistics from 1995-2005 on the distribution of observed blood-lead concentrations in counties across the nation, based on information from CDC's national surveillance database. It should be noted that there likely exists significant variation and differences in the sampling and analytical methodologies employed in performing childhood blood-lead testing among the different counties that contributed to the CDC dataset, and within counties over time. Sampling methods include both capillary and venous tests, and different laboratory methods likely are represented within the data with varying reporting limits or limits of detection. Variation in reporting limits and limits of detection could introduce significant biases into statistical models of any continuous measures of blood-lead concentration that could be used in statistical models, such as the geometric or arithmetic mean blood-lead concentration. Alternatively, there was agreement among the research team and the CDC that measures whether a testing result was found above or below certain threshold values (5, 10, 15, and 25 μg/dL) would be more robust to these potential reporting and detection biases. Therefore, the National (Low-Resolution) Models focus on the proportion of screened children found above these threshold values using a logistic regression modeling approach.

After summarizing the test-level data by year, quarter, and county, counties that contained less than five test records in a quarter were excluded for confidentiality reasons. The time series of summary statistics within select counties were initially investigated to determine appropriate exclusion criteria to ensure that the data retained for analysis represented blood-lead concentrations that were universally reported (i.e., there were periods of time in which some state or local childhood lead poisoning prevention programs only reported elevated blood-lead concentrations – and these data needed to be eliminated from the analysis). Thus, the number of quarterly summary statistics varied from county to county within the analysis dataset.

As a prelude to developing the screening algorithm for elimination of data from counties that were not following universal reporting protocols, a subset of data from counties with obvious non-universal reporting was identified from within the National quarterly aggregate summary database. The algorithm was developed based on application to this subset of data prior to being utilized on the remainder of the National Surveillance database. The algorithm is based on the following:

Let

- n_{ij} represent the number of children observed in the i^{th} county during the j^{th} quarter
- $P90(n_i)$ represent the 90^{th} percentile of observed n_{ij} within the i^{th} county
- AM_{ij} represent the arithmetic mean blood-lead concentration observed in the i^{th} county during the j^{th} quarter
- $P50(AM_i)$ represent the 50^{th} percentile of observed AM_{ij} within the i^{th} county
- $P10_{ij}$ represent the proportion of children with blood-lead concentration observed at or above 10 μg/dL in the i^{th} county during the j^{th} quarter.

Then the following 3 exclusion/inclusion criteria are applied sequentially:

> Criterion #1: If $n_{ij} < \text{Max}(P90(n_i)/5, 15)$ and ($AM_{ij} \geq 2 * P50(AM_i)$ or $P10_{ij} \geq 0.75$), then exclude the data from the i^{th} county during the j^{th} quarter. This exclusion criterion essentially eliminates county/quarter combinations with relatively lower screening penetration (compared to when peak screening was achieved) and high blood-lead concentrations. The rationale for this exclusion criterion is that the periods of time in which a lead poisoning prevention program is not conducting universal reporting will involve fewer reported testing results that have higher blood-lead concentrations.
>
> Criterion #2: If $n_{ij} > 100$ and $AM_{ij} < 7$, then include the data from the i^{th} county during the j^{th} quarter. This criterion was added to include a small number of county/quarter combinations within the testing subset of data that were eliminated by the first exclusion criteria but did not appear to be inconsistent with the remainder of data that would be included in the analyses. This second criteria was inspected carefully upon application to the entire set of quarterly county summary statistics from CDC's National Surveillance database, to ensure that it was reintroducing data into the analysis in a manner consistent with the data analysis goals.

Criterion #3: If $n_{ij} < 100$ and $AM_{ij} > 10$, then exclude the data from the i^{th} county during the j^{th} quarter. This third criteria was established to exclude a small amount of data that was not captured by the first exclusion criteria (mostly representing counties with a median observed blood-lead concentration slightly above 5 μg/dL)

Within the quarterly county summary statistics from CDC's National Surveillance database, there were 72,466 county/quarter combination-level records. Application of Criterion 1-3 above eliminated an total of 2,351 records (3.25%) from the final analysis dataset.

To enable analyses at a finer level of geographic detail than the county level, the MDPH provided blood-lead surveillance data on specific testing results for individual children (with confidential identification information excluded) so that data could be summarized and reported by census tract. The Massachusetts blood-lead surveillance data represents all children aged 6-36 months tested from the period 2000-2006. As with the national data, quarterly census-tract-level records were created for analysis.

Due to selection bias, it is expected that the CDC National Surveillance dataset as well as the Massachusetts surveillance data may show higher proportions of elevated blood-lead concentrations than found in the general population. For this reason, the proportion of children with elevated blood-lead concentrations as well as the distribution of the potential continuous summary measure derived from the surveillance data were compared with those reported by the most recent six years of available CDC National Health and Nutrition Examination Survey (NHANES). Results of this comparison are presented in Section 7.2. In the future, to account for differences between the surveillance and NHANES data, modifications could be made to the models to calibrate the surveillance data to better match the national distribution of childhood blood-lead concentrations as appropriate (Strauss, 2001a).

3.2 Demographic Data

Demographic information from the 2000 U.S. Census was utilized in both the high- and low-resolution models, with data being acquired at the county level for the entire nation and at the census-tract level for Massachusetts. The Census 2000 data gathered by the Census Bureau includes over 1,000 variables. To narrow the scope of the project, 43 variables within 9 general categories were selected and explored, most of which had been used previously by the project team in a CDC-sponsored study to predict risk of elevated blood-lead concentrations at the census tract level (Strauss, 2001b). In many cases, the census variables are constructed from counts or summary statistics published in the detailed Census 2000 tables. For example, within each geographic area, the Census Bureau reported the number of houses that were built before 1950 and the median income of all households. In order for the analysis to draw comparisons from tract to tract and/or county to county, however, the Census variables needed to be manipulated in a fashion that depended upon the format of the variable. For example, count variables, such as the number of housing units built before 1950, were changed to percentages. Summary statistic variables describing income on the other hand, may be standardized within state to adjust for between-state differences in the cost of living. Table 3-1 supplies the list of the variables investigated within the nine categories and notes how they were calculated.

Table 3-1. Initial Variables for Analysis Created From the 2000 Census

Variable Group	Census Variable*	Format	Calculation	Analyzed Variable
Density	Persons	Count	Land Area (Units = .001 km^2)	Population Density
	Housing units	Count	Land Area (Units = .001 km^2	Housing Density
Race	White population	Count	Persons	Pct White
	Black population	Count	Persons	Pct Black
	Indian, Eskimo, and Aleut population	Count	Persons	Pct American Indian and Alaskan Native
	Asian Pacific population	Count	Persons	Pct Asian
	Other Race population	Count	Persons	Pct Other Race
	Native Hawaiian and Other Pacific Islander population	Count	Persons	Pct Native Hawaiian and Other Pacific Islander
	Multiple Race population	Count	Persons	Pct Multiple Race
	Hispanic population	Count	Persons	Pct Hispanic
Age	Children Less than or Equal to 6 Years Old	Count	Persons	Pct le 6 years
	Median Age*	Statistic		Median age of persons
	Median Age of Children Less than or Equal to 6 Years Old*	Statistic		Median age of persons LE 6 years
Family Structure	Single Parent* = Single Male with Children + Single Female with Children	Count	Household with Children Less than or equal to 18 years old = Married Couple with children + Single Male with Children + Single Female with Children	Pct Single Parent
Education	Less than a 9th grade Education	Count	Persons 18 years old and over	Pct less than 9th grade
	Less than high school* = #13 + persons with 9th to 12th grade education without obtaining a high school diploma	Count	Persons 18 years old and over	Pct no HS degree
	Less than college* = #14 + persons with high school diploma, but no college experience	Count	Persons 18 years old and over	Pct no college
	Less than college degree* = #15 + persons that attended college without obtaining a college diploma	Count	Persons 18 years old and over	Pct no college degree

Table 3-1. (continued)

Variable Group	Census Variable*	Format	Calculation	Analyzed Variable
Income	Household Median Income	Statistic		Standardized Median Income for Households
	Family Median Income	Statistic		Standardized Median Income of Families
	Per Capita Income	Statistic		Standardized per capita income of persons
	Households without earnings	Count	Households	Pct No Earnings
	Households without wages	Count	Households	Pct No Wage or Salary
	Households that obtain public assistance	Count	Households	Pct With Public Assistance
Poverty Level	Persons below poverty level	Count	Persons for whom poverty status is determined	Pct Persons Below Poverty
	Persons who are less than or equal to five years old that are below poverty level*	Count	Persons who are less than or equal to five years old for whom poverty status is determined	Pct Persons Below Poverty of Age LE 5 Below
	Families with total income below the poverty level	Count	Families	Pct Families Below Poverty
	Families with total income below the poverty level that have children under 5 years old.	Count	Families with children under five years old	Pct Poverty of Families with Children LT 5
Housing Units	Vacant	Count	Housing Units	Pct Vacant
	Housing Units Built before 1940	Count	Housing Units	Pct Pre 1940 Housing
	Housing Units Built before 1950	Count	Housing Units	Pct Pre 1950 Housing
	Housing Units Built before 1960	Count	Housing Units	Pct Pre 1960 Housing
	Housing Units Built before 1970	Count	Housing Units	Pct Pre 1970 Housing
	Housing Units Built before 1980	Count	Housing Units	Pct Pre 1980 Housing
	Median Year that Housing Units were Built	Statistic		Median Year Built
	Median Year that Housing Units were Built - Calculated by the Project Team	Statistic		Calculated Median Year Built

Table 3-1. (continued)

Variable Group	Census Variable*	Format	Calculation	Analyzed Variable
Occupied Housing Units	Housing Units that are rented	Count	Occupied Housing Units	Pct Renter Occupied
	Occupied Housing Units Built before 1940	Count	Occupied Housing Units	Pct Pre 1940 Occupied Housing
	Occupied Housing Units Built before 1950	Count	Occupied Housing Units	Pct Pre 1950 Occupied Housing
	Occupied Housing Units Built before 1960	Count	Occupied Housing Units	Pct Pre 1960 Occupied Housing
	Occupied Housing Units Built before 1970	Count	Occupied Housing Units	Pct Pre 1970 Occupied Housing
	Occupied Housing Units Built before 1980	Count	Occupied Housing Units	Pct Pre 1980 Occupied Housing
	Median Year that Occupied Housing Units were Built	Statistic		Median Year Built - Occupied Only
Housing Value	Median Rent	Statistic		Standardized Median Gross Rent
	Value of Owner Occupied Housing Units	Statistic		Standardized Median Housing Unit Value

*Variables that were created by combining different pieces of information from the 2000 Census

Income and Poverty

Median income per household, family, and person were calculated. Additionally, the proportion of households that do not receive any wages, do not receive any earnings, and do receive public assistance were investigated. The census defines earnings and wages as follows:

- "Earnings" represent the amount of income received regularly before deductions for personal income taxes, Social Security, bond purchases, union dues, Medicare deductions, etc.
- "Wages" include total money earnings received for work performed as an employee during the calendar year 1999. It includes wages, salary, Armed Forces pay, commissions, tips, piece-rate payments, and cash bonuses earned before deductions were made for taxes, bonds, pensions, union dues, etc.

Similar to the income variables described above, the poverty level of individuals and families within each county were summarized as the variables Percent Persons and Percent Families Below the Poverty Level. In order to focus on the poverty level of the children within each county, however, Percent Persons Five Years and Under and Percent Families with Children Under Five Years Below Poverty Level variables were created. Note that in calculating the various percentages for each of the variables, the denominator changes. Also note that for some of the multivariate models presented later in the report, some of the income variables may have been rescaled to represent income in thousands of dollars, to allow the parameter estimates for the regression models to be discernable within the first 3 significant digits.

Race
The Census Bureau presents five general race groups; (1) White, (2) Black, (3) Indian, Eskimo, and Aleut, (4) Asian Pacific and (5) Other, each of which was included and explored separately. Additional variables were included on percent of Native Hawaiians and Other Pacific Islanders, percent of the population reporting multiple races (Percent Multiple Races), and percent of the population reporting that they are Hispanic (Percent Hispanic).

Housing Cost
Two variables were constructed to investigate housing cost – Median Rent and Median Housing Value. Median Housing Value includes the value of all housing units (owned and rented). Both of these variables were standardized to account for state-to-state differences in the cost of living. Note that for some of the multivariate models presented later in the report, some of the housing cost variables may have been rescaled to represent housing costs in thousands of dollars, to allow the parameter estimates for the regression models to be discernable within the first 3 significant digits.

Occupancy
Occupied housing units are more likely to have lead paint removed than vacant homes. Thus, the percent of housing units that are vacant potentially indicates the level of care taken to maintain buildings within the area. Buildings that are not occupied are more likely to accumulate dust or debris to which the children of an area may be exposed upon reoccupancy. Percent of vacant housing units was explored for those reasons. Similarly, the standard of care could be different between rental properties and owner-occupied properties. Thus, the percent of rental units in an area also was explored. The percent of occupied housing units that are rented, rather than owned, was calculated by dividing the number of rented occupied housing units within an area by the total number of occupied housing units.

Family Structure
The Census Bureau does not supply a unique variable that indicates the number of single parent households within an area. Therefore, this variable was created by combining Census variables as follows:

M = Number of Households with a male householder (no wife present) whose own children are under 18 years old
F = Number of Households with a female householder (no husband present) whose own children are under 18 years old
T = M + F + Number of married couples with own children under 18 years.

The Percent of Single Parent Households variable used represented (M+F)/T.

Housing Age
During the 1950s, as the United States started to become aware of the consequences associated with the exposure of lead in paint, the use of lead paint within homes began to decrease. In 1977, however, the use of lead paint in homes became illegal. Thus, the years during which the housing units were built within each area is important to characterize; older homes are more likely to contain lead paint than newer homes. A number of variables related to housing age by county were investigated to identify those that best predict children's blood-lead levels. Census

data on the full population of housing units as well as the population of occupied housing units were investigated. Note that for some of the multivariate models presented later in the report, the median age of house variable was centered at 1950 to provide stability to the intercept term in the models.

Children's Age

The Census Bureau does not report all data by single years of age. More typically the agency reports the total number of people that fall into various age categories. The variable, "Pct LE 6 years" was created to identify the number of children within each geographic area less than or equal to six years of age at the time of the 2000 Census. Additionally, the median age of the total population and of those less than or equal to six years old was calculated by taking a weighted average of the midpoint of each age category (the counts are used as the weights).

Education

A series of variables pertaining to the proportion of adults with various levels of education were created as follows:

- L9 = Number of people older than 18, that have less than a 9th grade education
- L12 = Number of people older than 18, that have 9th though 12th grade experience, but do not have a high school diploma
- 12 = Number of people older than 18, that obtained a high school diploma or GED
- C = Number of people older than 18, that have some college experience but did not receive a college degree
- T = Number of People that are over than 18 years old

Percentage variables were created from the L9 through C variables by dividing them by the total number of people over 18 years old. Exploratory analyses were conducted upon the four percentage variables.

Population Variables

Because both counties and census tracts vary with respect to spatial area and population, and previous work suggests that risk of childhood lead poisoning differs between rural and urban areas, a population density variable was used as a potential explanatory variable or effect modifier in the statistical models. Population density was explored in two ways. The first divides the number of people within the tract by the amount of land area measured in .001 square kilometers. The second divides the number of housing units by the amount of land area measured in .001 square kilometers. Housing units include the following: a house, an apartment, a mobile home, a group of rooms, or single room that is occupied as separate living quarters.

3.3 Environmental Data

Environmental data acquired for this project include air and groundwater monitoring data aggregated at the county level for the low-resolution model and at higher resolutions for the Massachusetts analyses. In cases where the data were available for a limited number of air-monitoring stations or drinking water samples available for the region(s) being investigated, geo-spatial modeling techniques might be used as appropriate to develop predictions across the entire region. Existence of industrial sources of lead within each county and census tract, as indicated

by the Toxics Release Inventory (TRI), also were included as an environmental data source. Each of theses data sources are discussed in further detail below.

3.3.1 Concentrations of Lead in Air

EPA maintains a number of ongoing air monitoring programs that collect data over time on concentrations of various criteria air pollutants, air toxics, constituents of particulate matter, and other airborne chemicals. Each of these monitoring programs have multiple air monitoring stations that are deployed throughout the country to meet various goals associated with the Clean Air Act and other federal and state regulations and programs. For example, some of the monitoring stations are placed in close proximity to industrial sources of pollution and major populations centers, while other stations are placed in remote areas to assess background chemical concentrations. While many of these monitoring sites provide information on the concentration of lead in air over time, a quick assessment of the spatial coverage of these monitoring networks suggested that making use of these data would be problematic for this study due to time and resource constraints. Lead concentrations in air from the monitoring networks are not available in the majority of counties that will be covered in the low-resolution model, or the census tracts that will be covered in the high-resolution models – as shown at the following EPA Website (http://www.epa.gov/airtrends/lead.html).

Rather than using air monitoring data as described above, the study used modeled predictions of concentrations of lead in air from EPA's 1999 National Scale Air Toxics Assessment – in which county and census-tract-level predictions are available throughout the entire country based on the use of predictive models. Documentation for the 1999 National Scale Air Toxics Assessment, as well as the predicted air concentration data can be found at http://www.epa.gov/ttn/atw/nata1999/tables.html. The predictions were generated using the Assessment System for Population Exposure Nationwide, or ASPEN. This model is based on the EPA's Industrial Source Complex Long Term model (ISCLT), which simulates the behavior of the pollutants after they are emitted into the atmosphere. ASPEN uses estimates of toxic air pollutant emissions and meteorological data from National Weather Service Stations to estimate air toxics concentrations nationwide.

The ASPEN model takes into account important determinants of pollutant concentrations, such as:

- rate of release
- location of release
- the height from which the pollutants are released
- wind speeds and directions from the meteorological stations nearest to the release
- breakdown of the pollutants in the atmosphere after being released (i.e., reactive decay)
- settling of pollutants out of the atmosphere (i.e., deposition)
- transformation of one pollutant into another (i.e., secondary formation).

The model estimates toxic air pollutant concentrations for every county and census tract in the continental United States; however, these data are only available for 1999. Both the Broad-Based National Model and the High-Resolution Model within Massachusetts considered the integration of information from the ASPEN Model. The National Model investigated the

median, average, and 95th percentile predicted air lead concentration within each county, while the High-Resolution Model only considered the average predicted air lead concentration within each census tract. Within the National Model, the median, average and 95th percentile predicted air-lead concentrations were mostly distributed near zero. For this reason, zero/one indicator variables were created to indicate that the observed value of these ASPEN Model predictions were observed at or above the 95th and 99th percentile within the analysis dataset for potential use within the predictive models. In addition, EPA collaborators identified a subset of 20 counties with observed elevated air-lead concentrations, and an indicator variable was used to assess whether these 20 counties had higher risk of childhood lead poisoning in the predictive models.

The second air-lead variable investigated is based on predictions from the HAPEM5 (Hazardous Air Pollutants Exposure Model, Version 5) model. According to the EPA website, "the HAPEM5 model has been designed to predict the 'apparent' inhalation exposure for specified population groups and air toxics. Through a series of calculation routines, the model makes use of census data, human activity patterns, ambient air quality levels, climate data, and indoor/outdoor concentration relationships to estimate an expected range of 'apparent' inhalation exposure concentrations for groups of individuals."[2] Because air quality concentrations in indoor environments can be quite different than those in the outdoor environment, an exposure model generally is employed to predict the apparent inhalation exposure. The Air Exposure (HAPEM5) model variable captures the predicted exposure data from this model.

The third air-lead variable considered, Air Hazard Quotient (HQ), is derived from the 1999 National Scale Air Toxics Assessment data. This variable represents lifetime exposure for children at the centroids of each census tract or county. Lifetime exposure is calculated based on considering annual exposures and yearly activity patterns. The HAPEM5 and HQ air-lead variables were only considered within the context of the High-Resolution Model within Massachusetts.

3.3.2 Toxic Release Inventory Variables

EPA's Toxics Release Inventory (TRI) catalogs various sources of lead, based on information provided by industrial facilities. This data source was used to generate county- and census-tract-level estimates of the total amount of lead and/or lead-containing compounds that are released by industrial facilities into the environment via air, surface water, or underwater injection. Although the above-described ASPEN modeling results are based on the (airborne) emissions data and how they would theoretically translate into average ambient air-lead concentrations, the data from the TRI are available for multiple years and for other types of emissions (such as surface water). Thus, this information has the potential to add predictive power to the models.

Three types of TRI variables were utilized – total compounds, lead only, and total lead. Within each type, five pollution variables were explored – total lead in the air, lead in fugitive air, lead from smokestacks, lead in surface water, and lead in water by injection. Thus, 15 total TRI data variables were evaluated.

[2] http://epa.gov/ttn/atw/nata1999/ted/teddraft.html

Within the National Model, the distributions of the TRI emissions variables were mostly concentrated near zero. For this reason, additional zero/one indicator variables were created to indicate that the observed value of these TRI emissions were observed at or above the 95th and 99th percentile within the analysis dataset for potential use within the predictive models.

3.3.3 Water Quality Data

The plumbing system inside a home and the service line from the street to the home may contain lead and can contribute to drinking water contamination. To address this potential source, EPA obtained data from their Safe Drinking Water Information System that includes the 90th percentile result of tap water lead levels for public water systems. Public water suppliers must monitor at customer's taps every 6 months. Public water systems can reduce monitoring to annually, triennially, or every 9 years (if granted a monitoring waiver) if the 90th percentile value from previous monitoring is at or below the action level of 15 parts per billion. The number of customer's taps, or monitoring sites, that a system is required to sample is based on the population served by the system. Further, systems are required to select sites that are most likely to have the highest lead levels (i.e., older homes, homes with copper pipes with lead solder or homes served by a lead service line). Therefore, the 90th percentile value of samples collected during a monitoring period is not reflective of individual exposure to lead in drinking water. Data available from this monitoring program include 90th percentile water lead values for public drinking water systems serving greater than 3,300 persons (systems serving less than 3,300 persons are required to report the 90th percentile level only if they exceed the action level), the population size served by each facility, the start and end date for the monitoring period, and the county in which the facility is located. These data were used to construct a population-size weighted average 90th percentile water-lead concentration variable within each county/quarter combination. However, it is important to note that most public water systems do not remain within county lines. Large water systems may serve multiple counties or a county may be served by several small public water systems.

Because there were some county/quarter combinations with no observed data from EPA's Safe Drinking Water Information System, an indicator variable was developed to indicate whether the county/quarter included a monitored facility (or not) – allowing an intercept to be fit among those county/quarters with no drinking water monitoring, and a slope estimate to be fit for the effect of the weighted average 90th percentile drinking water-lead concentration among reported facilities.

EPA's Safe Drinking Water Information System data were not geocoded to the census-tract level, and therefore these data were only available for use in supporting the Broad-Based National Model at this time.

3.4 Programmatic Data

Most of the explanatory variables being explored in this project are considered risk factors for childhood lead poisoning. Among factors that might mitigate these risks, it was anticipated that the level and characteristics of programmatic support from either federal, state, or local sponsors may contribute toward meaningful reductions in the prevalence of childhood lead poisoning. The level of financial support available within each county served as a proxy for programmatic

support in the low-resolution (National) models. In the high-resolution models run for Massachusetts, information from housing inspections also were explored within the statistical models. The following sections detail the specific characteristics of the variables used within the models.

3.4.1 Programmatic Funding Variables

The goal of this variable is to construct a longitudinal history of current and cumulative per-capita dollars allocated to each county and census tract to combat childhood lead poisoning. For use in both the national and state models, data were obtained from HUD's Office of Healthy Homes and Lead Hazard Control on grants funded since the inception of the Lead-Based Paint Hazard Control Grant Program in 1992. Data also were obtained from CDC's Lead Poisoning Prevention Branch on their program's grant funding approximately three weeks prior to the end of this project, and therefore these data were only able to be integrated into the Massachusetts models due to time constraints.

Four variables were generated from these data and analyzed – current and cumulative funding allocated to each county or census tract to combat childhood lead poisoning, both Standardized by number of children per tract and Not Standardized. The Standardized variable is a funding per child variable while the Not Standardized versions are funding for geographic area variables. For the high-resolution model in the Commonwealth of Massachusetts, information on within-state funding levels was obtained and analyzed. Within-state funding data were available down to the township level. The state, HUD, and CDC funding data also were combined to create Total Funding variables, including both current and cumulative levels and both Standardized and Not Standardized versions. The total funding variables also were only investigated as part of the Massachusetts analyses.

Becuase there may be delays in the effects of programmatic funding on risk of lead poisoning, time-lagged versions (at 6-, 12-, 18-, 24-, 30-, and 36-months) of the programmatic funding variables in the National Model also were investigated.

3.4.2 EPA Region

The EPA region was investigated as a potential predictor of children's blood-lead levels to determine if that high-level geographic indicator should be included as a stratification variable in the national multivariate models.

3.4.3 Housing Inspection Data (Massachusetts)

The Commonwealth of Massachusetts maintains an extensive database on all lead-based paint inspections conducted over time (dating back to the early 1990s). The MDPH provided a database that contains a single record for each inspection, with the following information: housing-unit id, census tract, date of inspection, and result of inspection (whether the housing unit was found to be in compliance with Massachusetts standards). The database contains records on over 200,000 housing units – with many housing units having multiple inspections

over time. Note that for units with multiple records, time periods in which the units were both in and out of compliance with the Massachusetts standards were identified.

These data can be used in the Massachusetts high-resolution models in two ways. First, a longitudinal summary measure of the proportion of housing within each census tract that was known to be in compliance with the Massachusetts standards was developed. It was anticipated that within a census tract, as this proportion increases over time, the risk of childhood lead poisoning will decrease. Second, due to the fact that individual blood-lead records from Massachusetts with linkable housing-unit identification variables were available, a determination could be made regarding whether a housing unit was in compliance at the time of the blood-lead test for each child in the database (with potential outcomes of the determination being yes, no, and unknown).

The first approach described above is consistent with the methods for exploring aggregated summary blood-lead information over time within each census tract. The second approach allows utilization of some predictive information at the individual child level. This information may help improve prediction, and also may help assess what information might be lost when transitioning from individual-level data to aggregate summary data in the analyses. Unfortunately, due to time and resource constraints, only the first method was explored within this project. Thus, the three measures listed below were calculated using four different methods. The three measures are:

- P - represents the Proportion of Housing Units within a census tract that are assumed to Meet the Massachusetts Standard of Care at any given time
- F - represents the Proportion of Housing Units within a census tract that are assumed to Not Meet the Massachusetts Standard of Care at any given time
- N - represents the Proportion of Housing Units within a census tract with Housing Inspection Information at any given time.

As noted, the measures were generated in four different ways, each handling the longitudinal information in a slightly different manner. The four measures, numbered in the model results from 1 to 4, are listed below.

1. Naïve Method 1 – Create a longitudinal history for each housing unit inspected, and treat the first inspection observation as being representative for time periods preceding that inspection.
2. Naïve Method 2 – Create a longitudinal history for each housing unit inspected, and assume missing information for time period preceding the first test on each unit.
3. Naïve Method 3 – Create a longitudinal history for each housing unit inspected, and treat the first inspection observation as being representative for time periods preceding that inspection if the housing unit failed, and assume missing information for time period preceding the first test if the unit passed.
4. MDPH Approved Method – Create a longitudinal history for each housing unit, with different rules for the treatment of the time-period preceding the first test based on (a) the housing inspection result and (b) the reason for ordering the inspection.

Note that for housing units with multiple inspections, each housing inspection result is assumed to be representative of the house (either pass or fail) until the next result. The last result is carried forward over time (e.g., if the last observed inspection on a house passed in November of 1998 – that particular house is assumed to be meeting the Massachusetts standard of care over all subsequent time periods in the dataset). If multiple inspections occur on the same house within a particular quarter (3-month interval), the maximum result (with pass being coded as a 1, and fail being coded as zero) is used to represent the house. The 0/1 results are then summed across all observed housing units within each census tract over time (quarters). The summed results are then divided by the number of housing units reported within each census tract from the 2000 Census.

While all of the above described housing inspection variables were investigated in the exploratory data analyses, only the P4, F4, and N4 variables associated with the MDPH-approved method of constructing the longitudinal history within each housing unit observed was considered within the context of the multivariate models.

3.5 Data Linkages

The primary objective of this pilot study was to utilize combined information from different sources at various levels of geographic and temporal specificity to more accurately target geographic areas at high risk for not meeting the 2010 goal of eliminating childhood lead poisoning. As such, work on the study required careful integration of a variety of data sources with various characteristics and documentation. Data to support this study were gathered from a variety of sources, including federal, state, and local lead poisoning prevention programs, as well as publicly available data downloaded from the internet (e.g., Census data, EPA's Toxics Release Inventory), as detailed in the previous sections.

Upon receipt of each data source, the data and supporting documentation was reviewed to gain knowledge on the structure, relationship, and quality of the data. Database managers worked with the project team (including collaborators providing data to the project, as well as EPA) to determine the final format for each database, desired uses of the databases, as well as the requirements for maintaining the databases. Based on this information, separate master databases were constructed for the national model and for the high-resolution Massachusetts model that integrate the various environmental, demographic, and programmatic variables, and facilitate statistical analyses of the combined data. These databases were constructed by combining data from a variety of formats including MS SQL Server, MS Access, Excel, ACSII, Access, ArcView, and SAS® electronic databases. In order to combine the various data sets, they were merged on key fields, including state, county, census tract, and time period. The data being used for analyses of a particular geographic level (e.g., county) are comparable because they are representative of that geographic area.

Throughout the development process, checks for completeness were conducted on all study databases, and the project team worked with data-sharing collaborators and EPA to attempt to complete missing data as necessary to support the proposed statistical analyses. Any changes to the databases (corrections, additions, deletions, etc.) were documented in appropriate metadata files. Documentation of the combined master databases is included in Appendix H.

Standard Operating Procedures (SOPs) were followed to ensure the proper storage, backup, and retrieval of datasets created and analyzed for this study. The various databases were backed up to tape nightly via automated backup routines, and were only accessible to members of the project team. CD-ROM backups were made on a regular basis to serve as a safeguard in case the backup system failed for any reason.

Microsoft Access and SQL server were the primary software tools used for data management. The SAS® System was the primary statistical data analysis tool used on this project. ArcView software was used to translate results into maps, as seen in Appendices F and G.

The data utilized for the study were maintained in a manner that preserved the confidentiality of all the data and prevented its unauthorized release. As data files were received from EPA, the original data (e.g., data with personal identifiers) were handled as though they were classified as confidential business information (CBI) under the Toxic Substances Control Act (TSCA), even though EPA may not specifically classify these data as "CBI." The data files were not shared with anyone outside of the project team.

4.0 EXPLORATORY DATA ANALYSES

Because the goal of this study was to develop a series of statistical models that predict the risk of childhood lead poisoning at the geographic level across multiple response variables (proportion of children screened at or above 5, 10, 15 and 25 ug/dL), all potential predictor variables first were explored individually to determine their predictive ability. Results from these bivariate analyses were assessed to identify the set of variables to include in the multivariate model that predicts how the risk of childhood lead poisoning changes over time among the various census tracts and counties included in the analysis.

This section of the report provides the results of the series of exploratory analyses described in Section 2.2, which were performed to assess the potential predictive power of various candidate demographic, environmental, and programmatic variables for potential use in the multivariate models. These exploratory analyses initiated with an assessment of the study sample, i.e., the proportion of counties in the sample with complete and reliable data for both the response variable and the explanatory variables.

Each candidate predictor variable was reviewed with particular attention focusing on the manner in which the county-level predictor variables would be merged with the quarterly summary blood-lead information prior to fitting the statistical models. In preparation for developing longitudinal statistical models, univariate summaries of each predictor variable as a function of time were produced. Comparisons of these distributions were made using side-by-side box-plots for continuous data or bar-charts for categorical data. This helps verify that the data are clean and ready for analysis and helps identify cells with sparse data. Such descriptive analyses were conducted on each predictor variable database to characterize the distributions of all observed variables using frequency distributions for categorical variables, and simple summary statistics (mean, median, mode, minimum, maximum, and select percentiles) for continuous variables.

The univariate descriptions then were followed by fitting a series of cross-sectional bivariate relationships between the blood-lead response variable(s) and each candidate explanatory variable. These cross-sectional relationships were explored as a function of time to better understand the stability of these relationships, and whether they change over time, so that they can be modeled appropriately in the more sophisticated longitudinal analyses. These analyses also help identify which explanatory variables are most predictive of the blood-lead response variable.

4.1 Relationship between National Blood-Lead Data and Explanatory Variables

The response variable for the national data analysis consisted of quarterly summary statistics from 1995-2005 on the distribution of observed blood-lead concentrations in counties across the nation, based on information from CDC's National Childhood Lead Poisoning Surveillance Database. The time series of summary statistics within select counties were initially investigated to determine appropriate exclusion criteria to ensure that the data retained for analysis represented blood-lead concentrations that were universally reported (i.e., there were periods of time in which some state or local childhood lead poisoning prevention programs only reported elevated blood-lead concentrations – and these data needed to be eliminated from the analysis).

Thus, the number of quarterly summary statistics varied from county to county within the analysis dataset.

The national blood-lead data were categorized into four time periods – (1) January 1, 1995 to December 31, 1999; (2) January 1, 2000 to December 31, 2001; (3) January 1, 2002 to December 31, 2003, and (4) January 1, 2004 to December 31, 2005 – so that change over time could be evaluated. Using the specified four time periods split the dataset of quarterly county-level records into roughly similar sizes. Presented below are the exploratory analysis results for the demographic, environmental, and programmatic variables investigated. Detailed figures and tables containing results are included in Appendix A. A detailed discussion of the results seen in Appendix A is contained in Appendix D.

To allow comparison of the different variables explored within each variable type, Tables 4-1 through 4-4 present the log-likelihood statistic from each single covariate model presented in Appendix A for each of the four blood-lead threshold values, respectively. Each explanatory variable was investigated in four different ways with respect to how the effect of the variable might vary over time within the longitudinal analysis dataset:

1. Investigate the explanatory variable on its own, assuming that the effect remains stable over time.
2. Investigate the explanatory variable with a linear interaction with time, assuming that the effect of the variable on risk of childhood lead poisoning either increases or decreases linearly over time (on the logit scale).
3. Investigate the explanatory variable with a quadratic interaction with time, assuming that the effect of the variable on risk of childhood lead poisoning either increases or decreases as a quadratic function in time (on the logit scale).
4. Investigate the interaction between the explanatory variable and four select time periods, which is helpful for diagnosing whether the effect remains stable (or changes) over time – but is not particularly useful for the final multivariate model where the application of the model might be to forecast how risk of lead poisoning might extend into future years.

Within each variable category, the variable that provided the best fit across the four time variables is indicated with a double asterisk (**). For example, within the income category, the Categorical time variable achieved the best fit for seven of the eight income variables in the model of proportion of children with blood-lead levels above 5 μg/dL. Within that category, Percent No Household Wage achieved the best model fit across the 8 income variables. Those variables (indicated with the double asterisk) were the most likely to become candidate predictors for the multivariate statistical models.

Table 4-1. Summary of Exploratory Analysis Fit as shown by -2 Log Likelihoods for Pr(PbB >= 5 µg/dL) Models

Parameter Category	Variable Name	X Only	Linear Time	Quadratic Time	Categorical Time
Income	Median_Family_Income		235522.9		233514.6
	Median_HH_Income	235495.9			233439
	Median_Per_Capita_Income		235600.3	234626.6	233735.4
	Pct_HH_No_Earnings	235611.3	235601.3	234404.8	233260.3
	Pct_HH_No_Wage	235579.7	235567.8	234323.0**	233183.8*
	Pct_LT_Poverty		235763.3	234363.2	233608.5
	Pct_Family_Income_LT_Poverty	235761.5	235744.8	234446.2	233751.1
	Pct_LE_5Yrs_LT_Poverty		235794.2	234373.9	
Race	Pct_Asian	235769	235681.3	234627.7	234511.6
	Pct_Black	235867.4		234612.0**	
	Pct_White	235915.1		235259.3	233688.1
	Pct_NHOPI	235888.9		235536.9	235918.5
	Pct_Other_Race	235862.1	235848.1	235559.1	235250.1
	Pct_Multi_Race		235817.2	234679.6	234595
	Pct_Hispanic	235851.8			233502.6*
Housing Cost	Median_Rent	235265.4	235189.1	234046.8**	233401.0*
	Housing_Value	235571.9	235562.9	235128.2	234857.9
Occupancy	Pct_Rented	235913.4	235887.7		
	Pct_Vacant	235884.8	235872.2	235427.2**	234479.4*
Single Parent	Pct_Single_Parent	235878.1	235884	234527.9**	233660.7*
Home Age	Median_Yr_Built		235421.5	235483.4	234415.3
	Median_Yr_Occ_Built	235412.7	235432.6	234214.3	233119.1
	Pct_Built_Pre_1940	235268.9	235277.3	233803.4	233243.7
	Pct_Built_Pre_1950		235241.8	233574.5	232990.3
	Pct_Built_Pre_1960	235344.5	235357.7	233484.5**	232839.7*
	Pct_Built_Pre_1970	235449	235463.1	233650.7	232904.6
	Pct_Built_Pre_1980	235487.1	235503	233834.3	232946.2
	Pct_Occ_Built_Pre_1940	235277.5	235285.1	233834	233291
	Pct_Occ_Built_Pre_1950	235233.3	235245.1	233601.1	233030.9
	Pct_Occ_Built_Pre_1960		235359.2	233499.6	232866.4
	Pct_Occ_Built_Pre_1970	235458.6	235471.3	233666	232928.4
	Pct_Occ_Built_Pre_1980	235495.4	235510.7	233840.8	232959.7
Children	Pct_LE_Six	235827.7	235829.8	234448.4**	233449.9
	Num_LE_Six	235905.5	235924.2		227590.8*
Education	Pct_LT_9th_Grade	235850.1	235848.4	234398.6	233744
	Pct_No_HS_Degree		235715.8	234321.2	233411.6
	Pct_No_College		235509.4	234237.9	
	Pct_No_College_Degree	235556.5	235505.7	234226.0**	233106.0*
Population	Total_Housing_Units	235906.3**	235922.2		
	Total_Pop	235908.5	235928.6		228102.6*
	Housing_Density	235920.4	235925.8		235699.2

Parameter Category	Variable Name	X Only	Linear Time	Quadratic Time	Categorical Time
Air Lead	air_avg	235905.9	235906.7		235836
	air_med	235905	235905.5		235921
	air_p95	235907.6	235909.8	235815.6**	235735.7*
	air_avg_p95	235906.4	235909.6		233727.0*
	air_med_p95	235907.2	235135.4**	air_med_p95	235914.4
	air_p95_p95	235911.9	235918.5	236090.1	235244.7
	air_avg_p99		235906.2		
	air_med_p99	235903.5		235766.3**	235718.3
	air_p95_p99	235905.8			235695.8
TRI	TRI Compounds air_fug		235943.8	235684.6	235325.2
	TRI Compounds air_tot	235922.9	235940.6		235199.3
	TRI Compounds air_stk	235921.5	235940.7	235406.7	235355.6
	TRI Compounds under_inj	235925.3		235968.5	235986.4
	TRI Compounds water_surf	235921.3	235942.9	235910.3	235912.1
	TRI Lead Only air_fug	235927.2		235312.9	235155.4
	TRI Lead Only air_tot	235929.4	235955.8	235385.1	235169.4
	TRI Lead Only air_stk	235928.7	235955.2		235437.5
	TRI Lead Only under_inj	235928.5		235985.8	235997.1
	TRI Lead Only water_surf	235926.6	235952.3		235942.9
	TRI Lead Total air_fug	235927.3	235950.7	235244.9**	234850.4*
	TRI Lead Total air_tot	235930	235957.3		235036
	TRI Lead Total air_stk		235955.9	235478.1	235321.1
	TRI Lead Total under_inj	235928.4	235956.1	235984.3	235999.2
	TRI Lead Total water_surf	235926.7			235933
	tri_as1_p95	235909		234966.5	235011.7
	tri_as2_p95	235912.6	235917.9	234262.7	
	tri_as3_p95	235907.5	235910.7		
	tri_af1_p95	235914	235919.2	235132.3	235125.1
	tri_af2_p95	235910.9	235916.8	234837.1	233669.7
	tri_af3_p95		235915.8	234396.5	233058.4
	tri_at1_p95			234716.4	234606.5
	tri_at2_p95	235911.7	235915.7	233993.9	232572.7
	tri_at3_p95	235910.2		233685.3**	232564.5*
	tri_ws1_p95		235919.1	235370.9	235422
	tri_ws2_p95	235908.1	235915.4	233861.9	233253.8
	tri_ws3_p95	235908			
	tri_ui1_p95	235904.1	235904.1	234661.9	233559.3
	tri_ui2_p95	235904.1	235904.1	234661.9	233559.3
	tri_ui3_p95	235904.1	235904.1	234661.9	233559.3
	tri_as1_p99		235908.5	235519.5	
	tri_as2_p99	235908.3	235914.4	235104.8	
	tri_as3_p99	235906.6	235912		
	tri_af1_p99	235906.1		235735.5	
	tri_af2_p99	235907.5	235912.6	234680	234116.8
	tri_af3_p99	235905.7		235128.3	234668.4
	tri_at1_p99		235909.1	235711.7	235364.2
	tri_at2_p99		235915.6		234704.6

Parameter Category	Variable Name	X Only	Linear Time	Quadratic Time	Categorical Time
	tri_at3_p99	235907.7	235913.8	235213.9	234780.2
	tri_ws1_p99	235907	235912.3	235863.4	235847.8
	tri_ws2_p99	235907.6	235913	235667.2	235712.8
	tri_ws3_p99	235907.6	235913.9	235706.5	235748.5
	tri_ui2_p99	235904.1	235904.1	234661.9	233559.3
	tri_ui3_p99	235905	235910.1	235921.5	235887.1
Funding	CDC_cur_lag6		235415.6	234816.9	234090.5
	CDC_cur_lag12	235427.3	235275.1	234678.5	233854.6
	CDC_cur_lag18		235270.6	234599.8	233781.8*
	CDC_cur_lag24	235847	235039.4	234566.7	
	CDC_cur_lag30	235897.1	235246		
	CDC_cur_lag36	235727.6	235050.5	234702.5	234162.9
	HUD_cur_lag6	235914	235833.5	235432.4	235494.8
	HUD_cur_lag12	235761.9	235691.3	235358.3	235381.6
	HUD_cur_lag18	235762.7	235673.8	235437.5	235352.7
	HUD_cur_lag24	235707.6	235525.1	235434.3	235211.7
	HUD_cur_lag30	235672.8	235487.1	235445.4	235175.9
	HUD_cur_lag36	235626.3	235273		234912.4
	CDC_cum_lag6	235611.1	234603.5	234200.7	233952.6
	CDC_cum_lag12	235762.8	234668.8	234304.8	
	CDC_cum_lag18	235859.9	234760.3	234472.7	234162.5
	CDC_cum_lag24	235898.2	234851.6	234646.1	
	CDC_cum_lag30		234933.1	234793.8	234433.2
	CDC_cum_lag36	235883	234920.4	234824.7	234466.2
	HUD_cum_lag6	235908.5	234878.9	234725.3	234382.8
	HUD_cum_lag12	235924.3			234413.4
	HUD_cum_lag18	235961.9	235051	234650.3	
	HUD_cum_lag24	235952.8	235099.6	234558.3	234503.5
	HUD_cum_lag30	235897.3	235132.9	234450.4	234499.3
	HUD_cum_lag36	235794	235148.6	234334.5	
	tot_cur_lag6	235918.7	235841.5	235210	235306.5
	tot_cur_lag12		235727.4	235169.7	235194.9
	tot_cur_lag18	235799.2	235654.7	235245.5	235101.6
	tot_cur_lag24	235742.9	235439.1	235226.8	234897.9
	tot_cur_lag30	235657.8	235355.1	235216.4	234839.7
	tot_cur_lag36	235546.3	235043.1	235054.3	234518.3
	tot_cum_lag6	235928.9		234463	
	tot_cum_lag12		234735	234380.2	234067.2
	tot_cum_lag18	235958	234825.3	234340.7	234123.9
	tot_cum_lag24	235944.2	234876.7		234168.7
	tot_cum_lag30	235888.1		234190.6	234192.1
	tot_cum_lag36	235774.3	234937.9	234115.8**	234217.6
	HUD_cur	252824.5	252845.5		252876.3
	HUD_cum	252864.5	252670.1	252673.7	252679.3
	CDC_cur	252679.7	252170.2	251830.5	251799.9
	CDC_cum	252586.4	252213.5	252226.6	252202.3

Parameter Category	Variable Name	X Only	Linear Time	Quadratic Time	Categorical Time
Funding	Current – HUD+CDC		235788	235296.1	235214.1
	Cumulative – HUD+CDC	235940.8		234583.8	234050.2
Screening	screen_penetration	232188	232220.8	231442.3**	230769.8*

** Variable factor(s) showed best fit when adjusted for degrees of freedom and were thus chosen to represent parameter category in multivariate analysis.

* Variable factors showed best fit; however, were not included in multivariate analysis because the time categorical variable had less than ideal prediction properties.

Table 4-2. Summary of Exploratory Analysis Fit as shown by -2 Log Likelihoods for Pr(PbB >= 10 µg/dL) Models

Parameter Category	Variable Name	X Only	Linear Time	Quadratic Time	Categorical Time
Income	Median_Family_Income	252403.4	252372.3	251685	251406
	Median_HH_Income	252382.8		251605.3	251222.3
	Median_Per_Capita_Income	252486.2	252425	251792.5	251439.9
	Pct_HH_No_Earnings	252446.7	252444	251506.4	251091
	Pct_HH_No_Wage	252363.4	252336.2	251373.8**	250935.7
	Pct_HH_Public_Assist	252779.8	252806.7	251403.6	250521.9
	Pct_LT_Poverty	252743.7	252763.4	251392.3	250602.3
	Pct_Family_Income_LT_Poverty	252715.7	252736.5		250516
	Pct_LE_5Yrs_LT_Poverty	252839.5	252848.3		250483.1*
Race	Pct_Asian	252668.4	252569.2	252369.5	250374.8
	Pct_Black	253025.3	253012.9	252123.9	251966.4
	Pct_White	252930.5	252888		252285.1
	Pct_NHOPI		252837.9	253276.4	252846.7
	Pct_Other_Race	252775.6	252727.4	252540.6	252403.1
	Pct_Multi_Race	252788.2	252745.4	251886.2**	
	Pct_Hispanic	252813.9			248350.1*
Housing Cost	Median_Rent	252008.3	251790.4	251043.0**	250455.8*
	Housing_Value	252392.2	252324.9	252139.5	
Occupancy	Pct_Rented	253049.2	252943.3**		251135.1*
	Pct_Vacant	252968.6	252990.6		252368.7
Single Parent	Pct_Single_Parent	253124	253084.3	251927.9**	251596.8*
Home Age	Median_Yr_Built	252361	252361	252263.4	252227.6
	Median_Yr_Occ_Built	252391.1	252391.8	251439.8	250974.6
	Pct_Built_Pre_1940	252030	251996.7		251206.9
	Pct_Built_Pre_1950	252006.6	251977.8	251082.9	250788.7
	Pct_Built_Pre_1960	252250.6	252241.6	251073.7**	250590.1*
	Pct_Built_Pre_1970	252433.4	252431.3	251254.4	250672.3
	Pct_Built_Pre_1980	252481.8		251344.2	250756.3
	Pct_Occ_Built_Pre_1940	252035.3	252001.6	251398	
	Pct_Occ_Built_Pre_1950	252020.3	251992	251106	250805.4
	Pct_Occ_Built_Pre_1960		252267.3	251103.2	250611.8
	Pct_Occ_Built_Pre_1970	252477.8	252477.2		250713.2
	Pct_Occ_Built_Pre_1980	252508.2	252502	251371.5	250782.6
Children	Pct_LE_Six	252767	252725.9	251597.1**	251003.9*
	Num_LE_Six	252881.3	252835.7		

Parameter Category	Variable Name	X Only	Linear Time	Quadratic Time	Categorical Time
Education	Pct_LT_9th_Grade	252734.3	252691.3	251568.3	249792.4*
	Pct_No_HS_Degree	252574.1	252551.1	251197.8	250216.7
	Pct_No_College	252279.4	252216.7	251090.4**	250602.5
	Pct_No_College_Degree		252237.3	251142.9	250691.1
Population	Total_Housing_Units	252895.2	252818.7	251937	243207.7
	Total_Pop	252890.7	252834.7	251907.9**	242397.1*
	Housing_Density	252854.6	252838.1	252786.9	252547.2
Air Lead	air_avg	252880.2	252861.4	252906.8	252015.3
	air_med	252884.8	252867.1	253026.4	251966.2*
	air_p95	252876.9	252863.8	252740.9**	252187.9
	air_avg_p95	252912.5	252878.3		249593.0*
	air_med_p95	252910	252728.4**		252883.7
	air_p95_p95		252885	253156.2	
	air_avg_p99	252841.7	252844	252840.5	252814.5
	air_med_p99	252839.7	252842.6	252819.6	252786.5
	air_p95_p99	252848.7	252852.6	252832.5	
Tri	TRI Compounds air_fug	252867.9	252889.1	252914	252314.2
	TRI Compounds air_tot	252889.4	252901.7	252880.6	252313.3
	TRI Compounds air_stk	252884.9	252894.2	252815.7	252454.4
	TRI Compounds under_inj	252862.5	252885.6		252919.1
	TRI Compounds water_surf	252864.1	252885.1		252884.7
	TRI Lead Only air_fug	252869.4	252894.4	252747.2	252624.6
	TRI Lead Only air_tot	252884.3	252905.2	252653.5**	252138.6
	TRI Lead Only air_stk	252882.2	252900.4	252707.7	252298.7
	TRI Lead Only under_inj	252866.5	252889.9	252907.5	252931.1
	TRI Lead Only water_surf	252866.8	252892.3		252900.6
	TRI Lead Total air_fug	252871.7	252896.1	252798.3	252475
	TRI Lead Total air_tot	252891.5	252909.9	252702.6	252031.8*
	TRI Lead Total air_stk	252890.3	252905.6	252699.8	252203.4
	TRI Lead Total under_inj	252867.1	252889.4	252904.5	252929.7
	TRI Lead Total water_surf	252868.4	252893.9		252897.6
	tri_as1_p95	252887.8	252864.5		252435.7
	tri_as2_p95		252886.8		247577.4
	tri_as3_p95	252922.4	252903.3	251601.7	248106.4
	tri_af1_p95	252923.9	252910.9	252833	252941.9
	tri_af2_p95	252864.3	252862.8	252237.8	248140.5
	tri_af3_p95	252883.1	252872.3	251998.1	248650.8
	tri_at1_p95	252886.9	252855.8		252210.2
	tri_at2_p95	252911.6		251677.1	247510.1*
	tri_at3_p95		252893.2	251553.2	248305.6
	tri_ws1_p95	252904.6	252880.2	253019.4	252996.8
	tri_ws2_p95		252885.1	251407.9**	
	tri_ws3_p95	252892.3	252895.8	251422.9	248164.1
	tri_ui1_p95	252840.1	252840.1	251871.5	251377.4
	tri_ui2_p95	252840.1	252840.1	251871.5	251377.4
	tri_ui3_p95	252840.1	252840.1	251871.5	251377.4
	tri_as1_p99		252856.4		

Parameter Category	Variable Name	X Only	Linear Time	Quadratic Time	Categorical Time
	tri_as2_p99		252864.7	252768.7	252011.1
	tri_as3_p99	252867.3	252867.6	252781.5	
	tri_af1_p99	252855.1		252872.8	
	tri_af2_p99	252866.8	252862.5	252719	251667.2
	tri_af3_p99	252851.4	252852.6	252743.1	251999.2
	tri_at1_p99	252844.8	252850.4	252865.3	252235.8
	tri_at2_p99	252869.5	252869.6		252035.9
	tri_at3_p99	252863.2	252860.2	252780.2	252047.7
	tri_ws1_p99	252864.6	252865.7	252817.4	252839.7
	tri_ws2_p99	252853.8	252857.9	252679.5	252740.5
	tri_ws3_p99	252858.5	252859.9		252756.3
	tri_ui1_p99	252840.1	252840.1	251871.5	251377.4
	tri_ui2_p99	252840.1	252840.1	251871.5	251377.4
	tri_ui3_p99	252842.9	252851.7	252822.1	252824.4
Funding	CDC_cur_lag6	252593.4	252079	251795.6	
	CDC_cur_lag12	252468.6	251799	251576.9**	251723.5
	CDC_cur_lag18	252652.5	252011.3		251869.8
	CDC_cur_lag24	252698.2	252155.7	251983.6	251948.3
	CDC_cur_lag30		252579.9	252298.4	252201.5
	CDC_cur_lag36	252861.4	252659.6		
	HUD_cur_lag6	252848.9	252867.7	252599.3	252883.1
	HUD_cur_lag12	252836	252839.2	252674.2	252823
	HUD_cur_lag18	252857.7	252873.7	252693.1	252856.3
	HUD_cur_lag24	252847.7	252847.4	252785.4	252804.8
	HUD_cur_lag30	252821.7	252849.1	252709.2	
	HUD_cur_lag36	252800.2	252769.4	252762.3	252713.8
	CDC_cum_lag6	252638.8	252230.7	252255.1	
	CDC_cum_lag12		252277.9	252311.3	252258.1
	CDC_cum_lag18			252387.7	252319.1
	CDC_cum_lag24	252817.8	252422.6	252451.6	252377.5
	CDC_cum_lag30	252840.3	252479.6	252498.3	252424.6
	CDC_cum_lag36	252836.1	252501	252515.7	252446.4
	HUD_cum_lag6	252859.8		252650.4	252640.5
	HUD_cum_lag12	252856.9	252607.7	252639.5	252615.3
	HUD_cum_lag18	252856.7		252636	252602.2
	HUD_cum_lag24	252852.4	252583.7	252612.3	252582.4
	HUD_cum_lag30	252853.1	252570.4	252586.8	252559.1
	HUD_cum_lag36		252556.3	252544.6	252520.3
	tot_cur_lag6	252840.8	252857	252532.5	252837.6
	tot_cur_lag12	252848.3	252826.7	252615.3	252765.8
	tot_cur_lag18	252863.6	252859.5	252627.1	252788.9
	tot_cur_lag24	252855.8	252803.6	252686.2	252715.8
	tot_cur_lag30		252838.7	252608.6	252773.5
	tot_cur_lag36	252796.7	252732.3	252677.3	252653.3
	tot_cum_lag6	252863.1	252551.2	252577.9	252560.4
	tot_cum_lag12	252861	252531.9	252565.2	
	tot_cum_lag18	252858.7	252530.3	252565.1	252537.4

Parameter Category	Variable Name	X Only	Linear Time	Quadratic Time	Categorical Time
Funding	tot_cum_lag24		252518.4	252551.1	252528.4
	tot_cum_lag30	252854.9	252510.3	252535.2	252515.6
	tot_cum_lag36	252855.1		252503	252491.2
	HUD_cur	252824.5	252845.5		252876.3
	HUD_cum	252864.5	252670.1	252673.7	252679.3
	CDC_cur	252679.7	252170.2	251830.5	251799.9
	CDC_cum	252586.4	252213.5	252226.6	252202.3
	Current – HUD+CDC	252812.9	252831.6	252573.5	252837
	Cumulative – HUD+CDC	252861.2	252590.5	252602.3	252591.5
Screening	screen_penetration	243704.7	243043.2**		243416.3

** Variable factor(s) showed best fit when adjusted for degrees of freedom and were thus chosen to represent parameter category in multivariate analysis.

* Variable factors showed best fit; however, were not included in multivariate analysis because the time categorical variable had less than ideal prediction properties.

Table 4-3. Summary of Exploratory Analysis Fit as shown by -2 Log Likelihoods for Pr(PbB >= 15 µg/dL) Models

Parameter Category	Variable Name	X Only	Linear Time	Quadratic Time	Categorical Time
Income	Median_Family_Income	292772.5	292744.3	292659.9	292689.4
	Median_HH_Income	292732.9		292590.8	292579.8
	Median_Per_Capita_Income		292812.2	292732.2	292766.8
	Pct_HH_No_Earnings	292772.9	292756		292428.7
	Pct_HH_No_Wage	292673.9	292612.3	292449.8**	292310.0*
	Pct_HH_Public_Assist	293171.5	293226	292889.5	292588
	Pct_LT_Poverty	293148.6	293226.2	292933.3	292609.9
	Pct_Family_Income_LT_Poverty	293112.9		292898.4	292547.8
	Pct_LE_5Yrs_LT_Poverty	293290.6	293338.3	292997.4	292700.3
Race	Pct_Asian		292966.2**	293176.7	292343.9
	Pct_Black	293610.6	293556.2	293324.7	293411.1
	Pct_White	293448.5	293343.9	293334.4	293311.7
	Pct_NHOPI	293217.4	293230		293216.3
	Pct_Other_Race	293165.6	293130.1	293088.7	293085.7
	Pct_Multi_Race		293181.3	293035.3	292946.7
	Pct_Hispanic	293237.8		293524.2	292091.3*
Housing Cost	Median_Rent	292358.3	292076.9	291971.3**	292091.4
	Housing_Value	292727.1	292691.2	292732.6	292573.7
Occupancy	Pct_Rented	293905.5		293533.3**	
	Pct_Vacant	293703.2	293721.5	293765	293491.3*
Single Parent	Pct_Single_Parent	293863	293781.9**		
Home Age	Median_Yr_Built	292980.6	292980.6	293008.9	292998.6
	Median_Yr_Occ_Built	292999.8	293001.5	292856.3	292749.7
	Pct_Built_Pre_1940	292381.9	292335.9	292195.2	292161
	Pct_Built_Pre_1950	292404.2	292364.6		292078.8
	Pct_Built_Pre_1960		292815.6	292525.3	292400.7

Parameter Category	Variable Name	X Only	Linear Time	Quadratic Time	Categorical Time
Home Age	Pct_Built_Pre_1970	293100	293103.4	292826.6	292669
	Pct_Built_Pre_1980	293086.5	293089.8	292851.2	292684.6
	Pct_Occ_Built_Pre_1940	292356.8	292313.7	292177	292138.7
	Pct_Occ_Built_Pre_1950		292357.3	292139.4**	292071.1*
	Pct_Occ_Built_Pre_1960	292842.2		292546.1	292418.1
	Pct_Occ_Built_Pre_1970		293160.9	292884	292725.2
	Pct_Occ_Built_Pre_1980	293111.9			292711
Children	Pct_LE_Six	293214.5		292911.5**	292859.9
	Num_LE_Six	293389	293281.3	293584.6	290311.1*
Education	Pct_LT_9th_Grade	293199.4	293131.7	293000.2	292491.7
	Pct_No_HS_Degree	292960	292951	292701.2	292376
	Pct_No_College	292557.4	292434.4	292260.2**	292152.4*
	Pct_No_College_Degree	292613.7	292501.1	292328.1	292254.9
Population	Total_Housing_Units	293444.1	293298.7	293466.2	
	Total_Pop	293417.4	293297.9	293561.7	290521.5*
	Housing_Density	293273.5		293267.2**	293232.3
Air Lead	air_avg	293392.7	293361.7	293366.1	
	air_med	293410.8	293382.5	293418.4	292995.5*
	air_p95	293378.4	293353.2	293316.8**	293112.1
	air_avg_p95	293522.6	293462.9	293467.7	292547.1*
	air_med_p95	293524.1	293509.8**	292580.3*	
	air_p95_p95	293421.4		293566.7	292731.7
	air_med_p99	293289.2**	293293.8	293304.7	293262.7
	air_avg_p99			293312.1	293284.2
	air_med_p99	293289.2**	293293.8	293304.7	293262.7
	air_p95_p99	293307	293306.7		
Tri	TRI Compounds air_fug	293312	293329.5	293358.7	
	TRI Compounds air_tot		293391.4	293416.4	
	TRI Compounds air_stk	293379.4	293373.4	293382.6	293324
	TRI Compounds under_inj	293280.9**		293325.9	
	TRI Compounds water_surf	293302.9	293322.7		293339.6
	TRI Lead Only air_fug	293303.9	293327.8	293305.4	293302.9
	TRI Lead Only air_tot			293317.6	293193.9*
	TRI Lead Only air_stk	293339.1	293351.3	293337.7	293240.5
	TRI Lead Only under_inj		293317.4		293358.1
	TRI Lead Only water_surf	293295.9	293321.1		293362.7
	TRI Lead Total air_fug		293334.7	293331	
	TRI Lead Total air_tot	293361.2	293369.8	293351.5	293196.7
	TRI Lead Total air_stk	293364.2	293370.5	293355.6	293251.3
	TRI Lead Total under_inj		293315.8	293344.4	293356.7
	TRI Lead Total water_surf	293300.6	293326	293350.3	293369
	tri_as1_p95	293431.9	293374.3	293326.2	293354.1
	tri_as2_p95	293447.9			
	tri_as3_p95		293468.1	293184.4	292325.9
	tri_af1_p95	293511.4	293486.3	293488.2	293535.3
	tri_af2_p95	293346.9	293341.6		292046.2*
	tri_af3_p95	293406.8	293388.9	293217	292195.8

Parameter Category	Variable Name	X Only	Linear Time	Quadratic Time	Categorical Time
	tri_at1_p95	293445.5	293386	293309.2	
	tri_at2_p95	293490.2	293434.4	293188.8	
	tri_at3_p95	293524.3	293473.4	293159	292314
	tri_ws1_p95		293384.9	293682.5	
	tri_ws2_p95	293397.5	293388.1	293081.1**	292172.3
	tri_ws3_p95	293449.3	293433.9	293122.9	292238.7
	tri_ui1_p95	293260.7	293260.7	293101.5	
	tri_ui2_p95	293260.7	293260.7	293101.5	
	tri_ui3_p95	293260.7	293260.7	293101.5	
	tri_as1_p99	293300.7	293299.9		293278.9
	tri_as2_p99		293334.1	293325.5	
	tri_as3_p99	293352.7	293342.5	293339.9	293153
	tri_af1_p99	293317.7		293328.2	293169
	tri_af2_p99	293341.5	293329	293303.8	293016
	tri_af3_p99	293319.2	293311.8	293303.8	293101.5
	tri_at1_p99	293295.1	293296.7		293129.5
	tri_at2_p99	293359.2	293347.3		293169.3
	tri_at3_p99	293332.6	293315.4	293315.4	293153.9
	tri_ws1_p99		293338.4	293329.3	293345.2
	tri_ws2_p99	293304.9		293222.9	293282
	tri_ws3_p99	293308.9			
	tri_ui1_p99	293260.7	293260.7	293101.5	
	tri_ui2_p99	293260.7	293260.7	293101.5	
	tri_ui3_p99	293269.5		293293.4	293282.9
Funding	CDC_cur_lag6	293205.3	292977.5	292865.8	292890.8
	CDC_cur_lag12	293218.4	292865.8	292774.0**	292863.8
	CDC_cur_lag18	293255.1	292878.1	292799.5	292857.9
	CDC_cur_lag24	293274.7	292961.8	292890.1	292900.2
	CDC_cur_lag30		293176.8	293015.4	293003.6
	CDC_cur_lag36	293290.1	293206.1	293076.1	293048.7
	HUD_cur_lag6		293284.6	293128.1	293287.2
	HUD_cur_lag12	293253.2	293249.9	293149.3	293242.7
	HUD_cur_lag18	293258.5	293268.9	293117.4	
	HUD_cur_lag24	293253.7	293257.1	293203.1	293251.9
	HUD_cur_lag30	293238.4	293251.9	293106.6	293268.9
	HUD_cur_lag36	293208.4	293194.5	293190.6	293186.5
	CDC_cum_lag6	293422.1	293214.4	293231.5	293238
	CDC_cum_lag12	293393.4	293207.1		
	CDC_cum_lag18	293368.7	293212.7		293218.9
	CDC_cum_lag24		293218.2		293214.6
	CDC_cum_lag30		293218	293228.4	293207.6
	CDC_cum_lag36	293310.8	293205.4	293218.2	293193.2
	HUD_cum_lag6	293239.4		293145.3	
	HUD_cum_lag12	293234.4	293120.8	293146.5	293150.1
	HUD_cum_lag18	293243.7	293134.7	293163.9	293150.3
	HUD_cum_lag24	293257.1	293145.3		293148.9
	HUD_cum_lag30	293267.1		293164.4	293133.9

Parameter Category	Variable Name	X Only	Linear Time	Quadratic Time	Categorical Time
Funding	HUD_cum_lag36	293278.3	293149.3	293146.1	293110.8
	tot_cur_lag6	293267.9	293280.2	293102.1	293268.2
	tot_cur_lag12	293260.9	293241.6	293127.1	293218.6
	tot_cur_lag18	293262.4	293261.9		293237.1
	tot_cur_lag24	293257.8	293239.6	293160.8	293215.4
	tot_cur_lag30	293245.2	293258.1	293049.9	293252.3
	tot_cur_lag36	293206.8		293151.9	293164.7
	tot_cum_lag6	293257.7	293107.8	293124.1	
	tot_cum_lag12		293100	293123.9	293133.9
	tot_cum_lag18	293257.7	293114.6	293142.1	293138.5
	tot_cum_lag24	293267.9		293154.5	293141.8
	tot_cum_lag30	293275.4	293128.8	293153.2	
	tot_cum_lag36	293284.2	293130.5	293142.3	293117.5
	HUD_cur	293269.1	293287.6	293172.6	293293
	HUD_cum		293156.9	293156.8	293193.2
	CDC_cur	293240.8		292881.8	292906.9
	CDC_cum	293450.7	293231.1	293246.3	293256.8
	Current – HUD+CDC	293267.6	293284.9		
	Cumulative – HUD+CDC		293135.6		293170.4
Screening	screen_penetration			286744.8**	

** Variable factor(s) showed best fit when adjusted for degrees of freedom and were thus chosen to represent parameter category in multivariate analysis.

* Variable factors showed best fit; however, were not included in multivariate analysis because the time categorical variable had less than ideal prediction properties.

Table 4-4. Summary of Exploratory Analysis Fit as shown by -2 Log Likelihoods for Pr(PbB >= 25 µg/dL) Models

Parameter Category	Variable Name	X Only	Linear Time	Quadratic Time	Categorical Time
Income	Median_Family_Income	364225.4	364192.5	364302.8	364426
	Median_HH_Income	364098.6	364048	364139.1	364252.5
	Median_Per_Capita_Income	364504.4	364426.7	364532.3	364681.7
	Pct_HH_No_Earnings	363967.7	363893.9**	364071.9	364002.6
	Pct_HH_No_Wage	363920.7		363960.8	363949.4
	Pct_HH_Public_Assist	364638.9	364705.3		364637.4
	Pct_LT_Poverty	364611.6	364741.3	364847.1	364616.8
	Pct_Family_Income_LT_Poverty	364541.7	364696	364777.3	364527.4
	Pct_LE_5Yrs_LT_Poverty	364863.6	364929.7		
Race	Pct_Asian	364527.2	364399.3**		364394.5*
	Pct_Black	365490.4	365511.8		
	Pct_White	365248.9	365038.9		365365.2
	Pct_NHOPI		364794.8	364856	364770.1
	Pct_Other_Race	364698	364678.7		
	Pct_Multi_Race	364801.9	364772.8	364831.2	364838.6
	Pct_Hispanic	364789.2		364703.1	
Housing	Median_Rent	363769.6		363524.9**	363876.3

Parameter Category	Variable Name	X Only	Linear Time	Quadratic Time	Categorical Time
Cost	Housing_Value	364228.4	364217.5	364266.9	364326.7
Occupancy	Pct_Rented		366003.6**	366069.6	366361.5
	Pct_Vacant	366381.1	366351.7	366647.4	366332.7
Single Parent	Pct_Single_Parent	365919.4	365878.4**	365993.3	366022.9
Home Age	Median_Yr_Built	364858.6	364893.4	364874.1	364997.8
	Median_Yr_Occ_Built	364890.1	364927.1	365077.9	365041.9
	Pct_Built_Pre_1940	363573.4	363524.8	363636.7	363617.9
	Pct_Built_Pre_1950	363716.9	363694.8	363795.8	363754.9
	Pct_Built_Pre_1960	364677.4	364715.4		364720.3
	Pct_Built_Pre_1970	365255	365322.4	365404.2	365310.9
	Pct_Built_Pre_1980	365157.7	365231.3	365333.9	365231.9
	Pct_Occ_Built_Pre_1940		363502.8**	363615.6	363591.1
	Pct_Occ_Built_Pre_1950	363710.5	363694.8	363795.4	
	Pct_Occ_Built_Pre_1960	364710.4			364754.6
	Pct_Occ_Built_Pre_1970	365363.2	365429.2	365509.7	365420.4
	Pct_Occ_Built_Pre_1980	365200.7		365373.4	365276
Children	Pct_LE_Six	364874.1		364787.6**	365003.5
	Num_LE_Six	365297.1			
Education	Pct_LT_9th_Grade	364694.7	364636.7	364676.4	364634.3
	Pct_No_HS_Degree	364365.8		364490.3	364359.9
	Pct_No_College	363888.4	363710.1**	363897.3	363941.3
	Pct_No_College_Degree	364172.5	363983.5	364160.7	
Population	Total_Housing_Units	365469.1			365116.6
	Total_Pop	365377.3	365172.2		
	Housing_Density	364890.0**	364908.1	364921.6	364925.2
	air_avg	365109.3	365098.9	365098.8	365022.8*
	air_med	365149.3		365142.8	
	air_p95	365066.7	365050.9**	365053.1	365028.1
	air_avg_p95	365506.4			365342.9
	air_med_p95	365605.6**		365444.5*	
	air_p95_p95	365231.4	365186.4	365205.3	
	air_med_p99	364910.2**		364934.1	364912.5*
	air_avg_p99	364950.2	364949.7	364959	364957.4
	air_p95_p99	364974.5	364974.2	364984.1	
Tri	TRI Compounds air_fug	365029.5	365019.8		365019.4
	TRI Compounds air_tot	365254.2	365215.5		365259.6
	TRI Compounds air_stk	365170.3	365147.7	365176.2	365183.3
	TRI Compounds under_inj	364946	364964.8	365107.4	
	TRI Compounds water_surf	364940.3	364958	364980.3	364994
	TRI Lead Only air_fug		364920.3		365037.8
	TRI Lead Only air_tot	364918.4		364960.3	364976.7
	TRI Lead Only air_stk	364950.5	364975.5	364992.6	365001.8
	TRI Lead Only under_inj	364931.6	364947.7	364984.2	364995.3
	TRI Lead Only water_surf	364934.6	364960.9		365008.7
	TRI Lead Total air_fug	364905.3**	364928.5		364977.9
	TRI Lead Total air_tot		364996.6	365013.8	365014.8

Parameter Category	Variable Name	X Only	Linear Time	Quadratic Time	Categorical Time
Tri	TRI Lead Total air_stk	365012.9	365029.9	365049.1	365054.7
	TRI Lead Total under_inj	364926.7	364942.8	364980	364990.4
	TRI Lead Total water_surf	364946.9	364973.4	365004.6	365022.5
	tri_as1_p95	365325.7	365249.4	365249.7	
	tri_as2_p95		365223.1	365193.4	365152.7
	tri_as3_p95	365559.3	365489.7	365459.1	365423.2
	tri_af1_p95		365498	365506.5	365620
	tri_af2_p95	365113.2	365113.5		365005.7
	tri_af3_p95	365283.2	365267	365225.7	365145.6
	tri_at1_p95	365406.5	365320.9	365306.8	365423.7
	tri_at2_p95	365435.5	365358.5	365319.4	365270.2
	tri_ws1_p95	365307.8		365270.9	365266.6
	tri_ws2_p95	365239.4	365230.3	365195	
	tri_ws3_p95	365338.3	365316.1	365285.3	
	tri_ui1_p95	364871.7	364871.7	365011.6	364990
	tri_ui2_p95	364871.7	364871.7	365011.6	364990
	tri_ui3_p95	364871.7	364871.7	365011.6	364990
	tri_as1_p99	364965.4	364961.3	364961.2	364964.3
	tri_as2_p99	365027.6	365021.2	365024.1	364974.7
	tri_as3_p99	365079.1	365059.8	365067.5	365020.9
	tri_af1_p99	365068	365051.1		
	tri_af2_p99	365053.7	365042.3	365038.6	364985.1
	tri_af3_p99	365022.6	365006.5	365010.2	364974
	tri_at1_p99	365048.8	365036.8	365050	364994.8
	tri_at2_p99	365034.9	365011.9	365019.8	365005.5
	tri_at3_p99	364999.7	364972.1	364983	
	tri_ws1_p99	364994.9	364977.9	364985.8	364990.4
	tri_ws2_p99	364991.4	364981.6	364976.5	
	tri_ws3_p99			364951.1	364993.6
	tri_ui1_p99	364871.7	364871.7	365011.6	364990
	tri_ui2_p99	364871.7	364871.7	365011.6	364990
	tri_ui3_p99		364867.2	364864.5**	364857.2*
Funding	CDC_cur_lag6	365029.5	364956.9	364956.9	364945.6
	CDC_cur_lag12	365008.6	364858.7	364871.8	364857.2
	CDC_cur_lag18	364967	364840.3	364881.1	364852
	CDC_cur_lag24	364989	364897	364960.1	
	CDC_cur_lag30	364937.9	364919.4	364963.5	364906.9
	CDC_cur_lag36	364893.3	364894.9	364991.3	364904.2
	HUD_cur_lag6	364916	364927.7	364808.6	364920
	HUD_cur_lag12		364889.1	364804.6	364900.9
	HUD_cur_lag18		364869.8	364768.1**	
	HUD_cur_lag24	364850.4	364863.1	364824	364891.8
	HUD_cur_lag30	364824.3	364841.4	364789.4	
	HUD_cur_lag36	364796.1	364782.8	364802.3	364768.8*
	CDC_cum_lag6	365040.7	365009.7	365084.8	365040
	CDC_cum_lag12	365006.7	365000.8	365059.1	365027.3
	CDC_cum_lag18	364981.8		365047.7	365021.5

Parameter Category	Variable Name	X Only	Linear Time	Quadratic Time	Categorical Time
Funding	CDC_cum_lag24	364957	364993.3	365033.1	365011.9
	CDC_cum_lag30	364934.4	364984.2	365017.3	
	CDC_cum_lag36	364920.3	364978	365004.7	
	HUD_cum_lag6	364860.6	364793.5	364808.3	364816.4
	HUD_cum_lag12	364850.7	364788	364807.5	
	HUD_cum_lag18	364851.8	364791.1	364810.2	364794.2
	HUD_cum_lag24	364867.1		364816.1	364790.5
	HUD_cum_lag30	364877.2	364807.6	364803.6	364772.6
	HUD_cum_lag36	364892.6	364826.6	364809.5	364775.8
	tot_cur_lag6	364930.2	364941.5	364842.7	364926.8
	tot_cur_lag12			364840.7	364900
	tot_cur_lag18	364862.5	364872.6	364795.3	364887.3
	tot_cur_lag24	364862	364873	364853.4	
	tot_cur_lag30	364830.9	364848.9	364794.7	364868.3
	tot_cur_lag36	364793.7	364783.3	364816.2	364776.2
	tot_cum_lag6	364890.4	364819.5	364839.4	364845.2
	tot_cum_lag12	364876.1	364814	364833.9	364836.1
	tot_cum_lag18	364873.5	364819	364837.1	364830.7
	tot_cum_lag24	364884.2	364833.1	364846	364832.5
	tot_cum_lag30	364889.6		364839.5	364821.7
	tot_cum_lag36	364902.3		364849.2	364829.3
	HUD_cur	364930.7	364948	364886.1	364936.8
	HUD_cum	364883.2	364815.7		364835.9
	CDC_cur	365027.8	364962.7	364969.9	364980.6
	CDC_cum	365081.4	365028.2	365131.3	365060.7
	Current – HUD+CDC	364946.5	364966.1	364919.2	
	Cumulative – HUD+CDC		364843	364865.1	364864.6
Screening	screen_penetration	360376	360256.7**		360321.8

** Variable factor(s) showed best fit when adjusted for degrees of freedom and were thus chosen to represent parameter category in multivariate analysis.

* Variable factors showed best fit; however, were not included in multivariate analysis because the time categorical variable had less than ideal prediction properties.

4.2 Relationship between Local Blood-Lead Data and Explanatory Variables

Many of the variables investigated for the National (Low Resolution) model also were explored for the local modeling using Massachusetts data. All of the census data were used in both models, although at the census-tract level rather than at the county level. The various demographic, environmental, and programmatic variables were explored using the same techniques as the national data, which were described in Section 2.2. Detailed figures and tables containing exploratory results are included in Appendix B. A detailed discussion of the results seen in Appendix B is contained in Appendix E. Table 4-5 presents the log-likelihood statistics that resulted from the bivariate modeling. Variables presenting the best model fit within each variable category are highlighted in yellow.

Table 4-5. Summary of Log-likelihood Ratios from each Model Fit to all Potential Explanatory Variables, Massachusetts Data

Variable Category	Variable	Model 1	Model 2	Model 3	Model 4	Model 5
Income	Median Family Income ($)	51727.2	48154.7	86501.8	139627.0	178071.4
	Median Household Income ($)	**51689.2**	**48114.6**	**86375.5**	**139433.1**	177919.9
	Median Per Capita Income ($)	51917.2	48345.5	86812.3	140036.2	178467.2
	Percent No Household Earnings	52020.1	48445.8	86853.4	139531.9	177346.7
	Percent No Household Wage	52039.6	48467.3	86858.3	139459.8	**177184.8**
	Percent Household on Public Assistance	51963.8	48385.7	86854.4	139836.6	178218.1
	Percent Below Poverty Line	51974.1	48389.9	86778.4	139558.4	177838.6
	Percent Family Income Below Poverty Line	51991.3	48412.1	86847.5	139686.4	177952.4
	Percent Less than 5 Years in Poverty	52025.2	48446.0	86924.8	.	178034.4
Race	Percent Amer. Indian and Alaskan Native Alone	52259.2	48692.5	87120.6	139739.5	177392.1
	Percent Asian Alone	52264.2	48699.0	87139.8	.	177384.8
	Percent Black Alone	52170.1	48599.1	87051.0	139781.5	177691.4
	Percent White Alone	52123.5	48547.4	86960.9	139784.2	178066.6
	Percent Native Hawaiian and Other Pacific Islander Alone	52273.4	48706.6	87135.6	139740.3	**177378.7**
	Percent Other Race Alone	52202.6	48633.4	87055.1	139688.4	177606.3
	Percent Multiple Races	**52042.2**	**48470.6**	**86889.1**	**139661.3**	178104.6
	Percent Hispanic	52181.3	48609.2	87019.6	139808.6	177754.4
Housing Costs	Median Rent ($)	**52004.0**	**48440.1**	**86942.8**	**139972.2**	**177952.2**
	Housing Value ($)	52094.5	48520.9	87032.6	140053.9	178202.0
Occupancy	Percent Rented	**52006.7**	**48421.1**	**86681.5**	**139426.7**	177818.5
	Percent Vacant	52186.5	48617.5	87003.7	139451.9	**177021.2**
Single Parent	Percent Single Parent	**51747.7**	**48155.6**	**86542.8**	**139654.1**	**178338.5**
Home Age	Year Built	51949.7	48360.5	86621.1	139739.2	178275.2
	Year Occupied Unit Built	51966.8	48377.9	86641.4	139748.6	178258.9
	Percent Built Before 1940	51923.9	48335.3	86505.1	139476.8	178110.9
	Percent Built Before 1950	**51897.9**	**48308.1**	**86483.1**	139547.1	178188.5
	Percent Built Before 1960	51959.8	48374.0	86653.5	139697.8	178061.5
	Percent Built Before 1970	52052.3	48469.4	86806.3	139775.8	177977.1
	Percent Built Before 1980	52073.4	48490.9	86850.4	139771.3	177896.0
	Percent Occupied Units Built Before 1940	51931.0	48342.4	86512.2	**139475.5**	178078.2
	Percent Occupied Units Built Before 1950	51910.8	48321.2	86497.8	139549.3	178149.3
	Percent Occupied Units Built Before 1960	51975.9	48389.9	86675.0	139713.0	178044.5

Variable Category	Variable	Model 1	Model 2	Model 3	Model 4	Model 5
	Percent Occupied Units Built Before 1970	52070.4	48486.2	86829.5	139799.3	177979.1
	Percent Occupied Units Built Before 1980	52083.6	48501.0	86860.0	139776.7	**177867.1**
Children	Percent Less than 6 Years of Age	52280.4	48713.7	87126.9	.	**177615.0**
	Number Less than 6 Years of Age	**52243.2**	**48678.2**	**86913.2**	**138995.8**	.
Education	Percent Less than 9th Grade	52006.5	48435.7	86828.7	**139726.7**	**177859.2**
	Percent without High School Degree	51852.7	48279.9	**86709.8**	139857.7	178346.7
	Percent without any College	**51787.2**	**48218.2**	86729.0	140104.0	178689.0
	Percent without College Degree	51822.1	48251.2	86759.6	140084.9	178625.6
Population	Total Housing Units	52286.9	48720.8	87090.7	139461.7	**176701.5**
	Total Population	**52223.0**	**48660.6**	**86909.7**	**138948.0**	.
	Housing Density	52272.6	48697.2	87068.2	139579.8	177260.2
Air	Air Dispersion (ASPEN) Model	52275.1	48708.3	87136.9	139740.1	177377.1
	Air Exposure (HAPEM5) Model	52273.1	48706.3	87134.9	139737.8	177375.0
	Air Hazard Quotient (HQ)	**52272.3**	**48705.5**	**87134.0**	**139737.0**	**177374.2**
HUD Funding	Current HUD Funding ($ per Child)	52290.9	48722.4	87140.4	139755.6	177426.7
	Cumulative HUD Funding ($ per Child)	52287.3	48723.3	87163.4	139804.9	177444.3
	Current State Funding ($ per Child)	**52162.4**	**48582.7**	**87014.3**	139706.6	177503.7
	Cumulative State Funding ($ per Child)	52200.7	48617.7	87044.0	139740.2	177459.9
	Current CDC Funding ($ per Child)	52288.3	48720.8	87151.7	139760.5	177456.0
	Cumulative CDC Funding ($ per Child)	52292.6	48725.0	87136.0	139706.1	177330.9
	Current Total Funding ($ per Child)	52282.8	48711.9	.	.	177377.3
	Cumulative Total Funding ($ per Child)	52292.2	48721.9	87145.9	.	.
	Current HUD Funding ($ per Census Tract)	52302.9	48737.1	87167.5	139723.5	177172.3
	Cumulative HUD Funding ($ per Census Tract)	52306.8	48740.4	87097.6	.	177127.9
	Current State Funding ($ per Census Tract)	52232.2	48659.5	87178.1	140076.2	178005.1
	Cumulative State Funding ($ per Census Tract)	52210.6	48638.3	87222.7	140117.0	178018.7
	Current CDC Funding ($ per Census Tract)	52297.5	48729.5	87162.7	.	177462.6
	Cumulative CDC Funding ($ per Census Tract)	52303.7	48737.2	87121.6	**139570.8**	**176972.0**
	Current Total Funding ($ per Census Tract)	52303.1	48735.6	87173.9	139790.6	177370.8
	Cumulative Total Funding ($ per Census Tract)	52300.1	48732.3	87179.7	139780.5	177447.9

Variable Category	Variable	Model 1	Model 2	Model 3	Model 4	Model 5
TRI	TRI Compounds (Total Air)	52295.5	48728.8	87153.5	139761.4	177395.8
	TRI Compounds (Fugitive Air)	52287.7	48720.7	87146.1	139750.0	177399.8
	TRI Compounds (Stacks)	52295.6	48728.9	87154.3	139761.8	177395.5
	TRI Compounds (Water Surface)	52285.5	48718.8	87147.0	139753.0	177391.9
	TRI Lead Only (Total Air)	52291.0	48724.0	87153.1	139762.9	177405.1
	TRI Lead Only (Fugitive Air)	52289.6	48722.6	87151.4	139759.1	177404.2
	TRI Lead Only (Stacks)	52291.7	48725.0	87153.7	139757.2	177390.5
	TRI Lead Only (Water Surface)	**52274.4**	**48708.5**	**87138.1**	**139746.5**	**177386.0**
	TRI Total Lead (Total Air)	52295.8	48729.0	87155.2	139759.7	177392.4
	TRI Total Lead (Fugitive Air)	52291.5	48724.6	87154.0	139761.5	177401.9
	TRI Total Lead (Stacks)	52295.6	48728.9	87154.3	139760.6	177393.4
	TRI Total Lead (Water Surface)	52285.4	48718.6	87147.4	139753.3	177393.0
Housing Inspection	P1: Proportion of Housing Units Passing MA Standard of Care: Naïve Method 1	52214.0	48643.0	87070.2	139774.6	177816.1
	F1: Proportion of Housing Units Failing MA Standard of Care: Naïve Method 1	52108.5	48548.2	86916.6	139803.2	178520.9
	N1: Proportion of Housing Units Assessed: Naïve Method 1	52131.7	48554.8	86963.0	139849.2	178224.8
	P2: Proportion of Housing Units Passing MA Standard of Care: Naïve Method 2	52240.8	48671.5	87101.3	139767.8	177800.8
	F2: Proportion of Housing Units Failing MA Standard of Care: Naïve Method 2	52199.5	48645.5	87046.4	139861.5	178375.2
	N2: Proportion of Housing Units Assessed: Naïve Method 2	52208.4	48641.6	87066.2	139826.1	178110.9
	P3: Proportion of Housing Units Passing MA Standard of Care: Naïve Method 3	52240.8	48671.5	87101.3	**139767.8**	**177800.8**
	F3: Proportion of Housing Units Failing MA Standard of Care: Naïve Method 3	52108.5	48548.2	**86916.6**	139803.2	178520.9
	N3: Proportion of Housing Units Assessed: Naïve Method 3	52160.8	48586.2	86996.1	139865.4	178257.4
	P4: Proportion of Housing Units Passing MA Standard of Care: MDPH Method	52240.5	48671.1	87098.8	139769.1	177809.5
	F4: Proportion of Housing Units Failing MA Standard of Care: MDPH Method	**52106.5**	**48545.8**	86919.8	139808.9	178518.4
	N4: Proportion of Housing Units Assessed: MDPH Method	52160.3	48585.5	86994.8	.	178259.8

5.0 STATISTICAL MODELING RESULTS

As described in Section 2.3, for each statistical model within each of the two broad model types (Low and High Resolution) the variables that led to the best model fits were initially included in a multivariate statistical model and assessed jointly to determine which variables were predictive of children's blood-lead levels. If higher order interactions with time were not significant within the multivariate model and did not negatively impact the fit of the model upon removal, they were subsequently removed. As results of each model run were reviewed, some variables were dropped from the model if they were not significant predictors of the outcome variable and were not improving the fit of the model by being included. Thus, each model was run and results were assessed multiple times until a final model was reached. The sections below present the final model results for the national risk models (Section 5.1) and the local risk models for Massachusetts (Section 5.2). Maps of the predicted results are included in Section 6 and in Appendix G.

5.1 Low-Resolution Modeling Results

Table 5-1 presents the full set of variables included in the final multivariate models for Models 1 through 4. Across all four models, the time and space variables were important predictors of the various outcomes. The same three variables related to time and space were included in all four models:

- EPA region
- the interaction between EPA region and a continuous measure of time (in years, centered at the year 2000)
- the interaction of EPA region and quarter of the year with the 3rd quarter (July-September) associated with the highest predicted lead levels.

Notes on the other variable types explored and a summarization of the set of variables included in the final models are presented below.

- *Income* – Percent of Units with No Household Wages was included in Models 1 to 3 with all interactions with time included. Percent of Units with No Household Earnings was included in all Model 4 although the interaction with time squared was dropped.
- *Race* – Percent Black was included in Model 1, Percent Multiple Races in Model 2, and Percent Asian in Models 3 and 4, although the interaction with time squared was dropped in Model 3 and both interactions were dropped in Model 4. The best-fitting race variables were included in Models 1 through 5.
- *Housing Cost* – Median Rent was only included in all models and all interaction terms appeared to be strong predictors.
- *Occupancy* – Percent Vacant was included in Model 1 and Percent Rented in the other three. The interaction with time squared was dropped in Models 2 and 4.
- *Single Parent Status* – The percent of single parent households was included in all models with the interaction with time squared was dropped in Models 3 and 4.
- *Housing Age* – Percent Built Pre-1960 was included in Models 1 and 2, Percent Built Pre-1950 in Model 3, and Percent Built Pre-1940 in Model 4.

- *Children's Age* – The percent of children less than six years old was included in all models, although the p-values for each term in Model 2 are high.
- *Education Level* – Percent Without a College Degree was included in all models, although the interaction with time squared was dropped in Model 3 and both interaction terms were dropped in Model 4.
- *Population* – Total Housing Units was included in Model 1 without either interaction with time. Total Population was included in Model 2 with both interactions. Housing Density was included in Models 3 and 4, although the interaction terms were dropped from Model 4.
- *Air Lead* – Median Air Lead, 99th Percentile was included in Models 1, 2, and 4, although the interaction terms were dropped from Models 3 and 4. Air Lead 9th Percentile was included in Model 2.
- *TRI* –TRI Lead Total Air, 95th Percentile was included in Model 1 with all interaction terms. TRI Lead Water Surface 95th Percentile was included in Models 2 and 3 with both interaction terms. TRI Lead Underwater Injection 95th Percentile was included in Model 4 but the interaction with time squared was dropped.
- *Drinking Water* – The two Mean Water Lead Concentration variables were included in each model, although the interaction with time squared was dropped in Model 3 and both interactions were dropped in Model 4.
- *Funding* – Total Cumulative Funding 36-month Time Lag was included in Model 1 with both interactions with time. Current CDC Funding 12-month Time Lag was included in Models 2 and 3 with both interaction terms. Current HUD Funding 12-month Time Lag was included in Model 4 with all terms being significant.
- *Screening* – Screening penetration was included in each model, although the interaction with time squared was dropped in Models 2 and 4.

Thus, in Model 4 for probability of blood-lead level ≥ 25, most of the interactions with time squared were dropped from the model and a number of the interactions with time were dropped as well. Note that when the interaction with time and/or time squared were significant or improved the model, the lower order terms were kept in the model even if a particular term had a p-value above 0.05.

Tables 5-2 through 5-7 present the parameter estimates from each of the four multivariate national models. The standard error and p-value associated with each predictor also is included. Estimates also are presented for the three variance components that were included in the national models – $\sigma^2_{\delta_0}$, $\sigma^2_{\delta_0,\delta_1}$, and $\sigma^2_{\delta_1}$ (related to the random intercept (δ_{0i}) and slope (δ_{1i}) terms.

Following each table are two figures that provide information on the fit of the final models. The first, a histogram of the residuals from the final model fit, helps determine whether or not it is reasonable to assume that the random errors in a statistical process can be assumed to be drawn from a normal distribution. Figures 5-1, 5-3, 5-5, and 5-7 contain the residual histograms of the observed-predicted probabilities from each of the four logistic regression models. Please note that for these four histograms – the model was actually applied on the logit scale. However, because the logit is undefined for observed proportions at zero and one, the histograms were applied to the original scale of measure.

The second set of figures plot the observed values versus the predicted values for each model. If the multivariate model fitted is appropriate, predicted values obtained from regressing the observed values on the multivariate model's predicted values when plotted against observed values, one would expect all the points to be very close to the 45° line. Figures 5-2, 5-4, 5-6, and 5-8 contain these comparison plots for each of the four national models, respectively. The plots were conducted on both the observed probability and logit probability scales, with observed data points at zero and one censored in the logit scale plots.

In general, these plots suggest that the models are performing well. A weighted regression line (blue line) fit to the observed versus predicted plots shows a very high R^2 value in most of the models that mirrors the 45° line (shown in red) for the majority of the data. One trend observed in these plots that is important to consider is that the Broad-Based National Models tend to under-predict for county/quarter combinations with higher proportions that exceed the 5, 10, 15 and 25 μg/dL threshold values. Further exploration may be necessary to determine whether these higher values represent county/quarter combinations with fairly sparse data (i.e., few observations) – which might explain why they would have been less influential because the models are influenced by the number of observations associated with each observed value. For the higher blood-lead threshold categories, the model appears to over-predict the lower observed proportions – suggesting the possibility of a regression to the mean effect.

Table 5-1. Summary of Variables Included in Final National Multivariate Model

Variable Type	Model 1	Model 2	Model 3	Model 4
Area and Time	Region	Region	Region	Region
	Time*Region	Time*Region	Time*Region	Time*Region
	Region*Quarter	Region*Quarter	Region*Quarter	Region*Quarter
Income	Percent of Units No HH Wages	Percent of Units No HH Wages	Percent of Units No HH Wages	Percent of Units No HH Earnings
Race	Percent Black	Percent Multiple Races	Percent Asian	Percent Asian
Housing Cost	Median Rent	Median Rent	Median Rent	Median Rent
Occupancy	Percent Vacant	Percent Rented	Percent Rented	Percent Rented
Family Structure	Percent Single Parent	Percent Single Parent	Percent Single Parent	Percent Single Parent
Housing Age	Percent Built Pre-1960	Percent Built Pre-1960	Percent Occupied Built Pre-1950	Percent Occupied Built Pre-1940
Children's Age	Percent < Six Years Old	Percent < Six Years Old	Percent < Six Years Old	Percent < Six Years Old
Education	Percent without College Degree	Percent No College	Percent No College	Percent No College
Population	Total Housing Units	Total Population	Housing Density	Housing Density
Air Lead	Median Air Lead, 99th percentile	Air Lead, 95th percentile	Median Air 99th percentile	Median Air 99th percentile
TRI	TRI Lead Total Air, 95th percentile	TRI Lead Water Surface 95th, percentile	TRI Lead Water Surface 95th, percentile	TRI Lead UI 95th, percentile
Drinking Water Lead	Mean Water Lead (water=1)	Mean Water Lead (water=1)	Mean Water Lead (water=1)	Mean Water Lead (water=1)
	Mean Water Lead (water=2)	Mean Water Lead (water=2)	Mean Water Lead (water=2)	Mean Water Lead (water=2)
Funding	Total Cumulative Funding 36-month Time Lag	Current CDC Funding 12-month Time Lag	Current CDC Funding 12-month Time Lag	Current HUD Funding, 12-month Time Lag
Screening	Screening Penetration	Screening Penetration	Screening Penetration	Screening Penetration

Table 5-2. Model 1 (Proportion ≥ 5 μg/dL) Parameter Estimates for Multivariate National Model

Region 1			
Effect	Estimate	StdErr	P-Value
Region	-2.200	0.291	< 0001
Time*Region	0.115	0.060	0.0546
Region*Quarter-1	-0.149	0.008	< 0001
Region*Quarter-2	-0.054	0.008	< 0001
Region*Quarter-3	0.243	0.007	< 0001
Region*Quarter-4	0.000	.	.

Region 2			
Effect	Estimate	StdErr	P-Value
Region	-2.567	0.299	< 0001
Time*Region	0.134	0.061	0 0295
Region*Quarter-1	-0.256	0.005	< 0001
Region*Quarter-2	-0.142	0.005	< 0001
Region*Quarter-3	0.209	0.005	< 0001
Region*Quarter-4	0.000	.	.

Region 3			
Effect	Estimate	StdErr	P-Value
Region	-2.250	0.286	< 0001
Time*Region	-0.023	0.059	0.6958
Region*Quarter-1	-0.271	0.008	< 0001
Region*Quarter-2	-0.125	0.008	< 0001
Region*Quarter-3	0.156	0.008	< 0001
Region*Quarter-4	0.000	.	.

Region 4			
Effect	Estimate	StdErr	P-Value
Region	-2.003	0.280	< 0001
Time*Region	-0.065	0.058	0.2628
Region*Quarter-1	-0.195	0.006	< 0001
Region*Quarter-2	-0.038	0.006	< 0001
Region*Quarter-3	0.068	0.006	< 0001
Region*Quarter-4	0.000	.	.

Region 5			
Effect	Estimate	StdErr	P-Value
Region	-2.546	0.284	< 0001
Time*Region	0.018	0.058	0.758996
Region*Quarter-1	-0.334	0.004	< 0001
Region*Quarter-2	-0.117	0.004	< 0001
Region*Quarter-3	0.280	0.004	< 0001
Region*Quarter-4	0.000	.	.

Region 6			
Effect	Estimate	StdErr	P-Value
Region	-2.196	0.283	< 0001
Time*Region	0.016	0.058	0.7801
Region*Quarter-1	-0.114	0.008	< 0001
Region*Quarter-2	-0.018	0.008	0 0334
Region*Quarter-3	0.102	0.008	< 0001
Region*Quarter-4	0.000	.	.

Region 7			
Effect	Estimate	StdErr	P-Value
Region	-2.406	0.279	< 0001
Time*Region	0.048	0.057	0.4064
Region*Quarter-1	-0.291	0.009	< 0001
Region*Quarter-2	-0.006	0.009	0.5220
Region*Quarter-3	0.184	0.008	< 0001
Region*Quarter-4	0.000	.	.

Region 8			
Effect	Estimate	StdErr	P-Value
Region	-2.621	0.285	< 0001
Time*Region	0.076	0.059	0.2013
Region*Quarter-1	-0.277	0.057	< 0001
Region*Quarter-2	-0.250	0.055	< 0001
Region*Quarter-3	0.057	0.051	0.2686
Region*Quarter-4	0.000	.	.

Region 9			
Effect	Estimate	StdErr	P-Value
Region	-2.125	0.318	< 0001
Time*Region	0.118	0.066	0.0749
Region*Quarter-1	-0.016	0.021	0.4596
Region*Quarter-2	-0.214	0.020	< 0001
Region*Quarter-3	-0.006	0.019	0.7635
Region*Quarter-4	0.000	.	.

Region 10			
Effect	Estimate	StdErr	P-Value
Region	-2.516	0.311	< 0001
Time*Region	0.052	0.065	0.4231
Region*Quarter-1	-0.325	0.052	< 0001
Region*Quarter-2	-0.196	0.050	0 00009
Region*Quarter-3	0.113	0.049	0 020451
Region*Quarter-4	0.000	.	.

Variance Components			
Effect	Estimate	StdErr	P-Value
UN(1,1)	0.210	0.007	
UN(2,1)	-0.020	0.001	
UN(2,2)	0.009	0.000	

Other Predictors									
Effect	X			X*Time			X*Time2		
	Est.	StdErr	P-Val	Est.	StdErr	P-Val	Est.	StdErr	P-Val
Screening Penetration	-1.010	0.066	<0.0001	0.117	0.014	<0.0001	-0.087	0.004	<0.0001
Pct. Units Built Before 1960	1.792	0.109	<0.0001	-0.209	0.023	<0.0001	0.026	0.001	<0.0001
TRI Lead Total Air ≥ 95th Percentile	0.250	0.051	<0.0001	-0.013	0.010	0.1944	0.000	0.000	<0.0001
Median Rent	-0.039	0.019	0.0465	-0.013	0.004	0.0009	-0.002	0.000	<0.0001
Total Cumulative Funding: 36-Month Lag	-0.002	0.000	<0.0001	0.001	0.000	<0.0001	-0.0002	<0.0001	<0.0001
Pct of Residents Without CollegeDegree	0.544	0.254	0.0321	0.077	0.053	0.1431	-0.014	0.002	<0.0001
Pct Units with No Household Wage	0.270	0.300	0.3687	-0.120	0.064	0.0592	0.005	0.004	0.2343
Pct < 6 Yrs of Age	-1.569	1.094	0.1515	-0.449	0.237	0.0585	0.061	0.015	0.0001
Pct Single Parent	0.139	0.243	0.5662	-0.101	0.052	0.0511	-0.002	0.002	0.3862
Pct Black Alone	0.697	0.125	<0.0001	0.072	0.027	0.0069	0.003	0.001	0.0409
Pct. Vacant Units	0.102	0.161	0.5258	0.004	0.035	0.9063	-0.012	0.003	<0.0001
Mean Water Lead Conc.(water =1)	0.056	0.011	<0.0001	0.002	0.004	0.6776	-0.003	0.001	<0.0001
Mean Water Lead Conc.(water =2)	-0.151	0.027	<0.0001	-0.013	0.010	0.1800	0.013	0.002	<0.0001
Median Air ≥ 99th Percentile	0.307	0.100	0.0022	0.012	0.010	0.2031	-0.001	0.000	0.0007
Total Housing Units	-0.001	0.001	0.6118	.	.	.	.	.	.

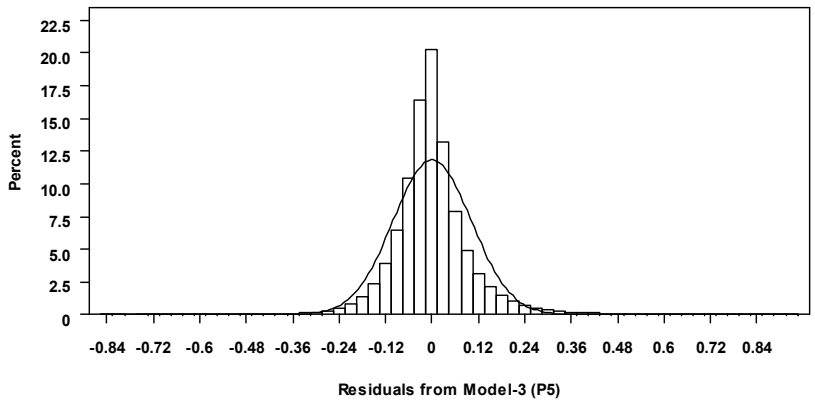


Figure 5-1. Histograms of Residuals from Fitted National Multivariate Model 1

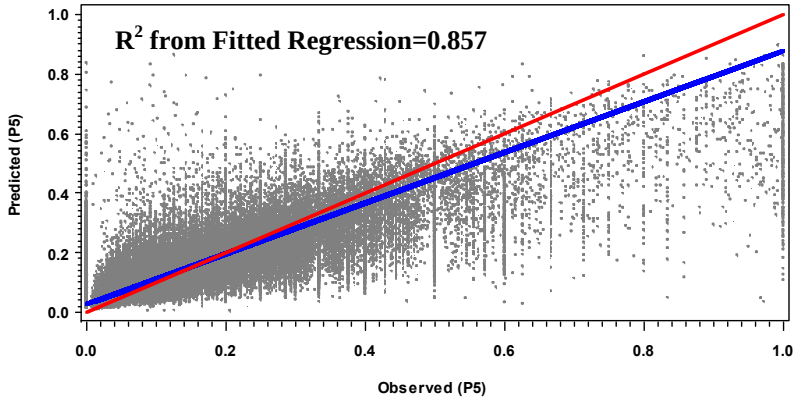


Figure 5-2a. Plot of National Multivariate Model Predicted Values versus Observed with Fitted Regression Line and 45° Reference Line for Proportion of Children with BLL ≥ 5 µg/dL

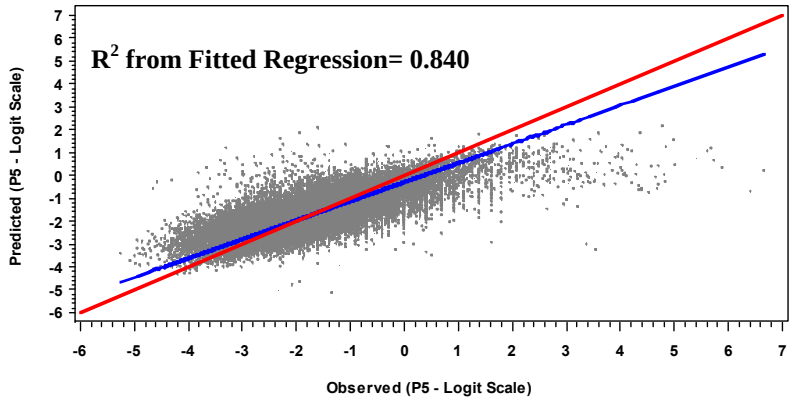


Figure 5-2b. Plot of National Multivariate Model Predicted Values versus Observed with Fitted Regression Line and 45° Reference Line for Proportion of Children with BLL ≥ 5 µg/dL (Logit Scale)

Table 5-3. Model 2 (Proportion ≥ 10 μg/dL) Parameter Estimates for Multivariate National Model

Region 1			
Effect	Estimate	StdErr	P-Value
Region	-4.973	0.277	<.0001
Time*Region	0.117	0.051	0.0216
Region*Quarter-1	-0.195	0.015	<.0001
Region*Quarter-2	-0.012	0.015	0.3983
Region*Quarter-3	0.342	0.014	<.0001
Region*Quarter-4	0.000	.	.

Region 2			
Effect	Estimate	StdErr	P-Value
Region	-5.505	0.286	<.0001
Time*Region	0.185	0.053	0.0005
Region*Quarter-1	-0.354	0.012	<.0001
Region*Quarter-2	-0.145	0.011	<.0001
Region*Quarter-3	0.304	0.010	<.0001
Region*Quarter-4	0.000	.	.

Region 3			
Effect	Estimate	StdErr	P-Value
Region	-5.173	0.266	<.0001
Time*Region	0.041	0.050	0.4085
Region*Quarter-1	-0.333	0.014	<.0001
Region*Quarter-2	-0.119	0.013	<.0001
Region*Quarter-3	0.248	0.012	<.0001
Region*Quarter-4	0.000	.	.

Region 4			
Effect	Estimate	StdErr	P-Value
Region	-5.170	0.260	<.0001
Time*Region	0.049	0.049	0.3080
Region*Quarter-1	-0.231	0.015	<.0001
Region*Quarter-2	0.017	0.015	0.2443
Region*Quarter-3	0.133	0.014	<.0001
Region*Quarter-4	0.000	.	.

Region 5			
Effect	Estimate	StdErr	P-Value
Region	-5.417	0.259	<.0001
Time*Region	0.046	0.048	0.3447
Region*Quarter-1	-0.467	0.007	<.0001
Region*Quarter-2	-0.138	0.007	<.0001
Region*Quarter-3	0.370	0.006	<.0001
Region*Quarter-4	0.000	.	.

Region 6			
Effect	Estimate	StdErr	P-Value
Region	-5.175	0.264	<.0001
Time*Region	0.098	0.050	0.0491
Region*Quarter-1	-0.166	0.019	<.0001
Region*Quarter-2	-0.048	0.018	0.0098
Region*Quarter-3	0.090	0.018	<.0001
Region*Quarter-4	0.000	.	.

Region 7			
Effect	Estimate	StdErr	P-Value
Region	-5.071	0.259	<.0001
Time*Region	0.061	0.048	0.2048
Region*Quarter-1	-0.463	0.017	<.0001
Region*Quarter-2	-0.008	0.015	0.5943
Region*Quarter-3	0.225	0.014	<.0001
Region*Quarter-4	0.000	.	.

Region 8			
Effect	Estimate	StdErr	P-Value
Region	-5.624	0.286	<.0001
Time*Region	0.042	0.056	0.4465
Region*Quarter-1	-0.482	0.136	0.0004
Region*Quarter-2	-0.164	0.120	0.1734
Region*Quarter-3	-0.061	0.113	0.5930
Region*Quarter-4	0.000	.	.

Region 9			
Effect	Estimate	StdErr	P-Value
Region	-4.747	0.310	<.0001
Time*Region	0.249	0.059	<.0001
Region*Quarter-1	0.088	0.035	0.0111
Region*Quarter-2	-0.108	0.033	0.0011
Region*Quarter-3	0.050	0.032	0.1223
Region*Quarter-4	0.000	.	.

Region 10			
Effect	Estimate	StdErr	P-Value
Region	-5.677	0.315	<.0001
Time*Region	0.076	0.061	0.2119
Region*Quarter-1	-0.572	0.105	<.0001
Region*Quarter-2	-0.385	0.099	0.000106
Region*Quarter-3	0.153	0.091	0.0939
Region*Quarter-4	0.000	.	.

Variance Components			
Effect	Estimate	StdErr	P-Value
UN(1,1)	0.304	0.012	
UN(2,1)	-0.012	0.002	
UN(2,2)	0.007	0.000	

Other Predictors									
Effect	X			X*Time			X*Time2		
	Est.	StdErr	P-Val	Est.	StdErr	P-Val	Est.	StdErr	P-Val
Screening Penetration	-3.605	0.093	<0.0001	-0.519	0.024	<0.0001	.	.	.
Median Rent	0.030	0.022	0.1769	-0.014	0.004	0.0003	-0.002	0.000	<0.0001
Pct. Units Built Before 1960	2.028	0.133	<0.0001	0.026	0.024	0.2828	0.013	0.002	<0.0001
Pct of Residents No College	0.624	0.217	0.0040	-0.123	0.041	0.0026	0.030	0.004	<0.0001
Pct Units with No Household Wage	1.049	0.374	0.0050	-0.064	0.073	0.3791	-0.092	0.008	<0.0001
TRI Lead Water Surface ≥ 95th Percentile	0.105	0.059	0.0773	-0.004	0.010	0.6734	0.005	0.001	<0.0001
Current CDC Funding: 12-Month Time Lag	0.050	0.015	0.0009	0.053	0.003	<0.0001	0.003	0.000	<0.0001
Pct < 6 Yrs of Age	1.154	1.407	0.4118	-0.126	0.277	0.6493	-0.003	0.028	0.9192
Pct Multiple Races	-3.486	1.230	0.0046	-0.884	0.235	0.0002	0.132	0.031	<0.0001
Total_Pop	0.001	0.000	0.0320	0.000	0.000	0.0006	0.000	0.000	<0.0001
Pct Single Parent	1.770	0.258	<0.0001	-0.043	0.050	0.3842	0.010	0.003	0.0044
Mean Water Lead Conc.(water =1)	0.008	0.021	0.6943	0.019	0.008	0.0189	-0.004	0.001	0.0027
Mean Water Lead Conc.(water =2)	-0.019	0.048	0.6858	-0.071	0.019	0.0001	0.014	0.003	<0.0001
Air Lead: 95th Percentile	2.552	0.609	<0.0001	-0.086	0.101	0.3948	-0.065	0.008	<0.0001
Pct. Rented Units	0.021	0.262	0.9363	-0.015	0.048	0.7615	.	.	.

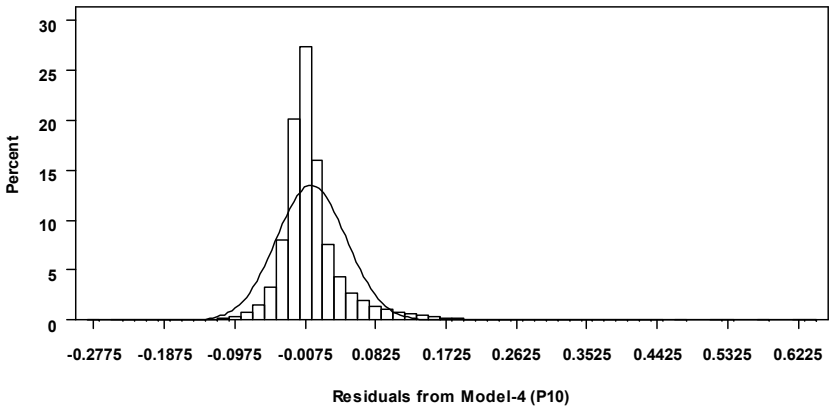


Figure 5-3. Histograms of Residuals from Fitted National Multivariate Model 2

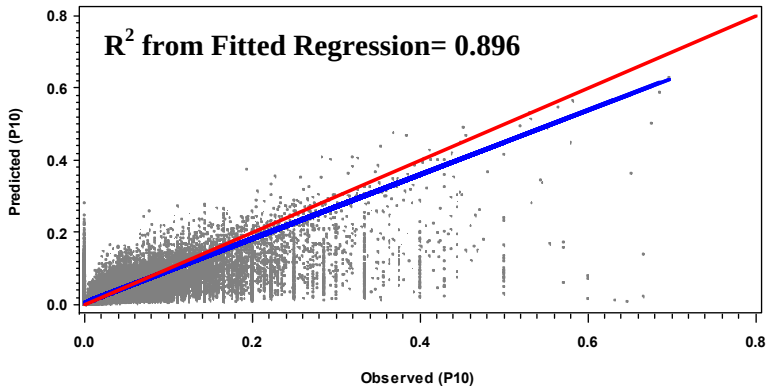


Figure 5-4a. Plot of National Multivariate Model Predicted Values versus Observed with Fitted Regression Line and 45° Reference Line for Proportion of Children with BLL ≥ 10 µg/dL.

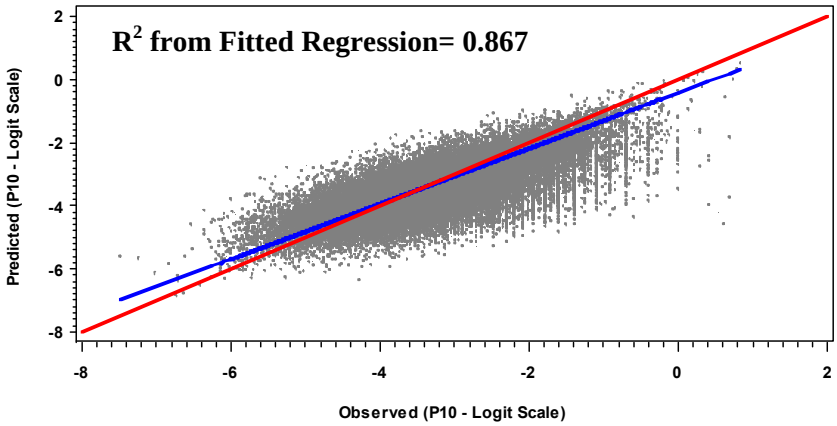


Figure 5-4b. Plot of National Multivariate Model Predicted Values versus Observed with Fitted Regression Line and 45° Reference Line for Proportion of Children with BLL ≥ 10 µg/dL (Logit Scale)

Table 5-4. Model 3 (Proportion ≥ 15 μg/dL) Parameter Estimates for Multivariate National Model

Region 1			
Effect	Estimate	StdErr	P-Value
Region	-7.800	0.339	<.0001
Time*Region	0.186	0.066	0.004948
Region*Quarter-1	-0.197	0.027	<.0001
Region*Quarter-2	0.028	0.025	0.265927
Region*Quarter-3	0.411	0.023	<.0001
Region*Quarter-4	0.000	.	.

Region 2			
Effect	Estimate	StdErr	P-Value
Region	-8.243	0.346	<.0001
Time*Region	0.262	0.068	0.000123
Region*Quarter-1	-0.352	0.021	<.0001
Region*Quarter-2	-0.073	0.019	0.000164
Region*Quarter-3	0.373	0.018	<.0001
Region*Quarter-4	0.000	.	.

Region 3			
Effect	Estimate	StdErr	P-Value
Region	-7.620	0.325	<.0001
Time*Region	0.097	0.064	0.13021
Region*Quarter-1	-0.361	0.021	<.0001
Region*Quarter-2	-0.093	0.020	<.0001
Region*Quarter-3	0.279	0.018	<.0001
Region*Quarter-4	0.000	.	.

Region 4			
Effect	Estimate	StdErr	P-Value
Region	-7.743	0.316	<.0001
Time*Region	0.144	0.062	0.02099
Region*Quarter-1	-0.262	0.029	<.0001
Region*Quarter-2	0.089	0.027	0.000873
Region*Quarter-3	0.169	0.026	<.0001
Region*Quarter-4	0.000	.	.

Region 5			
Effect	Estimate	StdErr	P-Value
Region	-7.958	0.315	<.0001
Time*Region	0.118	0.062	0.057593
Region*Quarter-1	-0.512	0.011	<.0001
Region*Quarter-2	-0.106	0.010	<.0001
Region*Quarter-3	0.440	0.009	<.0001
Region*Quarter-4	0.000	.	.

Region 6			
Effect	Estimate	StdErr	P-Value
Region	-7.702	0.321	<.0001
Time*Region	0.186	0.064	0.003527
Region*Quarter-1	-0.146	0.032	<.0001
Region*Quarter-2	-0.089	0.032	0.005426
Region*Quarter-3	0.032	0.031	0.304026
Region*Quarter-4	0.000	.	.

Region 7			
Effect	Estimate	StdErr	P-Value
Region	-7.750	0.317	<.0001
Time*Region	0.139	0.063	0.026081
Region*Quarter-1	-0.492	0.027	<.0001
Region*Quarter-2	0.079	0.024	0.001043
Region*Quarter-3	0.303	0.023	<.0001
Region*Quarter-4	0.000	.	.

Region 8			
Effect	Estimate	StdErr	P-Value
Region	-8.192	0.363	<.0001
Time*Region	0.121	0.075	0.104597
Region*Quarter-1	-0.730	0.241	0.002463
Region*Quarter-2	-0.104	0.193	0.589189
Region*Quarter-3	-0.093	0.185	0.613293
Region*Quarter-4	0.000	.	.

Region 9			
Effect	Estimate	StdErr	P-Value
Region	-6.830	0.382	<.0001
Time*Region	0.149	0.077	0.053885
Region*Quarter-1	-0.171	0.049	0.000454
Region*Quarter-2	-0.220	0.046	<.0001
Region*Quarter-3	0.016	0.045	0.722024
Region*Quarter-4	0.000	.	.

Region 10			
Effect	Estimate	StdErr	P-Value
Region	-8.170	0.393	<.0001
Time*Region	0.172	0.079	0.028832
Region*Quarter-1	-0.892	0.186	<.0001
Region*Quarter-2	-0.350	0.156	0.025118
Region*Quarter-3	0.019	0.147	0.895229
Region*Quarter-4	0.000	.	.

Variance Components			
Effect	Estimate	StdErr	P-Value
UN(1,1)	0.365	0.016	
UN(2,1)	-0.018	0.002	
UN(2,2)	0.009	0.001	

Other Predictors									
Effect	X			X*Time			X*Time2		
	Est.	StdErr	P-Val	Est.	StdErr	P-Val	Est.	StdErr	P-Val
Screening Penetration	-4.185	0.187	<0.0001	-0.542	0.039	<0.0001	0.024	0.011	0.0273
Median Rent	0.178	0.027	<0.0001	-0.020	0.005	0.0002	-0.004	0.000	<0.0001
Pct. Occupied Units Built Before 1950	2.993	0.185	<0.0001	-0.004	0.037	0.9228	0.017	0.003	<0.0001
Pct of Residents No College	0.799	0.260	0.0021	-0.032	0.053	0.5384	.	.	.
Pct Units with No Household Wage	1.400	0.454	0.0021	-0.195	0.094	0.0387	-0.021	0.007	0.0048
Current CDC Funding: 12-Month Time Lag	0.026	0.025	0.2810	0.059	0.005	<0.0001	0.004	0.001	<0.0001
Pct < 6 Yrs of Age	3.599	1.718	0.0361	-0.877	0.363	0.0156	0.255	0.027	<0.0001
Pct Asian Alone	-2.253	0.953	0.0180	0.316	0.191	0.0991	.	.	.
TRI Lead Water Surface ≥ 95th Percentile	0.223	0.066	0.0008	-0.006	0.012	0.6063	-0.001	0.001	0.2423
Mean Water Lead Conc.(water =1)	-0.025	0.032	0.4313	0.008	0.007	0.2212			
Mean Water Lead Conc.(water =2)	0.110	0.075	0.1392	-0.044	0.016	0.0053			
Housing_Density	-0.006	0.002	0.0001	-0.001	0.000	0.0266	0.000	0.000	<0.0001
Median Air ≥ 99th Percentile	0.665	0.143	<0.0001						
Pct. Rented Units	0.949	0.325	0.0035	0.052	0.066	0.4366	-0.025	0.005	<0.0001
Pct Single Parent	2.135	0.313	<0.0001	-0.196	0.065	0.0024			

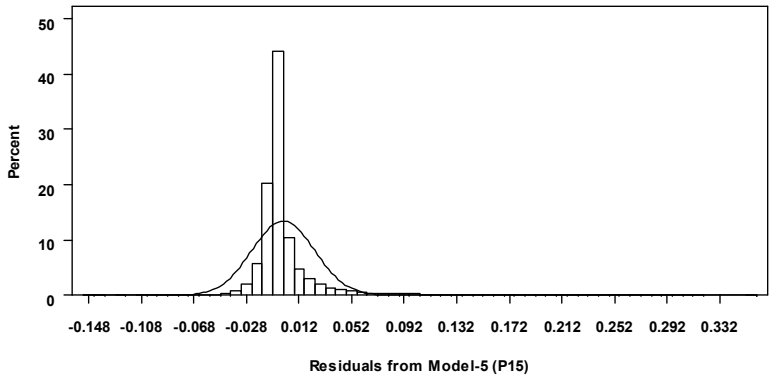


Figure 5-5. Histograms of Residuals from Fitted National Multivariate Model 3

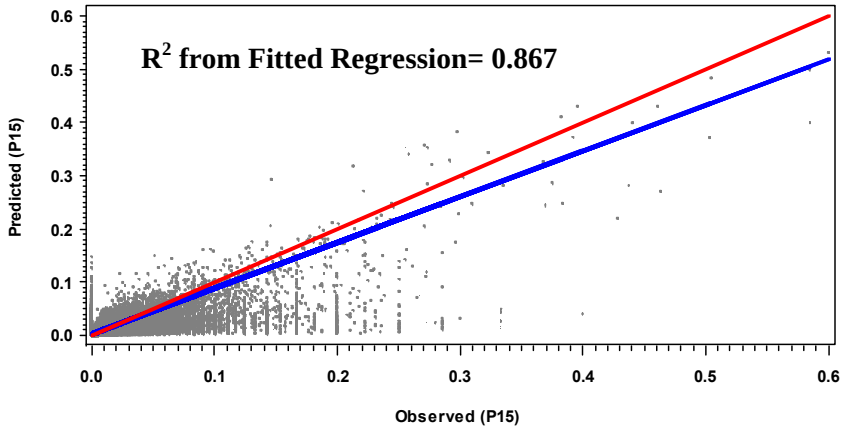


Figure 5-6a. Plot of National Multivariate Model Predicted Values versus Observed with Fitted Regression Line and 45° Reference Line for Proportion of Children with BLL ≥ 15 µg/dL.

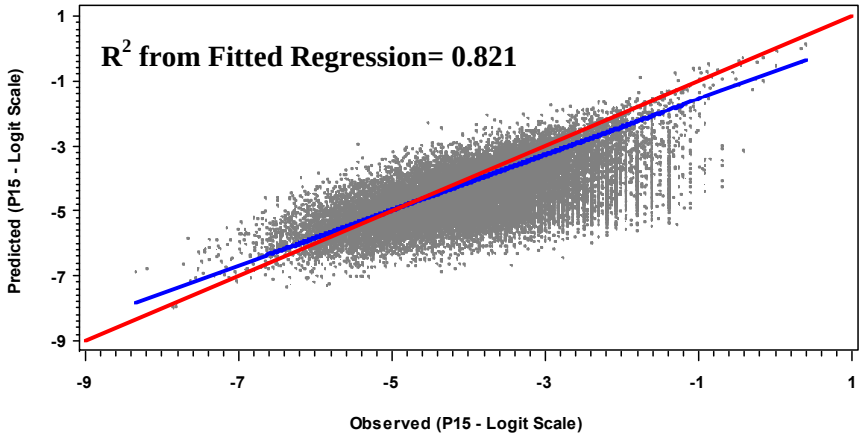


Figure 5-6b. Plot of National Multivariate Model Predicted Values versus Observed with Fitted Regression Line and 45° Reference Line for Proportion of Children with BLL ≥ 15 µg/dL (Logit Scale)

Table 5-5. Model 4 (Proportion ≥ 25 μg/dL) Parameter Estimates for Multivariate National Model

Region 1			
Effect	Estimate	StdErr	P-Value
Region	-9.307	0.403	<.0001
Time*Region	0.250	0.074	0.000684
Region*Quarter-1	-0.270	0.055	<.0001
Region*Quarter-2	0.111	0.050	0.026461
Region*Quarter-3	0.441	0.046	<.0001
Region*Quarter-4	0.000	.	.

Region 2			
Effect	Estimate	StdErr	P-Value
Region	-9.797	0.421	<.0001
Time*Region	0.207	0.081	0.01021
Region*Quarter-1	-0.287	0.043	<.0001
Region*Quarter-2	0.003	0.039	0.930399
Region*Quarter-3	0.466	0.036	<.0001
Region*Quarter-4	0.000	.	.

Region 3			
Effect	Estimate	StdErr	P-Value
Region	-9.083	0.393	<.0001
Time*Region	0.090	0.073	0.218218
Region*Quarter-1	-0.452	0.045	<.0001
Region*Quarter-2	-0.048	0.042	0.248085
Region*Quarter-3	0.319	0.037	<.0001
Region*Quarter-4	0.000	.	.

Region 4			
Effect	Estimate	StdErr	P-Value
Region	-9.481	0.380	<.0001
Time*Region	0.156	0.070	0.025874
Region*Quarter-1	-0.122	0.063	0.051773
Region*Quarter-2	0.204	0.058	0.0005
Region*Quarter-3	0.313	0.056	<.0001
Region*Quarter-4	0.000	.	.

Region 5			
Effect	Estimate	StdErr	P-Value
Region	-9.426	0.379	<.0001
Time*Region	0.131	0.071	0.064377
Region*Quarter-1	-0.541	0.023	<.0001
Region*Quarter-2	-0.008	0.020	0.705084
Region*Quarter-3	0.537	0.018	<.0001
Region*Quarter-4	0.000	.	.

Region 6			
Effect	Estimate	StdErr	P-Value
Region	-9.396	0.386	<.0001
Time*Region	0.165	0.073	0.024058
Region*Quarter-1	-0.117	0.071	0.100954
Region*Quarter-2	0.031	0.070	0.658998
Region*Quarter-3	0.136	0.067	0.04341
Region*Quarter-4	0.000	.	.

Region 7			
Effect	Estimate	StdErr	P-Value
Region	-9.474	0.380	<.0001
Time*Region	0.172	0.072	0.016446
Region*Quarter-1	-0.409	0.059	<.0001
Region*Quarter-2	0.299	0.050	<.0001
Region*Quarter-3	0.500	0.048	<.0001
Region*Quarter-4	0.000	.	.

Region 8			
Effect	Estimate	StdErr	P-Value
Region	-9.634	0.493	<.0001
Time*Region	0.170	0.103	0.098117
Region*Quarter-1	-0.984	0.531	0.064023
Region*Quarter-2	0.087	0.373	0.814912
Region*Quarter-3	-0.349	0.395	0.376556
Region*Quarter-4	0.000	.	.

Region 9			
Effect	Estimate	StdErr	P-Value
Region	-8.770	0.463	<.0001
Time*Region	0.157	0.088	0.074789
Region*Quarter-1	-0.115	0.098	0.241363
Region*Quarter-2	-0.036	0.092	0.691996
Region*Quarter-3	0.025	0.092	0.78802
Region*Quarter-4	0.000	.	.

Region 10			
Effect	Estimate	StdErr	P-Value
Region	-9.959	0.540	<.0001
Time*Region	0.206	0.098	0.035646
Region*Quarter-1	-0.225	0.388	0.561648
Region*Quarter-2	0.180	0.345	0.602567
Region*Quarter-3	0.194	0.350	0.580418
Region*Quarter-4	0.000	.	.

Variance Components			
Effect	Estimate	StdErr	P-Value
UN(1,1)	0.370	0.022	
UN(2,1)	-0.018	0.003	
UN(2,2)	0.010	0.001	

Other Predictors									
Effect	X			X*Time			X*Time2		
	Est.	StdErr	P-Val	Est.	StdErr	P-Val	Est.	StdErr	P-Val
Screening Penetration	-4.491	0.275	<0.0001	-0.478	0.073	<0.0001	.	.	.
Pct. Occupied Units Built Before 1940	3.344	0.268	<0.0001	-0.005	0.057	0.9231	.	.	.
Median Rent	0.183	0.033	<0.0001	-0.013	0.006	0.0178	-0.003	0.001	<0.0001
Pct of Residents No College	0.351	0.324	0.2787	.	.	.	.	.	.
Pct Units with No Household Earnings	2.475	0.619	0.0001	-0.111	0.131	0.3951	.	.	.
Pct Asian Alone	-1.135	1.082	0.2943	.	.	.	.	.	.
Current HUD Funding: 12-Month Time Lag	-0.015	0.005	0.0010	-0.006	0.002	0.0004	0.002	0.000	<0.0001
Pct < 6 Yrs of Age	6.920	2.254	0.0021	-0.928	0.462	0.0448	0.173	0.028	<0.0001
TRI Lead Underground Injection ≥ 95th Percentile	-0.089	0.183	0.6295	-0.021	0.040	0.5914			
Mean Water Lead Conc.(water =1)	0.020	0.021	0.3446	.	.	.	.	.	.
Mean Water Lead Conc.(water =2)	0.030	0.053	0.5689	.	.	.	.	.	.
Housing_Density	-0.006	0.002	0.0008	.	.	.	.	.	.
Median Air ≥ 99th Percentile	0.383	0.183	0.0365	.	.	.	.	.	.
Pct Single Parent	2.060	0.428	<0.0001	-0.355	0.093	0.0001	.	.	.
Pct. Rented Units	1.102	0.408	0.0069	0.037	0.078	0.6330	.	.	.

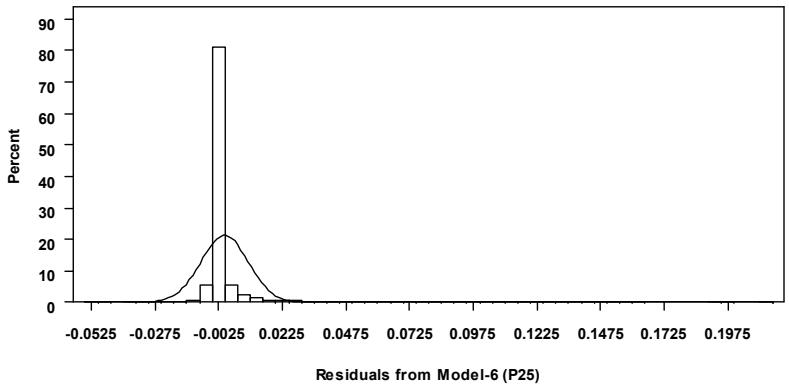


Figure 5-7. Histograms of Residuals from Fitted National Multivariate Model 4

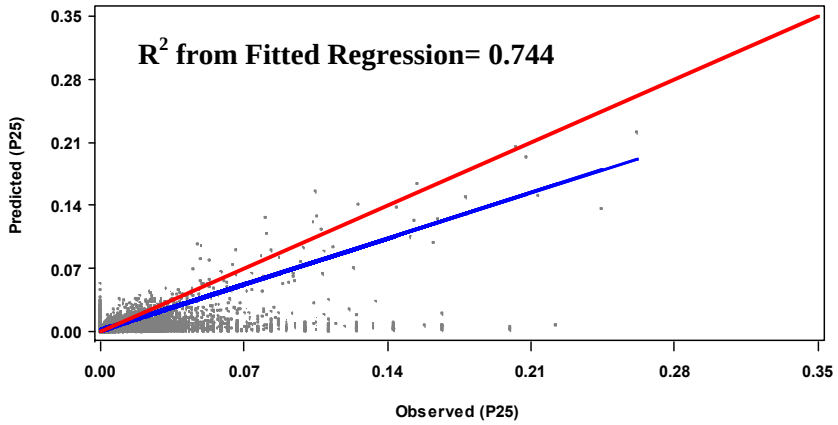


Figure 5-8a. Plot of National Multivariate Model Predicted Values versus Observed with Fitted Regression Line and 45° Reference Line for Proportion of Children with BLL ≥ 25 µg/dL.

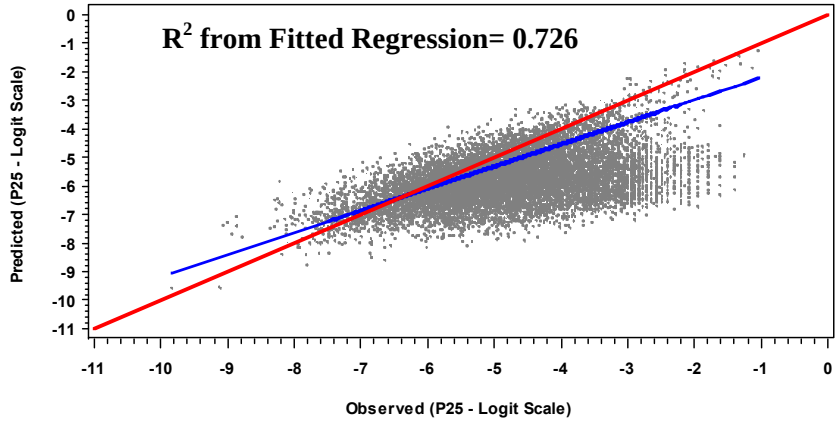


Figure 5-8b. Plot of National Multivariate Model Predicted Values versus Observed with Fitted Regression Line and 45° Reference Line for Proportion of Children with BLL ≥ 25 µg/dL (Logit Scale)

5.2 High-Resolution Modeling Results

The Massachusetts final multivariate models were constructed similarly to the national models. One basic difference in the Massachusetts models is that there was no EPA region. Thus, there was no area variable included other than census tract. As in the national model, time period and quarter were significant predictors; however, in the Massachusetts model they are not interacted with an area variable. Table 5-6 presents the full set of variables included in the final multivariate models for Models 1 through 5. Model 6 was not fit for the Massachusetts data because of the scarcity of data above 25 µg/dL.

Among the demographic variables, housing cost, occupancy, family structure, and housing age were significant predictors in all five models. Median Rent was the selected housing cost variable in all five models. For occupancy, Percent Rented was the selected variable in four of the five models. Percent of Single Parent Households is the family structure variable in all models. Three housing age variables were included across the five models, but Percent Built Pre-1950 was the included variable in Models 1 to 3.

Race and Income variables were included in four of the five final models. Median Household Income and Percent Multiple Races were the two variables used in all four models. Children's Age, Education, and Population each had a variable included in one of the final models. Number of Children less than or equal to six years old and Total Population were included in Model 3. Percent Without 9th Grade Education was included in Model 4.

Unlike the national models, none of the environmental variables were included in the final multivariate models for Massachusetts. On the other hand, the housing inspection data from Massachusetts were predictive and included in all of the final models. The percentage of units passing the Massachusetts standard of care (calculated using the MDPH method) was included in all five models. Additionally, the percentage of units failing the Massachusetts standard of care (calculated using the MDPH method) was included in Models 4 and 5.

The selected programmatic funding variable was included in Models 1, 2, and 5. Current State Funding ($ per Child) was used in the GM models and Cumulative CDC Funding ($ per tract) was used in Model 5.

As with the national models, parameter estimates and associated standard errors and p-values are presented for all models in Table 5-7. Figures 5-9 to 5-18 contain the histograms of residuals and plots of observed versus predicted values that allow assessment of the various model fits.

These plots suggest that models 1-3 are performing well, with Models 4 and 5 providing a somewhat suboptimal fit (perhaps due to fewer children being observed above the 10 and 15 µg/dL threshold values in Massachusetts). The weighted regression line fit to the observed versus predicted plots (shown in blue) also demonstrates a systematic degradation in model performance from Models 3 through 5, with the R^2 value diminishing as the blood-lead threshold value increases. Similar to the National Models, the High-Resolution Multivariate Models in Massachusetts tend to under-predict for census-tract/quarter combinations with higher geometric mean blood-lead concentrations and higher exceedance proportions. Further exploration may be necessary to determine whether these higher values represent county/quarter combinations with

fairly sparse data (i.e., few observations) – which might explain why they would have been less influential in Models 2 through 6, which are influenced by the number of observations associated with each observed value.

Appendix F presents predictions of areas of the country estimated to have the highest children's blood-lead levels. These predictions were generated by averaging predicted values across the four quarters of 2006. Table F-1 lists the 150 counties/townships in the United States with the highest predicted GM blood-lead levels (using Model 2) and proportion of children above 5, 10, 15, and 25 μg/dL. Table F-2 lists the 10 counties in each state with the highest levels of those same five outcomes. Table F-3 lists the 150 Massachusetts census tracts with the highest predicted GM blood-lead levels. Figure F-1 provides a map of these 150 Massachusetts census tracts.

Table 5-6. Summary of Variables Included in Final Massachusetts Multivariate Model

Variable Type	Model 1	Model 2	Model 3	Model 4	Model 5
Time	Time Period, Quarter	Time Period, Quarter	Time Period, Quarter	Time Period, Quarter	Time Period, Quarter
Income	Median Household Income	Median Household Income	Median Household Income	Median Household Income	
Race	Percent Multiple Race	Percent Multiple Race	Percent Multiple Race	Percent Multiple Race	
Housing Cost	Median Rent	Median Rent	Median Rent	Median Rent	Median Rent
Occupancy	Percent Rented	Percent Rented	Percent Rented	Percent Rented	Percent Vacant
Family Structure	Percent Single Parent	Percent Single Parent	Percent Single Parent	Percent Single Parent	Percent Single Parent
Housing Age	Percent Built Pre-1950	Percent Built Pre-1950	Percent Built Pre-1950	Percent Occupied Built Pre-1940	Percent Occupied Built Pre-1980
Children's Age			Number less than 6 years old		
Education				Percent without 9th Grade education	
Population			Total Population		
Housing Inspection	P4 - % Passing Standard of Care, MDPH Method	P4 - % Passing Standard of Care, MDPH Method	P4 - % Passing Standard of Care, MDPH Method	P4 - % Passing Standard of Care, MDPH Method	P4 - % Passing Standard of Care, MDPH Method
				F4 - % failing standard of care, MDPH Method	F4 - % failing standard of care, MDPH Method
Funding	Current State Funding ($ per Child)	Current State Funding ($ per Child)			Cumulative CDC Funding ($ per tract)

Table 5-7. Massachusetts Multivariate Model Estimates

Model	Effect	Levels	Estimate	Standard Error	P-Value
1 (Geometric Mean)	Intercept	_	2.290	0.033	<.0001
	Time	_	-0.088	0.002	<.0001
	Quarter (Season)	1	-0.187	0.007	<.0001
		2	-0.143	0.007	<.0001
		3	0.130	0.006	<.0001
		4	0.000	.	.
	Median Household Income	_	-0.008	0.001	<.0001
	Percent Multiple Races	_	3.909	0.536	<.0001
	Median Rent ($):	_	-0.032	0.005	<.0001
	Percent Rented Units	_	-0.568	0.069	<.0001
	Percent Single Parent Households	_	0.702	0.096	<.0001
	Percent Units Built Before 1950	_	0.849	0.047	<.0001
	p4	_	-0.983	0.166	<.0001
	Current State Funding	_	0.028	0.006	<.0001
	$\sigma^2_{\delta_0}$	0.229	.		
	$\sigma^2_{\delta_0,\delta_1}$	-0.026	.		
	$\sigma^2_{\delta_1}$	0.004	.		
	σ^2_{Error}	0.191	.		
2 (Weighted Geometric Mean)	Intercept	_	2.249	0.033	<.0001
	Time	_	-0.087	0.002	<.0001
	Quarter (Season)	1	-0.185	0.006	<.0001
		2	-0.137	0.006	<.0001
		3	0.127	0.006	<.0001
		4	0.000	.	.
	Median Household Income	_	-0.008	0.001	<.0001
	Percent Multiple Races	_	3.826	0.531	<.0001
	Median Rent ($):	_	-0.032	0.005	<.0001
	Percent Rented Units	_	-0.561	0.069	<.0001
	Percent Single Parent Households	_	0.735	0.096	<.0001
	Percent Units Built Before 1950	_	0.866	0.046	<.0001
	p4	_	-0.933	0.164	<.0001
	Current State Funding	_	0.031	0.006	<.0001
	$\sigma^2_{\delta_0}$	0.218	.		
	$\sigma^2_{\delta_0,\delta_1}$	-0.024	.		
	$\sigma^2_{\delta_1}$	0.004	.		
	σ^2_{Error}	3.986	.		

Model	Effect	Levels	Estimate	Standard Error	P-Value
3 (Proportion of Children with Blood Lead ≥ 5 µg/dL)	Intercept	_	-2.312	0.054	<.0001
	Time	_	-0.146	0.003	<.0001
	Quarter (Season)	1	-0.195	0.010	<.0001
		2	-0.117	0.010	<.0001
		3	0.187	0.009	<.0001
		4	0.000	.	.
	Median Household Income	_	-0.010	0.001	<.0001
	Percent Multiple Races	_	3.420	0.610	<.0001
	Median Rent ($):	_	-0.044	0.006	<.0001
	Percent Rented Units	_	-0.724	0.083	<.0001
	Percent Single Parent Households	_	0.817	0.119	<.0001
	Percent Units Built Before 1950	_	1.468	0.056	<.0001
	Number Residents Less than Six Years of Age	_	0.000	0.000	0.0114
	Total Population	_	0.000	0.000	0.0292
	p4	_	-0.649	0.218	0.0029
	$\sigma^2_{\delta_0}$	0.129	0.007		
	$\sigma^2_{\delta_0,\delta_1}$	-0.008	0.001		
	$\sigma^2_{\delta_1}$	0.004	0.000		
4 (Proportion of Children with Blood Lead ≥ 10 µg/dL)	Intercept	_	-4.235	0.052	<.0001
	Time	_	-0.136	0.005	<.0001
	Quarter (Season)	1	-0.282	0.022	<.0001
		2	-0.130	0.021	<.0001
		3	0.247	0.020	<.0001
		4	0.000	.	.
	Median Household Income	_	-0.010	0.001	<.0001
	Percent Multiple Races	_	4.326	0.781	<.0001
	Median Rent ($):	_	-0.057	0.009	<.0001
	Percent Rented Units	_	-0.562	0.119	<.0001
	Percent Single Parent Households	_	0.697	0.159	<.0001
	Percent Occupied Units Built Before 1980	_	1.758	0.083	<.0001
	Percent Residents with Less than Ninth Grade Education	_	-0.581	0.279	0.0372
	f4	_	1.802	0.504	0.0004
	p4	_	-1.339	0.321	<.0001
	$\sigma^2_{\delta_0}$	0.175	0.016		
	$\sigma^2_{\delta_0,\delta_1}$	-0.013	0.003		
	$\sigma^2_{\delta_1}$	0.004	0.001		

Model	Effect	Levels	Estimate	Standard Error	P-Value
5 (Proportion of Children with Blood Lead ≥ 15 µg/dL)	Intercept	_	-6.278	0.144	<.0001
	Time	_	-0.093	0.008	<.0001
	Quarter (Season)	1	-0.307	0.042	<.0001
		2	-0.093	0.040	0.019
		3	0.332	0.036	<.0001
		4	0.000	.	.
	Median Rent ($):	_	-0.049	0.010	<.0001
	Percent Vacant Units	_	1.039	0.335	0.0019
	Percent Single Parent Households	_	1.002	0.169	<.0001
	Percent Occupied Units Built Before 1980	_	1.187	0.168	<.0001
	f4	_	4.047	0.677	<.0001
	p4	_	-1.476	0.440	0.0008
	Cumulative CDC Funding	_	0.000	0.000	0.0127
	$\sigma^2_{\delta_0}$	0.249	0.035		
	$\sigma^2_{\delta_0,\delta_1}$	-0.020	0.008		
	$\sigma^2_{\delta_1}$	0.004	0.002		

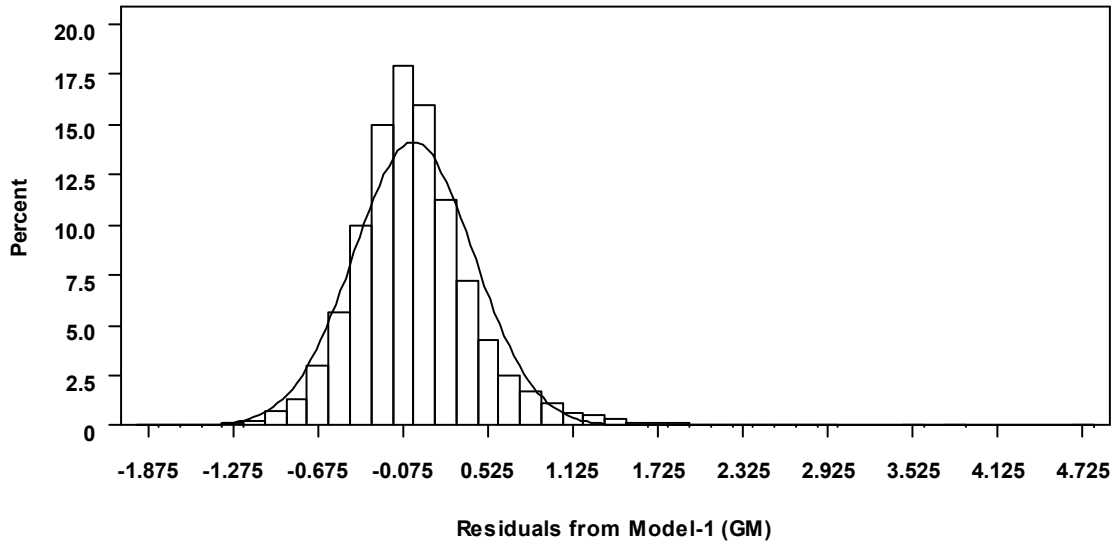
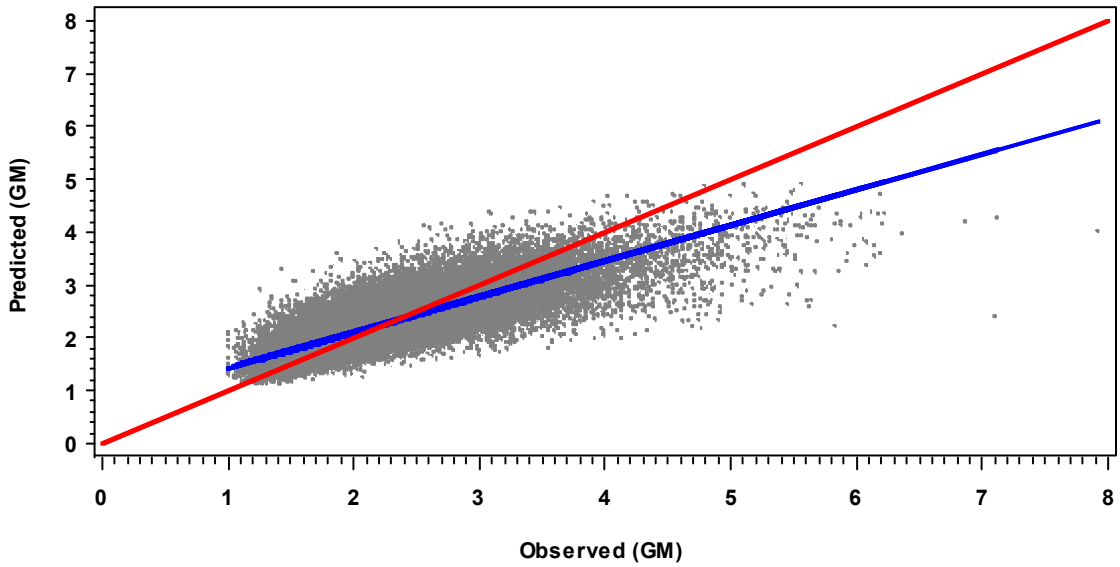


Figure 5-9. Histograms of Residuals from Fitted Massachusetts Multivariate Model 1



R^2 from Fitted Regression = 0.697

Figure 5-10. Plot of Massachusetts Multivariate Model Predicted Values versus Observed with Fitted Regression Line and 45° Reference Line for Unweighted Geometric Mean Response

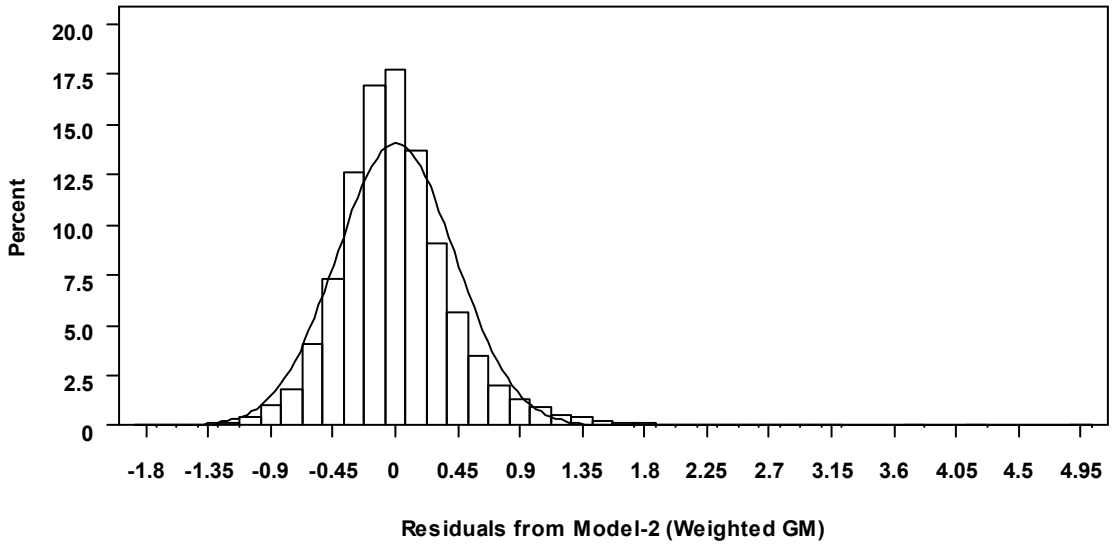
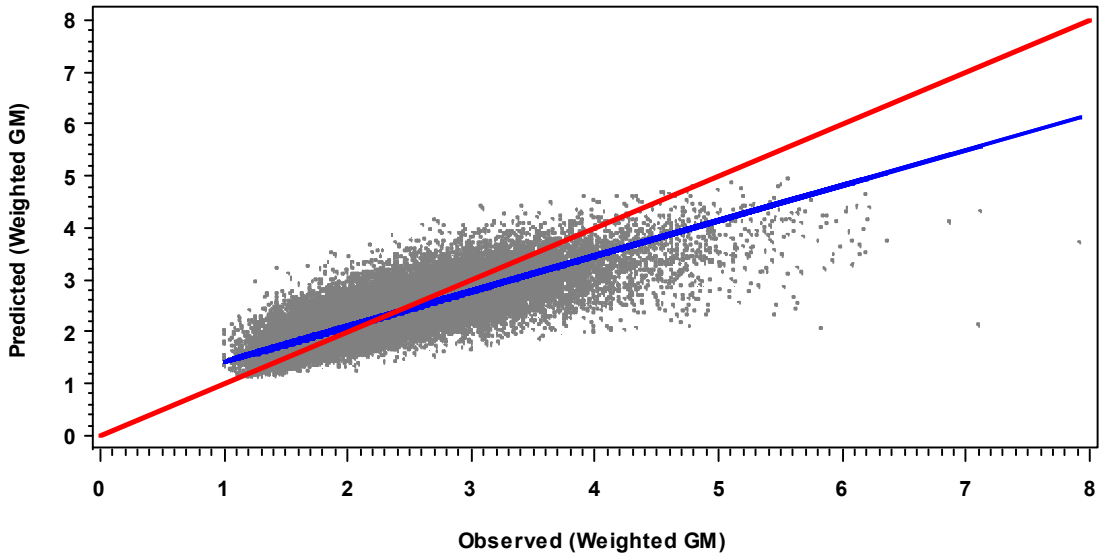


Figure 5-11. Histograms of Residuals from Fitted Massachusetts Multivariate Model 2



R^2 from Fitted Regression = 0.700

Figure 5-12. Plot of Massachusetts Multivariate Model Predicted Values versus Observed with Fitted Regression Line and 45° Reference Line for Weighted Geometric Mean Response

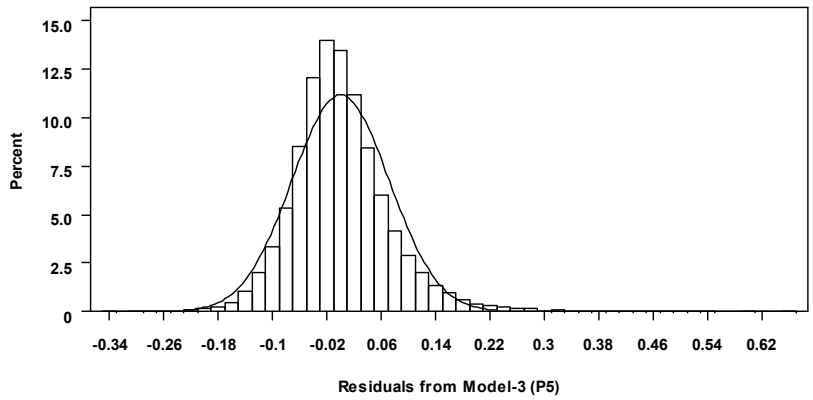


Figure 5-13. Histograms of Residuals from Fitted Massachusetts Multivariate Model 3

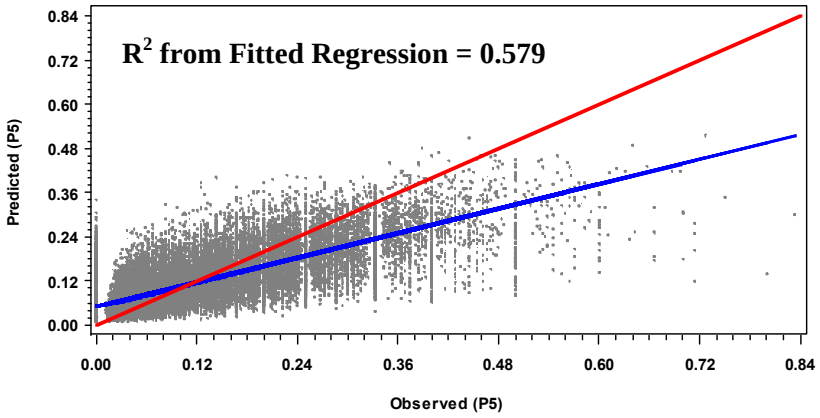


Figure 5-14a. Plot of Massachusetts Multivariate Model Predicted Values versus Observed with Fitted Regression Line and 45° Reference Line for Proportion of Children with BLL ≥ 5 µg/dL.

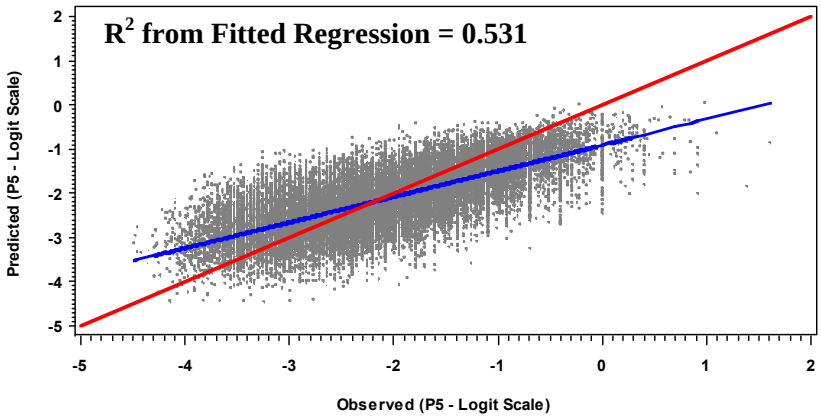


Figure 5-14b. Plot of Massachusetts Multivariate Model Predicted Values versus Observed with Fitted Regression Line and 45° Reference Line for Proportion of Children with BLL ≥ µg/dL (Logit Scale)

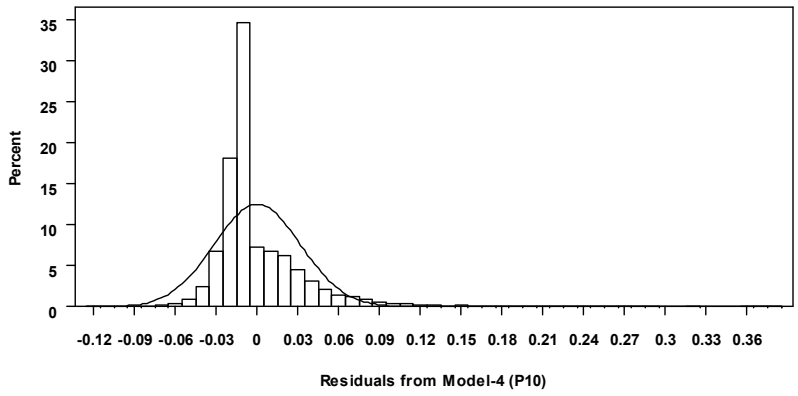


Figure 5-15. Histograms of Residuals from Fitted Massachusetts Multivariate Model 4

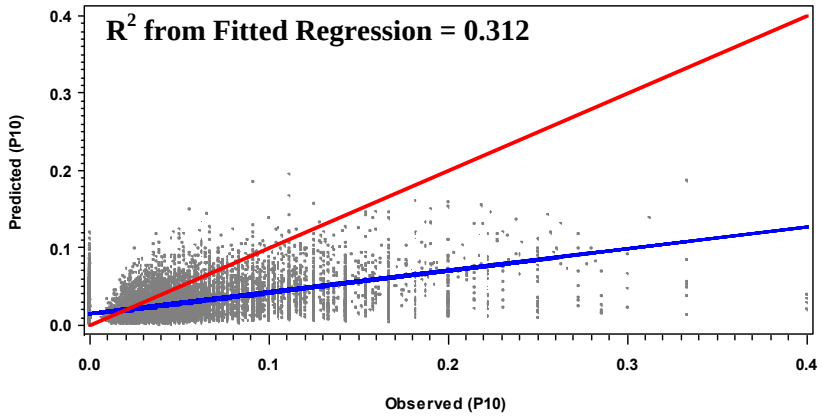


Figure 5-16a. Plot of Massachusetts Multivariate Model Predicted Values versus Observed with Fitted Regression Line and 45° Reference Line for Proportion of Children with BLL ≥ 10 µg/dL

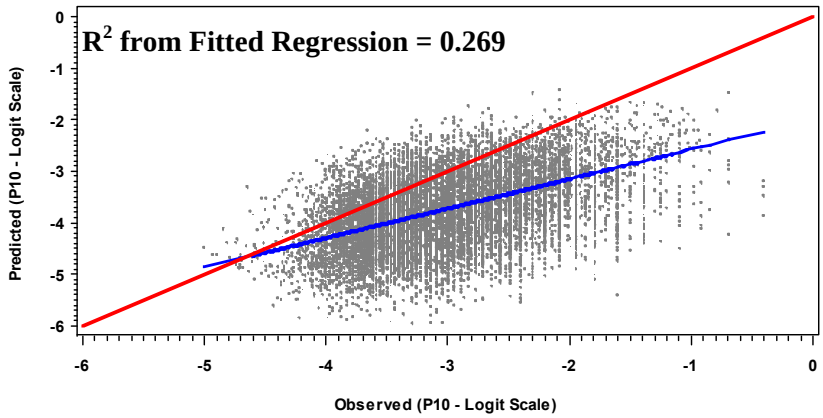


Figure 5-16b. Plot of Massachusetts Multivariate Model Predicted Values versus Observed with Fitted Regression Line and 45° Reference Line for Proportion of Children with BLL ≥ 10 µg/dL (Logit Scale)

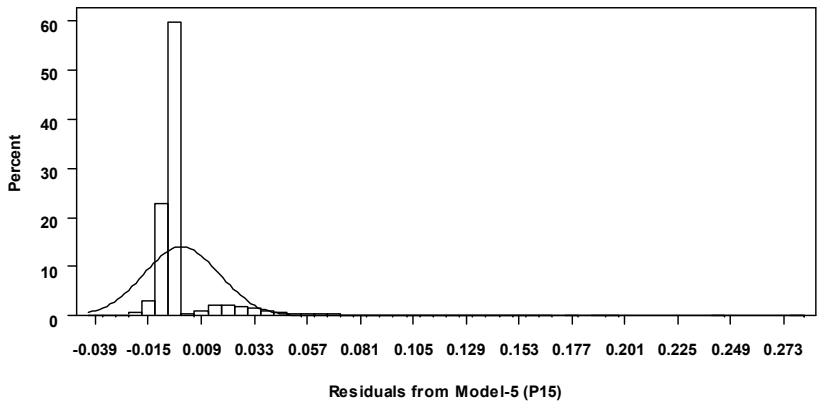


Figure 5-17. Histograms of Residuals from Fitted Massachusetts Multivariate Model 5

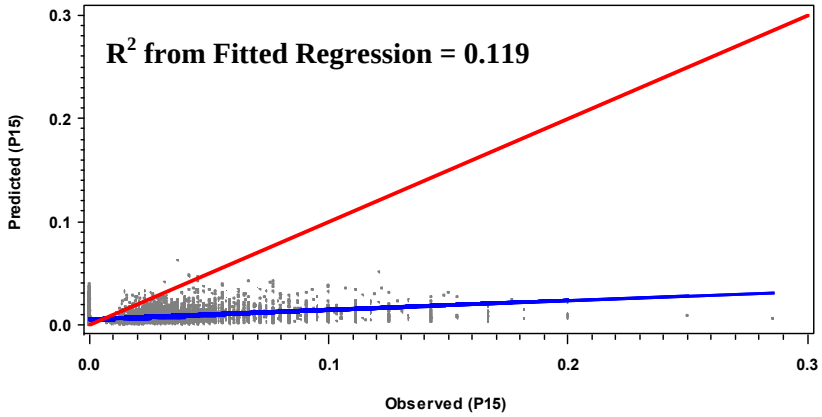


Figure 5-18a. Plot of Massachusetts Multivariate Model Predicted Values versus Observed with Fitted Regression Line and 45° Reference Line for Proportion of Children with BLL ≥ 15 µg/dL.

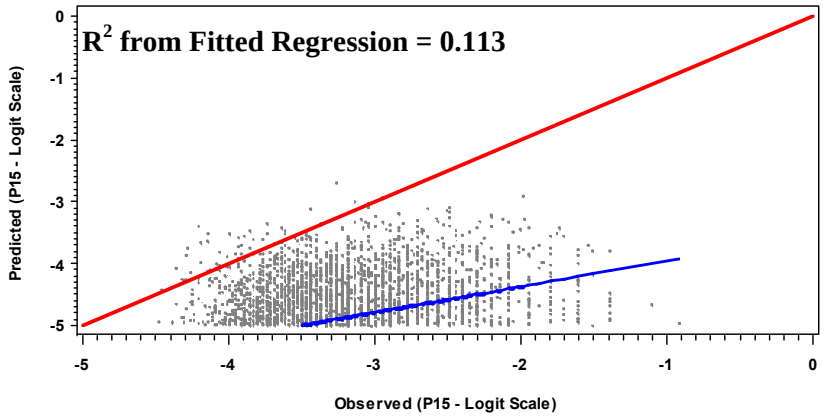


Figure 5-18b. Plot of Massachusetts Multivariate Model Predicted Values versus Observed with Fitted Regression Line and 45° Reference Line for Proportion of Children with BLL ≥ 15 µg/dL (Logit Scale)

6.0 GRAPHICAL PRESENTATION OF MODELING RESULTS

In addition to the discussion of the final multivariate modeling results in Section 5, it is informative to be able to view the results visually. Two methods were utilized for graphical presentation of the results – mapping and via use of an interactive software tool. Section 6.1 presents a subset of the maps generated, while Sections 6.2 and 6.3 discuss the interactive software tool.

6.1 Maps of Observed and Predicted Blood-Lead Outcomes

Mapping is an informative method to graphically present the results of the multivariate models. Figure 6-1 contains maps displaying the observed levels of GM blood-lead levels in 2000 and 2005 based on CDC's national surveillance data, and the comparable predicted GM blood-lead levels in 2000 and 2005. A key difference is that maps of observed levels contain many counties with missing data either because they do not submit childhood lead surveillance data to CDC or they have too few test records to be included in the analysis, while the maps of predicted levels covers all counties in the country. Appendix G contains detailed maps from the national level models of GM blood-lead levels and proportion of children with BLLs $\geq$ 10 μg/dL.

Because it is difficult to view many of the individual counties within the U.S.-level maps, regional-level maps also were produced. Figure 6-2 contains examples of these for EPA Region V. Comparable maps for all regions are included in Appendix G. With darker colors representing areas of higher lead levels, it appears that lead levels are declining across EPA Region V from the 2000 to 2005 time period. Figure 6-3 contains maps of observed and predicted proportion of children's blood-lead levels in Massachusetts at the census-tract level. The Boston area is enlarged to better show the tracts in that area.

6.2 Visualization Tool Development

In addition to generating maps, a software tool was developed to provide a flexible way for users to quickly view data for particular areas and to obtain information that led to the results being viewed. To do this, the project team utilized existing technology developed through internal research and modified this technology to meet the needs of this study. The software sews together a series of static maps so that they can be viewed dynamically. This allows users to view a movie of changes in surfaces over time and space.

The software is written in C++. Users interact with the software via a Windows GUI that is implemented using Microsoft Foundations Classes (MFC). The 3-dimensional graphics within the tool were implemented using an Open Graphics Library (OpenGL).

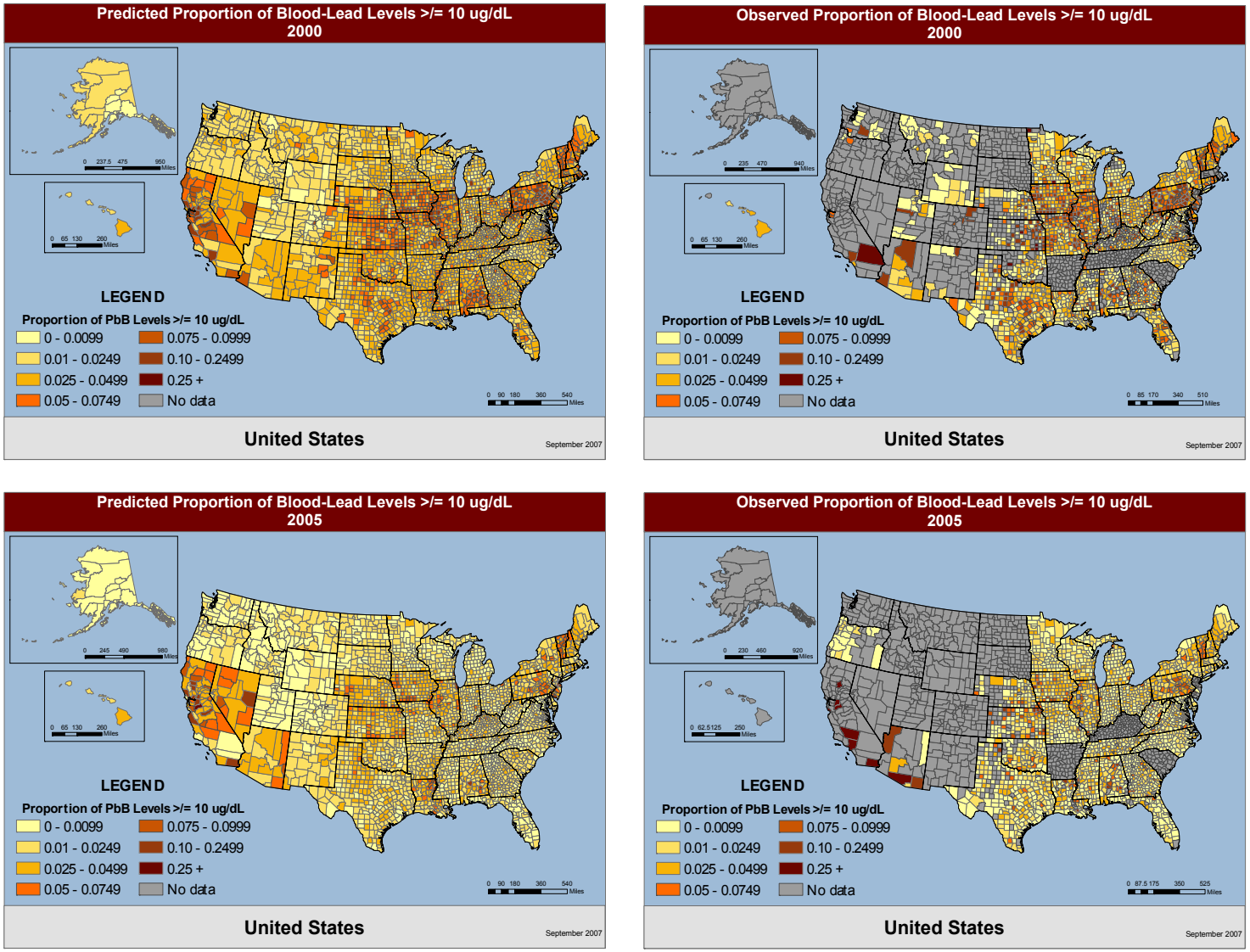


Figure 6-1. Observed and Predicted Proportion of Children with Blood-Lead Levels ≥ 10 µg/dL in the United States by County, 2000 and 2005

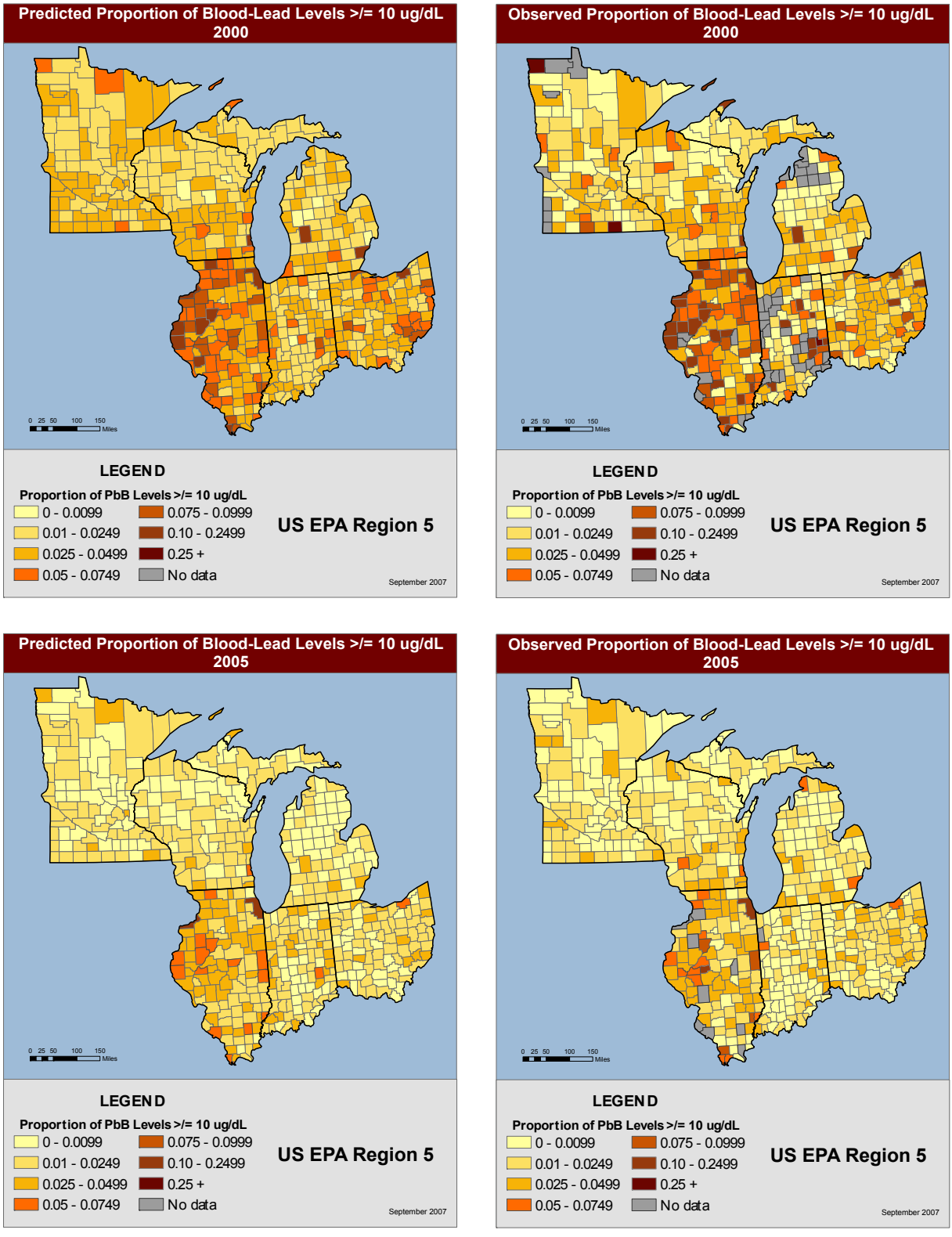


Figure 6-2. Observed and Predicted Proportion of Children with Blood-Lead Levels ≥ 10 µg/dL in Region V by County, 2000 and 2005

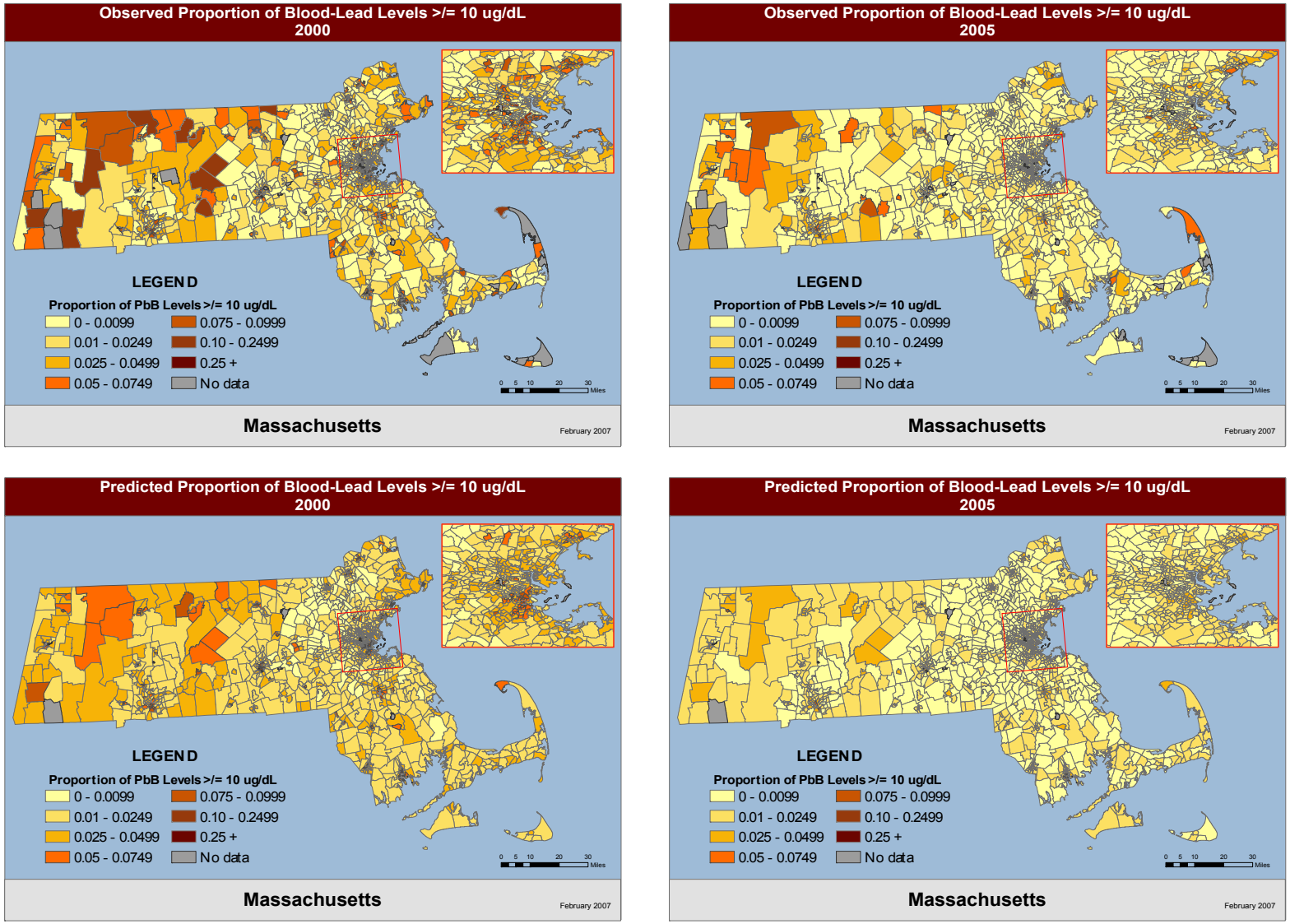


Figure 6-3. Observed and Predicted Proportion of Children with Blood-Lead Levels ≥ 10 µg/dL in Massachusetts by Census Tract, 2000 and 2005

The software visualizes the observed values and the predicted values for the response variable of each model (6 national models, 5 Massachusetts models). The software interpolates the predicted values spatially within each state using a squared inverse distance algorithm; it interpolates linearly in time. The predicted values are defined for each county. There are two visualization modes: (1) a spatial surface moving in time, and (2) a time series. The tool was built in a flexible way so that it can be easily adapted to accept updated data.

Figures 6-4 and 6-5 are screen shots from the visualization tool. Figure 6-4 provides an example of a response surface generated by the tool to illustrate predicted blood-lead levels across a geographic area. In this example, the area is the state of Illinois. Figure 6-5 provides an example of a method the visualization provides to plot predicted blood-lead levels in a given geographic area over time.

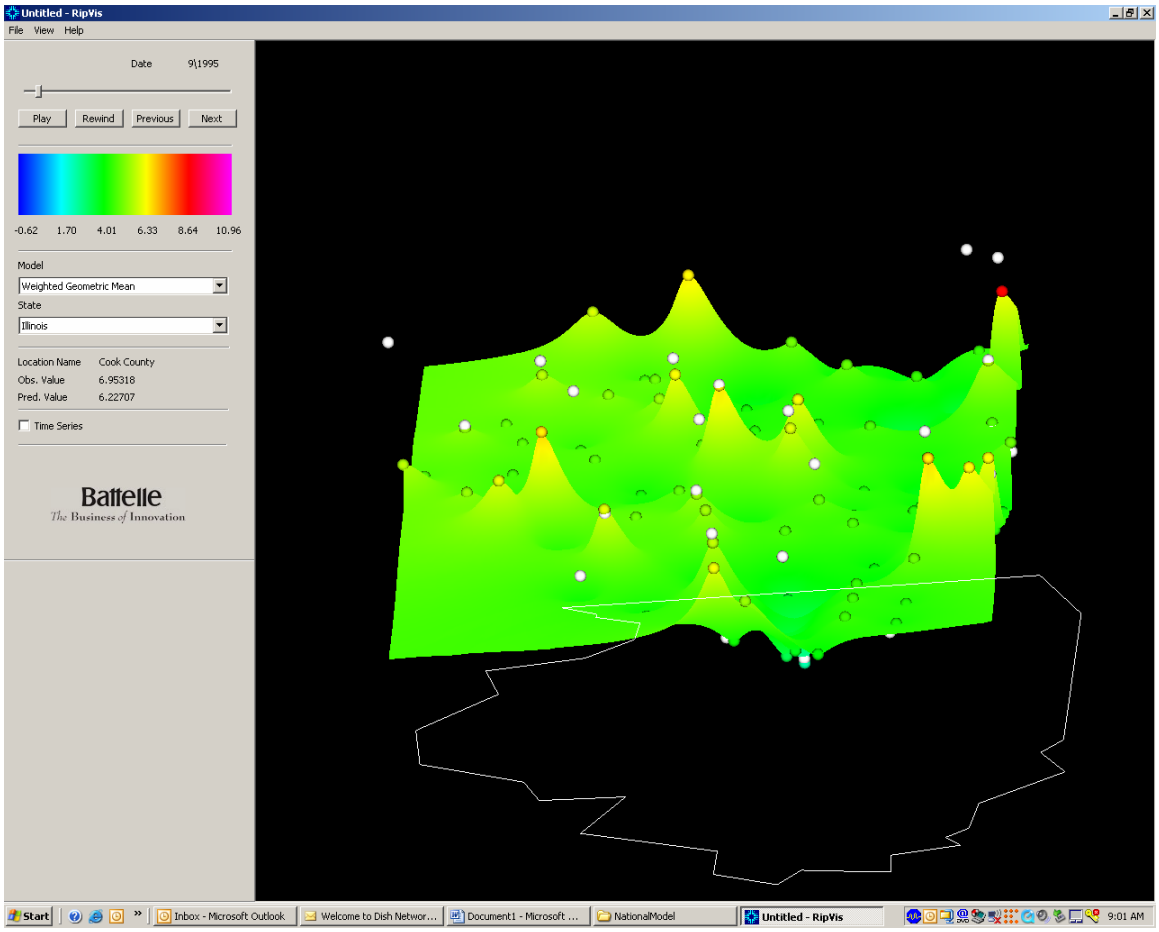


Figure 6-4. Response Surface of Predicted Geometric Mean Blood-Lead Concentration Across the State of Illinois from the Visualization Tool

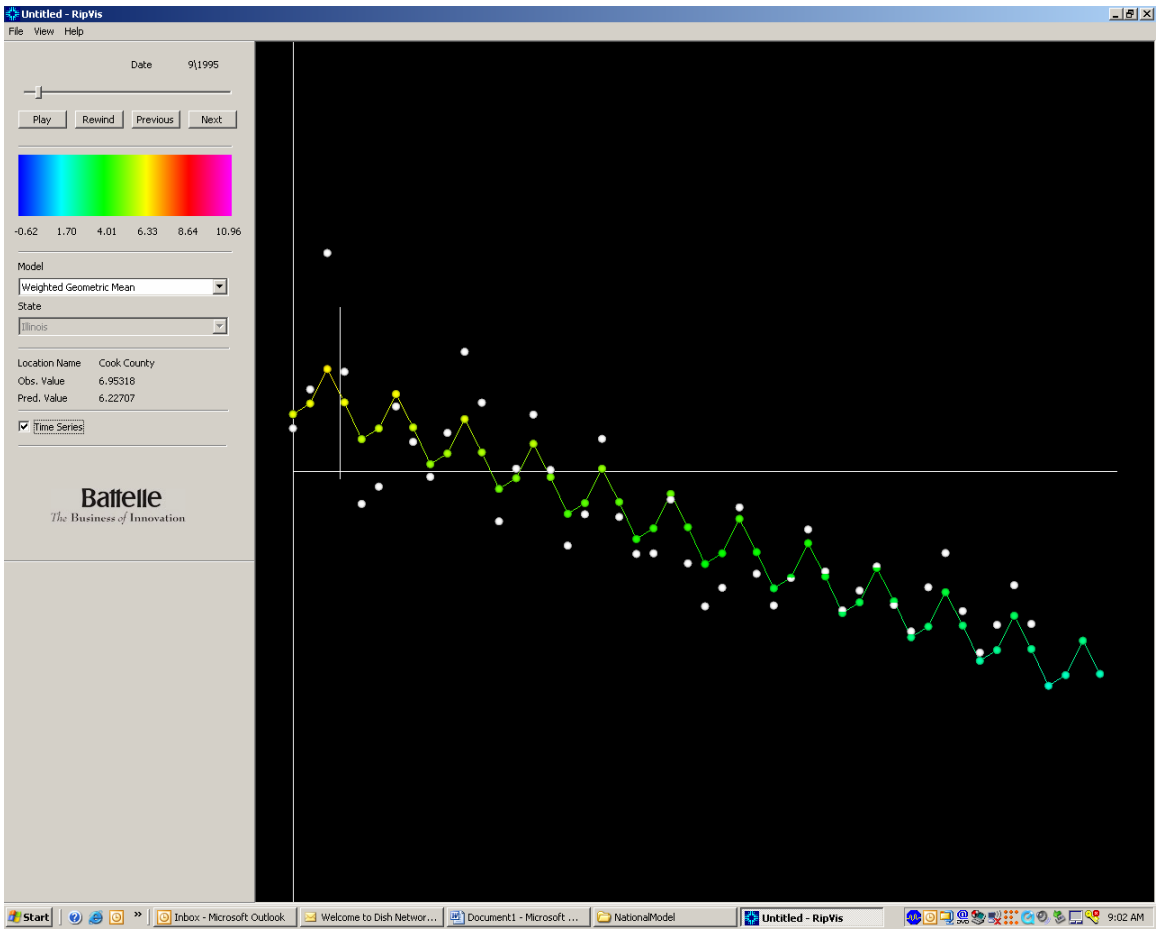


Figure 6-5. Time Series Plot of Observed and Predicted Geometric Mean Blood-Lead Concentration in Cook County Illinois from the Visualization Tool

7.0 DISCUSSION AND FUTURE WORK

The goal of this study was to determine whether tools could be developed to differentiate geographic areas (counties and census tracts), based on their predicted risk of containing children with elevated blood-lead levels. Statistical models were developed that link CDC's childhood blood-lead surveillance data to demographic predictor variables available in the 2000 U.S. Census. While earlier chapters of this report focus on the development and performance of these statistical models, this chapter provides a discussion of the factors that should be considered when using the models, and some preliminary ideas for improvement.

7.1 Major Findings

The results of this study suggest that longitudinal predictive models can be developed at the county level across the nation based on the use of quarterly summary information from CDC's National Surveillance Database, and at the census-tract level within states that have a long history of universal screening and reporting, such as Massachusetts. These models can be used to describe how risk of childhood lead poisoning changes over time within different regions of the country, as well as within small geographic areas within states (e.g., counties) and even smaller geographic areas within counties (e.g., census tracts). They can be used to predict the risk of childhood lead poisoning in counties (or census tracts) with little or no surveillance data, and also can be used to identify those counties (or census tracts) that are at highest risk at the end of the period of observation (see Appendix F for a list of the 150 counties across the country at highest risk predicted by each of the six models, as well as the top 10 counties within each state).

The statistical model chosen (a random-effects model with separate intercepts and slopes estimated within each county or census tract) also allows ranking of geographic areas based on the rate of decline over time after accounting for the fixed-effects variables of the model (although only among those areas that provided adequate surveillance data). Within the context of the Broad-Based National Model, these random effects would allow us to identify those counties that are experiencing a more rapid reduction in risk of childhood lead poisoning over time (to identify best practices) and those counties that are experiencing a significantly less rapid decline over time (to identify areas in need of additional attention and resources for combating lead poisoning), after already accounting for the demographic, programmatic, and environmental factors included in the multivariate model.

Within the context of the series of Broad-Based National Models, the data suggest that there are significant differences in the distribution of childhood blood-lead concentrations among the different regions of the country, and that the manner in which these distributions change over time and are impacted by seasonality also is regionally specific. After accounting for these regional differences, a number of demographic, environmental, and programmatic variables were found to be highly predictive of childhood blood-lead concentrations among the different response variables modeled within this project. The specific variables that were found to be predictive within the multivariate models varied based on the response variable; however, there were certainly some variables that were found to be selected in multiple models. In addition to various census demographic variables that were identified in previous risk modeling efforts (e.g., age of housing, percent single parent families, race/ethnicity), it was found that variables constructed from EPA's Safe Drinking Water Information System, time-lagged programmatic

funding information from HUD and/or CDC, and variables associated with high lead emmisions or predicted air concentrations were selected within the National (Low Resolution) multivariate statistical models.

Within the context of the High-Resolution Model developed using data from the Commonwealth of Massachusetts, the project team also found a highly significant downward trend in the risk of childhood lead poisoning among the five models developed. Due to a very small number of children observed at or above 25 µg/dL within Massachusetts over the 2000-2006 period of observation – this sixth model was not included. After accounting for the long-term reduction over time and seasonality using similar methods that were employed in the Broad-Based National Model, we found that only the demographic and programmatic variables were predictive of the risk of childhood lead poisoning at the census-tract level. Of particular interest were the variables that described the proportion of housing units within each census tract that were found to be in compliance and out of compliance with the Massachusetts Standard of Care. In all five of the multivariate models, the risk of childhood lead poisoning was significantly reduced as the proportion of housing units in compliance increased within a census tract. In addition, for the last two models (which predicted proportion of children at or above 10 and 15 µg/dL), the risk of childhood lead poisoning increased significantly as the proportion of housing units out of compliance increased within a census tract.

7.2 Comparison Between Results and NHANES

Due to selection bias associated with surveillance data, it is expected that the CDC National Surveillance dataset as well as the Massachusetts surveillance data may show higher proportions of elevated blood-lead concentrations than found in the general population. For this reason, the proportion of children with elevated blood-lead concentrations as well as the distribution of the potential continuous summary measure derived from the surveillance data were compared with those reported by the most recent six years of available CDC National Health and Nutrition Examination Survey (NHANES). Results of this comparison are presented graphically in Figure 7-1 – suggesting that there is a highly significant difference between the NHANES and CDC's National Surveillance Database with respect to the proportion of children observed at or above 5 µg/dL (with lesser differences observed for the proportion of children observed at or above 10, 15, and 25 µg/dL). In future work on this project, EPA might consider methods for calibrating the Surveillance data to better match the National Distribution of childhood blood-lead concentrations using methods similar to those employed by Strauss, et. al. 2001a.

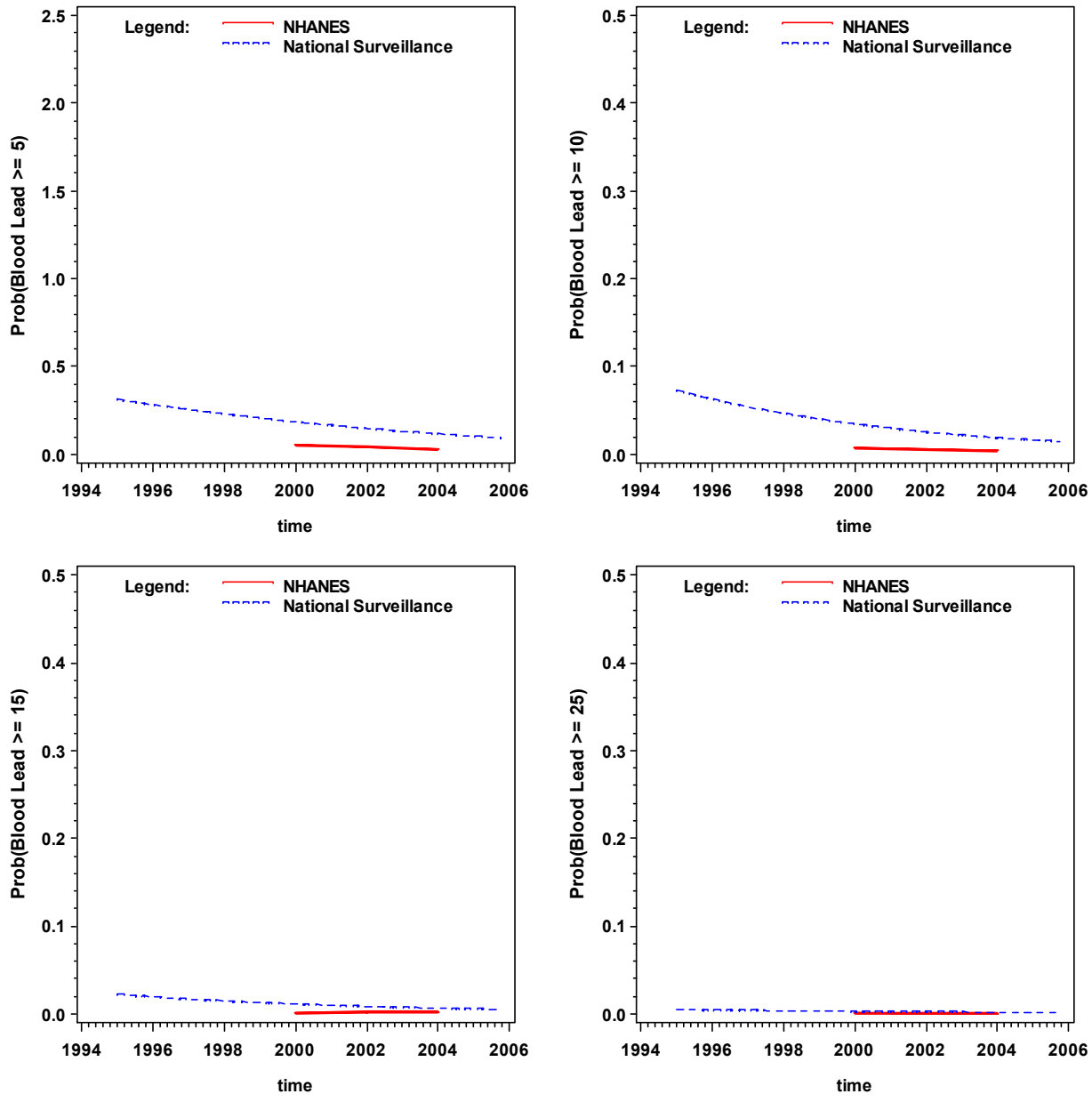


Figure 7-1. Comparison of National Surveillance Data to NHANES Data

7.3 Data Issues

The models that were developed as part of this project are based on data sources that have both strengths and limitations. In this section, four potentially limiting aspects of the data are considered – biases from the geocoding process, biases inherent in the surveillance data, reporting limits in surveillance data, predicting within-area relationships with ecological models, and use of Census data from 2000.

7.3.1 Biases from Geocoding

The quarterly summary statistics from CDC's National Surveillance Database utilized in these analyses were available at the county level of geographic specificity. CDC based this summarization on county FIPS codes reported by its grantees. This field is quite well reported in CDC's CBLS database. The Massachusetts surveillance data was summarized and analyzed at the census-tract level, with the geocoding of address data within the Massachusetts data being conducted by MDPH staff. While there is no reason to suspect lack of data quality within the Massachusetts surveillance data, experience shows that the process of geocoding can introduce some subtle biases into surveillance data. Thus, the following section is offered as a guide for EPA to consider for future modeling efforts in which state or local surveillance data are geocoded to the census-tract level:

The geocoding process is highly dependent on the quality of address data recorded by the local lead poisoning prevention programs with whom the blood-lead information originated. Several factors could prevent an address from being successfully geocoded, such as:

- Erroneous, illegible, or purposefully misleading address information being provided to the childhood lead poisoning prevention program
- Address data that contain either a P.O. Box or Rural Route as part of the street address, which typically cannot be successfully geocoded
- Errors in data entry.

While these problems with address data are likely to occur in all programs with a non-trivial frequency, there may be a systematic bias that programs introduce (albeit unintentionally) when correcting address data. It is likely that address data errors are identified and corrected with higher frequency for children who have an elevated blood-lead level and require follow up.

Given the potential bias introduced through the geocoding process, further research may be worthwhile to determine whether there are reasonable approaches that could be used to adjust the models for this bias.

7.3.2 Reporting Limits in Surveillance Data

Other naturally occurring biases in the surveillance data may influence the degree to which models are representative of the true trends in childhood lead poisoning. For example, within the context of the Broad-Based National Model, there may be differences between states and localities in the manner in which childhood blood-lead testing results are reported to CDC. Sections 2 and 3 included a discussion about a screening algorithm that was applied to the

surveillance data to supress county/quarter data combinations for areas that were not conducting universal reporting of blood-lead testing results. Additional scrutiny of these data by CDC and other members of the lead poisoning prevention community may reveal other county/quarter combinations that were not identified through the screening algorithm that should be excluded from these analyses. We are confident that the overall impact of including these data in the current work will not severely bias the fixed-effects parameter estimates in the series of generalized linear mixed models developed in this project.

7.3.3 Selection Bias in Surveillance Data

Selection bias is perhaps the most serious bias that is yet left unaccounted for in the models that have been developed, and may have severe impact on their predictive ability. Surveillance data are observational by nature, and are not designed to be representative of the general population. There are many competing forces that influence whether or not a child is screened at an appropriate age, and recorded in the blood-lead surveillance database. Some have hypothesized that surveillance data in the urban environment are representative of the affluent (who have private health insurance) and the poor (who receive Medicaid or other medical assistance), while under-representing the working poor (who may have no health insurance, and no mechanism for receiving appropriate preventive medical testing). While this may be true in general, many outstanding lead poisoning prevention programs currently are extending outreach, education, and screening services to areas with historically high incidence of childhood lead poisoning. These programs generally provide assistance to all members of the community, regardless of entitlement status. While these services are typically offered in high-risk urban areas with the infrastructure of a federally funded (CDC and/or HUD) or state-funded lead poisoning prevention program, they typically are less available in similar high-risk rural areas without similar infrastructure. In addition to outreach, education and screening activities, many childhood lead poisoning prevention programs (or partnering housing agencies) receive funding from HUD's Office of Lead Hazard Control to conduct environmental investigations and reduce lead hazards in the residential environment. Many of these activities generate targeted screening of children living in deteriorated, older housing – which also is a non-trivial source of selection bias in the surveillance data.

An important question for EPA to address is how selection bias is likely to influence the relative rankings of counties within a region or census tracts within a more localized area, as well as the predictive ability of the models themselves.

7.3.4 Limitations of Ecological Models for Predicting Within-Area Relationships

The models that were developed within this project are ecological models that describe quarterly distributional summary statistics within geographic areas as a function of predictor variables assessed within those same geographic areas. It also may be the case that some of these predictor variables have significant variation within a county (or census tract) – and that this within-area variation is highly predictive of risk of childhood lead poisoning within these geographic areas. Unfortunately, the data limitations within this study (for both the blood-lead response variables as well as many of the predictor variables) prohibit us from ascertaining these important person-level relationships. This type of relationship can be established only by linking

individual blood-lead concentration data with individual-level environmental, demographic, and/or programmatic information (which usually is not available).

Within the context of the High-Resolution Massachusetts Model – it may be possible to link individual blood-lead records with the longitudinal housing inspection information to assess the loss of information associated with going from an individual-level model to an area-based ecological model. This type of assessment could be introduced in later stages of this project.

7.3.5 Use of 2000 Census Data and Other Time Invariant Data as Predictors

One potential criticism of the modeling effort is that we are linking blood-lead surveillance data collected between 1995 and 2006 to census data that were collected in 2000. Is the demographic information collected in 2000 likely to remain unchanged over the course of time? The answer probably depends on the variable under consideration. For example, age of housing in census tracts or proportion of housing built prior to 1950 is not likely to change dramatically in census tracts, unless there is a lot of demolition or new construction occurring. On the flip side, average income is likely to change substantively over time.

Even though the demographic information contained in the 2000 Census is likely to change over time, the more important question is what effect will that change have on our model predictions? While the models likely would be improved with the use of more current census data for use as predictors, we do not believe that the use of older (less current) information will result in poor or inaccurate prediction. In fact, for the purpose of predicting current or future trends in childhood lead poisoning, we are more concerned with the age of the surveillance data that are being used as the response variable in this modeling exercise than with the age of the predictor variables.

Similar arguments can be made for the use of static air modeling data, and averaged information from EPA's Toxics Release Inventory.

7.4 Model Validation Issues

The risk index models developed as part of this project may require validation before being used by childhood lead poisoning prevention programs throughout the country. The following four issues might be considered by EPA as being important to address as part of this validation exercise:

1. Within counties and/or census tracts that contribute blood-lead information to the models, how representative is the screened population of children (on which the models are based) of the general population of children?

2. Within counties and/or census tracts that do not contribute much information to the models (e.g., counties with low screening penetration), how well does the model perform at predicting relative risk and blood-lead distributions?

3. Can risk index models based on historical blood-lead data from 1995 through 2005 accurately predict risk and blood-lead distributions in future years (e.g., can it be used to forecast towards the federal 2010 goal)?

4. Can the High-Resolution Model developed in Massachusetts be generalized to predict risk and blood-lead distributions in other states across the Nation (or even within EPA Region 1)?

If EPA is to provide childhood lead poisoning prevention programs with a risk characterization tool based on these models, a comprehensive validation should be pursued to address the above four issues.

Validation of the Surveillance Data

The first issue is related to the quality of the data supporting model development. For example, if CDC's surveillance data are biased toward inclusion of high-risk children (as shown in the comparisons to NHANES), the risk index models also will be biased and tend to over-predict children at high risk. Note that if the bias is consistent among all counties and census tracts (i.e., it over-represents high risk children everywhere), the model predictions for the proportion of children in each blood-lead category likely will be biased, while the ability for risk indices to differentiate between high- and low- risk areas will be preserved. If the biases occur differently in different areas, non-trivial adjustments to the model would need to be pursued prior to use by childhood lead poisoning prevention programs.

Because the unit of analysis in the development of the Broad-Based National Model is at the county level, the goal of a validation exercise would be to determine whether the distribution of children's blood-lead concentrations that are included in the surveillance data for a sample of census tracts are representative of the general population of children found within those census tracts. One possible approach, would be to develop a field testing validation survey, in which a stratified random sample of counties are selected for a short-term outreach campaign in which eligible children are sampled in a representative manner. Stratification variables to be considered would be Rural/Suburban/Urban, predicted level of risk from the model, and possibly levels of socio-economic status. Obviously, development of such a survey would be costly, difficult to implement, and likely beyond the scope of this project. Alternatively, CDC might be able to reveal the specific counties that participated in various waves of NHANES – with comparisons being made in those specific counties. Access to the identification of the specific counties from which NHANES study subjects were sampled (within the NHANES analysis dataset) would provide this project with the best foundation to address the serious biases identified in Section 7.2 and calibrate the model to ensure that it is more reflective of the U.S. population.

Validation of the Models in Areas with Low Screening Penetration

This second issue relates to the performance of the risk index models in predicting both relative risk and the number of children in different blood-lead categories in the census tracts that historically had low screening penetration. Due to the fact that there is little to no data in these geographic areas to determine the fit of the risk index models, some field studies similar to the one described in the previous section would need to be conducted to address this issue. The major difference between the two field studies is that the census tracts chosen for this validation exercise would be tracts in which the screening penetration is low.

A similar approach could be used to conduct this field validation exercise in which a stratified sample of counties would be identified for the study, and a representative sample of children's blood-lead levels would be obtained within those census tracts using an intense, brief outreach effort. The counties again would be chosen using a stratified random sampling approach, to obtain a sample of tracts that represents a combination of high-, medium- and low-risk areas in the rural, suburban, and urban environments. This is an area for potential future collaboration with CDC and perhaps some of their lead poisoning prevention grantees.

Validation of the Models in Predicting Future Blood-Lead Concentrations

Validation of this third issue can be performed to a certain extent using data that are already available as part of the modeling process. For example, in the national model where data are available from 1995 through 2005, data for a state or set of states can be removed for one or more years and the missing data predicted by the model. If all 2005 data were removed, models would be developed using the data from 1996 through 2004, and then the "future" predictive ability of those models can be assessed by applying them to the data from 2005.

Validation of the Models in Predicting Blood-Lead Concentrations in Other Geographic Areas

The last type of validation involves the determination of synergies (or lack thereof) in prediction between the Broad-Based National Model and the High-Resolution Model. Conceptually, we should be able to aggregate the modeling predictions from multiple census tracts within a county from the low-resolution model and match the county-level predictions from the Broad-Based National Model. Due to the fact that the National Model and Massachusetts Models were developed independently, using different data sources for the surveillance data (CDC and MDPH), and utilizing different predictor variables – these synergies may not exist.

Further work on integrating the Broad-Based National Model with the High-Resolution Model (or multiple high resolution models if EPA is successful at expanding this project to include multiple additional programs) can be done by fitting these two types of models jointly under the concept of hierarchical linear modeling. This type of model, while more sophisticated and computer-intense, can be developed using specialized software under a Monte-Carlo Markov Chain Bayesian formulation.

7.5 Other Recommendations for Immediate Future Work

The previous sections within Chapter 7 focus on various important issues related to the development of models to predict risk of childhood lead poisoning at the geographic level, including calibration to the nationally representative trends over time observed in NHANES, assessment of the potential impact of a variety of important biases and other data quality issues, and various model validation exercises that can be explored. EPA also has been including other state and local lead poisoning prevention programs as part of the project conference calls in anticipation of developing additional High-Resolution Models as part of follow-up work to this project. While these are all worthy tasks to pursue as part of future work, there are some additional analyses that the project team would recommend pursuing on the Broad-Based National Model as well as the High-Resolution Model within Massachusetts prior to approval of this report as a final report. These activities include the following:

- Broad-Based National Model
    - CDC grantee relationship managers may have insight into data quality issues (such as the previously discussed laboratory minimum reporting values, and not following universal reporting guidelines) for specific geographic areas and periods of time. Additional scrutiny of these data could be used to improve the quality of the blood-lead response data that serves as a basis for these models. The maps and visualization tool should help foster this review of the data.
    - In addition to the above data review –further investigation into using urban vs. rural status as a potential effect modifier in the analyses also is recommended. Differentiating between urban and rural areas can be conducting in numerous ways, including:
        - Determining whether the county is part of a Metropolitan Statistical Area within the 2000 US Census
        - Identification of the counties that contain the U.S. top 100 (or 200) cities based on population size
        - Use of a population density score (with a cut-off value).

      Use of this variable as a potential effect modifier might include fitting separate intercepts and slopes for the effects of time and seasonality within the different regions of the country, as well as the potential for using different environmental, programmatic, and demographic predictor variables in these two area types in the multivariate predictive models.
    - Once the proper way of handling the potential effect modifier for rural versus urban areas – the exploratory analyses that assess the predictive ability of each candidate environmental, programmatic, and demographic variable could be refit in a manner consistent with the baseline effects that will be included in the model. Thus – rather than assessing the predictive ability of a candidate variable after adjusting for the downward trend of time, it should be assessed after adjusting it for region, region*time, region*seasonality, and potentially region*urban/rural.

High-Resolution Model in Massachusetts

- Due to the fact that we know that Massachusetts followed universal screening and reporting guidelines during the entire period of observation (2000-2006), and the fact that these data have been used previously to support federally funded research projects – there is less concern about some of the previously mentioned data quality issues. This does not mean that the Massachusetts data are not potentially biased or flawed, as there are still probable selection biases and potential geocoding biases that were introduced into the analysis dataset that supports the High-Resolution Model. Our collaborators at the Massachusetts Department of Public Health are invited to review and comment on this work, and add their insight and experience in making recommendations on additional ways of handling the various data sources that were integrated into this model.

- It also is recommended that comparisons be made between the observed and predicted data from the Broad-Based National Model for counties in Massachusetts (based on the input data received from CDC) with the observed and predicted data from the High-

Resolution Model (based on the input data received from MDPH) by aggregating the observed and predicted census tract data within Massachusetts to the county level.

- Finally, pursuit of some additional analyses of the individual-level data from MDPH is recommended – by linking individual blood-lead testing results on children over time to the housing inspection results (as well as other census-tract level predictors that were used in the current High-Resolution Model). This will help identify the degree of information loss experienced by pursuit of the ecological models of aggregate summary data.

8.0 REFERENCES

24 CFR Part 35; 40 CFR Part 745, Lead;. Requirements for Disclosure of Known Lead-Based Paint and/or Lead-Based Paint Hazards in Housing; Final Rule (3/6/1996). Accessed at http://www.leadsafehomes.info/pdfs/all_titleten_fulltext_english.pdf#search=%22HUD%201018%20Rule%22

40 CFR Part 745, Lead;. Identification of Dangerous Levels of Lead; Final Rule (1/5/2001). Accessed at http://www.epa.gov/fedrgstr/EPA-TOX/2001/January/Day-05/t84.pdf

Battelle. Draft Quality Management Plan for the Targeting Elevated Blood-lead Levels in Children Pilot Study. February 2007.

CDC. 1997. Screening Young Children for Lead Poisoning: Guidance for State and Local Public Health Officials, edited by U.S. Department of Health and Human Sevices. Atlanta GA: Public Health Sevices, CDC.

HUD. 1995. The Relation of Lead Contaminated House Dust and Blood-Lead Levels Among Urban Children. Washington DC: U.S. Department of Housing and Urban Development.

Lanphear, BP, TD Matte, J Rogers, RP Clickner, B Dietz, RL Bornschein, P Succop, KR Mahaffey, S Dixon, W Galke, M Rabinowitz, M Farfel, C Rohde, J Schwartz, P Ashley, and DE Jacobs. 1998. The contribution of lead-contaminated house dust and residential soil to children's blood-lead levels. A pooled analysis of 12 epidemiologic studies. *Environ Res* 79 (1):51-68.

Miranda, ML, DC Dolinoy, and MA Overstreet. 2002. Mapping for Prevention: GIS Models for Directing Childhood Lead Poisoning Prevention Programs. *Environmental Health Perspectives* 110 (9):947-53.

Miranda, ML, JM Silva, MA Overstreet Galeano, JP Brown, DS Campbell, E Coley, CS Cowan, D Harvell, J Lassiter, JL Parks, and W Sandele. 2005. Building Geographic Information System Capacity in Local Health Departments: Lessons from a North Carolina Project. *Am J Public Health* 95 (12):2180-5.

Spivey, Angela. The Weight of Lead: Effects Add Up in Adults. *Environmental Health Perspectives* Volume 115, Number 1, January 2007.

Strauss, Warren, R Carroll, Steve Bortnick, John Menkedick, and B Schultz. 2001a. Combining Datasets to Predict the Effects of Regulation of Environmental Lead Exposure in Housing Stock. *Biometrics* 57:203-210.

Strauss, Warren, Ramzi Nahhas, Leanna House, Amy Kurokawa, and Bradley Skarpness. 2001b. Development of Models to Predict Risk of Childhood Lead Poisoning at the Census Tract Level. Columbus OH: Technical Report to the U.S. Centers for Disease Control and Prevention under Contract No. 200-98-0102.

Strauss, Warren, Tim Pivetz, P Ashley, John Menkedick, E Slone, and S Cameron. 2006. Evaluation of Lead Hazard Control Treatments in Four Massachusetts Communities through Analysis of Blood-lead Surveillance Data. *Environmental Research* 99 (2):214-223.

U.S. Department of Housing and Urban Development. September 15, 1999. Final Rule, Requirements for Notification, Evaluation and Reduction of Lead-Based Paint Hazards in Federally Owned Residential Property and Housing Receiving Federal Assistance. Washington DC: Federal Register, 50140-50231.