IV-A-2
4.

The MITRE Corporation
Metrek Division
1820 Dolley Madison Boulevard
McLean, Virginia 22102

WP- 81W00541 (No.) ___ (Vol.) ___ (Series) 1 (Rev.) ___ (Supp.) ___ (Corr.)

Subject: Statistical Analysis of Blood Lead Levels Among Children in Louisville, Kentucky

To: L. Duncan

From: J. Watson

Dept.: W53

Date: 30 June 1981

Contract No.: 68-01-5051

Sponsor: EPA

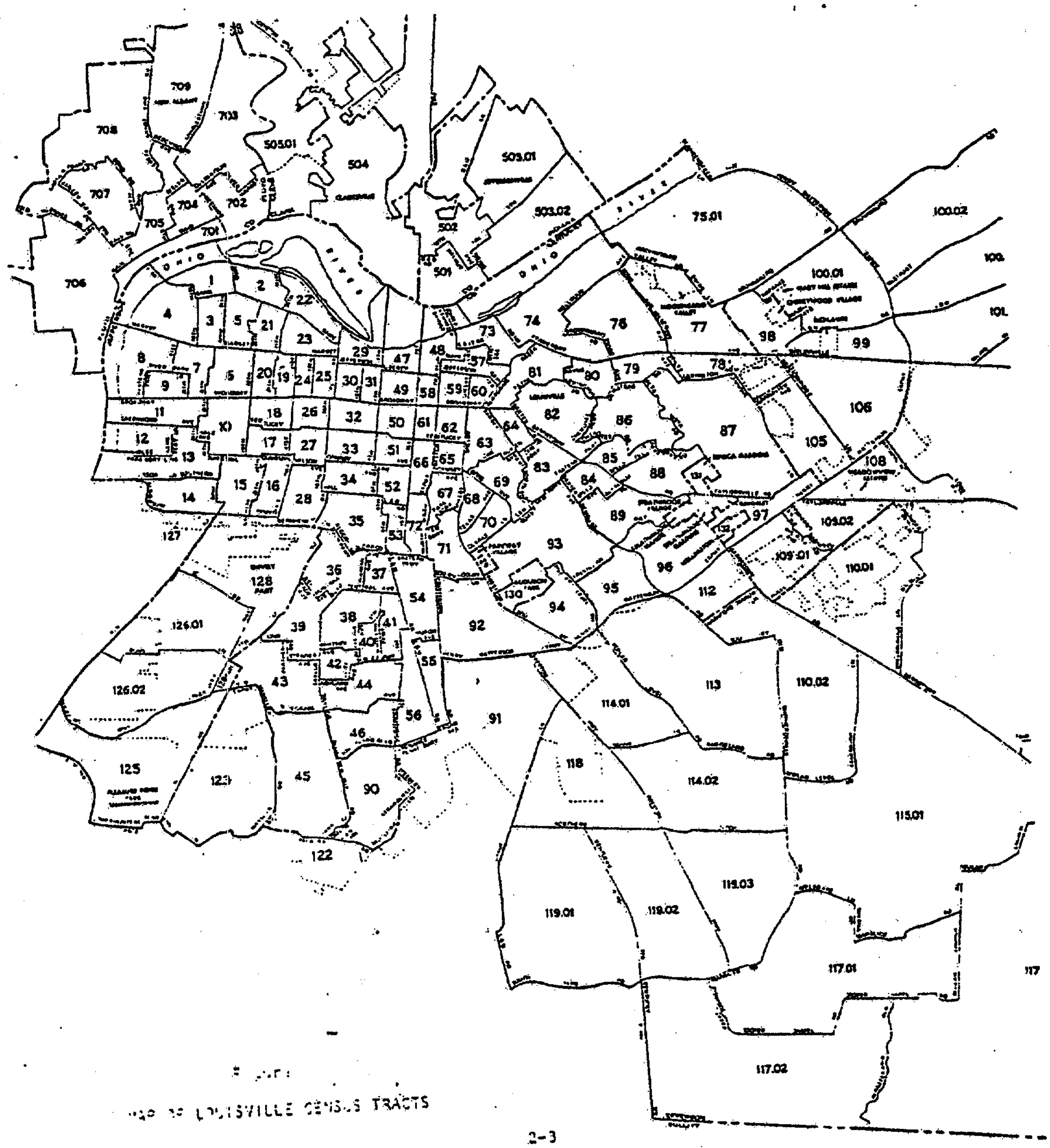
Project No.: 1672Y

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF TABLES (Concluded)



Map of Louisville Census Tracts

1.0 INTRODUCTION

1.1 Purpose and Scope

This paper reports the results of statistical analysis performed by The MITRE Corporation on blood lead levels indicated in initial tests of children in Louisville, Ky., during the period 1974 to 1979. The work represents a follow-up of that reported earlier in MITRE Working Paper WP-80W00466 and is based on more comprehensive test data provided to MITRE by Dr. Irwin Billick of the U.S. Department of Housing and Urban Development (HUD). The study was supported jointly by HUD and the Environmental Protection Agency.

As in the preliminary study, the data were formatted by MITRE and then MITRE made computer runs at the MITRE Washington Computer Center to analyze the data. For these FORTRAN was used to a limited extent; chiefly, however, the Statistical Package for the Social Sciences (SPSS) was used (Nie, et al., 1975). A second line of analysis (being reported separately) investigated blood-lead levels in children repetitively tested over a period of up to 5 years.

The objective of the analysis was to investigate the relationship of variations in geometric mean blood lead levels in the children tested to variations in other factors considered possibly associated. Of particular interest was variation in the quantities of gasoline lead consumed in the Louisville area, this being of especial concern to EPA's Mobile Sources Enforcement Division. Other independent variables investigated in the analysis were race (black and white); residence (as indicated by census tract); age; and both time period and quarter or season of the year during which the blood tests were made. Specifically, the following information was developed:

a. A complete set of correlation and regression coefficients relating blood lead levels quantitatively to the independent variables investigated;

b. Assessment from statistical evidence of the possible role that the various factors might play in variations observed in blood lead levels;

c. Comparison of the statistical results for the Louisville data with results previously obtained from less complete Louisville data and from very extensive data from New York City (Billick et al., 1979; MITRE 1980, 1980a).

Major conclusions of the statistical analysis are summarized in the following section of this report. Section 2 explains the nature of the data and of the computer runs. The principal statistical results from which the conclusions were derived are then presented and discussed (Section 3). Other statistical data are given in the appendix.

1.2 Major Conclusions

a. Variations in gasoline lead account for up to 45 percent of the observed variations in geometric mean blood lead levels in the Louisville children tested. This is by far the strongest interrelationship which any of the variables examined show with blood lead but the correlation is less that that observed in New York City test data.

b. As with the New York City data, the statistical results for Louisville are more consistent with the hypothesis that variation in gasoline lead levels principally explains variations in geometric mean blood lead level observed during a particular period of the year, rather than with a theory that some other seasonal factor largely accounts for blood lead variation.

c. The complete set of variables incorporated into the analysis statistically explains at most about 55 percent of the variations observed in geometric mean blood lead level. This finding implies that compared to the regression model derived for New York City (MITRE, 1980), relatively poor results would be expected from trying to fit a regression equation to available data so as to project possible future blood lead levels as a function of gasoline lead and other relevant variables.

d. A striking feature of this more comprehensive set of Louisville data was the fact that both gasoline lead and blood lead averages declined markedly after 1977 (a year for which no usable test results were available). Statistical findings from the post-1977 data exclusively were consistent with an hypothesis that after this time gasoline lead exerted a strong influence on blood-lead variations, rather than that both levels had declined independently.

e. Were obtained from the user of census tract data in that this additional information failed to provide any significant basis for interpreting the role which factors related to place of residence might have in accounting for blood-lead. Incorporation of residence (as indicated by groupings of contiguous census tracts) into the regression models did not increase the fraction of observed variations in geometric mean blood-lead levels thereby explained. Only a few of the census tract groupings showed a significant statistical interrelationship with geometric mean blood levels; most correlated very weakly. These findings indicate that in general whatever factors besides gasoline lead are associated with blood-lead variations cannot be determined from place of residence within the available data base. A small number of census tracts were identified as associated with unknown factors related to blood-lead levels that tend to be significantly either higher or lower than averages for the metropolitan area of the whole, other things (i.e., age, race, and season of the year) being equal.

2.0 NATURE OF THE COMPUTER RUNS

2.1 General Characteristics of the Data Used

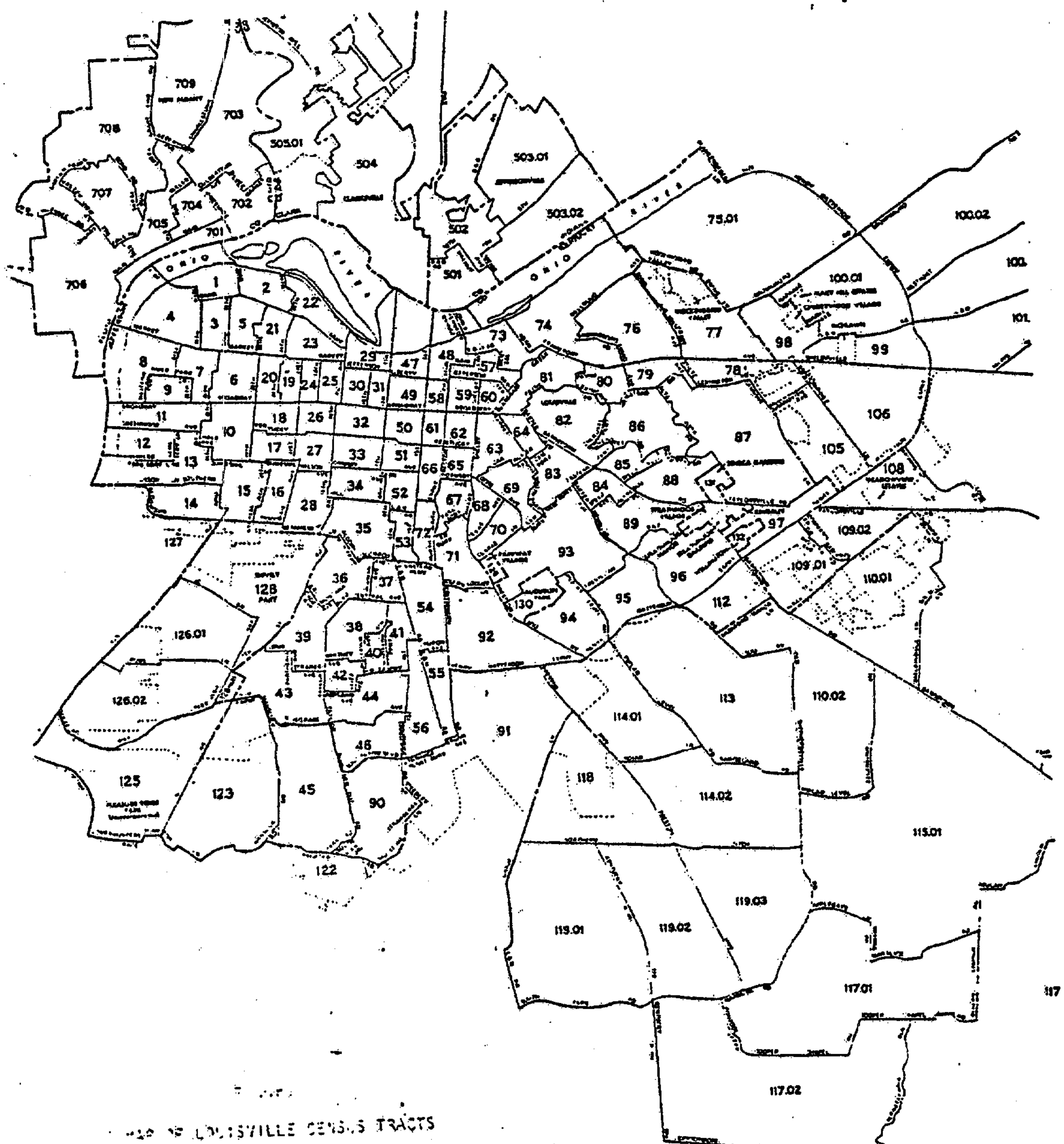
The statistical analysis reported here was based on computer runs made by MITRE at the Washington Computer Center. Dr. Billick of HUD provided a tape containing records of blood tests taken in Louisville, Ky., during the period 1974 through 1979. MITRE developed a computer program (using FORTRAN) to convert the basic data to a format convenient for subsequent organization into the files required for the SPSS programs. The FORTRAN program enabled the test data to be initially screened to eliminate incomplete results and/or those found to be in error on the basis of validation criteria deterministically applied (e.g., reported birth date not earlier than the test date; reported test results outside the possible range of values, etc.).

In all, 12,913 records of blood-lead levels from the years 1974-76 and 1978-79[1] were found initially valid for analysis. Each record represented the initial test performed on a different child. (Repeated test results were handled separately in a study reported elsewhere.) This number of unique test results comprised an increase of nearly 50% over the body of similar test data used in the previous

[1]Results for 1977 were omitted because changes in selection of the sample for testing made the blood-lead data unsuitable for statistical analysis with data from other years.

analysis of Louisville data (MITRE, 1980a). About 70 percent of the children tested were black and nearly 30 percent white, while records of other and unknown races constituted less than 1 percent of the total. Tests were heaviest in the years 1975, 1976 and 1978 (comprising over 70 percent of all records). Ninety-eight percent of the children tested were 6 years of age or younger, with age ranges 1 and 4 (i.e., children up to 12 months and children between 37 and 48 months respectively) making a combined total of 53 percent. Testing was heaviest in the first quarter of the years (36.9 percent for the months of January-March) and lightest in the third quarter (16.3 percent). Residence of each test subject as to census tract was indicated on each record (see map reproduced as Figure 1). Those census tracts within the inner city (i.e., towards the north and west up to the Ohio River) contained the heaviest concentration of test results.

Data within many of the tracts were too sparse to permit a comprehensive analysis by individual census tract. Therefore, two sets of runs were made, distinguished as to whether or not census tract was considered. In the first set of runs, this variable was ignored; all records were treated without regard to residence of the test subject. However, in the second facet of analysis, location as indicated by tract number was taken into account (as is discussed in Section 2.4 below).



MAP OF LOUISVILLE CENSUS TRACTS

2.2 Data Structure for the Runs

The unique test records were first aggregated into subgroups homogeneous as to year and quarter; race (black or white, there being too few records of "other" race to permit analysis); and age range by 12 month period. As sex had been found in all previous analyses to be insignificant, it was ignored here. In the first set of runs, census tract was also ignored. Thus each subgroup record which was input to the first SPSS runs represented test results taken in a single quarter of a particular year for members of a specific race and within one 12-month age range. A single value for gasoline lead (as determined by average content per gallon and gallons of gasoline consumed for that quarter and measured in hundred millions of grams) was inserted into each subgroup record. The number of test results thus aggregated into a single subgroup record varied from 1 to over 200. For each subgroup the geometric mean of the blood lead level (GMBL) in micrograms of lead per 100 milliliters of blood (μg/100 ml) was computed and entered.[1] Geometric mean was used because blood lead levels had previously been found (Billick, et al., 1979) to approximate closely the log-normal distribution. Table 2-1 illustrates the format of the aggregated

[1] Where a record consisted of only one test result, the "geometric mean blood lead level" obviously could not be obtained and the corresponding entry denoted blood lead level for the single individual. Only about 1.5 percent of the subgroup records consisted of as few as 2 test results. Records with fewer than 3 test results were not used in the analysis.

TABLE 2-1

ILLUSTRATION OF AGGREGATED QUARTERLY RECORD FORMAT
(NO CENSUS TRACT INFO)

Race	Age Range	Test Year	Test Qtr.	Gas Lead	GMBL	Nr in Aggregate Record	Census Tract
1	3	76	2	0.737	3.204	14	--
1	4	76	2	0.737	3.243	26	--
.	.	.	.	.	.	.	
.	.	.	.	.	.	.	
.	.	.	.	.	.	.	
1	1	78	1	0.434	2.751	12	--
1	2	78	1	0.434	2.928	5	--
.	.	.	.	.	.	.	
.	.	.	.	.	.	.	
.	.	.	.	.	.	.	
2	3	75	4	0.701	3.321	10	--
2	4	75	4	0.701	3.344	18	--
ILLUSTRATION OF AGGREGATED QUARTERLY RECORD FORMAT, CENSUS TRACT INCLUDED							
1	1	76	2	0.737	3.224	2	2
1	2	76	2	0.737	3.448	6	2
1	3	76	2	0.737	3.119	4	2

records for the first set of runs. In addition to the basic variables shown there (which were available for use in all runs), additional values were provided for particular runs.

Quarter of the year was handled as a "dummy variable." That is, a designated variable had its value set as either one or zero depending on which 3-month period of the year the record represented. Dummy variables represent a standard means by which each category of a nominal variable (such as quarter of the year) can be assigned arbitrary metric values of 0 and 1 and inserted into a regression equation where it represents the fraction of the sample within that category. Since, however, the number of degrees of freedom is always one less than the number of categories represented by dummy variables (i.e., the value of the kth category or fraction of the sample which it represents is completely determined by that of the first k-1), one of the dummies must be excluded and "becomes a sort of reference point...referred to as the reference category" (Nie, et al., 1975). In the runs, fourth quarter served as the reference category.

Age range was also handled as a dummy variable in most of the runs. The reference category was provided by the range of 7 years and older. Other treatments of the age factor and additional dummy variables used in particular runs are explained in the following sections.

Runs were made separately for the two races, white and black. The data were organized into sub-files according to race in order to facilitate implementation of the SPSS programs. The small number of "other" races precluded meaningful analysis and the few records available were ignored.

2.3 Runs Ignoring Census Tract

As noted, the first analysis grouped all records homogeneous as to time (i.e., year and quarter or "season" of the year), race, and age-range, irrespective of census tract. Aggregated records consisting of fewer than 3 test results were excluded. The SPSS subprogram REGRESSION (Nie, et al., 1975) was used to calculate regression coefficients for ln(GMBL) as the dependent variable. This subprogram also provides as output mean values of all the variables and a complete matrix of correlation coefficients. Print-out of residuals from some of the regression runs was also provided for analysis, largely by eye. Higher-order coefficients between ln(GMBL) and selected variables were also calculated as is detailed in Section 3.

Experimentation was made with various ways of representing age. It had been found in previous analysis that blood-lead levels tend to increase up to about age 3 and thereafter to decline. In one set of runs a variable termed "age deviation" was computed. This variable expressed (as a rational number) the square of the difference (in months) between the subgroup or aggregate mean age and age 40 months (taken to represent the approximate peak). Another set of runs used both mean age (in months) and its square in an attempt to represent the apparent curvilinear relationship between blood-lead level and age.

Runs were also made using dummy variables to separate test results into two major time groupings. This approach took note of the observation that gasoline lead levels tended to decrease over time; the dummy variables were used as a means of representing the relative weight of test results before and after selected years. In one set of runs, the dummy variables separated results before and after the year 1977 (a year for which, as noted, no test results were usable). In another, the separation was between the years 1974 and 1975 on the one hand, and 1976 and following on the other hand. Another approach used with the time factor in one set of runs was to compute "quarterly difference" as a real value expressing the difference between 31 and the sequential number of the test quarter for the aggregated record; the number 31 was that assigned to a quarter during the year 1977, assumed in this set of runs as the breakpoint between relatively high and low gas lead levels.

Tabulation of the correlation coefficients is shown in Section 3, which also discusses their interpretation. The complete set of correlation coefficients obtained can be seen from the listing of variables included in a run, as explained next.

Each separate grouping of independent "carrier" variables in a particular run represented a distinct "model" by means of which the dependent variable, ln(GMBL), was expressed as a function of these variables. Different groupings were tested in an attempt to find the set of independent variables which best explained variation in ln(GMBL). The distinct models tested in each arbitrarily-numbered set of runs is shown in Table 2-2. Here the column heading represents a specific set of runs while the rows denote the factors included as independent variables. An "x" placed at the row-column intersection indicates that a particular factor (or set of related factors, such as age-group or season of the year) was used in a specific set of runs. Gasoline lead was included in all runs. For example, in Run Set #1, the independent variables were the age-groups and season of the year (represented by dummy variables), along with gasoline lead, whereas in Run Set #2 age-grouping was replaced by age deviation. Results of the regression runs are given and discussed in Section 3 below.

2.4 Runs Using Census Tract Information

The second major area of analytic effort comprised an investigation of the role of different residential census tracts in statistically accounting for variation in ln(GMBL).

Aggregation of individual records for these runs differed from that shown in Table 2-1 for the previous runs in that records were

TABLE 2-2

VARIABLES USED IN REGRESSION - CORRELATION RUN/SETS
LOUISVILLE, KENTUCKY, DATA

(Run Sets Arbitrarily Numbered)

Census Tract Ignored

Variables Used	Run Set Number 1	2	3	4	5	6	7
Mean Age						X	
Age Groups 1 to 6[1]	X		X	X	X		X
Gasoline Lead	X	X	X	X	X	X	X
Year Less than 1976[1]					X		
Year Less than 1977[1]				X			
Age Deviation		X					
Age Aquare						X	
Test Qtr Deviation			X				
Gas Lead Squared		X					
LN (GMBL)	X	X	X	X	X	X	X
Pre-1977 Results (separate)							X
Post 1977 Results (separate)							X

Census Tract Used to Produce Aggregated Records

Variables Used \ Run Set No.	CBLK 1	CBLK 2	CBLK 3[2]	CBLK 4[2]
Age Groups 1 to 6[1]	x	x	x	x
Season of Year, 1 to 3[1]	x	x	x	x
Gasoline Lead	x	x	x	x
Census Clusters 1 to 13[1]	x			
Major Census Blocks 1 to 3[1]			x	x
Non-Clustered Tracts		x	x	x
Ln (GMBL)	x	x	x	x

[1]Dummy Variable

[2]The difference between Run Sets CBLK 3 and 4 was that in the former the test records were aggregated first by Census tract and their dummy variables denoting major blocks were assigned. In the latter, 1 Run Set (CBLK 4) test records were aggregated by each of the 3 major blocks with no further distinction among the Census Tracts comprising each block. In both run sets using major blocks, 1, 2 and 3 test records aggregated for other census tracts were included.

grouped by census tract as well as by race, age and test quarter. There were thus fewer individual test results for each aggregate record but many more aggregated records in these runs. The number of census tracts was too great and the test results within any single tract were too few to permit distinguishing each tract individually in the analytic runs. Therefore, census tracts were grouped into clusters and into major blocks for the sets of runs in this area of analysis. Dummy variables were used to denote a particular cluster or major block.

Demographic characteristics of the Louisville census tracts had been investigated by Dr. Billick and in particular by his associate, Prof. Robert Earickson, of the University of Maryland. The results showed similarities among various tracts - largely contiguous - sufficient to warrant forming several clusters comprising two or more census tracts. These clusters are shown on the map in Figure 1. In all, 14 such clusters were defined. Test records aggregated for each census tract were in effect combined into a single grouping or cluster according to this definition and a dummy variable, CB1 to CB13, inclusive, was assigned to denote that the results represented a particular cluster. The fourteenth cluster served as the reference category. Allocation of individual census tract to a particular cluster is listed in Table 2-3. Aggregated records for these 14 census blocks were selected from each subfile (black and white) and

TABLE 2-3

ALLOCATION OF CENSUS TRACTS TO

CLUSTERS AND MAJOR BLOCKS

Cluster No.	Census Tracts	Major Block	Census Tracts
1	1,2	1	1-11,16,17,21,28
2	3,4		
3	6,7,8	2	20,22-27,29-34, 47-52,57-62
4	10,11,12,17		
5	13,15,16		
6	18,26,27,33,34		
7	19,20,24,25	3	37,40,41,44,53-55, 63-74,80-83
8	21-23		
9	36-39		
10	41-44		
11	51,52		
12	59-62		
13	64-70		
14	82,83		

regression runs made. Those records aggregated by census tract (as illustrated in Table 2-2) for tracts NOT included in any of the clusters were also investigated for each racial subfile in a separate running of the SPSS program REGRESSION (Nie, et al., 1975).

There was also evidence of some connection between year of blood testing and major area of the Louisville city represented. Testing in the early years was heaviest in Western Louisville and in the northwest corner of the city at the bend of the Ohio River, as shown in Figure 2. Three principal target areas were identified and the contiguous census tracts within each were treated as a major block. Two sets of runs were made with records from these tracts. In one, aggregated records from these tracts were designated by dummy variables for BLK1, BLK2, and BLK3. Other census tract represented the reference category. The three blocks represented some 45 percent of the records aggregated by census tract in the subfile of whites and nearly two thirds of the black records. The tracts and their grouping are also given in Table 2-3. In the other set of runs, individual test results were first grouped according to which of the major blocks they represented. Aggregated records were then made so that each represented all the tests for a given year and quarter, age group, and census block. Gasoline lead value was inserted and the ln(GMBL) for the test results in the aggregate was computed.

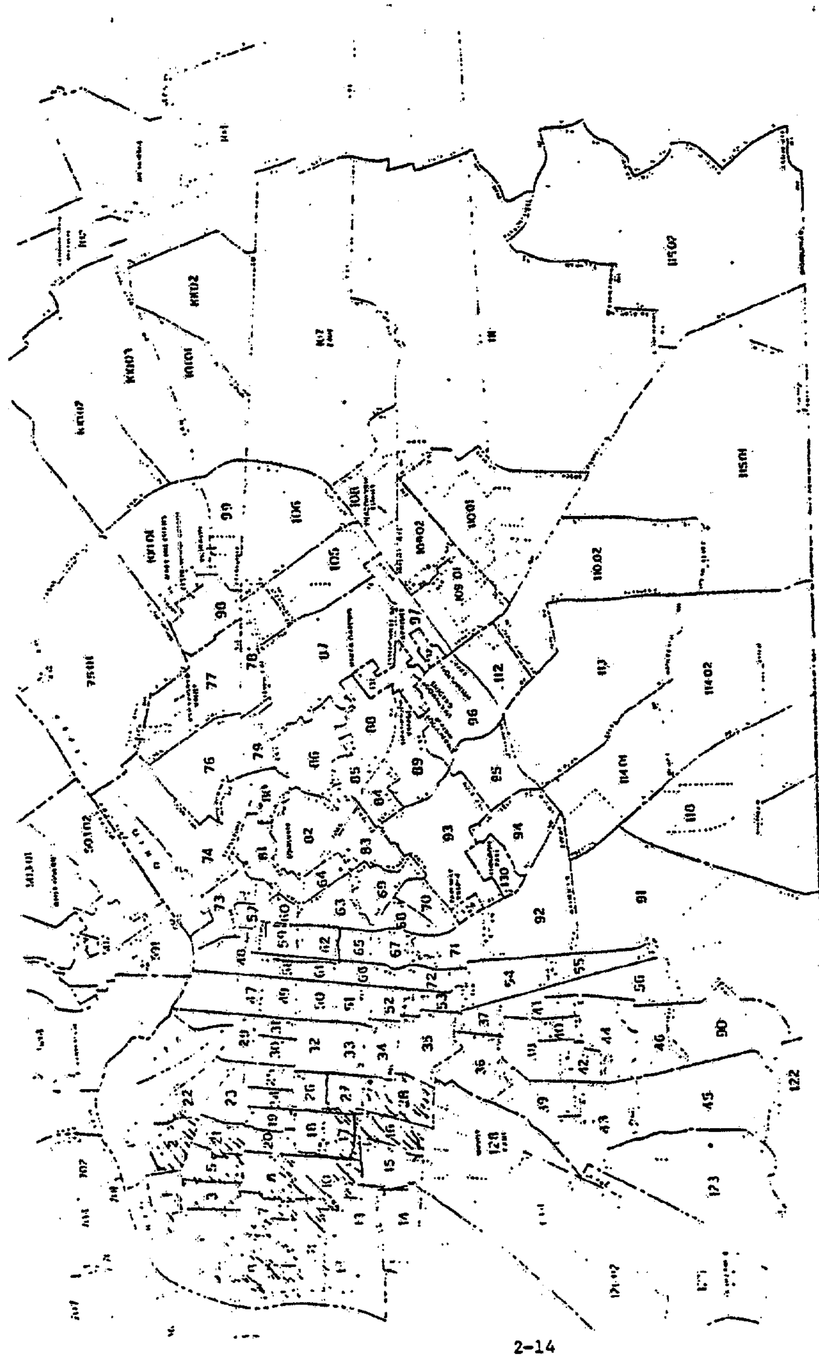


FIGURE 2

MAJOR CENSUS BLOCK MAP

Regression runs with calculation of correlation coefficients between every pair of variables included in the runs were made for these census tract clusters and major blocks. Again aggregated records with less than 3 test results were excluded. The variables within each set of runs are listed in Table 2-2. Coefficients of correlation and of regression obtained are presented and discussed in Section 3.

3.0 STATISTICAL RESULTS OF THE ANALYSIS

3.1 Analysis With No Use of Census Data

3.1.1 Correlation Results

The runs in which dummy variables for age group and for season of the year were the only factors included with ln(GMBL) and gasoline lead (Run Set #1 in Table 2-1) are regarded as the basic set. The runs afforded a basis for statistically examining the effect which three types of factors have on variations in ln(GMBL): age of the subjects tested; season or quarter of the year during which the tests were made; and gasoline lead.

Results for the white and black racial groups[1] were quite similar. The interrelationship between blood lead and gasoline lead was by far the highest statistical association of any found in this set of runs. In contrast with results from any other similar analyses, the interrelationship was very nearly the same for the two races; studies of previous data have always reflected a higher correlation between blood and gasoline lead for blacks than for whites (MITRE, 1980, 1980a).

Interdependence of all variables analyzed with geometric mean blood lead levels and with gasoline lead is shown in Table 3-1 as

---

[1] The number of test subjects not classified as in one of these two racial groups was, as noted, too small to permit meaningful analysis. The total of such tests was less than one percent of all results.

TABLE 3-1

CORRELATION COEFFICIENTS (ZERO-ORDER) BETWEEN GEOMETRIC MEAN BLOOD LEVEL AND INDEPENDENT VARIABLES AND BETWEEN GASOLINE LEAD AND OTHER INDEPENDENT VARIABLES (CENSUS TRACT NOT CONSIDERED)

Louisville, Ky., Comprehensive Data

	WHITES				BLACKS			
	With ln (GMBL)		With Gasoline Lead		With ln (GMBL)		With Gasoline Lead	
	Coefficients of		Coefficients of		Coefficients of		Coefficients of	
	Correlation	Determination	Correlation	Determination	Correlation	Determination	Correlation	Determination
[1]Age Group 1	-.212	.045	-.050	.003	-.207		-.010	*
[1]Age Group 2	.038	.001	.012	*	.012	*	.008	*
[1]Age Group 3	.139	.019	.012	*	.128	.016	-.010	*
[1]Age Group 4	.065	.004	.012	*	.070	.005	-.010	*
[1]Age Group 5	.012	*	.012	*	-.019	*	-.010	*
[1]Age Group 6	.022	*	.012	*	.011	*	-.010	*
[1,2]Quarter 1	-.094	.009	-.233	.054	-.146	.021	-.247	.061
[1,2]Quarter 2	.170	.029	.023	*	.197	.039	.018	*
[1,2]Quarter 3	.045	.002	.035	.001	.015	*	.030	*
Gasoline Lead	.664	.441	—	—	.673	.453	—	—
Age Deviation	-.136	.019	-.076	.006	-.018	*	-.033	.801
Gas Lead Squared	.613	.376	.990	.980	.622	.387	.990	.980
Mean Age	.016	*	-.005	*	.075	.006	.015	*
Mean Age Squared	-.028	*	-.027	*	.043	.002	-.004	*
[1]Period Before 1977	.641	.411	.873	.762	.663	.440	.866	.750
[1]Period Before 1976	.591	.349	.634	.402	.574	.330	.628	.394
Quarterly Deviation	-.692	.479	-.881	.776	-.682	.465	-.874	.764

[1]Dummy Variable

[2]i.e., Jan.-March = Quarter 1, Apr.-June = Quarter 2, etc.

*Less than 10^{-3}

measured by the coefficient of correlation, r_{1j}, and the square of that value, r^2_{1j} (where the subscript "1" denotes blood lead and the j denotes the jth "carrier" or "independent variable"). The square of the coefficient of correlation, often called the "coefficient of determination," expresses the fraction of the variation in blood lead levels accounted for by variation in a specific independent variable. On this basis, changes in gasoline lead account for about 45 percent of the geometric mean blood lead variation observed in the sample for both races, i.e., when the gasoline lead is the same, the variations in blood lead are 55 percent of what they are overall. By contrast, the role of other factors is insignificant, no other variable accounting for as much as 4 percent of blood lead variation.

These correlation coefficients express interrelationships within the sample. To estimate the corresponding values for the underlying populations, confidence intervals at the 5 percent level of significance were also calculated for coefficients of correlation between blood lead level and gasoline lead for each race, based on Fisher's z-transform (Guttman and Wilks, 1965; Kendall and Stuart, 1967). From these calculations, confidence intervals for the coefficients of determination were then derived. These intervals represent a range within which the coefficients for the underlying populations may be presumed to lie with a probability of 95 percent.

On this basis it is seen that the variations in geometric means of blood lead levels accounted for by gasoline lead variation is very unlikely to be less than 30 percent and may be as much as 57 percent.

Racial Group	Lower Limit		Upper Limit	
	r	r^2	r	r^2
Black	.568	.323	.756	.572
White	.551	.304	.752	.566

These results are somewhat different from those obtained in the analysis of the first set of Louisville data. In particular, the more comprehensive test results available in the present analysis show a substantially higher correlation of ln(GMBL) with gasoline lead for whites (.664 as compared with .498) and one slightly higher for blacks (.673 against .605). A somewhat surprising difference is that with the larger data base gasoline lead correlated much more weakly with quarter of the year than in the first Louisville analysis (MITRE, 1980a). Blood lead also correlated somewhat less strongly with season in the present results. The implication of these facts in regard to seasonal effects on blood lead levels is discussed below. The New York City data also showed both blood lead and gasoline lead correlating more strongly with quarter of the year (MITRE, 1980).

Obviously, correlation results do not establish a causal relationship. However, the results reported here are highly consistent with an hypothesis that changes in gasoline lead represent a stronger influence in determining increase or decrease in blood lead levels than do any of the other factors analyzed in the first set of runs, i.e., those associated with season of the year or age of children tested. It is also clear that these factors are associated to a lesser extent statistically with changes in blood lead level. Hence they may be presumed to exert some influence on blood lead, as discussed next.

3.1.2 Evidence of the Effect of Other Factors

The statistical results are quite incompatible with any theory that the correlation between gasoline lead and blood lead levels could reflect the accidental result of some other causal factor associated with age or season of the year. These conclusions are indicated by Table 3-2 for which the higher-order correlation coefficients (and the corresponding coefficients of determination) were calculated for selected variables showing a high statistical interrelationship with changes in ln(GMBL). These coefficients measure the extent of variation between two variables when other variables are in effect held constant (i.e., their effect is subtracted out). As can be seen in Table 3-2, in general the statistical correlation between gasoline lead and blood lead is slightly higher under conditions such that these other factors do not vary than when they do. In particular, removing the effect of ALL three of the other variables with which ln(GMBL) correlated other

TABLE 3-2

PARTIAL CORRELATION OF SELECTED VARIABLES COMPARED WITH ZERO-ORDER COEFFICIENTS RUNS IGNORING CENSUS TRACT, LOUISVILLE, KY., COMPREHENSIVE DATA

Variables Correlated	Other Variable(s) Controlled	Order of Correlation Coefficient	WHITES Coefficient of Correlation	WHITES Coefficient of Determination	BLACKS Coefficient of Correlation	BLACKS Coefficient of Determination
Blood Lead, Gasoline Lead	None	0	.664	.441	.673	.453
Blood Lead, Age Range 3	None	0	.139	.019	.128	.016
Blood Lead, 1st Quarter of Yr.	None	0	-.094	.009	-.146	.021
Blood Lead, 2nd Quarter of Yr.	None	0	.170	.029	.197	.039
Blood Lead, 3rd Quarter of Yr.	None	0	.045	.002	.015	
Gasoline Lead, Age Range 3	None	0	.013	<.001	-.010	<.001
Gasoline Lead, 1st Quarter of Yr.	None	0	-.233	.054	-.247	.061
Gasoline Lead, 2nd Quarter of Yr.	None	0	0.23	<.001	.018	<.001
Gasoline Lead, 3rd Quarter of Yr.	None	0	.035	.001	.030	.001
Blood Lead, Gasoline Lead	2nd Quarter of Yr.	1	.670	.449	.683	.461
Blood Lead, 2nd Quarter of Yr.	Gasoline Lead	1	.207	.043	.250	.063
Blood Lead, Age Range 2	Gasoline Lead	1	.175	.031	.163	.033
Blood Lead, Gasoline Lead	1st Quarter of Yr.	1	.663	.440	.664	.441
Blood Lead, Gasoline Lead	Age Range 3	1	.668	.443	.679	.461
Blood Lead, Gasoline Lead	2nd & 3rd Qts. of Yr.	2	.670	.462	.683	.469
Blood Lead, Gasoline Lead	Age Range 3, 1st Qtr.	2	.668	.443	.671	.450
Blood Lead, Gasoline Lead	Age Range 3, 2nd Qtr.	2	.675	.456	.690	.476
Blood Lead, Gasoline Lead	1st & 2nd Qtr., Age Range 3	3	.685	.469	.693	.480
Blood Lead, Gasoline Lead	2nd & 3rd Qtr. Age Ranges 2,3	4	.675	.450	.690	.476
Gasoline Lead, 2nd Quarter	Blood Lead	1	-.122	.038	-.158	.038
Gasoline Lead, 3rd Quarter	Blood Lead	1	.007	<.001	.027	<.001

than trivially (as measured by the zero-order correlation coefficient) results in a third-order correlation coefficient which is about two percentage points higher than the coefficient of zero order. The very small changes between the zero-order and higher-order coefficients for blood lead and gasoline lead are exactly what would be expected from the fact that gasoline lead correlates quite weakly with these other variables. Therefore, removing the effect of this interrelationship has little effect on the measured correlation between ln(GMBL) and gasoline lead.[1]

The correlation coefficients between ln(GMBL) and both second quarter and age group 2 also increase slightly when the effects of gas lead are removed as can be seen in Table 3-2. This result is as expected from the fact that gas lead is virtually uncorrelated with either of these other two variables, as shown by the zero-order coefficients which are .012 and .022. The statistical evidence points to the role which some (unknown) factors associated with this particular age and season of the year have in accounting to a lesser extent for some changes in ln(GMBL) independent of gasoline lead.

---

[1]As Kendall and Stuart in their monumental work on statistics (1963, p. 334) write: "If we find that holding another variable fixed reduces the correlation between two variables, we infer that the interdependence arises in part through the agency of that other variable.... Conversely, if the partial correlation is larger than the original correlation between the variables we infer that the other variable was obscuring the strong connection...."

Attempts to find one or more real-valued variables to represent age differences were not successful. As can be seen from Table 3-1, age duration showed a weak correlation with ln(GMBL). The correlation was negative, as expected; the farther away a test subject's age was from the age at which blood lead level manifested a peak, the lower ln(GMBL) was, other things equal. Mean age and mean age squared showed only a trivial correlation. As is discussed in the next section, these variables also contributed little when used in linear regression models.

3.1.3 Linear Regression of ln(GMBL)

3.1.3.1 Regression Coefficients. Regression coefficients were calculated on the assumption of a linear relationship between the dependent variable (natural logarithm of GMBL) and the other so-called independent variables; specifically, that the observed value of ln(GMBL)(y) can be closely approximated by the value ($\hat{y}$) calculated from the following expression:

$$\hat{y} = \alpha \sum_i + BiXi + e$$

where x_i = value of the ith independent variable.

The constant factor, α, is presumed to account for those elements which affect variation in blood lead but which are not included in the regression analysis. The value of each element, Bi, termed a regression coefficient, expresses the number of units of variation in y expected for each unit variation in x_i, holding all other independent variables constant. For example, in blacks, the coefficient of regression for gasoline lead has a value of 0.921 (Table 3-3) measured in hundred millions of grams. It would therefore be expected that when all other independent variables (age group and quarter of the year) are identical, a difference of 1 unit in gasoline lead would be reflected as a difference of about 0.92 units in ln(GMBL); i.e., when gasoline lead increases by one unit with no change in the other independent variables the natural logarithm of the GMBL would rise by 0.92 and the geometric mean blood level would increase about 2.51 μg/100 ml. These values as calculated are shown in Table 3-3. They represent the regression coefficients for the sample population.

The SPSS subprogram REGRESSION routinely calculates values for the F-test and these are also given in Table 3-3. The F-statistic is used in testing the hypothesis that the sample comes from a population for which the regression coefficient is actually zero and the calculated non-zero value results from sampling fluctuations (or measurement errors). In other words, the F-statistic tests the hypothesis that the independent variable for which the regression coefficient is calculated contributes nothing to predicting the value of the dependent variable, here ln(GMBL).

The F-statistic is calculated as a ratio and expresses the probability of obtaining a value larger than that calculated for a sample of given size from a population in which the value of the

TABLE 3-3

REGRESSION COEFFICIENTS FOR LOG GEOMETRIC MEAN BLOOD-LEAD LEVEL, IGNORING CENSUS TRACT AND USING AGE RANGE LOUISVILLE, KENTUCKY, COMPREHENSIVE DATA

Independent Variable	Whites (Data Sets = 125) Regression Coefficient	Whites F-Ratio Value	Blacks (Data Sets = 136) Regression Coefficient	Blacks F-Ratio Value
Age Range 1	-0.078	1.917	-0.098	4.836
Age Range 2	0.059	1.119	0.016	0.138
Age Range 3	0.120	4.715	0.094	4.439
Age Range 4	0.075	1.820	0.061	1.874
Age Range 5	0.043	0.593	0.010	0.051
Age Range 6	0.049	0.769	0.031	0.485
Quarter 1	0.131	8.561	0.083	4.962
Quarter 2	0.170	14.756	0.142	15.637
Quarter 3	0.111	6.653	0.069	3.757
Gas Lead	1.051	112.562	0.921	131.817
Constant Term	2.297	—	2.507	—
R^2:	.555		.565	
Adjusted R^2:	.515		.530	
Critical F Values:				
5% level		3.93		3.917
2.5% level		5.31		5.145

regression coefficient is actually zero. Hence the hypothesis of a zero value is rejected (at a given level of significance corresponding to a set probability of error) if the calculated value ratio exceeds the critical or threshold value. These critical values, which have been compiled in standard tables (e.g., Mood and Graybill, 1963) depend on both number of independent variables and number of data points (sample size) used in the regression calculations. The critical value for the F-ratio for each set of regression coefficients is also given in Table 3-3.

With sample size and number of independent variables involved here, the hypothesis of a zero value for a population coefficient of regression can be rejected at the 5-percent level of significance for a calculated F-value exceeding about 3.9 for either race individually. As can be seen from Table 3-3, the hypothesis can on this basis be rejected ONLY for gas lead, age range 3, seasons of the year, and age range 1 for blacks. For the other variables, the hypothesis <u>cannot</u> be rejected with only 5 percent chance of error: these regression coefficients may in the underlying populations actually be zero. In other words, virtually as good results would be obtained if age were ignored in the model.

3.1.3.2 <u>Coefficient of Multiple Correlation</u>. From the coefficients of correlation expressing the interrelationship between each pair of variables, it is possible to calculate a coefficient of multiple correlation, usually denoted as $R_{1.23...n}$. The square of this value, R^2, often termed the coefficient of multiple determination, measures the extent of variation remaining in the dependent variable when the values of the independent variables are held constant. Thus, for a value of R=0.5, R^2 = 0.25, indicating that only 25 percent of the variation in the dependent variable is accounted for by variations in the independent variable. Perfect correlation would of course be indicated by a value of R (and hence R^2)=1, which is the situation when the dependent variable can be completely predicted from knowing the independent variables; i.e., for a given set of values for the independent variables, the dependent variable always has the same value (for example the relationship that exists between cost of a bag of groceries and cost of each item in the bag).

Values of R and R^2 obtained for blood lead level in each race as a function of all the variables considered in the analysis (gasoline lead, race, age range and quarter of the year) are given in Table 3-3. About 45 percent of the ln(GMBL) variation is statistically indicated as resulting from unknown factors not used in the regression analysis. These influences could reflect other features in the environment, anomalies in the sample data, or both.[1] By contrast, in the New York City data as much as 80 percent of the blood lead variation was explained by the independent variables considered. On the basis of these indications, results of using the regression coefficients to predict for Louisville expected values of ln(GMBL) as a function of a given set of values for the independent variables would be much poorer than expected for New York City. In summary, the regression model developed by the Louisville analysis does not appear to be as useful for Louisville as does the corresponding New York model.

In general, the regression results obtained from Run Set #1 were as good as or better than any as is discussed next. The use of age deviation in Run Set #2 proved of no consequence as indicated by the F-ratio value for the regression coefficient. Nor did mean age and age square (Run Set #6) result in a more useful regression equation; the coefficients for these two variables were minute, the F-ratio values very low as computed for blacks and slightly below the critical level for whites. Regression coefficients computed for quarter or season of the year differed very little between Run Sets #1 and #6. In Run Set #2, the coefficients computed for these variables were smaller and the F-ratio values lower than in Run Set #1.

---

[1] These results compare closely with those obtained in the earlier Louisville analysis, where the R^2 values were .590 for blacks and only .501 for whites. Although with the much larger (by an order of magnitude) sample size, the New York City data resulted in a much closer fit of the regression model, in Louisville increasing the sample about 50 percent resulted in no improvement.

Complete sets of regression coefficients and accompanying F-ratio values are provided in tables in the Appendix.

Introduction of the squared term for gas lead in Run Set #2, with the coefficient of correlation with gas level near unity, resulted in multicollinearity in the model. No gains to offset this drawback were observed in the equation which used both gas lead variables. In particular, the difference in sign of the regression coefficients (positive for gas lead and negative for gas lead squared), which is counter intuitive, indicated poor results from the double rate assigned gasoline lead.

3.1.3.3 Runs Using a Variable to Reflect Time Trends. Run Sets 3, 4, and 5 were designed to examine the possible effect on blood-lead levels of long-term time trends. That is, it was conceivable that the measured geometric mean of blood lead might reflect a general tendency either to rise or to fall over the testing period (1974-1979) for any of a variety of reasons; possible influences thus manifested could include changes over time in gasoline lead levels, in the population from which the samples were drawn, or in extent of exposure to lead from other sources. Accordingly, as noted in Section 2, three sets of runs were made, each of which used a different variable to distinguish test period: dummy variable reflecting whether the test occurred before or after 1976; dummy variable similarly distinguishing tests before and after

1977; and "quarterly deviation" - an interval variable representing the difference between the sequentially-numbered quarter in which the test occurred and the sequential number of a mid-quarter in 1977 (as will be recalled, no test results from that year were used in the analysis).

3.1.3.4 Correlation Results. All three of the above variables correlated strongly with both gasoline lead and ln(GMBL) as can be seen from Table 3-1. The dummy variable for year earlier than 1977 showed higher correlation coefficients with both than did that for year prior to 1976; this result was consistent with the fact that quarterly deviation (measured from 1977) also correlated quite strongly with both gasoline and blood lead. It was subsequently observed (in another connection) that all quarterly gasoline lead values prior to 1977 were well above 0.5 (measured in billions of grams), whereas after 1977 in no quarter did it exceed 0.493 and tended to drop off even further. Indeed its mean values were about 0.74 for the earlier period and 0.45 for the later (as shown in Run Set #7 discussed below). Overall mean values of ln(GMBL) also declined from around 3.2 before 1977 which indeed provided a dividing line in the context of the study.

These correlation results were, however, all quite ambiguous as to the reason for the decline in blood lead levels and in particular as to the possible role of gasoline lead therein. As already noted, statistical evidence would in no event establish a causative relationship. But while consistent with an hypothesis that declining blood gasoline lead consumption over time was the reason for a drop in blood-lead levels, the evidence was also consistent with an opposite theory that both gasoline and blood lead dropped independent of one another during the period of testing. In either event, one would expect the strong 3-way correlation that was observed among variables representing time, gasoline lead, and blood lead; the first-order coefficient of blood lead with either of the others, controlling for the third, would be very much lowered because of the strong mutual interrelationship.

To attack this ambiguity, Run Set #7 was made in which test results were separated into those prior to 1977 and those comprising test years 1978 and 1979. The results were quite interesting and appear significant. Although the reduced sample size results in much wider confidence intervals for the coefficients so that the numerical values must be treated with some caution, Table 3-4 leaves little room for doubt as to the significance which variations in gasoline lead consumption have in accounting for variations in ln(GMBL) after 1977. The correlation coefficient of over 0.62 between gasoline lead and ln(GMBL) for whites is more than twice the value of any other correlation coefficient with ln(GMBL); test results from blacks are very nearly as strong in this regard. Interestingly enough, in the

TABLE 3-4

CORRELATION COEFFICIENTS OF LN(GMBL) AND GASOLINE LEAD WITH OTHER VARIABLES, TEST RESULTS BEFORE AND AFTER 1977 LOUISVILLE, KENTUCKY COMPREHENSIVE DATA

Other Variables	Pre-1977 Test Data								Post-1977 Test Data							
	Whites N = 64				Blacks N = 74				Whites N = 59				Blacks N = 62			
	With Ln(GMBL)		With Gas Lead		With Ln(GMBL)		With Gas Lead		With Ln(GMBL)		With Gas Lead		With Ln(GMBL)		With Gas Lead	
	r_{ij}[a]	r_{ij}^2[a]	r_{ij}	r_{ij}^2	r_{ij}	r_{ij}^2	r_{ij}	r_{ij}^2	r_{ij}	r_{ij}^2	r_{ij}	r_{ij}^2	r_{ij}	r_{ij}^2	r_{ij}	r_{ij}^2
Age Group 1	-.263	.069	-.103	.011	-.253	.064	*	*	-.270	.073	-.017	*	-.282	.080	-.011	*
Age Group 2	.055	.033	.023	*	1.082	.007	.003	*	.036	.001	-.017	*	1.067	.005	-.011	*
Age Group 3	.287	.084	.023	*	.200	.040	*	*	.096	.009	-.017	*	.162	.026	-.011	*
Age Group 4	.109	.012	.023	*	.145	.021	*	*	.057	.003	-.017	*	.064	.004	-.011	*
Age Group 5	.121	.015	.023	*	.029	*	*	*	-.075	.006	-.017	*	-.002	*	-.011	*
Age Group 6	.116	.014	.023	*	.014	*	*	*	-.048	.002	-.017	*	.053	.003	-.011	*
Quarter 1[1]	-.212	.045	-.587	.345	-.359	.129	-.575	.331	-.026	*	-.187	.035	.019	*	-.195	.038
Quarter 2[1]	.215	.046	.034	.001	.291	.085	.006	*	.291	.085	.490	.240	.295	.087	.501	.251
Quarter 3[1]	-.023	*	.159	.025	-.025	*	.126	.016	.179	.032	.199	.040	.120	.014	.161	.026
Gas Lead	.239	.057	1.000	1.000	.253	.064	1.000	1.000	.623	.388	1.000	1.000	.520	.270	1.000	1.000

[1]Dummy variable.

[a]r_{ij} = *coefficient of correlation.*

[b]r_{ij}^2 = coefficient of determination.

*Less than 10^{-3}.

pre-1977 data gasoline lead is far less closely associated statistically with variations in ln(GMBL). Indeed among both whites and blacks other factors - age and season of the year - show a stronger interrelationship.

3.2 Analysis of Data Using Census Tract Information

As noted in Section 2, four sets of runs (designated arbitrarily CBLK1 through CBLK4) were made with data aggregated on the basis of census tract. In the first three, each subgroup or aggregated record consisted of test results from a single census tract (and of course homogeneous as to race, age group, gasoline lead, and test quarter). The aggregated records were then further designated as to whether or not the census tract belonged to a particular one of 14 clusters (runs CBLK1 and CBLK3) or one of 3 major census blocks (CBLK2). The clusters and blocks in Runs CBLK1 and CBLK3 were distinguished by the fact that the former selected ONLY aggregated records from one of the 14 clusters, whereas the latter selected those from other census tracts not included in the cluster set. In the fourth set of runs (CBLK4), individual test records were aggregated on the basis of each major census block without regard for individual tract designation. Thus each aggregated subgroup record was homogeneous as to race, test quarter, age group, particular census block (or no block), and gasoline lead consumption.

The results of these groupings were that in the fourth run there were fewer total aggregate records than in the other three but the number of individual test results in each aggregated grouping was larger. The computed ln(GMBL) value for each could therefore be expected to represent more nearly the underlying population, with the effect of fluctuations due to small sample size smoothed out.

All of these runs used the SPSS subprogram REGRESSION (Nie, et al., 1975), with printout of mean values for the variables and a complete matrix of correlation coefficients.

3.2.1 Objective of the Runs

The objective of all these sets of runs was to investigate what statistical interrelationships(s) existed between variations in blood-lead levels and differences in residence of the subjects tested. Despite the strong correlation between ln(GMBL) and gasoline lead noted, it was clear (as discussed above) that other factors also accounted for the variation in blood lead. In particular, it was hoped that these unknown factors might be related to place of residence (e.g., condition of home as to lead pipes and lead paint, proximity to lead from industrial discharges, etc.) and that this situation would be reflected in the statistical results of the runs. Because the data in these runs represented different aggregated groupings, it would not be expected that the correlation coefficients of blood and gasoline lead with variables reflecting age and quarter or season of the year (used both in these runs and in the runs which ignored census tract) would have the same values. For explanatory purposes, ideal results would have included: correlation coefficients between blood and gasoline lead that were similar to those found in the noncensus tract runs; high correlation of blood lead with variables denoting specific census clusters or blocks; and greater R^2 values than were obtained in the noncensus tract runs, thus reflecting additional statistical interrelationships with ln(GMBL) not found in these first runs.

3.2.2 Correlation Results

From the above point of view, however, the results obtained from the use of census tract data were disappointing. Coefficients of correlation between both gasoline lead and ln(GMBL) on the one hand and the other variables used on the other hand are given in the columns of Table 3-5. These tables also show R^2 values. Regression coefficients are given in Table 3-6. As can be seen from Table 3-5 no coefficient of correlation between a census variable and ln(GMBL) exceeded an absolute value of .236 so that none of the former variables accounted statistically for as much as 6 percent of the blood-lead variation. Highest positive values were observed for census-tract clusters (CB) 5 and 7 and major block 2, among whites, and for CB 7 and major block 2 among blacks. There is considerable commonalty among the census tracts comprising these census clusters

TABLE 3-5

CORRELATION COEFFICIENTS OF BLOOD LEAD AND GASOLINE LEAD WITH OTHER VARIABLES, USING CENSUS TRACT INFORMATION LOUISVILLE, KY., COMPREHENSIVE DATA (RUNS CBLK 1, CBLK 2, CBLK 3)

		Whites						Blacks					
		CBLK 1, N=133		CBLK 2, N=2223		CBLK 3, N=1503		CBLK 1, N=611		CBLK 2, N=2978		CBLK 3, N=1307	
Other Variables*		ln (GMBL)	Gas Lead	ln (GMBL)	Gas Lead	ln (GMBL)	Gas Lead	ln (GMBL)	Gas Lead	ln (GMBL)	Gas Lead	ln (GMBL)	Gas Lead
Age Group	1	-.080	.098	-.057	-.025	-.076	-.026	-.095	-.049	-.087	-.043	-.062	-.041
	2	-.156	-.119	.002	-.012	.031	.003	-.153	-.186	-.026	-.045	-.032	-.042
	3	.120	.001	.097	.022	.073	.024	.077	-.107	.072	-.014	.070	-.008
	4	.047	-.021	.011	-.022	.029	-.016	.097	.076	.031	.004	.038	-.008
	5	.045	.058	-.062	-.008	-.083	-.017	-.020	.082	-.038	.032	-.038	.026
	6	.028	.066	.011	.004	.015	-.023	.060	.148	.033	.036	.031	.033
[1]Qtr. of Yr.	1	-.179	-.329	-.023	-.227	.019	-.190	-.073	-.310	-.034	-.250	-.010	-.242
	2	.101	.019	.084	.097	.085	.116	.112	.038	.090	.024	.111	.015
	3	.060	.072	.053	.056	.050	.059	-.024	-.063	.010	-.038	-.012	-.023
[1]CB	1	-.109	.039	--	--	--	--	--[3]	--[3]	--	--	--	--
CB	2	-.227	.046	--	--	--	--	-.155	.031	--	--	--	--
CB	3	--[3]	--[3]	--	--	--	--	-.065	-.064	--	--	--	--
CB	4	--[3]	--[3]	--	--	--	--	-.033	-.126	--	--	--	--
CB	5	.145	-.037	--	--	--	--	-.063	-.050	--	--	--	--
CB	6	.069	-.014	--	--	--	--	.070	.073	--	--	--	--
CB	7	.175	.072	--	--	--	--	.236	.062	--	--	--	--
CB	8	.100	-.228	--	--	--	--	.104	.032	--	--	--	--
CB	9	-.030	.147	--	--	--	--	--[3]	--[3]	--	--	--	--
CB	10	-.024	.203	--	--	--	--	--[3]	--[3]	--	--	--	--
CB	11	.004	.026	--	--	--	--	--[3]	--[3]	--	--	--	--
CB	12	.151	.238	--	--	--	--	-.006	.098	--	--	--	--
CB	13	-.178	-.050	--	--	--	--	--[3]	--[3]	--	--	--	--
Major Block	1	--	--	.084	.025	--	--	--	--	-.010	-.044	--	--
" "	2	--	--	.223	.113	--	--	--	--	.158	.108	--	--
" "	3	--	--	.043	.060	--	--	--	--	.040	$<10^{-4}$	--	--
Gas Lead		.391	1.000	.372	1.000	.394	1.000	.465	1.000	.354	1.000	.352	1.000
R^2		.418		.222		.020		.339		.185		.168	

[1]Census Block comprising track as defined in Table 2-

[3]Incomputable, owing to small sample size

*Dummy Variables represented all values in this column except Gas Lead.

and major block 2 as can be seen from Table 2-3. The obvious inference is that some (unknown) conditions associated with residence in these census tracts are related slightly to an increase in blood-lead, other factors being equal.

Analogously, highest negative values among both blacks and whites were found for coefficients of correlation between ln(GMBL) and census cluster #2; a relatively large negative value was also found for the correlation coefficient between cluster number 13 and ln(GMBL) for whites, there being too few black subjects tested from this cluster to permit computation of the coefficient. For no major block was there a significant negative correlation coefficient with ln(GMBL). Indeed, the coefficients are so low as to imply that residence in one of the tracts comprising major blocks 1 and 3 has little if any significance in explaining blood-lead variations.

Most disappointing of all, from the standpoint of interpreting blood-lead variations, is the fact that R^2 values for all of these sets of runs were substantially lower than in the sets which ignored census tract information. This observation may be related to the fact that the coefficients of correlation between blood gasoline lead and ln(GMBL) - while the highest of any observed in the runs - were significantly reduced over those obtained in the run sets previously discussed. It would be tempting to infer that among children in the selected census tracts factors other than gasoline lead played a somewhat greater role in accounting for blood-lead variations than among Louisville children generally. While the results of run sets CBLK1 and 2 are consistent with such an hypothesis, those of CBLK3 are not. It can be observed that for test subjects outside the tracts grouped into 14 clusters, aggregation on the basis of census tract results in lower values both for overall R^2 and for the coefficient of correlation between gasoline lead and ln(GMBL) than when data are not so aggregated (Run Sets 1 - 6). A more satisfactory explanation - in that it accounts for more of the significant observations - would be that aggregation by census tract, which resulted in a smaller number of test results per aggregated record, somehow obscured the relationships manifested with ln(GMBL) computed for each aggregated record from a larger sample size.

In any event, the correlation results of the last four sets of runs provide little basis for attempting to explain additional casual factors in blood-lead variation.

3.2.3 Regression Results

The R^2 values obtained in all of the runs using data aggregated by census tract were so low as to indicate that the models would be virtually worthless for predictive purposes. However, the regression results are of some interest in other regards. The regression coefficients and the F-ratio values are listed in Table 3-6.

TABLE 3-6

REGRESSION RESULTS FOR RUNS USING CENSUS TRACT INFORMATION (CBLK1-CBLK4)
LOUISVILLE, KENTUCKY DATA, DEPENDENT VARIABLE: LN(CMBL)

Independent Variables	Whites								Blacks							
	CBLK1 N = 133		CBLK2 N = 1503		CBLK3 N = 2223		CBLK4 N = 480		CBLK1 N = 611		CBLK2 N = 1307		CBLK3 N = 2978		CBLK4 N = 482	
	Regression Coefficient	F-Ratio Value	Regression Coefficient	F-Ratio Value	Regression Coefficient	F-Ratio Value	Regression Coefficient	F-Ratio Value	Regression Coefficient	F-Ratio Value	Regression Coefficient	F-Ratio Value	Regression Coefficient	F-Ratio Value	Regression Coefficient	F-Ratio Value
Age Group 1	-0.171	0.912	-0.057	0.876	-0.019	0.156	0.153	12.558	-0.072	0.750	-0.048	0.640	-0.095	7.328	0.127	8.207
Age Group 2	0.060	0.239	0.105	5.535	0.097	7.349	.233	28.623	0.027	0.260	.025	0.331	.017	0.428	0.202	19.520
Age Group 3	0.182	2.063	.147	9.590	0.180	22.951	.150	11.894	0.141	6.589	.099	4.569	.102	13.356	0.134	8.864
Age Group 4	0.142	1.235	.117	6.204	0.117	9.450	.144	10.732	0.108	3.864	.083	3.132	.060	4.678	0.109	6.040
Age Group 5	0.123	1.077	.029	0.384	0.056	2.305	.183	16.841	0.055	1.054	.016	0.132	Less Than 10^{-3}		0.155	11.841
Age Group 6	0.122	0.857	.107	4.715	0.112	7.872	.081	2.910	0.053	0.830	.065	1.740	.051	2.969	0.116	5.678
Quarter 1	-0.006	0.008	.179	39.358	0.136	32.644	-.110	0.090	0.109	18.002	.179	34.769	.133	49.264	-0.080	4.141
Quarter 2	0.120	3.667	.145	21.286	0.130	26.602	.059	2.571	0.147	28.849	.199	42.404	.159	68.821	0.085	4.732
Quarter 3	0.106	2.752	.142	19.891	0.135	27.235	Not Used		0.092	9.144	.117	12.306	.122	35.057	Not Used	
Cluster 1	-0.084	0.224	Not Used		Not Used		Not Used		No Tests		Not Used		Not Used		Not Used	
Cluster 2	-0.422	4.495	Not Used		Not Used		Not Used		-0.117	8.998	Not Used		Not Used		Not Used	
Cluster 3	No Tests		Not Used		Not Used		Not Used		0.011	0.092	Not Used		Not Used		Not Used	
Cluster 4	No Tests		Not Used		Not Used		Not Used		0.041	1.566	Not Used		Not Used		Not Used	
Cluster 5	0.303	1.644	Not Used		Not Used		Not Used		0.005	0.024	Not Used		Not Used		Not Used	
Cluster 6	0.198	0.768	Not Used		Not Used		Not Used		0.048	1.879	Not Used		Not Used		Not Used	
Cluster 7	0.270	1.461	Not Used		Not Used		Not Used		0.195	29.312	Not Used		Not Used		Not Used	
Cluster 8	0.090	0.271	Not Used		Not Used		Not Used		-0.364	7.604	Not Used		Not Used		Not Used	
Cluster 9	-0.088	0.154	Not Used		Not Used		Not Used		No Tests		Not Used		Not Used		Not Used	
Cluster 10	-0.294	0.972	Not Used		Not Used		Not Used		No Tests		Not Used		Not Used		Not Used	
Cluster 11	0.053	0.084	Not Used		Not Used		Not Used		No Tests		Not Used		Not Used		Not Used	
Cluster 12	0.081	0.180	Not Used		Not Used		Not Used		Too Few Tests		Not Used		Not Used		Not Used	
Cluster 13	-0.072	0.148	Not Used		Not Used		Not Used		No Tests		Not Used		Not Used		Not Used	
Gas Lead	0.759	22.985	1.213	291.827	1.043	321.831	.815	103.327	0.840	181.451	0.939	222.464	0.910	452.572	0.968	139.805
Constant Term	2.535	—	2.091	--	2.145	--	2.369	--	2.535	—	2.331	--	2.408	—	2.403	—
Major Block 1	Not Used		Not Used		0.178	48.829	.206	36.277	Not Used		Not Used		0.073	18.043	0.071	4.452
Major Block 2	Not Used		Not Used		0.255	128.943	.283	69.546	Not Used		Not Used		0.144	75.858	0.153	20.814
Major Block 3	Not Used		Not Used		0.130	15.265	0.144	8.968	Not Used		Not Used		0.182	12.423	0.163	9.900
Quarter 4	Not Used		Not Used		Not Used		-0.089	6.248	Not Used		Not Used		Not Used		-0.094	6.402
Critical F Value (5% Level of Significance)	3.89		3.85		3.85		3.88		3.88		3.85		3.85		3.88	

TEH 0531370

DUP050032581

First of all, it will be noted that where the number of aggregated records was small - as in Run Set CBLK1, which had only 133 data points (n) for whites, by far the smallest number in any of these runs - low F-ratio values were obtained for virtually all regression coefficients. The only exception is for gas lead, which everywhere produced a high F-ratio value, indicating again the significance of this factor. The low F values, which are not surprising, suggest that a small sample size tends to be less representative in regard to the characteristics of interest here of the underlying population than does a larger sample. It will be noted in the results for whites in Run Set CBLK1 that the hypothesis of a zero population coefficient of regression can be rejected at the 5-percent level ONLY for Cluster #2, in addition to gas lead. For blacks, where the sample size is over 4 times as large (611 aggregated records was compared with 133), the number of significant regression coefficients is substantially greater. However, even here the hypothesis of a zero population coefficient can be rejected for only about one half of the census clusters and other independent or carrier variables. In other words, the regression models - poor as they are, as indicated by the low R^2 values - would not be significantly degraded if ALL independent variables except gas lead and the dummy variables for cluster 2 had been omitted for whites.

However, nearly all of the dummy variables for age and season are significant to the model derived for whites in Run Set CBLK 2 with a sample size an order of magnitude greater; this run set used aggregated records from census tracts other than those in the 14 blocks. With blacks, virtually as good results would have been obtained in Run Set CBLK1 if about half of the independent variables were ignored; the model needs to use - in addition to gas lead - only the dummy variables for age group 3; possibly for age group 4, the F-ratio value for which is marginal; season of the year; and census clusters 2, 7, and 8. It will also be noted that adequate test results were lacking for each race from some of the census tracts, reflecting a difference in population distribution.

On the other hand, where the sample size is very large - as with Run Set CBLK3 (grouping data into 3 major census blocks) where there were over 2000 aggregated records for each of the two races - hypothesis of a zero popolation coefficient can be rejected for virtually all variables used in the models. Except for age group 2 among blacks and age groups 1 and 5 among whites, the results indicate that all variables used make a significant contribution in fitting the observations to the regression surface. Unfortunately, however, even here the R^2 values are too low to support the model for predictive purpose. It is also of interest to note that the regression results in Run Set CBLK3 are consistent with the correlation coefficients calculated in supporting the hypothesis that children in the three major census blocks have slightly higher blood-lead levels than children elsewhere in the metropolitan area, other things being equal. The results in this regard are the same for both races.

Further, it can be seen that the regression results in Run Set CBLK1 are consistent with the correlation results in that they point to specific census clusters where residential conditions appear most interrelated with blood-lead levels. The negative regression coefficients and the F-value (above the 5 percent significance level) for cluster #2 with both races agree with the negative correlation coefficients obtained for this cluster with ln(GMBL). Children in this cluster tend to have somewhat lower blood-lead levels than children in the other clusters. Conversely, the high F-ratio values obtained for the positive regression coefficients in the results for blacks in clusters 7 and 8 imply higher blood-lead levels here than elsewhere. These are also the clusters for which the highest positive correlation coefficients with ln(GMBL) were obtained. As to what residential conditions may induce these deviations, the statistical results of course provide no evidence.

As a minimum, then, the regression results support the hypothesis that residence in the census tracts comprising the clusters and the major blocks is somehow associated with higher blood-lead levels. They also point to a few particular census tracts (within specific clusters) where it would be of greatest potential value to examine residential conditions for factors that might be causally related to elevated levels of blood-lead. Finally, the runs using census tract information are consistent with the evidence widely observed elsewhere in this study that variation in gasoline lead consumption is the factor most strongly associated with variations in ln(GMBL).

APPENDIX

This appendix lists results of regression models developed in runs which were not of statistical significance for purposes of the study. They are given here for the sake of completeness.

TABLE A-1

REGRESSION COEFFICIENTS FOR LN(GMBL) BY RACE, NO CENSUS TRACT DATA, RUN SET #2, USING AGE DEVIATION

Variable	White Regression Coefficient	White F-Ratio Value	Black Regression Coefficient	Black F-Ratio Value
Age Deviation	$-0.3\text{x}10^{-4}$	2.921	$-.9\text{x}10^{6}$	0.010
Quarter 1	0.081	3.229	0.038	0.770
Quarter 2	0.110	6.044	0.088	5.408
Quarter 3	0.070	2.380	0.025	0.461
Gas Lead	3.675	28.584	3.178	31.831
Gas Lead Squared	-2.022	14.785	-1.728	16.424
Constant Term	1.600	--	1.871	--
R^2 Value	.568		.562	
Critical F-Value	3.9		3.9	
# Aggregated Records	123		136	

TABLE A-2

REGRESSION COEFFICIENTS FOR LN(GMBL) BY RACE, NO CENSUS TRACT DATA RUN SET #3, USING TEST QUARTER DEVIATION

Variable	White Regression Coefficient	White F-Ratio Value	Black Regression Coefficient	Black F-Ratio Value
Age Group 1	-0.088	2.689	-0.102	5.651
Age Group 2	0.055	1.103	0.010	0.057
Age Group 3	0.117	4.938	0.090	4.427
Age Group 4	-0.071	1.840	0.057	1.785
Age Group 5	0.393	0.559	0.006	0.022
Age Group 6	0.452	0.739	0.027	0.405
Test Quarter Deviation*	-0.022	13.104	-0.017	11.483
Quarter 1	0.022	0.184	-0.008	0.032
Quarter 2	0.144	11.772	0.119	11.459
Quarter 3	0.112	7.435	0.070	4.085
Gas Lead	0.192	0.564	0.268	1.672
Constant Term	2.828	--	2.914	--
R^2 Value	.602		.602	
Critical F Value	3.9		3.9	
# Aggregated Records	123		136	

*Numeric deviation between sequential number of test quarter and 31 (a quarter in 1977 taken as dividing point).

TABLE A-3

REGRESSION COEFFICIENTS FOR LN(GMBL) BY RACE, NO CENSUS TRACT DATA, RUN SET #4, USING DUMMY VARIABLE TO DISTINGUISH TESTS BEFORE AND AFTER 1977

Variable	White Regression Coefficient	White F-Ratio Value	Black Regression Coefficient	Black F-Ratio Value
Age Group 1	-0.085	2.330	-0.098	5.136
Age Group 2	0.057	1.102	0.015	0.120
Age Group 3	0.119	4.748	0.093	4.619
Age Group 4	0.073	1.809	0.060	1.939
Age Group 5	0.041	0.575	0.010	0.049
Age Group 6	0.047	0.751	0.030	0.494
[1]Dummy Variable for Time	0.139	4.205	0.147	7.449
Quarter 1	0.101	4.778	0.048	1.561
Quarter 2	0.170	15.503	0.138	15.660
Quarter 3	0.114	7.182	0.067	3.728
Gas Lead	0.656	9.179	0.516	9.492
Constant Term	2.471	--	2.684	--
R^2 Value	.571		.589	
Critical F Value	3.9		3.9	
# Aggregated Records	123		136	

[1]Value of 1 denoted test before 1977, value of zero test after 1977.

TABLE A-4

REGRESSION COEFFICIENTS FOR LN(GMBL) BY RACE, NO CENSUS TRACT DATA, RUN SET #5, USING DUMMY VARIABLE TO DISTINGUISH TEST BEFORE AND AFTER 1976

Variable	White Regression Coefficient	White F-Ratio Value	Black Regression Coefficient	Black F-Ratio Value
Age Group 1	-0.086	2.647	-0.104	5.974
Age Group 2	0.051	0.964	0.007	0.027
Age Group 3	0.113	4.694	0.088	4.265
Age Group 4	0.067	1.668	0.055	1.670
Age Group 5	0.035	0.457	0.004	0.009
Age Group 6	0.041	0.624	0.025	0.346
[1]Dummy Variable for Time (1976)	0.170	15.926	0.128	13.305
Quarter 1	0.123	8.615	0.070	3.914
Quarter 2	0.182	19.420	0.148	18.595
Quarter 3	0.128	9.891	0.084	5.907
Gas Lead	0.746	38.239	0.690	48.172
Constant Term	2.435	15.926	2.614	—
R^2 Value	0.611		0.607	
Critical F Value	3.9		3.9	
# Aggregated Records	123		136	

[1]Value of 1 denoted test before 1976, value of zero test after 1976.

TABLE A-5

REGRESSION COEFFICIENTS FOR LN(GMBL) BY RACE, NO CENSUS TRACT DATA RUN SET #6, USING MEAN AGE AND AGE SQUARED

Variable	White Regression Coefficient	White F-Ratio Value	Black Regression Coefficient	Black F-Ratio Value
Mean Age	0.004	3.588	0.002	1.393
Age Square	-0.0005	3.913	-0.0001	0.907
Quarter 1	0.132	8.319	0.080	4.313
Quarter 2	0.166	13.727	0.141	14.195
Quarter 3	0.111	6.299	0.067	3.198
Gas Lead	1.054	107.820	0.913	119.425
Constant Term	2.258	—	2.486	—
R^2 Value	.517		.513	
Critical F-Value	3.9		3.9	
# Aggregated Records	123		136	

REFERENCES

Billick, Irwin H., Anita S. Curran, and Douglas R. Shier, 1979. Relation of Pediatric Blood Lead Levels to Lead in Gasoline. Submitted to Environmental Health Perspectives.

Coerr, Stanton and Irwin H. Billick, 1979. The effect of lead in gasoline on lead in the blood of urban populations. Memorandum to David Hawkins, Assistant Administrator for Air and Hazardous Materials, U.S. Environmental Protection Agency. April 3, 1979.

Dixon, Wilfrid J. and F.J. Massey, Jr., 1969. Introduction to Statistical Analysis. Third edition. McGraw-Hill.

Guttman, Irwin and S.S. Wilks, 1965. Introducing Engineering Statistics. John Wiley & Sons, Inc., New York, NY.

Kendall, Maurice G. and Alan Stuart, 1967. The Advanced Theory of Statistics. Vol. 2, Section Edition. Hafner Publishing Co.

Mood, A.M. and F.A. Graybill, 1963. Introduction to the Theory of Statistics. Second edition. McGraw-Hill.

Nie, Norman H., C.H. Hull, J.G. Jenkins, K. Steinbrenner, and D.H. Bent, 1975. Statistical Package for the Social Sciences. Second edition. McGraw-Hill.

The MITRE Corporation, 1980a. Statistical Analysis of Blood Lead Levels in Relation to Gasoline Lead, Louisville, Ky. WP-80W00466. McLean, VA. June 2, 1981.

The MITRE Corporation, 1980. Correlation and Regression Analysis of Blood Lead Levels. WP-80W00447. McLean, VA. December 1, 1980.

W-53
DISTRIBUTION LIST
WP# 81W00541

D-12
S. W. GOUSE

W-50
L. KUSHNER
R. P. OUELLETTE
L. W. THOMAS

W-80
M. M. SCHOLL
E. G. SHARP

W-53
T. J. WRIGHT
L. J. DUNCAN (4)

W-58
G. ERSKINE
J. WATSON (3)

Metrek Library (2)

Warehouse (Storage) (1)

Records Resources (1)

W-50 Data File (5)
(W187)

SPONSOR (6)

AUTHORIZED FOR DISTRIBUTION

T. J. Wright
Thomas J. Wright
Department Head, W-53

Linda O Duncan
Project Approval

TEH 0531382

DUP050032593