LBL-16214
c.2

Lawrence Berkeley Laboratory
UNIVERSITY OF CALIFORNIA

Physics, Computer Science & Mathematics Division

RECEIVED
LAWRENCE
BERKELEY LABORATORY

AUG 10 1983

LIBRARY AND
DOCUMENTS SECTION

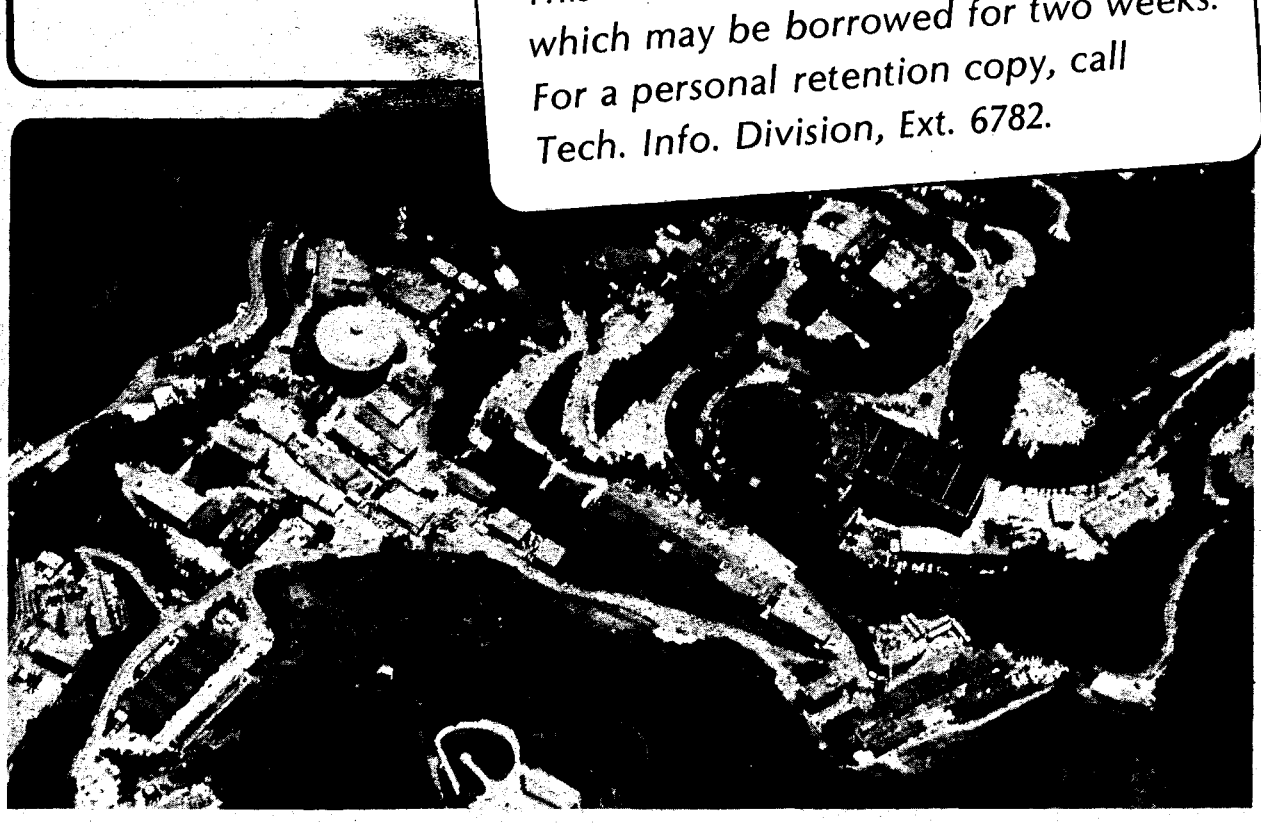
THE GEOGRAPHIC AND STATISTICAL ANALYSIS OF AIR QUALITY DATA IN THE UNITED STATES

L.D. Johnson
(Ph.D. Thesis)

June 1983

TWO-WEEK LOAN COPY

This is a Library Circulating Copy which may be borrowed for two weeks. For a personal retention copy, call Tech. Info. Division, Ext. 6782.



LBL-16214
c.2

Prepared for the U.S. Department of Energy under Contract DE-AC03-76SF00098

DISCLAIMER

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor the Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or the Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or the Regents of the University of California.

LBL-16214

The Geographic and Statistical Analysis of Air Quality Data in the United States

Laura Derelle Johnson
(Ph.D. Thesis)

Computer Science and Mathematics Department
Lawrence Berkeley Laboratory
University of California
Berkeley, California 94720

June 3, 1983

This research was supported by the Applied Mathematical Sciences Research Program of Office of Energy Research, U. S. Department of Energy under contract number DE-AC03-76SF00098.

TABLE OF CONTENTS

Chapter One: Air Quality Data 1

Introduction 1
Potential Applications of Air Quality Data 5
Data Limitations 7
Air Quality by Location 15
Two Pollutants in Ohio State 16
Methodological Perplexities 16
An Overview of the Following Chapters 19

Chapter Two: Geographic Estimation of Air Quality Data 36

The Two-Dimensional Moving Average Method 36
Combining Data from 1-Hour and 24-Hour Sampling Intervals for Sulfur Dioxide 43
Cross-Validation (CV) as an Integral Part of the Modeling 47
Utility of Surface Estimation to Detect Outliers 49
Choosing d_0 for Ohio State with Sulfur Dioxide and Suspended Particulate Data 53
Measuring Local Variability 62
Consistency in the Shape of Local Variability 77
Minimizing Local Variability as an Alternative to Cross-Validation 99

Chapter Three: Comparing Two Methods of Surface Estimation 101

Akima's Method of Bivariate Interpolation 101
Linear Interpolation of Sulfur Dioxide and Suspended Particulate Data in Ohio State 105
Contour Maps Given By The Two-Dimensional Moving Average 109
Comparing Contour Maps Obtained by Two Methods of Estimation 125
Comparison of Point Prediction from Two Methods of Geographic Estimation 137
The Occurrence of Bias in Cross-Validation Estimates 144
The "Better" Method 146

Chapter Four: Conclusions and Further Applications 160

Summary 160
Further Research 162

References 165

Acknowledgments

First of all, I would like to thank all the members of my thesis committee for their help and encouragement. Steve Selvin, the chairman, provided me with this topic and the confidence to explore my own ideas. Nick Jewell, acting as chairman for the last year provided me with intellectually motivating discussions from which original ideas and results arose. Next, I would like to mention Deane Merrill and thank him for his boundless energy and enthusiasm. Deane's general inquisitiveness often provoked explanations. Finally, I thank Orman Granger for sharing his insight into the meteorological aspects of this work.

This work would have been impossible without the Computer Science and Mathematics Department at Lawrence Berkeley Laboratory. I am extremely grateful to Carl Quong, head of this department, for his sustained interest and access to the facilities and staff. Members of the staff that have been especially helpful include: Bill Benson for his help with graphics, Wayne Graves for sharing his insight into the world of operating systems; and Carole Agazzi, Meri Jones, and Julie Snyder for answering numerous questions about word processing. I would also like to mention Susan T. Sacks for her continuous encouragement throughout the writing.

Bonnie Hutchings, the graduate secretary of The Biostatistics Group has helped with endless administrative duties throughout my graduate career and I thank her.

The Geographic and Statistical Analysis
of Air Quality Data in the United States
By
Laura Derelle Johnson
Abstract

This dissertation contains the development and the application of analytic procedures for examining and exploring some air quality data collected by the Environmental Protection Agency from 1974 through 1976. They are collected at monitoring stations most of which are in metropolitan areas. These data are irregularly distributed discrete point measurements. The techniques explored here may be useful in other disciplines with the same type of data.

The analysis is concentrated on two pollutants, suspended particulate and sulfur dioxide. There are two reasons for this restriction: (i) they are the most heavily monitored and (ii) they are of interest to the health field. The state of Ohio is utilized as an example in most of these analyses. This is because Ohio is the most thoroughly monitored state in the United States. A list of the limitations of these data is given.

Interpolation schemes are explored and a model is chosen which is a two-dimensional analogue of the moving average model in time series. The model is

$$e_i = \frac{\sum_{i=1}^{n} e^{(d_{ij}/d_0)^2} x_j}{\sum_{i=1}^{n} e^{(d_{ij}/d_0)^2}}$$

where

e_i = the estimated value at a point i.
x_j = a measured value at point j.
d_{ij} = the distance from the data point to the point of estimation.

d_0= the smoothing parameter.

The choice of d_0 has been explored in great detail. Cross-validation was used and several measures for the "best" d_0 were examined. This led to the development of a much more efficient method for choosing a smoothing parameter, the concept of local variability as a function of disk radius. Each disk radius corresponds to a d_0, so by minimizing the local variability function the most appropriate d_0 can be chosen. Local variability functions were calculated for Ohio, New York and Florida. This analysis as opposed to cross-validation makes the task of modeling the entire United States a much smaller one. This model combined with cross-validation has been useful in detecting outliers in these data.

The evaluation of the moving average model led to comparing to Akima's method of bivariate linear interpolation. A cross-validatory comparison for adequacy of estimation was done. Also, contour maps using each method are drawn and compared. The local variability function analysis allows for comparison by cross-validation to not be a two-deep cross-validatory choice. Some drawbacks to comparing cross-validation estimates are pointed out. How different goals may prescribe different estimation techniques is discussed

The potential for further research in this field is shown. Time, which may be important in these analyses, has not been included because of data availability limitations. Using a time parameter similar to d_0, the current distance parameter, has been suggested. Simulations may also be useful in evaluating the moving average model. The distributional theory of the local variability theory function is yet to be explored. Nicholas P. Jewell

CHAPTER ONE

Air Quality Data

Introduction

Air Quality has been a concern because of the potential hazardous effect of pollution on health. Because of this concern the Clean Air Act of 1970 and EPA Regulations for State Implementation Plans (SIP's) require ambient air quality data resulting from air monitoring operations of State, Local, and Federal networks to report each calendar quarter to the Environmental Protection Agency. These data must be transmitted to EPA Regional Offices within 45 days after each quarterly reporting period. Within 30 days, EPA Regional offices must give these data to the EPA Aerometric and Emissions Reporting System (AEROS) of which Storage and Retrieval of Aerometric Data System (SAROAD) has been an operational part. AEROS is managed by National Air Data Branch, Monitoring and Data Analysis Division of Air Quality Planning and Standards. The National Air Data Branch (NADB) formed AEROS which is a large unified data system with analysis capabilities, uniform procedures, and expanded storage capability. SAROAD is just one of the numerous subsystems of AEROS. SAROAD contains the information on air quality. Anyone who has an interest can request summaries of these data and they will be provided by EPA (1976).

EPA publishes an annual summary of all data submitted in an effort to provide these data to participating agencies as well as the public. The air quality data which are discussed in this thesis are directly from computer reports generated by EPA's on line computer system. The type of pollutants considered

in this report and their corresponding sampling intervals are listed in Table 1.1.

TABLE 1.1. Summary of monitored pollutants and their sampling intervals.

Pollutant	Sampling Interval
carbon monoxide	1-hour
total hydrocarbons	1-hour
nitrogen dioxide	24-hour
nitrogen dioxide	1-hour
non-methane hydrocarbons	1-hour
total oxidants	1-hour
ozone	1-hour
sulfur dioxide	24-hour
sulfur dioxide	1-hour
total suspended particulate	1-hour
sulfate	24-hour

The basic file used in this work was produced as part of the PARAP/PAREP (Populations at Risk to Air/Environmental Pollution) project. It was derived from two major sources: 1974-1976 air quality summaries by air quality monitoring stations, and an air quality monitoring station directory containing latitude and longitude coordinates.

The air quality file came directly from 1974-1976 yearly summary data in the Environmental Protection Agency's SAROAD databank. For each pollutant and sampling interval and each monitoring station, data were first averaged over all measurement methods (weighting by the number of observations). Then three-year average calculations were derived from the yearly averages and installed in SEEDIS (Socio-Economic-Environmental Demographic Information System).

The monitoring station directory was found to be incomplete and partially erroneous. Manual checking was performed against other files and pub-

lished maps until the station coordinates were believed to be accurate within a few kilometers in all cases.

At the outset of the research for this thesis these data were in SEEDIS with the following known errors and omissions:

1. Much of the monitoring of air quality that was actually done is not in SAROAD and thus is not available for this study.
2. Values for hydrocarbons seem systematically much too high for some known and some unknown reasons.
3. Some station values were discarded since they seemed so large as to be erroneous (Compare Tables 1.2 and 1.2a.).

TABLE 1.2. Stations which were discarded.

Pollutant	Station	Arithmetic Mean ($\mu g / m^3$)	Geometric Mean ($\mu g / m^3$)
Non-methane Hydrocarbons	MT Billings	2700.44	2191.00
Ozone	NY Tonawanda	316.615	300.265
Sulfur Dioxide	NM Farmington	404.909	351.461
Sulfate	TN Union City	152.800	152.800

TABLE 1.2a. Means and Standard Deviations of Pollutants Listed Above.

Pollutant	Mean	Standard Deviation
Non-methane Hydrocarbons	253.38	35.38
Ozone	34.58	25.89
Sulfur Dioxide	13.57	8.89
Sulfate	7.06	0.11

The averages that are available from the yearly summaries are:

geometric means, arithmetic means, geometric standard deviations, and arithmetic standard deviations The units are in micrograms per cubic meter. Since the distribution of the data appears to be lognormal, a geometric mean is a better measure of average concentration than an arithmetic mean. The geometric mean is estimated by the average of the logarithms of the values, exponentiated to maintain the correct units. The three-year arithmetic mean, a, and the geometric mean, g, were derived in the following manner:

$$a = \frac{\sum_{i=1}^{3} n_i a_i}{n}$$

$$g = e^{\frac{\sum_{i=1}^{3} n_i \ln(g_i)}{n}}$$

i = the ith year.

n_i = the number of observations in year i.

$n = \sum_{i=1}^{3} n_i$ = the total number of observations in three year period.

g_i = geometric mean at year i

$$g_i = e^{\frac{\sum_{j=1}^{n_i} \ln(o_j)}{n_i}}$$

a_i = the arithmetic mean at year i.

$$a_i = \sum_{j=1}^{n_i} \frac{o_j}{n_i}$$

o_j = the jth observed value.

j = the index for the jth observation.

The three year average geometric mean g was set equal to zero if the geometric mean for any year was zero.

Potential Applications of Air Quality Data

From a public health perspective, probably the most important use of these data is to measure and compare air quality across well-defined population centers so that levels of human exposure for geographic regions can be estimated. One obvious approach involves, using a statistical technique, calculating some average level of pollution for an area and correlating this average with disease incidence for that area. Air pollution contour maps are also very useful for examining high and low areas of pollution and their relationship to geographic clusters of disease. The problems of ecological correlations are well known, and do not indicate that there is a correlation at the individual level. Yet ecologic data are cheaper to analyze than undertaking a case-control study. In most cases, a study is not performed at the individual level without strong indication from other types of evidence. This is because of the large expense involved in such studies. Ecological studies are one way of indicating the potential utility of such a study.

Estimating an average level of pollution for an area based on discrete point observations from monitoring stations is is a challenging problem. Each observation has an uncertainty associated with it. The amount of uncertainty is not clear from examination of the data alone. The stations are clustered in areas of high population density and within clusters, they are not evenly distributed. Since the stations are not distributed geographically in a regular fashion and one would expect values from adjacent stations to be highly correlated with distance from one another, a simple non-weighted average is inadequate. Also, estimates of human exposure may need to weight observations in more popu-

lated areas more heavily than observations taken in less populated areas. So the sample mean is seriously biased as an estimate of human exposure. Often what is done in this type of situation is to fit a curve and then take the average of the interpolated points rather than the data itself. Sometimes, rather than an interpolation curve, a weighted average of the data is used. Finding the best method of interpolation or the best weights is a problem for which the statistical theory is not well described. These data from PAREP are used here to apply some of the existing statistical techniques. Thus the performance of these methods used on these data can be examined.

Another useful summary of these data is through contour maps. Contour maps can be used to show peaks and valleys in the pollutant concentrations. For contour maps many points are to be predicted for a region. Perhaps a different estimation procedure should be used to predict contours than that which predicts only one point in a region. This possibility will be explored in the third chapter.

There are basically four different methods of geographic surface estimation. These are: trend surfaces, moving averages, interpolation, and Kriging (Ripley, 1981). Techniques similar to moving averages and Kriging are two-dimensional generalizations of one-dimensional curve-fitting techniques (Grant, 1957). Interpolated surfaces go through every data point and are similar to spline-fitted curves in one-dimension (Akima, 1978). Kriging is a method which unlike the other methods explicitly considers the correlation between values of the surface (Matheron, 1965). Kriging results in the "weight" of each point in a cluster being reduced.

Data Limitations

Several factors associated with air pollution data cause a statistical treatment of the data to appear intractable. This section will mention some of the most important problems. Some of these can be solved at least partially. In these cases an attempt to solve them has been made. In other cases which cannot be solved a mention of them will be made so that future data collection can be improved.

For each station measurement, there is an uncertainty associated with its value. This uncertainty arises from several sources. For example, the machine readings may be in error depending on the time of day, air temperature, humidity, and the skill of the person reading the measurement. This problem has been recognized in other areas such as blood pressure readings and there are analytic techniques to take account of this. This uncertainty is usually viewed as being stochastic and its value is assumed to follow some known distribution. For air quality data, this same path could be followed. A further complication is that different stations might use different measuring methods. These differences may be another source of variation not due to actual air quality. Since the data in SEEDIS have already been averaged over measurement methods (for each pollutant and sampling interval), the assumption that these differences are random must be made in order to compare values across regions.

The fact that each data value is a 3-year average lends itself to two important data reliability problems. Since each station was run (i), for various amounts of time and (ii), at probably different times; each data value does not

measure exactly the same quantity. For example, an average which comes from an instrument which ran each day of the three years is far more reliable than that from an instrument running once a week, assuming everything else is equal about the two values.

Since the gases in the air are obviously affected by the weather, there is seasonal variation in air pollution. Averaging over three years may be less biased than one year averages because of seasonal variation and the fact that measurements are not taken uniformly over the year.

Fortunately data on the percentage of time active are available over the entire three years and this quantity is taken into account in these analyses. Table 1.3 shows the average time stations were active for all pollutants and for all stations in the United States (+territories) [1]. The percentage of time active is calculated from the total number of observations in the following way.

$$\text{Percentage of Time Active} = \frac{100N}{1096Y}$$

where,

N = Number of Observations and

Y = 24 or 1 for 1-hour or 24-hour measurement sampling intervals.

1096 = the number of days in the three year period, 1974-1976.

So if a station is run once a week the number of observations is 156 and the percentage of time active would be as follows:

[1] The territories with monitoring stations are Guam, Virgin Islands, and Puerto Rico.

	Percent Time Active
24-hour intervals	100×156/ 1096 = 14.2%
1-hour intervals	100×156/ 1096×24 = 0.59%

(Data sampled on a 1-hour interval basis were usually from stations that were run continuously.)

To see the empirical distribution of percent time active of the stations, Figures 1.1 thru 1.11 show their distributions by pollutant for the whole US (+ territories).[2] From the total number of stations in each distribution, notice that Sulfur Dioxide (24-hour interval) and Total Suspended Particulate (24-hour interval) are measured more often than the others. Sulfur Dioxide (1-hour interval) does not have the greatest mean percent time active but is one of the greater ones at 41.5%. Sulfur Dioxide (1-hour interval) has more stations than any other pollutant that is measured, which also enters into adequate monitoring. Thus percent time active statistics indicate that sulfur dioxide and total suspended particulate are the most extensively monitored.

Tables 1.3a. and 1.3b. show average activity times for all measured pollutants. Average and median number of days (hours) per year are derived from the percent time active formula. For daily measurements the possible number of observations in one year, is approximately 365; while for hourly observations, it is 24×365 = 8760. This number gives an idea of the frequency of measurement.

[2] A percent time active greater than 100 percent can occur if a pollutant was simultaneously measured by two or more instruments at a station.

TABLE 1.3a. Activity Time for Pollutants Measured at 24-hour Intervals for United States, 1974-76.

Pollutant	Percent Time Active		Days/Year		Number of Active Stations
	mean	median	mean	median	
Nitrogen Dioxide	9.5	8.7	34.7	31.6	1796
Sulfur Dioxide	10.4	10.5	38.0	38.3	2440
Suspended Particulate	11.0	10.5	40.0	38.3	5473
Sulfate	4.9	2.7	18.0	9.6	1045

TABLE 1.3b. Activity Time for Pollutants Measured at 1-hour Intervals for United States, 1974-76.

Pollutant	Percent Time Active		Hours/Year		Number of Active Stations
	mean	median	mean	median	
Carbon Monoxide	45.3	43.9	3968.3	3845.6	588
Total Hydrocarbons	40.1	31.2	3512.8	2733.1	205
Nitrogen Dioxide	43.5	38.2	3810.6	3346.3	353
Non-methane Hydrocarbons	24.2	22.3	2119.9	1953.5	80
Total Oxidants	52.1	52.1	4564.0	4559.6	121
Ozone	35.2	30.9	3083.5	2706.8	621
Sulfur Dioxide	41.9	36.1	3670.4	3162.4	1050

Another reliability problem is related to the fact that some stations were not run for the entire three years (Table 1.4). Note that almost 60% of the stations measuring any pollutant monitored suspended particulate at some time in the 3 years. This is the most actively measured of all pollutants. There is considerable variability in the number of years stations were active. Since the data available in SEEDIS do not indicate which stations were run in specific years, we could assume that average air quality does not change substantially over this time span, yet we know that meteorological and topographic factors do change. Then let us take this measure as an average measure of human exposure to air quality during this time frame. With this approach a 1974 measurement can be

compared to those taken in 1976. In the future it would be preferred to eliminate this assumption but the data, although available, are not easily accessible at this time.

TABLE 1.4. Frequency Distributions of Number of Years Active by Pollutant for the United States from 1974 thru 1976.

Pollutant	Sampling Interval (hours)	Percentage of Stations Active at some time within:		
		1 year	2 years	3 years
Carbon Monoxide	1	27.4	24.7	47.9
Total Hydrocarbons	1	33.2	21.0	45.8
Nitrogen Dioxide	1	33.1	19.8	47.0
Non-methane Hydrocarbons	1	32.1	28.4	39.5
Total Oxidants	1	24.0	19.8	56.2
Ozone	1	38.4	30.7	30.9
Sulfur Dioxide	1	33.4	31.3	35.3
Nitrogen Dioxide	24	24.3	26.5	49.2
Sulfur Dioxide	24	33.8	28.4	37.8
Suspended Particulate	24	23.3	19.4	57.3
Sulfate	24	33.2	50.5	17.3

Another important reliability issue is illustrated by the number of stations passing certain summary criteria. Table 1.5 shows these distributions by pollutant. The criteria consist of the following (EPA 1976).

A. For continuous observations with sampling intervals of less than 24 hours the criteria are:

1. Data representing quarterly periods must reflect a minimum of 75% of the total number of possible observations for the applicable quarter.

2. Data representing annual periods must reflect a minimum of 75%

of the total number of observations for the applicable year.

B. For noncontinuous observations with sampling intervals of 24 hrs or greater the criteria are as follows:

1. Data representing quarters must reflect a minimum of five observations for the applicable quarter. Should there be no measurements in one of the three months of the quarter, each remaining month month must have no less than two observations reported for the applicable period.

2. Data representing annual periods must reflect four quarters of observations that have satisfied quarterly criteria.

These criteria insure that samples are fairly well spread over the sampling period and that there are enough samples to be a reliable summary of pollutant level for the period of time covered. As is shown in Table 1.5, many stations did not pass these criteria. Sulfur dioxide and total suspended particulate have the highest percentage of stations which passed.

TABLE 1.5. Frequency Distribution of Number of Years that Active Stations Passed Summary Criteria by Pollutant for United States (+ territories) during 1974-76.

Pollutant	Sampling Interval (hours)	Percentage of Active Stations Passing For			
		0 years	1 year	2 years	3 years
Carbon Monoxide	1	46.4	25.3	16.3	11.9
Total Hydrocarbons	1	61.9	13.2	23.4	1.5
Nitrogen Dioxide	1	48.4	21.8	12.5	14.2
Non-methane Hydrocarbons	1	90.0	8.8	1.2	0.0
Total Oxidants	1	36.4	20.1	27.3	15.7
Ozone	1	158.1	26.6	12.2	3.1
Sulfur Dioxide	1	51.2	28.5	13.1	7.1
Nitrogen Dioxide	24	28.9	30.6	20.3	22.2
Sulfur Dioxide	24	26.9	24.7	20.8	27.6
Suspended Particulate	24	28.4	25.2	22.2	24.3
Sulfate	24	65.3	28.6	5.7	0.4

The irregular spatial distribution of the data points raises another data analysis problem. Certain areas are monitored more heavily than other areas as mentioned. The areas that are monitored most heavily are metropolitan areas where population density is highest. The data are therefore available in the most appropriate places for measuring human exposure. Yet for making general contours plots, this can be a problem. Contour estimates would be much more accurate if the data were regularly distributed.

Outliers are a typical problem for any data set. Air quality data are no exception. Basically there are two kinds of outliers in air quality data. One kind of outlier are those which are completely nonsensical such as a negative value of concentration. A more subtle kind of outlier is one that does not follow the pattern established by the other points. There are at least two ways these outliers can arise; (i) measurement error and (ii) accurate measurements that are measuring local values not representative of the region.

Another problem that is typical of any data set, especially a large one, is the number of missing values. In the PAREP air quality data under study the number of missing values is quite large and varies with the pollutant. Table 1.6 shows the amount of monitoring that took place in 1974-1976 for the whole US (+ territories) for all pollutant measurements under consideration. Total suspended particulate and sulfur dioxide are thought to be respiratory irritants (Landau, 1971), especially in combination, suggesting why they have been historically more heavily monitored than the other pollutants.

TABLE 1.6. Number of Active Stations By Pollutant.

Pollutant	Number of Active Stations
Carbon Monoxide (1-hour interval)	588
Total Hydrocarbons (1-hour interval)	205
Nitrogen Dioxide (24-hour interval)	1796
Nitrogen Dioxide (1-hour interval)	353
* Nitrogen Dioxide (Either or both intervals)	2006
Non-methane Hydrocarbons (1-hour interval)	80
Total Oxidants (1-hour interval)	121
Ozone (1-hour interval)	621
Sulfur Dioxide (24-hour interval)	2440
Sulfur Dioxide (1-hour interval)	1050
* Sulfur Dioxide (Either or both intervals)	3409
Total Suspended Particulate (24-hour interval)	5473
Sulfate (24-hour interval)	1045

*Some stations used both sampling intervals.

For the analyses here, Digital Equipment Corporation's VAX 11/780 computer was used. For the more complicated analyses, a smaller area with high monitoring density needed to be used in order not to use excessive amounts of computer time. This was mostly for the sake of applying developed statistical techniques and developing methods, not for extrapolating to the rest of the country. A more efficient technique as an alternative to that which is presently used is developed. Two different machines were used with two different

operating systems. The two operating systems were VMS and UNIX. In UNIX, access to the statistical package, system S (Becker, 1981) was available. System S was used for much of the analysis because of its interpolation and contour programs. The VMS machine is the machine which houses SEEDIS and all the air quality data.

Air Quality by Location

After checking for and removing most of the more obvious outliers there were 6625 monitoring stations active in 1974-76 in all. Not all of the stations have values for all pollutants. Of the 6625 monitoring stations, there are only 5777 distinct locations. In some cases the exact location of a station was not known, in which case the population centroid of the corresponding city was used. Assuming that monitoring stations are alike in the way they measure, the measurements can be averaged at a location to have only one value of pollution for each station with the same (lat,long) coordinate. This makes it easier to investigate interpolation strategies by cross-validation techniques. A new file was made with 5777 data points and installed in SEEDIS (Johnson (1982)). The Air Quality Monitoring Station by Location (AQMSLOC) File is the one that is used throughout the remainder of this study. Stations at the same location were considered as a single observation with a concentration equal to the geometric mean (weighted by the number of observations) of the individual values.

Two Pollutants in Ohio State

Tables 1.3 thru 1.6 showed that the most heavily monitored pollutants are sulfur dioxide and total suspended particulate. Even though some pollutants are monitored more often on the average, they have fewer stations and pass summary criteria less often. These pollutants are also the most suspected health hazards. We have restricted most of the statistical analyses to these two pollutants. There are two main reasons for this restriction. There are more data for these pollutants which is important for predicting air quality and these pollutants are also of interest to the health field. We also decided to analyze only the state of Ohio because of the difficulty of analyzing 5777 data points at once. Ohio has 455 stations within the boundary of the state. No reason existed to analyze the whole nation at once, since the monitoring stations within or near a homogeneous area can be used to estimate the population exposure in that area. So for solid statistical results, the analyses shall be confined largely to the state with the most monitoring stations and to the extensively monitored pollutants. Estimation techniques should do "best" in the most densely monitored areas, so the methodology will be examined and evaluated in these areas.

Methodological Perplexities

Statistically, there are two interesting problems which can be explored with these data. One is to map pollution gradients in detail for a region. The other is to estimate an average level of pollution for an region in order to compare with other geographic regions. These two ideas are connected through sur-

face estimation. The surface estimates points which can be used for making contours and for calculating averages. A question that arises is whether the same estimation procedure will do well for both purposes.

The statistical techniques for estimating a non-linear bivariate function are described by the general name of non-parametric regression. The desirable feature of non-parametric methods is that no distributional assumptions are required for their validity. In fact non-parametric regression techniques are used to estimate density. Commonly used non-parametric regression methods include nearest-neighbor techniques, splines, kernels, and partition-based methods. To estimate an arbitrary point in the plane by the technique of nearest neighbors, the k nearest neighboring data points are used. The parameter, k, usually remains constant and is the number the number of observations used for predicting any point in the plane. Splines or bivariate polynomial interpolation fit a polynomial of a fixed degree to each triangle in the plane. The triangles are defined by the data points. So in spline fitting to predict an arbitrary point in the plane, one sees which triangle it lies in and applies the appropriate degree polynomial, using the data at the vertices of the triangle, to determine the polynomial coefficients; and then the fitted polynomial is evaluated at the point of interest. Partition-based methods take a particular size partition; then all the points within this partition are used to evaluate arbitrary non data points in the partition. The kernel method uses a weighted average of the data points inside a window. The size of the window is chosen in various ways. For example, its size can vary with the number of data points or some other relevant parameter. The moving average which is discussed in chapter two is classified under the family of kernel techniques. Choosing the appropriate technique for a specific problem is of interest.

Suppose one has chosen a technique. Within a technique there are still choices to be made of certain control parameters. For example, in nearest-neighbor techniques the 'best' k must be chosen. With splines, it is the choice of the degree of the polynomial. In partition-based methods the size and type of the partition should be carefully chosen. In the kernel method, not only the weights need to be chosen, but also the size of the window.

There are a variety of methods for choosing techniques or control parameters of a technique. Many questions arise as to which method of choice should be used. Perhaps one method is better for choosing the estimation procedure overall while another is more useful for picking a control parameter. Cross-validation can be used for both these tasks and will be explored. We shall use cross-validation to choose between two methods of interpolation and also to choose the control parameters of one of the methods. We will try to choose between bivariate linear interpolation and a moving average method where the moving average requires a choice of parameter. Stone (1973) discusses "choice and assessment of a statistical prediction" applying cross-validation criteria.

Summarizing the assessment of the two-dimensional moving average model by the cross-validation technique can also be done in several ways. The issue is not clear-cut. A loss function must be chosen which does not favor non-optimal estimates. Ripley (1981) points out that squared error may be dominated by a few data points which are outliers. Ripley also mentions that weighting of the prediction errors by their standard errors and then forming a sum of squares "biases the comparison in favor of covariances which give high standard errors by means of small covariances at typical interpoint distances". In other words, outliers and inefficient estimates must be considered when using a squared error loss function. Perhaps some sort of trimmed squared error would

be the best to use. A trimmed squared error is the regular squared error with a certain proportion of extreme high and low values removed from the data. Different loss functions will be explored.

The concept 'local variability' which is a function of disk radius has been developed. The minimum local variability gives a disk radius which contains on the average the most homogeneous collection of points provided the disks are drawn around each station location. The disk radius has a correspondence with the control parameter in the moving average method. Thus local variability analysis is offered as an alternative to cross-validation for choosing the parameter. By analyzing local variability the problem which arises in choosing a loss function is eliminated.

An Overview of the Following Chapters

Chapter 2 gives a detailed description of the two-dimensional moving average model. Important properties of the method are discussed. In the SEEDIS implementation of this model, different values of the control parameter can be selected. General utility of this kernel method for detecting outliers will also be shown. The method for which sulfur dioxide 24-hour and 1-hour measurements are combined and the adaptability of our model to this linear combination is given. Cross-validation is discussed as an integral part of the two dimensional moving average model. Local variability is presented as an alternative to cross-validation. A control parameter is chosen using cross-validation for sulfur dioxide and suspended particulate in Ohio State. Local variability functions are calculated for the states of Ohio, New York and Florida to check con-

sistency in the shape of our function.

Chapter 3 gives an explanation of the polynomial interpolation method that is used. With sulfur dioxide and suspended particulate in a heavily monitored area, the state of Ohio, linear interpolation is done using system S. The moving average is also calculated for the same pollutants. Linear interpolation and the moving average model are compared for their usefulness in both contouring and estimation of a certain point in an area. Cross-validation is used to assess which interpolation technique might be "better". Some problems with cross-validation estimates are pointed out.

Finally, chapter 4 contains a summary including a synopsis of the problems that arise in this kind of research, conclusions which can be drawn from this study and the possibility for further research in this area.

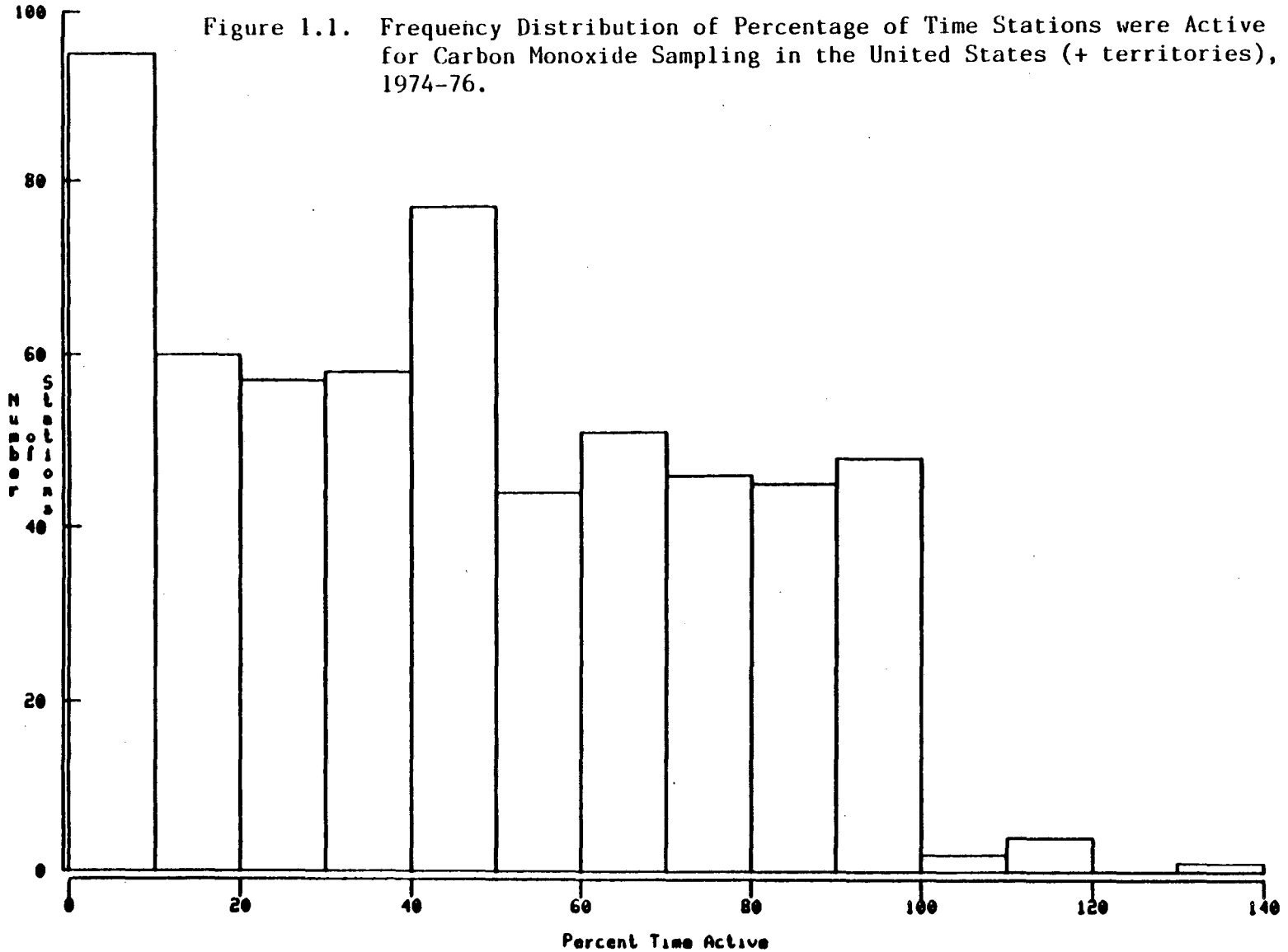


Figure 1.1. Frequency Distribution of Percentage of Time Stations were Active for Carbon Monoxide Sampling in the United States (+ territories), 1974-76.

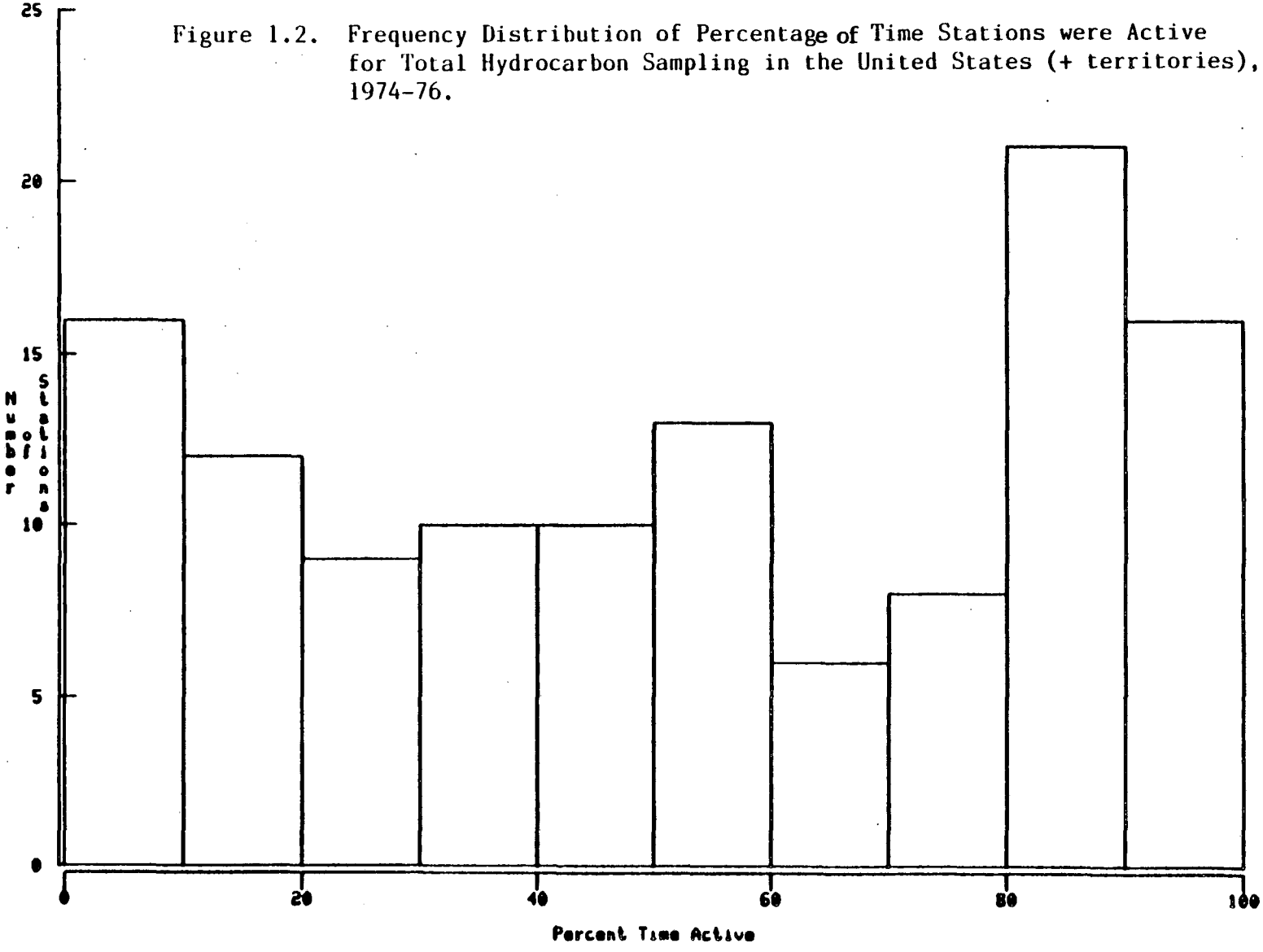
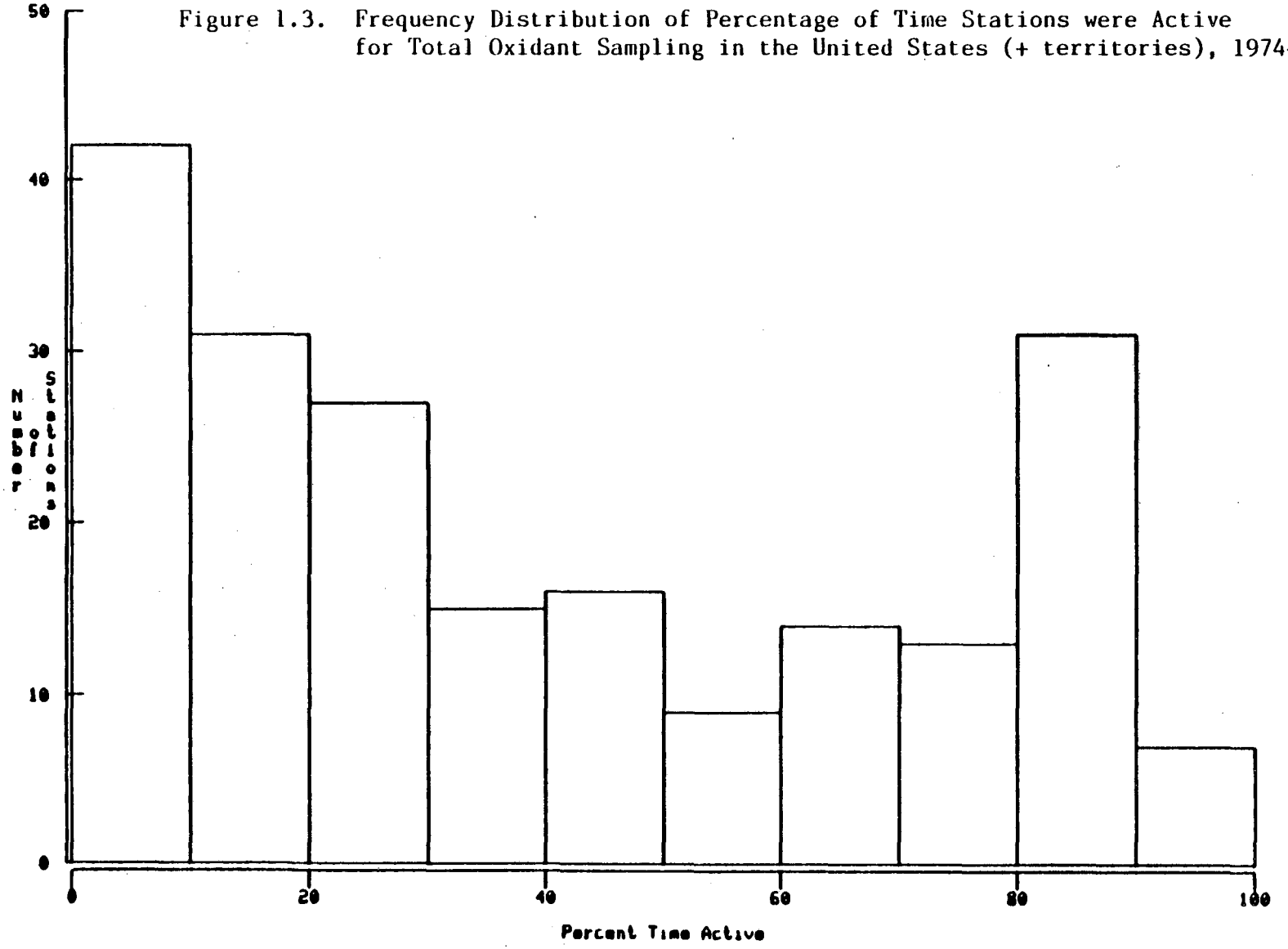


Figure 1.2. Frequency Distribution of Percentage of Time Stations were Active for Total Hydrocarbon Sampling in the United States (+ territories), 1974-76.

Figure 1.3. Frequency Distribution of Percentage of Time Stations were Active for Total Oxidant Sampling in the United States (+ territories), 1974-76.



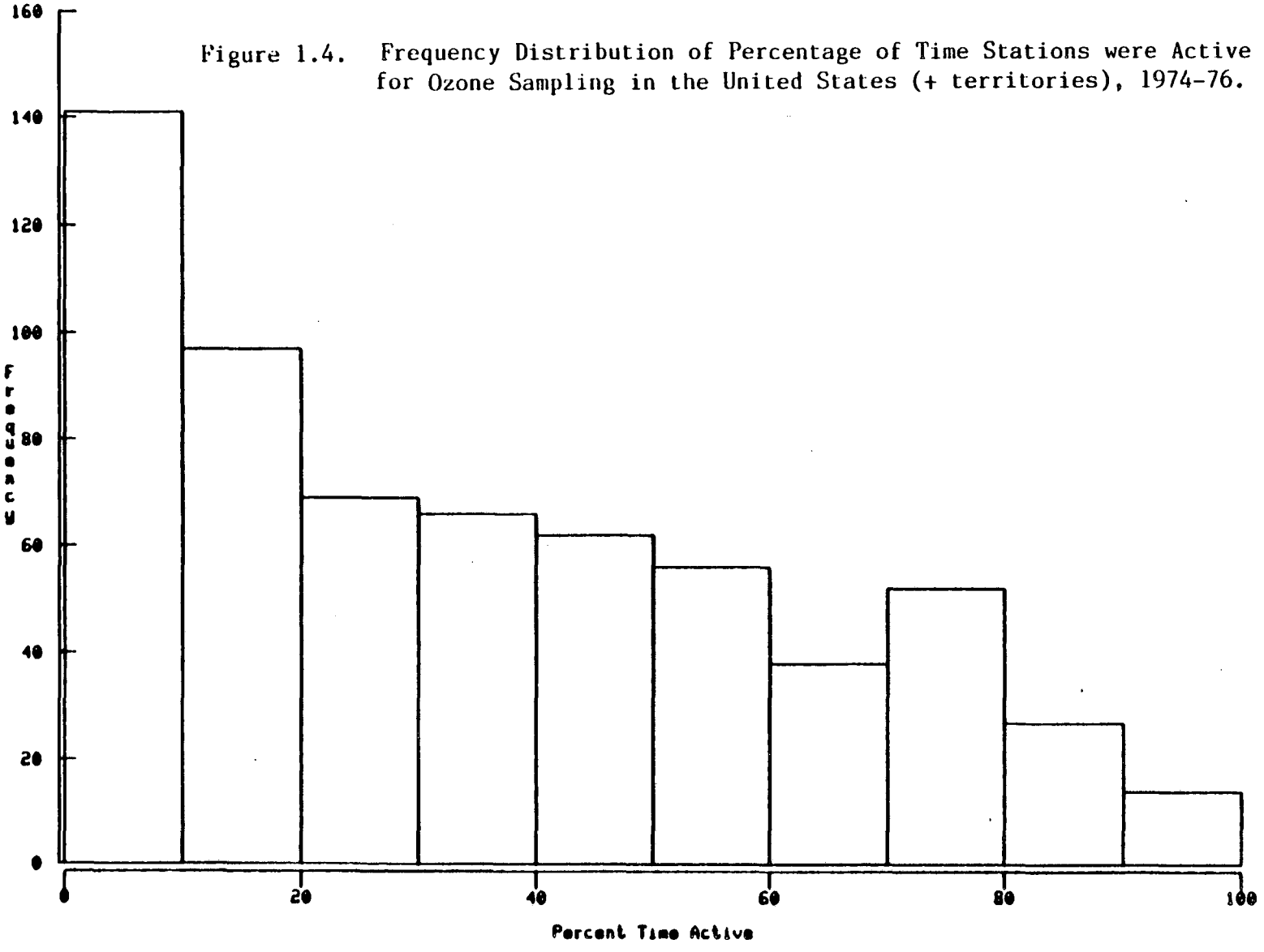


Figure 1.4. Frequency Distribution of Percentage of Time Stations were Active for Ozone Sampling in the United States (+ territories), 1974-76.

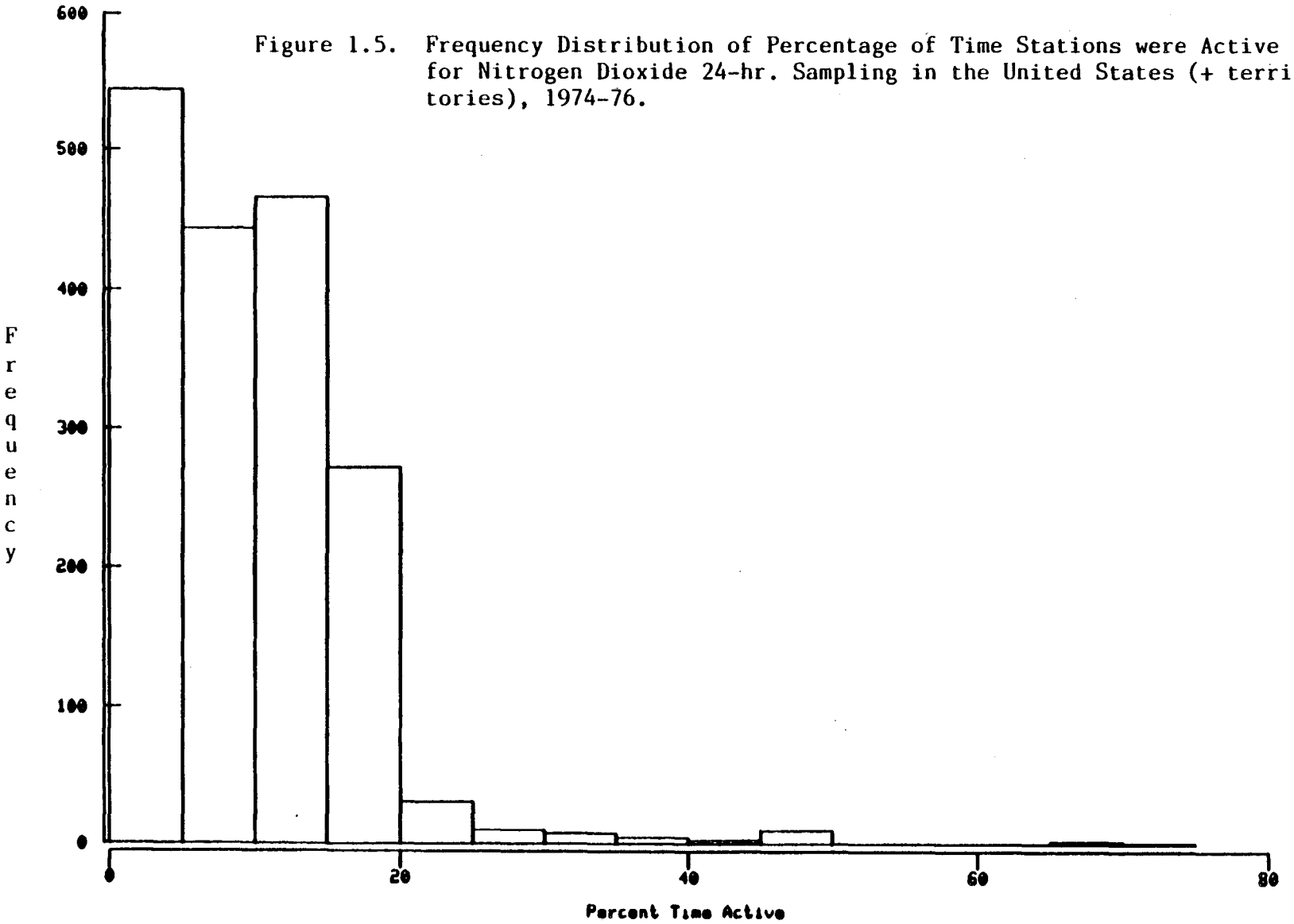


Figure 1.5. Frequency Distribution of Percentage of Time Stations were Active for Nitrogen Dioxide 24-hr. Sampling in the United States (+ territories), 1974-76.

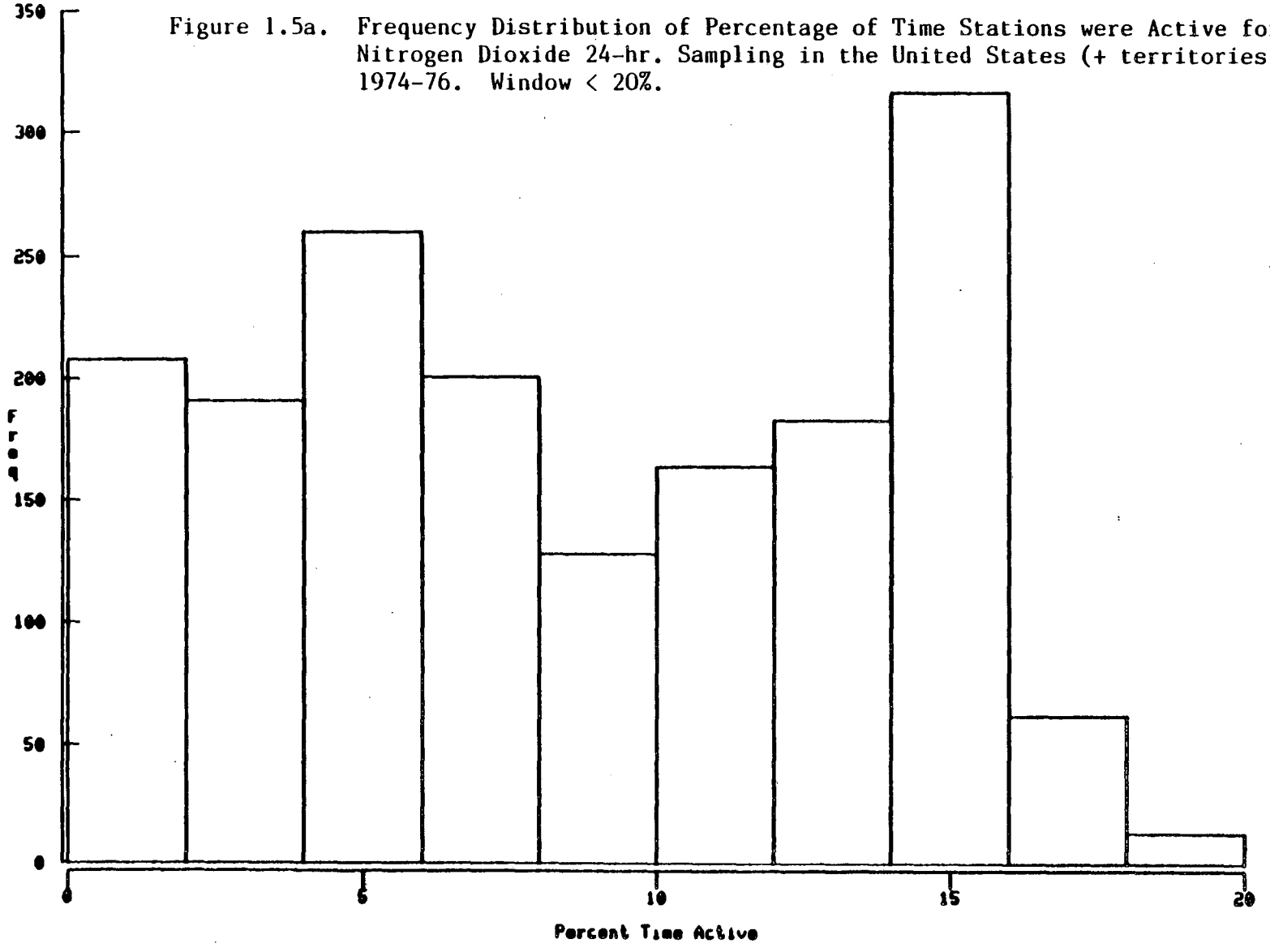
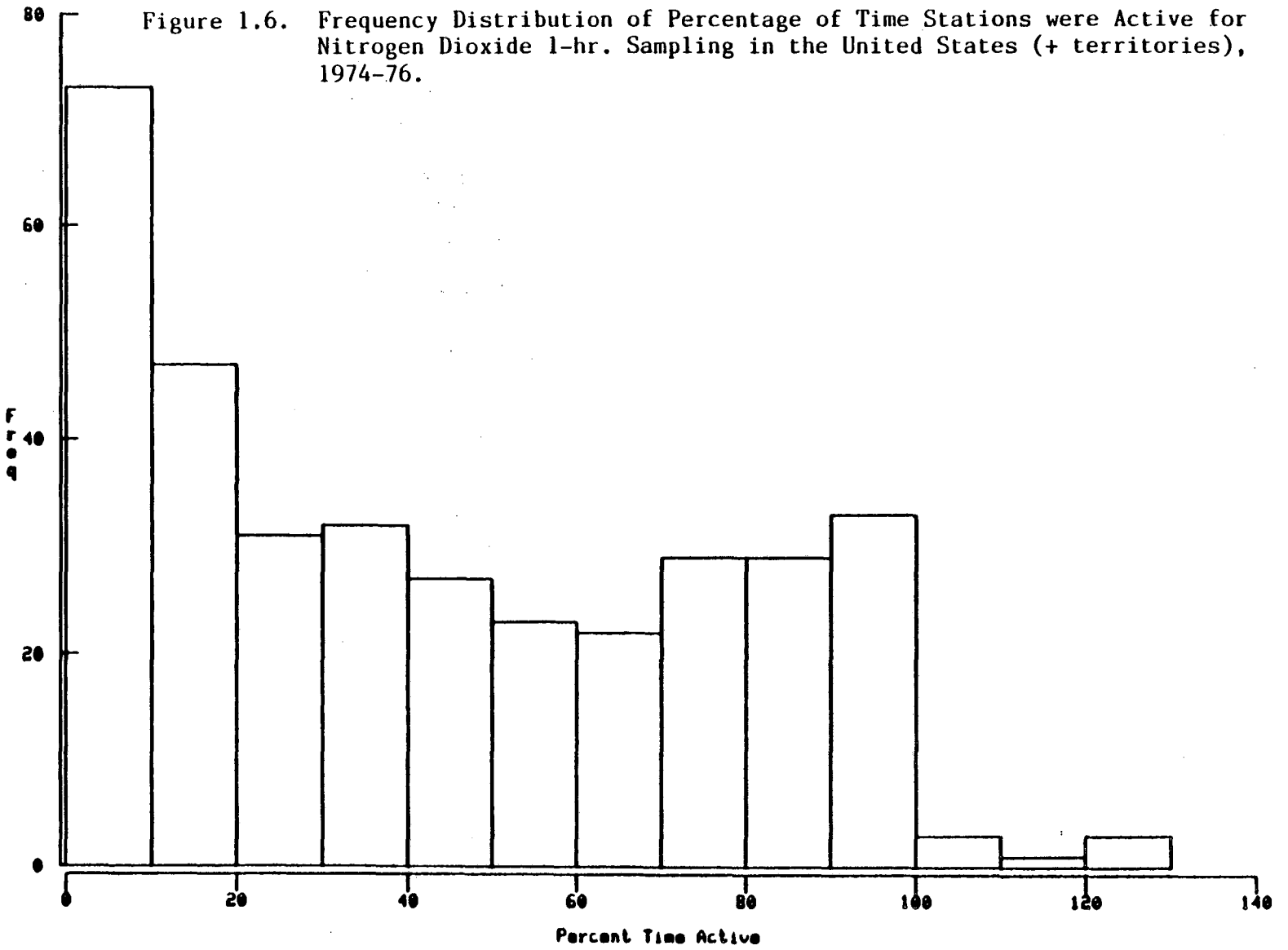


Figure 1.5a. Frequency Distribution of Percentage of Time Stations were Active for Nitrogen Dioxide 24-hr. Sampling in the United States (+ territories), 1974-76. Window < 20%.

Figure 1.6. Frequency Distribution of Percentage of Time Stations were Active for Nitrogen Dioxide 1-hr. Sampling in the United States (+ territories), 1974-76.



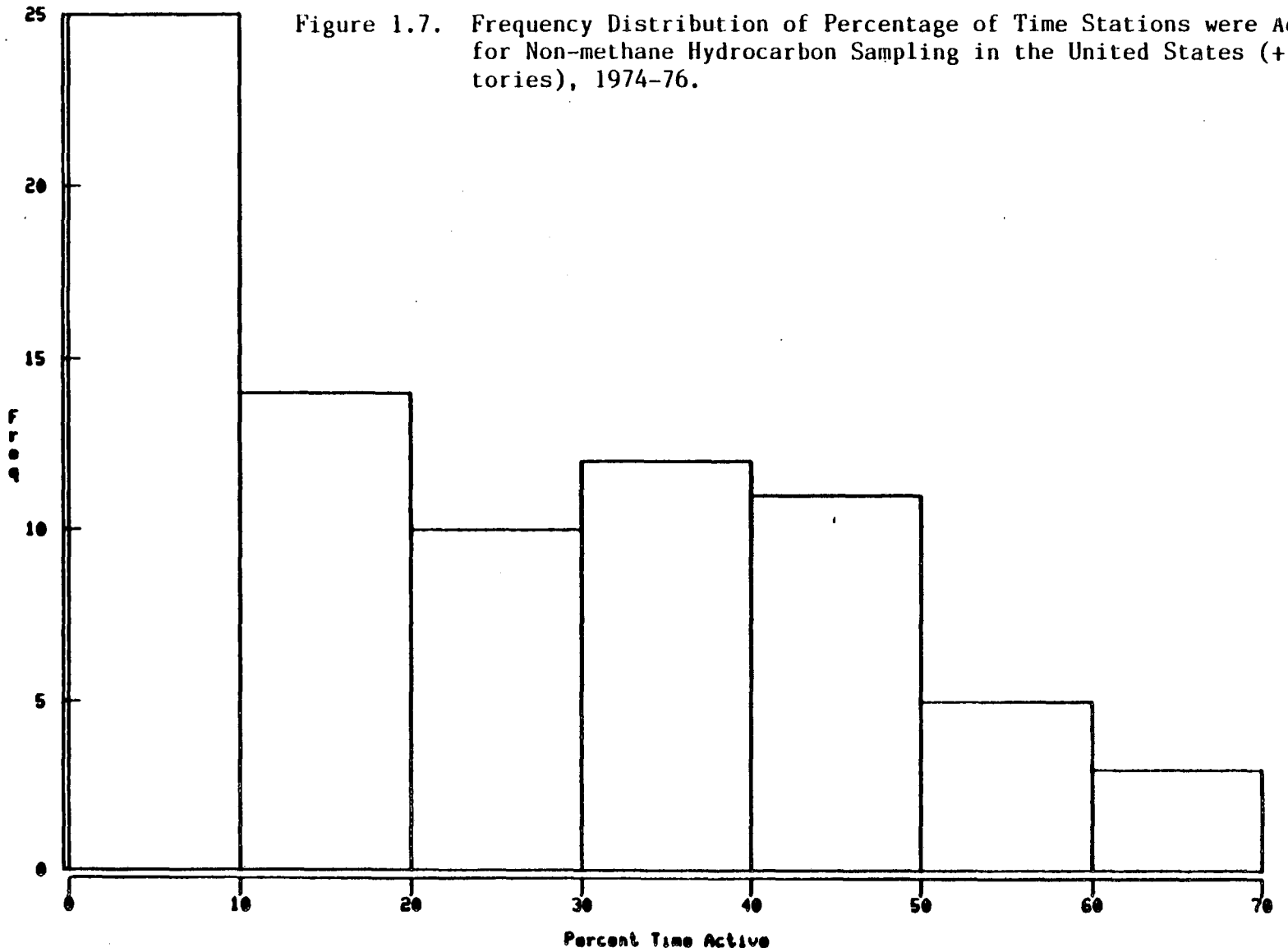
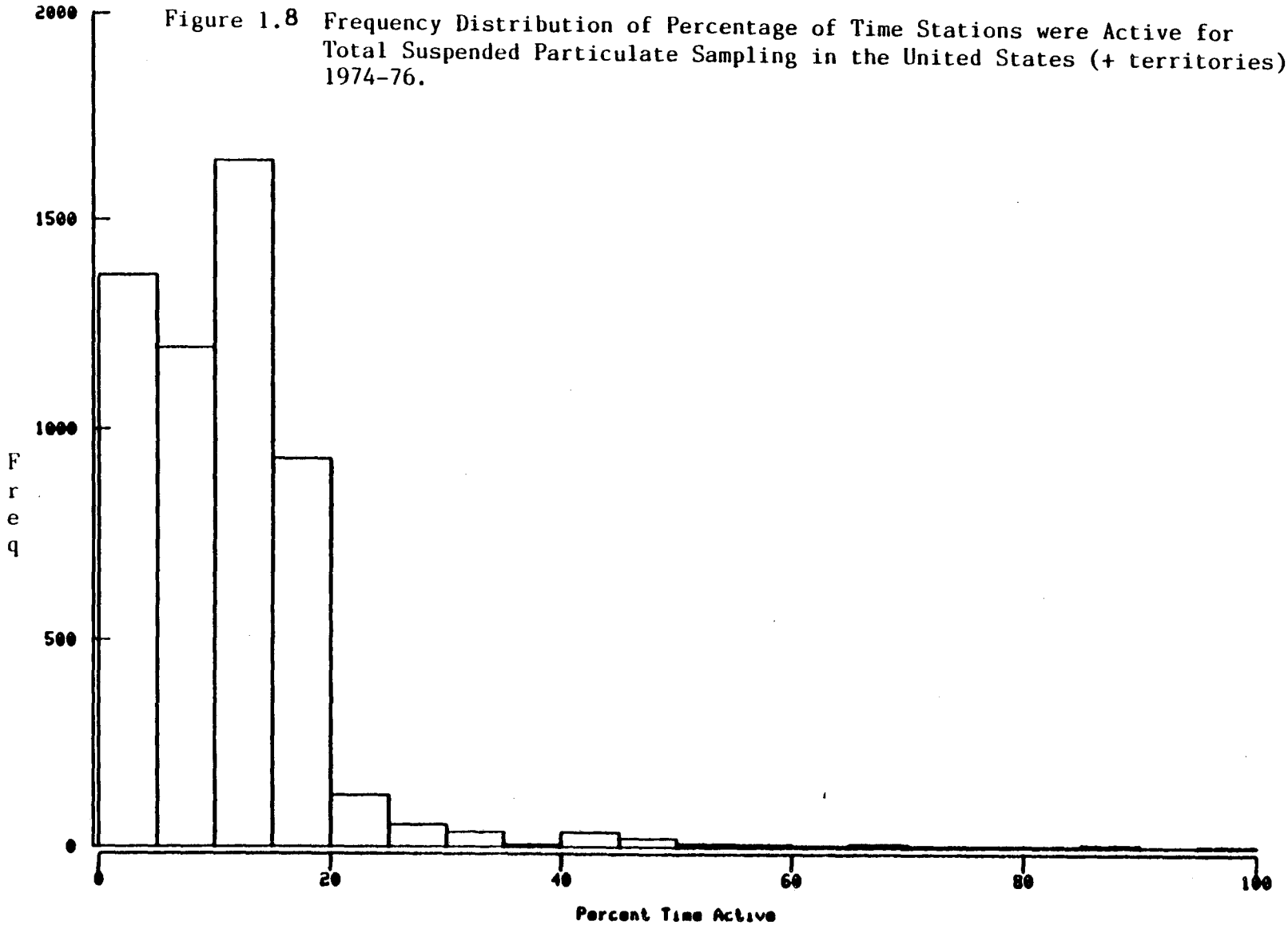


Figure 1.7. Frequency Distribution of Percentage of Time Stations were Active for Non-methane Hydrocarbon Sampling in the United States (+ territories), 1974-76.

Figure 1.8 Frequency Distribution of Percentage of Time Stations were Active for Total Suspended Particulate Sampling in the United States (+ territories), 1974-76.



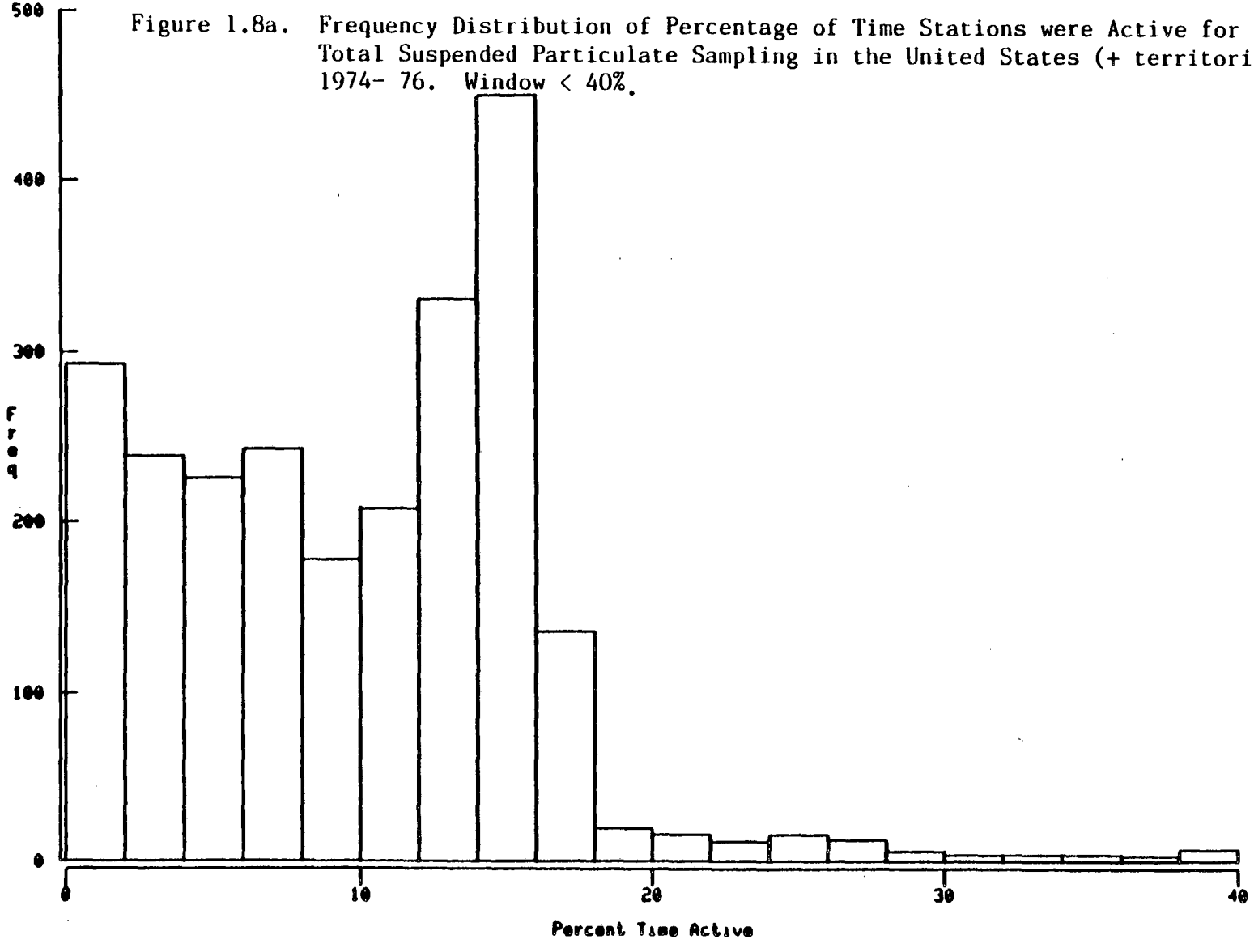
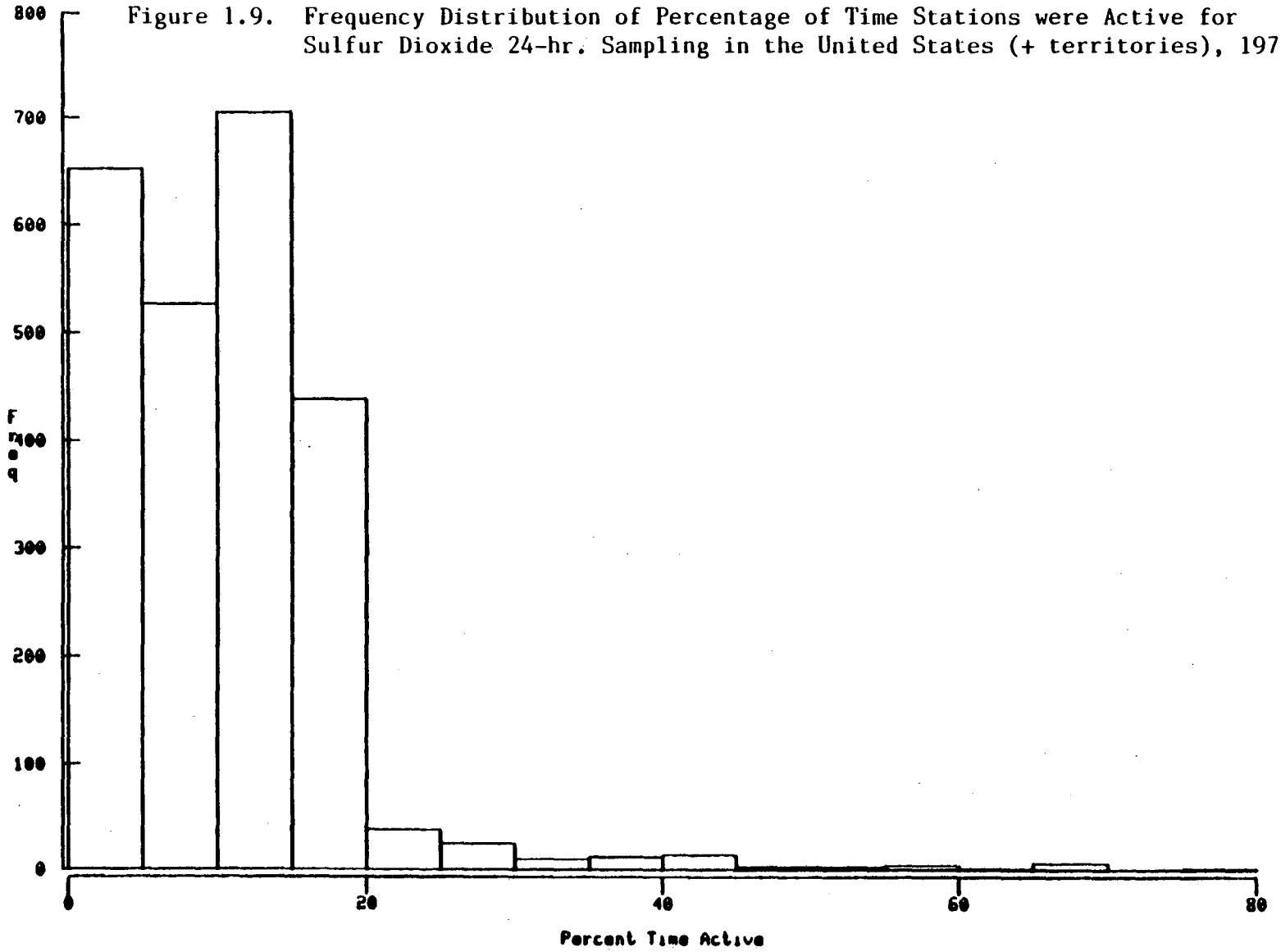


Figure 1.8a. Frequency Distribution of Percentage of Time Stations were Active for Total Suspended Particulate Sampling in the United States (+ territories), 1974- 76. Window < 40%.

Figure 1.9. Frequency Distribution of Percentage of Time Stations were Active for Sulfur Dioxide 24-hr. Sampling in the United States (+ territories), 1974-76.



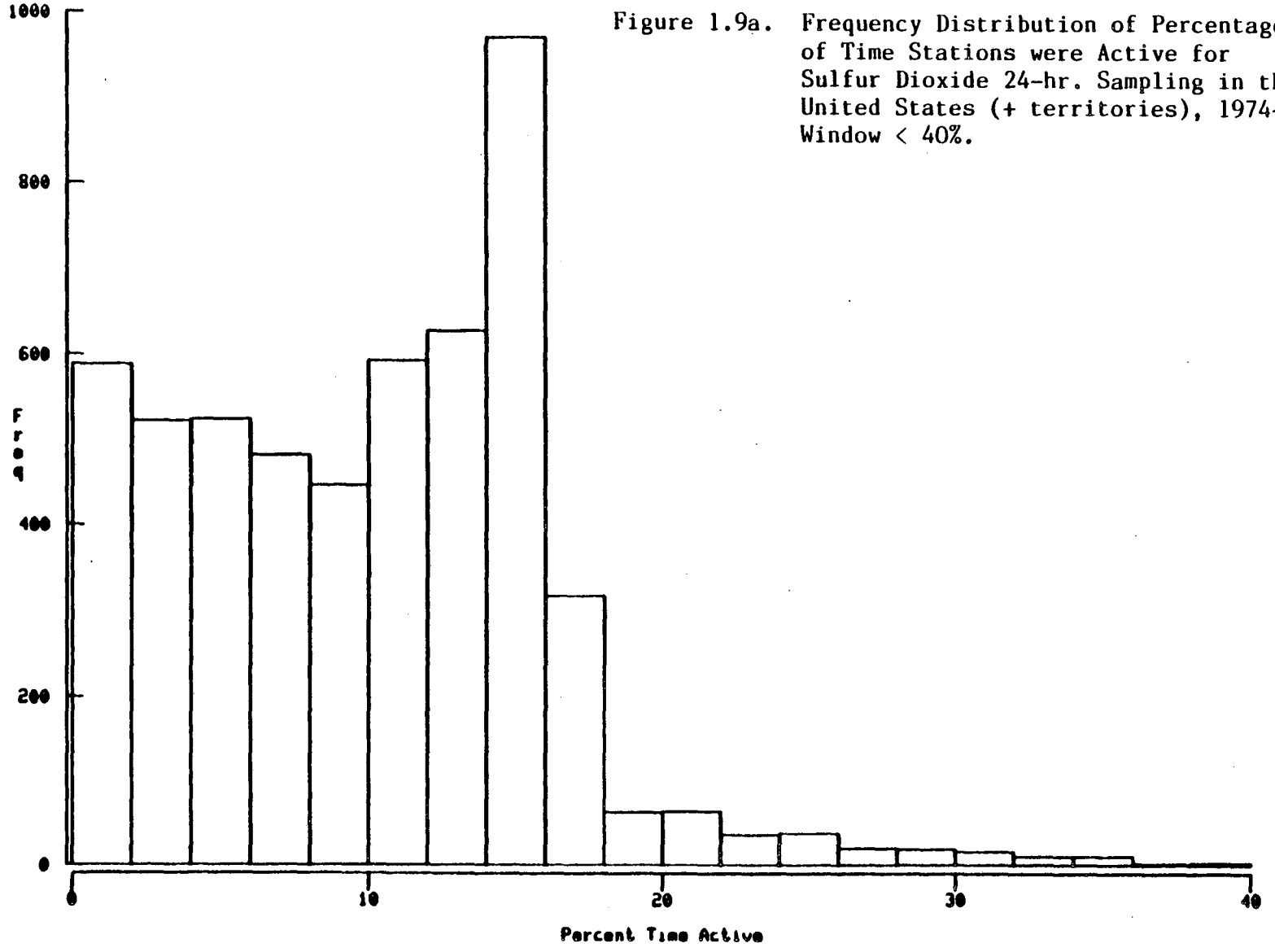


Figure 1.9a. Frequency Distribution of Percentage of Time Stations were Active for Sulfur Dioxide 24-hr. Sampling in the United States (+ territories), 1974-76. Window < 40%.

Figure 1.10. Frequency Distribution of Percentage of Time Stations were Active for Sulfur Dioxide 1-hr. Sampling in the United States (+ territories), 1974-76.

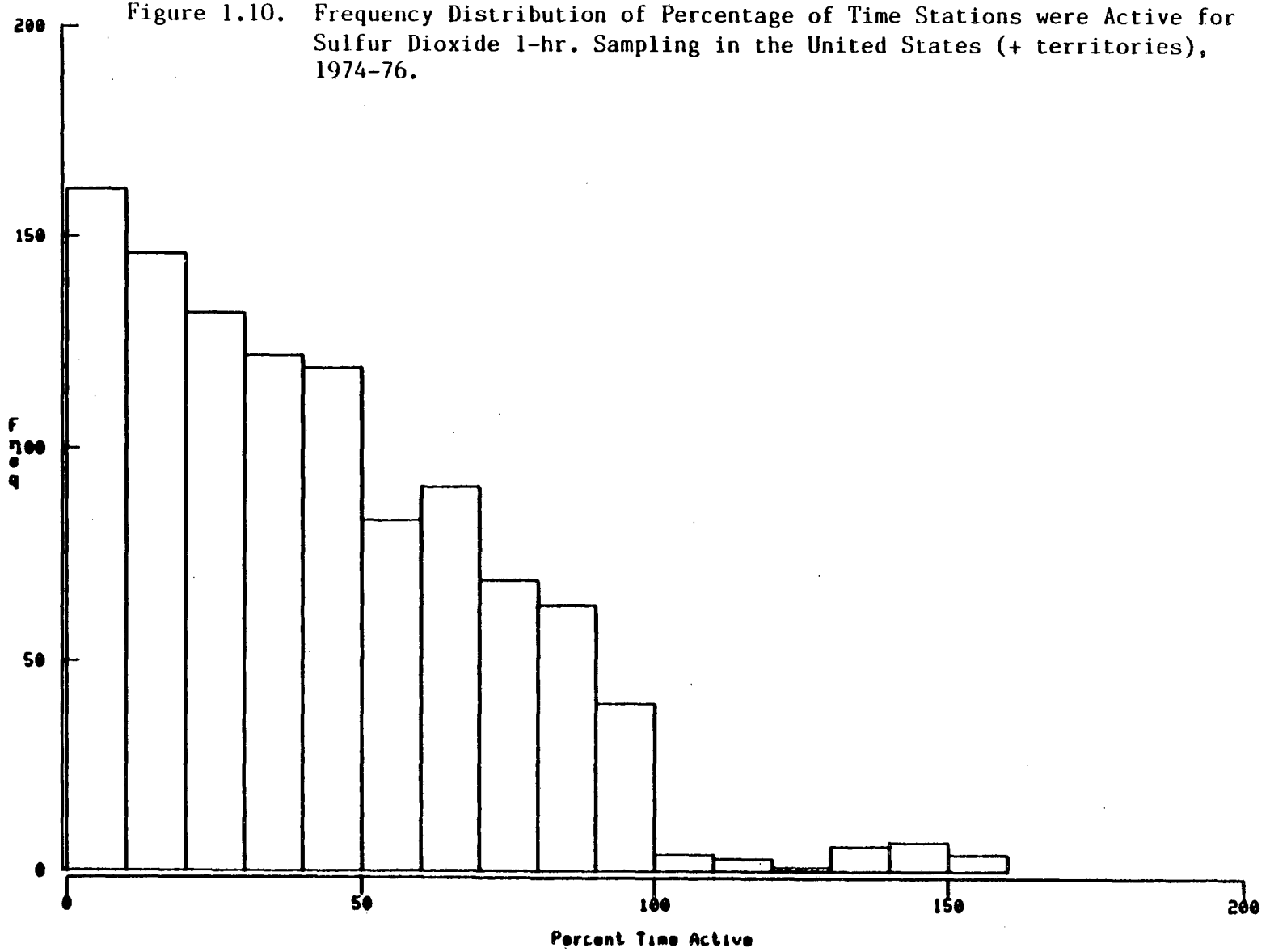
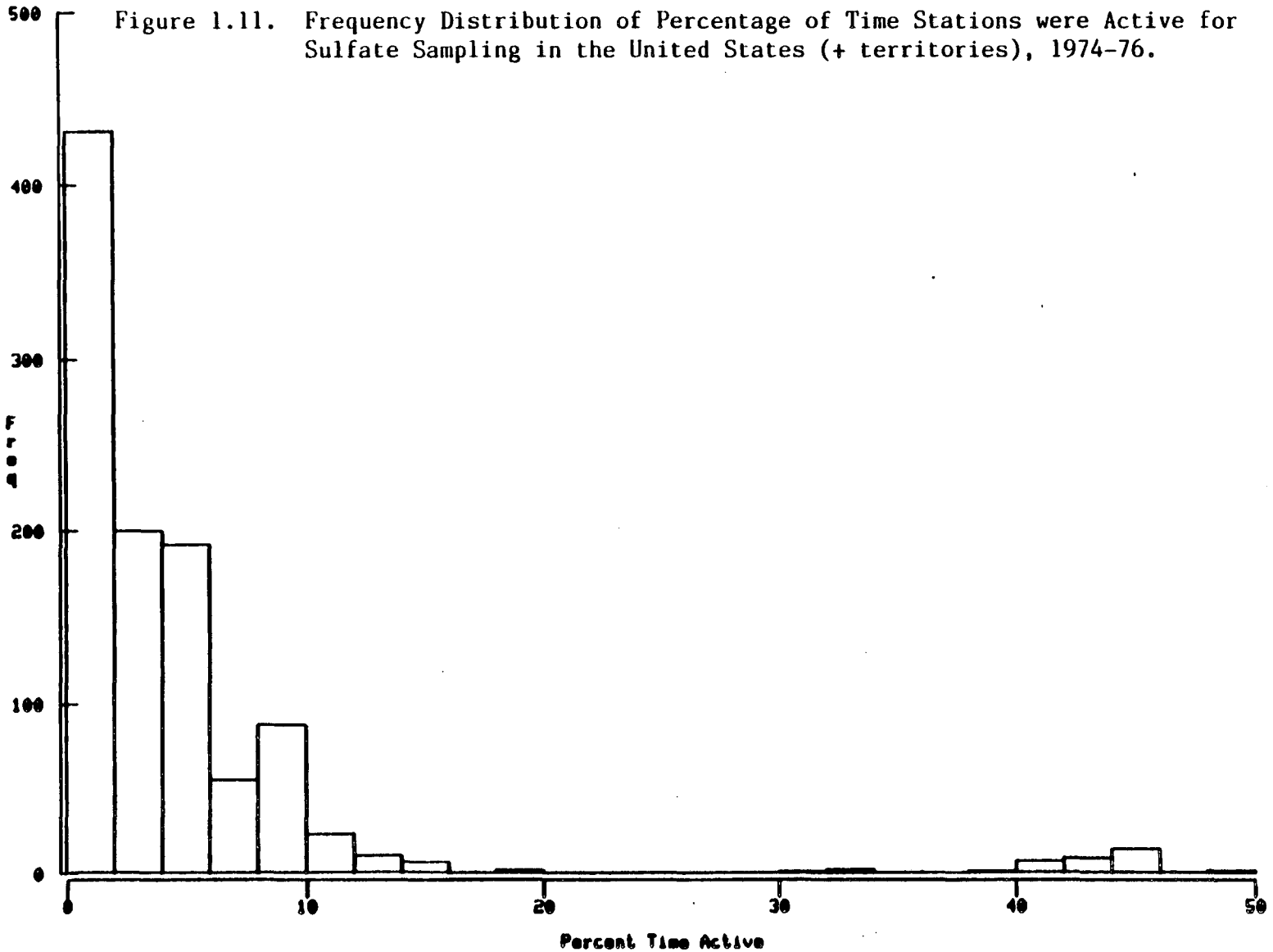


Figure 1.11. Frequency Distribution of Percentage of Time Stations were Active for Sulfate Sampling in the United States (+ territories), 1974-76.



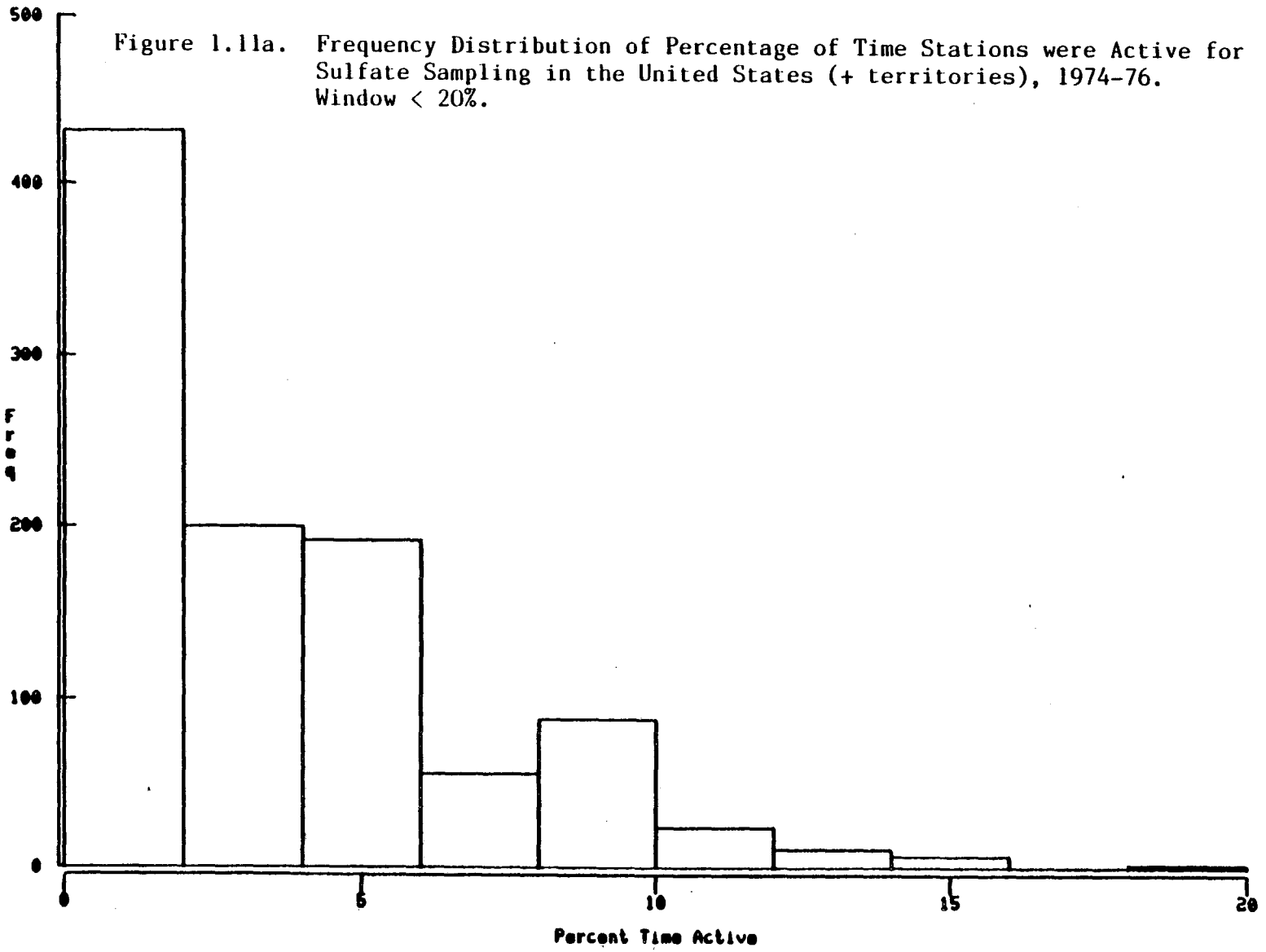


Figure 1.11a. Frequency Distribution of Percentage of Time Stations were Active for Sulfate Sampling in the United States (+ territories), 1974-76. Window < 20%.

CHAPTER TWO

Geographic Estimation of Air Quality

The Two-Dimensional Moving Average Model

Observations of air quality from monitoring stations are given and we wish to estimate the air quality for an entire geographic area. The sample points arise from stations which are not always in the area and which are not placed on a regular grid. The monitoring stations are placed where observers are available, usually in metropolitan areas rather than more rural or suburban regions. Under these circumstances, the sample mean may be seriously biased as an estimator of the mean level of pollution. An alternative is to fit an interpolating surface to the data. There are many methods of interpolating or smoothing data. Ripley (1981) gives an excellent review of the various smoothing techniques which apply to this problem. He briefly discusses trend surfaces, moving averages, splines, tessellations, triangulations, and Kriging. These methods estimate a surface from spatial data and use this surface to estimate values at given points within a geographic region.

The model we will explore is similar to the kernel method of nonparametric regression and is also like a moving average used in time series models. A time series moving average estimates a specific point as a weighted average of the points nearby in time where the weights change as a function of the distance to the point of estimation. The method described here is similar to this approach where instead of one-dimensional time we are using two-dimensional geography. The coordinates are the latitute and longitude of the

measurements. The weights change as a function of the geographic distance and the percentage of time a station was active. The weights can be written as

$$\omega_i = \omega(\mathbf{x}, \mathbf{x}_i) = p_i e^{-.5 d_i^2 / d_0^2}$$

where p_i = the percentage of time station i was active and d_i = distance between $\mathbf{x}$ and $\mathbf{x}_i$. The parameter d_0 is a control parameter. The units of d_i and d_0 are of distance which is expressed in kilometers. The choice of d_0 is discussed in detail in this chapter. Its value may change for different problems. These weights are used in the PAREP project for estimating pollution concentration at population centroids in geographic areas. According to Merrill (1982), this choice of ω_i was made for three reasons: "i) the estimated function should be smooth in the vicinity of the measured points; ii) the estimated function need not pass directly through all the measured points; iii) the area integral of the estimated function should be finite, so that distant points can be ignored in the calculation." Continuing with the notation of the section on moving averages in Ripley (1981), we can write

$$\lambda_i = \frac{\omega_i}{\sum_{i=1}^{i=N} \omega_i}$$

The λ_i's are the weights used in the moving average model.

N = the number of stations used in the average.

Then

$$\sum_{i=1}^{i=N} \lambda_i = 1 \quad and, \text{ for } a \ fixed \ d_0, \ as \ d_i \to 0, \ \omega_i \to p_i$$

These weights are relative weights and their sum is 1. For a fixed d_0, as the distance from the point of estimation to a station gets smaller, the weight in the weighted average for that station's value approaches the percentage of time that station was active.

The estimate at x is then given by

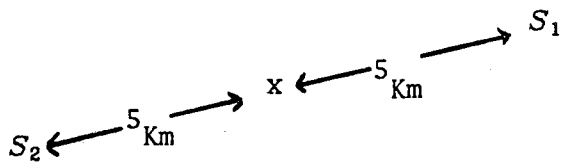
$$e(\mathbf{x}) = \sum_{i=1}^{i=N} \lambda_i z(\mathbf{x}_i)$$

where $z(\mathbf{x}_i)$ is the logarithm of the observed concentration at $\mathbf{x}_i$.

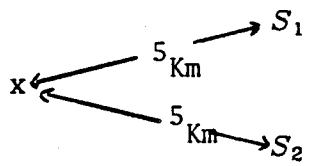
Strictly speaking this surface does not interpolate, i.e. pass through all the measured points, since λ_i does not tend to 1 as $d_i \to 0$. For any specific station i with a non-zero percentage of time active, p_i, $\frac{d_i}{\omega_i} \to 0$ as $d_i \to 0$, and therefore the function e is differentiable at $\mathbf{x}_i$. Ripley gives a more detailed discussion on the differentiability of these kinds of estimates and a complete review of this general estimation procedure.

A drawback to this estimation procedure is that the moving average is easily biased to clustered data points. If the point we are trying to estimate is equidistant from 10 clustered points on one side and 1 point on the other side, the estimate will be dominated by the 10 clustered points. Also let us examine the following situation,

a.



b.



where x is the point to be estimated and and S_i represents the station values. Assume that all stations are equidistant from the point of estimation and when the percentage of time active are all equal the weights for the estimate of x are the same in both situations. If we suppose that these two situations have the same atmospheric and pollution conditions, the example in a. where the estimated point has stations on both sides may be more accurately estimated because in a. the relevant point is surrounded by points while in b. there are only measurements on one side of the point of interest and therefore there is no information about possible pollution coming from the other side.

The parameter d_0 is a smoothing parameter which defines the shape of the weighting function. The shape should make the weights so that they are large for measurements within a radius of the estimated point where the variability of concentrations within that radius is assumed to be small. The value of d_0, should be large enough to smooth out insignificant variability among

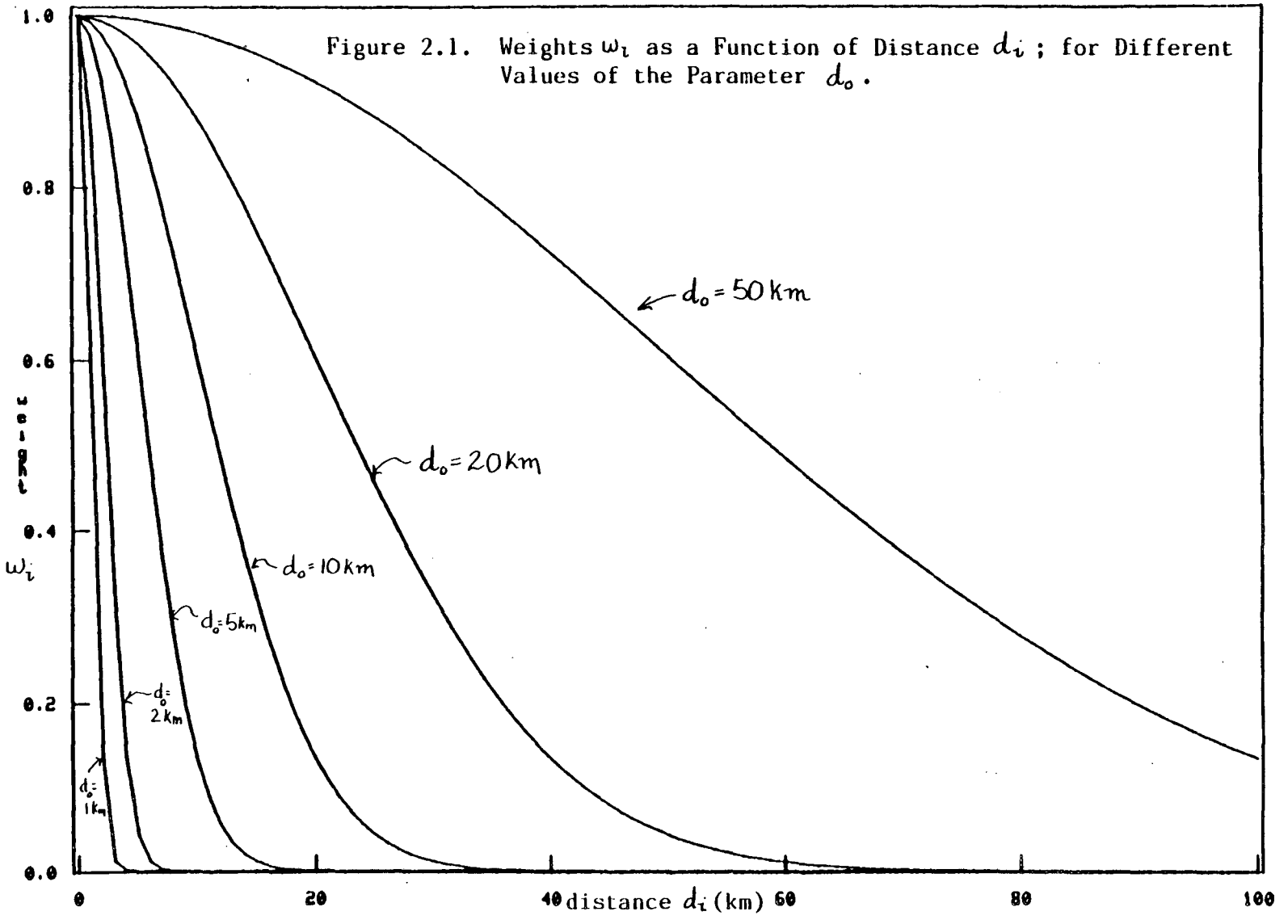


Figure 2.1. Weights w_i as a Function of Distance d_i ; for Different Values of the Parameter d_o .

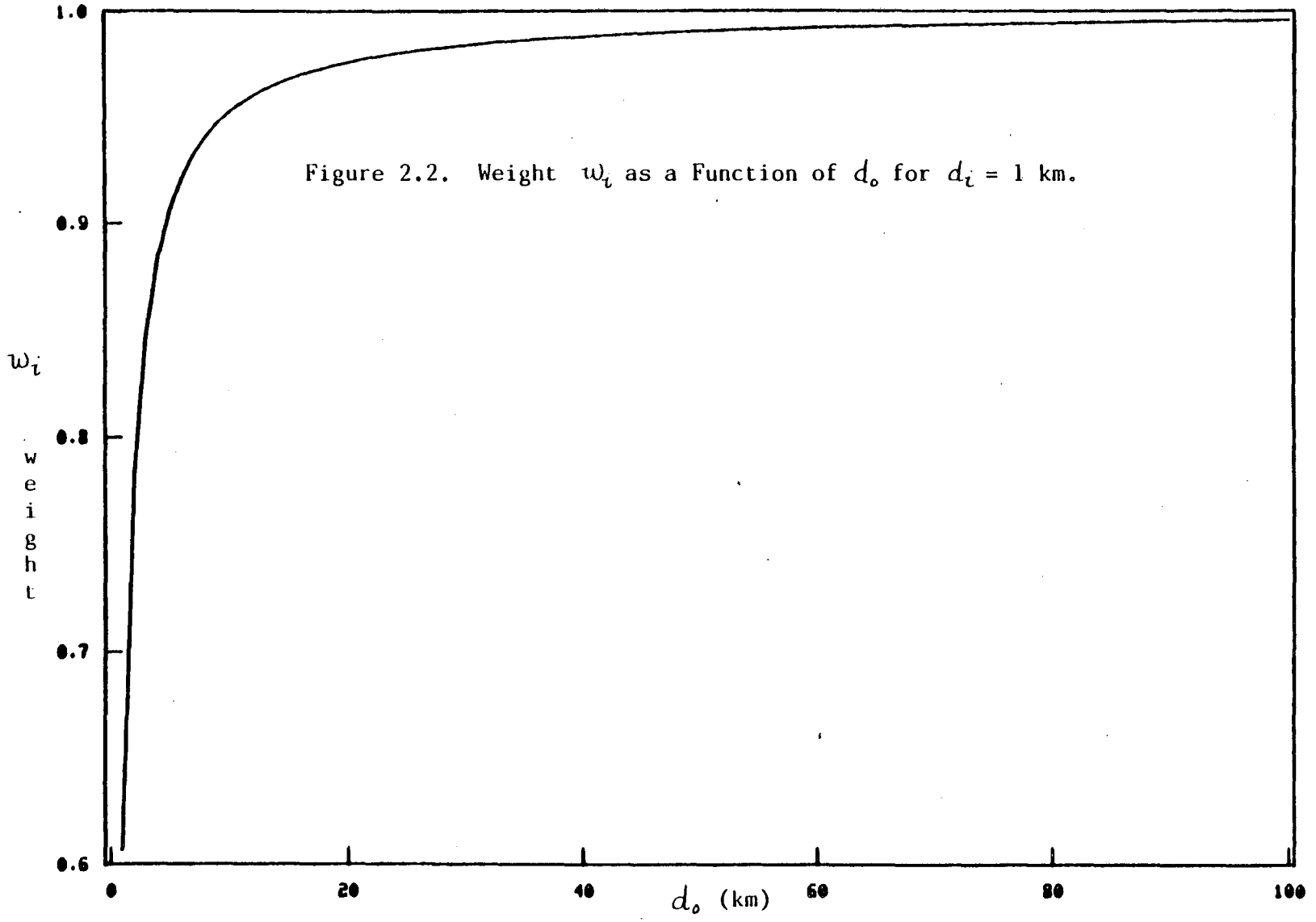


Figure 2.2. Weight w_i as a Function of d_o for d_i = 1 km.

adjacent stations, yet small enough to preserve significant trends over larger geographic areas. In order to avoid producing estimates for poorly monitored areas, we have excluded points more than $3d_0$ km away if they are the nearest points to our point of interest. If there are stations within 0 to $3d_0$ km of our point of interest, then measurements out to $4d_0$ km are used in the in the moving average model. The $4d_0$ criterion is used in order to avoid sharp discontinuities in the predicted surface. If we included only stations within $3d_0$ in the calculation, there would be a sharp discontinuity at the midpoint between 2 predicted points which are $6d_0$ km apart. Because of the shape of the weights and the number of points in the estimate the larger the d_0 the flatter the surface will be. See Figure 2.1 for curves of weight functions for constant p_i. Each curve corresponds to a different d_0 which covers 1, 2, 5, 10, 20, and 50 km. The flattest curve occurs for d_0=50 km and thus gives the smoothest contour map. Likewise the steepest curve is for d_0=1 km and this gives the least smooth map. See Figure 2.2 to see the weight as a function of d_0 for a d_i of 1 km. In the extreme case, as d_0 tends to ∞ the weights do not depend on d_i; all points in the area of interest, D, will be used and the predicted surface would be entirely flat. The absolute weights would all be 1 so our estimate would be simply the unweighted average of all points in D.

The two-dimensional moving average model has been implemented in SEEDIS so that given a particular set of points specified by latitude and longitude coordinates, estimates of air quality for these coordinates can be made. Values of d_0 the user can select are d_0 = 1, 2, 5, 10, 20, and 50 km. Deciding which d_0 to choose is not straightforward and an attempt to answer this question is discussed in detail in this chapter. A method for choosing d_0 will be given which is data dependent and could therefore be incorporated into SEEDIS and

provide automatic and rational choice of this critical parameter.

Just as air quality varies from point to point, it also varies over time. Air quality changes in response to changes in industrial activity, vehicle use, and weather conditions. The inclusion of a time parameter in the model, say t_0, in the same way as the distance parameter, d_0 is therefore a useful extension. Given a series of data values over time, one could could use data from year t to predict data in year $t+\Delta t$. If Δt is large then the weight should be small for that value. A possible choice for the weights w_i is

$$w_i = p_i e^{-.5(d_i^2/d_0^2 + t_i^2/t_0^2)}.$$

These are analogous to the weights given earlier where t_i = number of years difference between the year the measurement was taken and the year for which one wants to estimate. Time, t_i, is analogous to a third dimension of the two dimensional distance. This could be done if a time series of data were available; the time parameter, t_0, could be analyzed in the same way that we are exploring the distance parameter d_0.

Combining Data from 1-Hour and 24-Hour Sampling Intervals for Sulfur Dioxide

In order to illustrate the two dimensional moving average model that is discussed above, sulfur dioxide and suspended particulate in Ohio State are used for reasons of data density and availability. Using two pollutants also provides a comparison between two different kinds of air quality measurement which by their nature might require different estimation techniques. Sulfur dioxide generally is measured in two ways. Gas bubbler data from a 24-hour sampling inter

val are normally collected manually every sixth day; automatic air sampling devices usually provide continuous measurements for regular 1-hour sampling intervals. One way to estimate sulfur dioxide levels would be to proceed twice; first using data from 1-hour sampling intervals and second using measurements from 24-hour sampling intervals, yielding two independent estimates for the same variable, the concentration of sulfur dioxide. It is more useful to have a single estimate of sulfur dioxide for the area. If we just use 1-hour data for our estimate we have left out all the information obtained from the 24-hour data, and vice versa. If 1-hour data and 24-hour data are measuring the same aspect of sulfur dioxide, then the estimate which contains the most information about sulfur dioxide is clearly one that combines 1-hour and 24-hour data in a reasonable way.

Before applying the two dimensional moving average model, we combined the levels of sulfur dioxide at each station location. If a station measured sulfur dioxide only at 1-hour intervals then the combined measure was the average of the 1-hour measurements. Similarly, if a station had measured sulfur dioxide only with 24-hour sampling then the combined measure was the average of the 24-hour measurements. If a station had both 1-hour and 24-hour measures of sulfur dioxide, they were combined in the following manner.

Let

z_{ih} = the average of 1-hour measurements at monitoring station i.

z_{id} = the average 24-hour measurements at monitoring station i.

p_{ih} = percent time active of 1-hour monitor at station i.

p_{id} = percent time active of 24-hour monitor at station i.

(Note: If a station had neither a 1-hour nor a 24-hour measurement of sulfur dioxide, then it is regarded as missing and left out completely.)

Our combined measure is

$$z_{ic} = \frac{p_{ih} z_{ih} + p_{id} z_{id}}{p_{ih} + p_{id}}$$

Denote

$$p_{ic} = p_{ih} + p_{id}$$

The estimate of the combined observation at any point in a region where d_i is the distance from a station to the point of estimation is

$$e_c = \frac{\sum_{i=1}^{i=n} p_{ic}\, e^{(-.5 d_i^2 / d_0^2)} z_{ic}}{\sum_{i=1}^{i=n} p_{ic}\, e^{(-.5 d_i^2 / d_0^2)}}$$

where n = number of stations within the prescribed window.

Since

$$p_{ic} z_{ic} = p_{id} z_{id} + p_{ih} z_{ih}$$

then

$$e_c = \frac{\sum_{i=1}^{i=n} (p_{id} z_{id} + p_{ih} z_{ih}) e^{(-.5 d_i^2 / d_0^2)}}{\sum_{i=1}^{i=n} p_{ic}\, e^{(-.5 d_i^2 / d_0^2)}}$$

and thus

$$e_c = \frac{\sum_{i=1}^{i=n} p_{id}\, e^{(-.5d_i^2/d_\delta^2)} z_{id} + \sum_{i=1}^{i=n} p_{ih}\, e^{(-.5d_i^2/d_\delta^2)} z_{ih}}{\sum_{i=1}^{i=n} p_{id}\, e^{(-.5d_i^2/d_\delta^2)} + \sum_{i=1}^{i=n} p_{ih}\, e^{(-.5d_i^2/d_\delta^2)}}$$

which is equivalent to

$$e_c = \frac{num_{24} + num_1}{den_{24} + den_1}$$

where

num_j =numerator of estimate from data sampled at j-hour intervals.

and

den_j =denominator of estimate from data sampled at j-hour intervals.

Note that this is just a weighted average of the hourly and daily estimate where each is weighted by its respective monitoring density.

The above result makes it easy to combine estimates which can be obtained from the moving average model for 1-hour and 24-hour data separately. That is, one can either combine 1-hour and 24-hour data at each station before applying the model; or one can combine the model results obtained separately from 1-hour and 24-hour data. The results are the same. The monitoring density, expressed as the effective number of full time stations per unit area, quantifies the amount of data available around a particular point. It is defined as

$$monitoring\ density = \frac{\sum w_i}{2\pi d_0^2}$$

where the sum is over all active stations in the area of interest. The denominator, $2\pi d_0^2$, is a normalizing factor chosen so that the area integral of the density over the entire U.S., i.e.

$$\int_{US} monitoring\ density\ dA = \sum_{i=1}^{5777} p_i,$$

is the total number of effective full time stations. This was first discussed in Selvin, et. al. (1981). Since combining 24-hour and 1-hour data amounts to a weighted average and the model is also a weighted average, it is easy to generate estimates of a combined data point from the estimates of the uncombined data point as long as numerators and denominators are readily attainable.

Combining 1-hour and 24-hour data produces many more active stations for sulfur dioxide. The number for the combined data is only 2368 whereas for 24-hour alone the number is 1592 and for 1-hour alone the number of active stations is 202. Therefore sulfur dioxide is the second most well monitored in terms of the number of stations; if we combine nitrogen dioxide 1-hour and 24-hour data we have only 1218 stations (almost half as many). Nitrogen dioxide ranks third for number of active stations running in the United States.

Cross-Validation (CV) as an Integral Part of the Modeling

To choose the "optimum" value of d_0 and investigate the accuracy of these weighted averages, a predicted value for each station was produced from observations from other nearby stations (excluding the station's own value), and compared to the actual value observed at the selected station. By varying d_0,

several estimates were generated for comparison with each observed value. The method of choosing the "best" d_0 from all these estimates will be explained.

Two approaches were used to compare these estimates with their observed values in Johnson *et al.* (1982), employing correlation coefficients and squared error loss. Weighted and unweighted correlations were used in Johnson *et al.* (1982) and found to yield an inappropriate summary of the cross-validatory predictions because in small samples the correlation coefficient reflects a negative bias. In the extreme case of only two stations, the estimate made by leaving one out and predicting the other produces a correlation which is bound to be negative. Thus squared error loss was chosen as a function for measuring adequacy of various values of d_0. Squared error loss is defined as the sum of the squared residuals, where the residuals are defined as the estimted value minus the observed value. Yet squared error loss must be used very carefully for it can be heavily biased by outliers.

Using the method of cross-validation to choose d_0 is an expensive task. It requires $n(n-1)$ calculations, where n is the number of stations. Since n = 5777 stations for the whole United States, about 3 million calculations are required to cross-validate the entire United States. In order to facilitate cross-validatory estimates for any state, a file was created which has the cross-validatory estimates at each station location for d_0=1, 2, 5, 10, 20, and 50 km for total suspended particulate. Total suspended particulate data were used because these data are more abundant than those of any other pollutant. These estimates can be extracted without running the moving average model to get the cross-validation estimates.

So far we have only discussed using cross-validation as a method for choosing a parameter within a particular modeling technique. Cross-validation

can also be used for choosing among completely different methods. When we choose a different d_0, we are merely changing a control parameter within a given modeling technique. This is less difficult than choosing between two completely different models. Choosing a model by cross-validation is not quite as straightforward as choosing parameters and will be discussed in Chapter 3.

Utility of Surface Estimation to Detect Outliers

The model can be used not only for estimating pollutant concentrations for an area, but also as a method for detecting outliers. In fact it is the cross validatory feature which makes this possible. In a large data set such as this (n = 5777) it is extremely laborious to look for outliers in the usual way of visually perusing the data. The cross-validating estimates and graphic analysis allow one to analyze a large data set and quickly identify outliers.

With a relatively good method of estimation and a large sample, one would expect the estimates of an observed value to be correlated with the observed values. Graphically, we can compare the estimates with the corresponding observed values by means of a scatterplot. We would expect most of the points to fall on a 45 degree trend line. Gross outliers can be identified easily from one of these scatterplots. These techniques have often been used in regression analysis (Snedecor and Cochran, 1967).

Figure 2.3 shows, for individual stations, a scatterplot of the logarithm (base e) of the measured geometric mean of total suspended particulate concentration plotted against the estimate of the logarithm of the geometric mean. The smoothing parameter d_0, was chosen to be 50 kilometers. Fifty kilometers

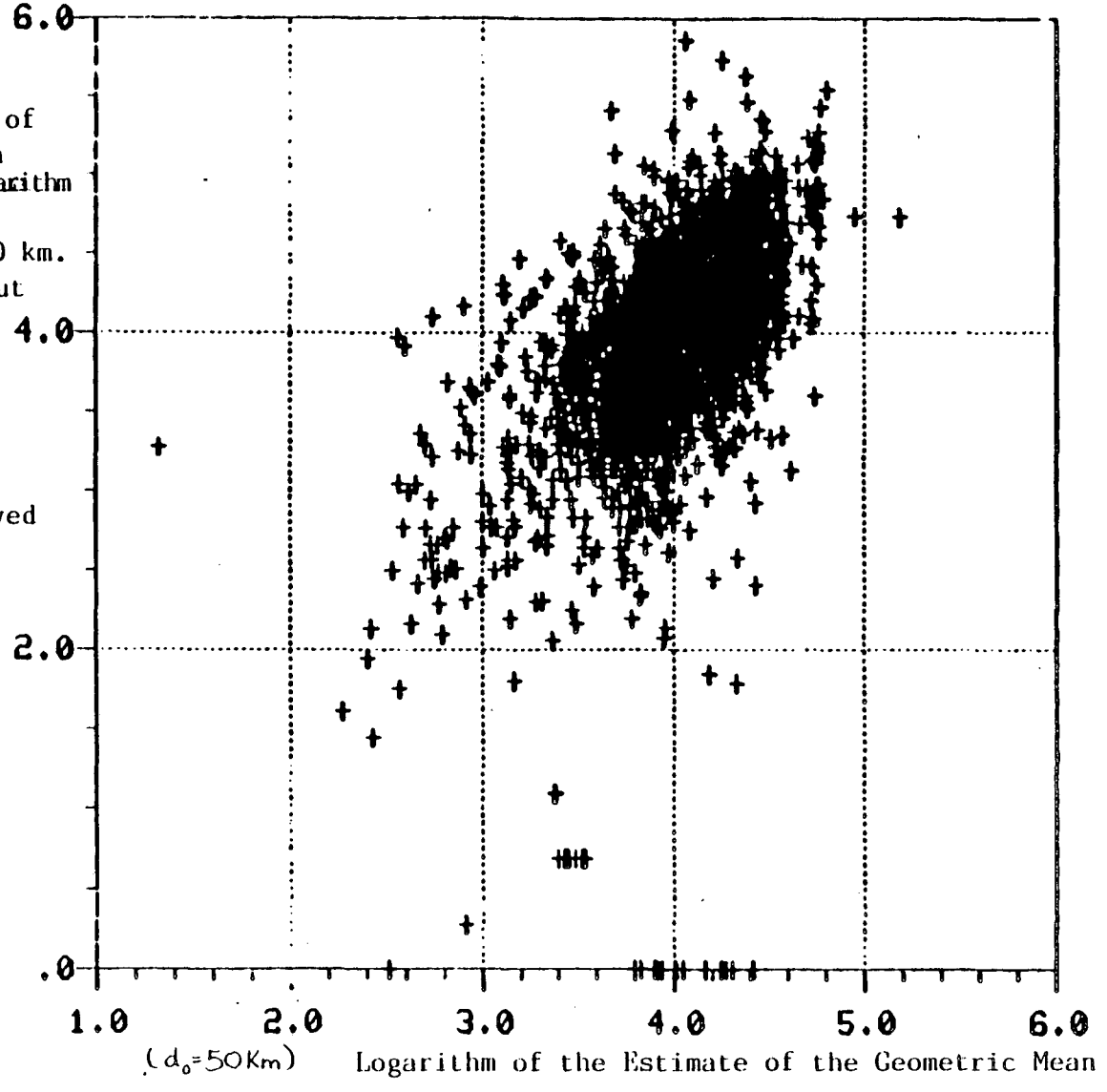


Figure 2.3. Logarithm of Observed Geometric Mean Plotted Against the Logarithm of the Estimate of the Geometric Mean, d_0 = 50 km. (Estimates have left out the observed value.)

was chosen for the sake of example and to be sure to estimate each observed value. Given each station's location, the moving average using the described weights was used to predict the value of pollution at each station location. Each station's estimate does not include its own measured value. Since the distribution of the geometric means are lognormal (a skewed distribution), logarithms of estimates and observed values were taken. Figure 2.3 shows some extraordinarily low observed values. The points that are particularly suspicious are those which have observed values (i.e. logarithm of geometric mean) exactly 0.00. In addition, other obvious outliers have observed values less than one. Also this plot provokes questions about the data; for example, in Figure 2.3. why is the estimate of one point so much smaller than the estimates of the others? Perhaps this point is near the points with the observed values less than one.

This plot suggests examining all the data where the observed value is less than or equal to 2.00 to see if there was some reason why these values are so low. It was found that all the stations with an observed value equal to 1.00 (log=0.00.) were active exactly 0.091 percent of the time, corresponding to a single 24-hour measurement during the 3-year period 1974-76. From now on, if a station had only one measurement its pollutant value is considered missing. Since most of these stations seem to be in the same geographic area it is reasonable to suggest that observers in this area designated missing values as 1.00. So these stations should be considered missing and not used in the estimation procedures. The value with the low estimate but a normal observed value is a station located next to those outliers whose observed values were designated 1.00. As soon as these spurious data points were removed from the sample, the remaining observed values had more reasonably sized estimates (See Figure 2.3a.). Figure 2.3a. shows the revised scatterplot obtained after the spurious

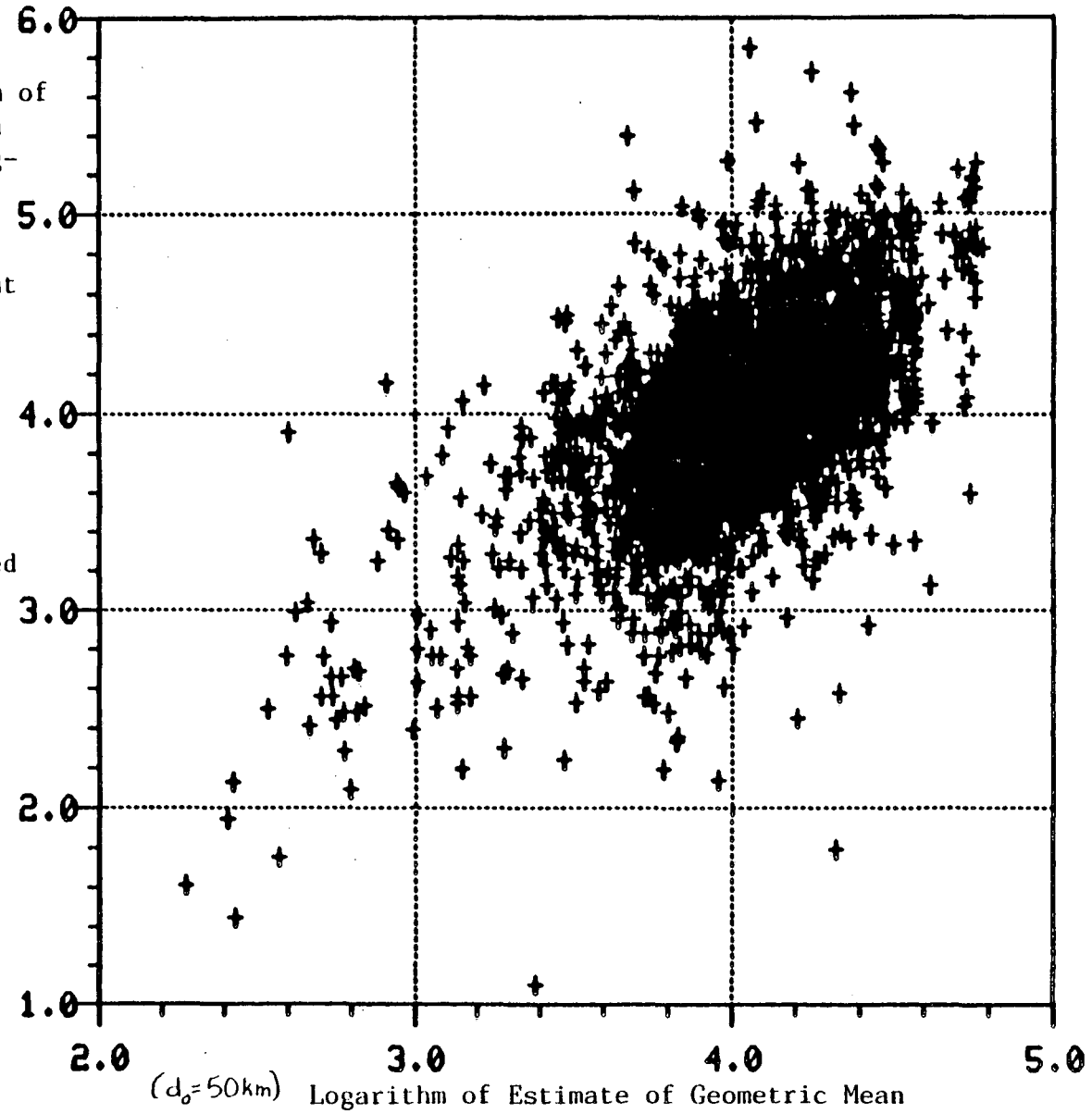


Figure 2.3a. Logarithm of Observed Geometric Mean Plotted Against the Logarithm of the Estimate of the Geometric Mean, d_o = 50 km. (Estimates have left out the observed value.)

values of Figure 2.3 were removed. Thus this model can be used to "launder" the data, a process which in large data sets is otherwise tedious.

Choosing d_0 for Ohio State with Sulfur Dioxide and Suspended Particulate Data

Sulfur dioxide and suspended particulate data for the state of Ohio were used to illustrate the selection of an optimal d_0. From 1974 through 1976 there were 185 active stations for sulfur dioxide and 398 active stations for total suspended particulate. For d_0= 1, 2, 5, 10, 20 and 50 km, the two dimensional moving average model was used to predict the observed value at the location of each of the active stations by leaving out the station's own value. The difference between the moving average model prediction and the observed value is the error in the estimate. We would like to choose the d_0 with the smallest prediction error.

Four composite measures of error were calculated. Let

z_i = logarithm of the pollutant concentration at location $\mathbf{x}_i$.

e_i = logarithm of the estimate at location $\mathbf{x}_i$ from surrounding data points.

Logarithms were taken because geometric means are assumed to have a lognormal distribution, i.e. the logarithm of the geometric mean is distributed normally. These measures of error are more easily interpreted for normally distributed data. Therefore the analysis will be done using logarithms of the observed values and the logarithms of the estimates throughout the thesis. When

values are discussed they are the logarithm of the geometric mean concentration unless otherwise mentioned.

The mean value is

$$\bar{z} = \sum_{i=1}^{i=n} z_i .$$

The variance of values is

$$S_z^2 = \frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})^2 .$$

The weighted variance of the values is

$$S_{zw}^2 = \frac{\sum_{i=1}^{n} p_i (z_i - \bar{z})^2}{\sum_{i=1}^{n} p_i}$$

In all these formulas,

n = the number of stations that could be estimated.

The error measures are:

(LF) Loss Function

$$LF = \frac{\sum_{i=1}^{n}(z_i - e_i)^2}{n}$$

(WLF) Weighted Loss Function

$$WLF = \frac{\sum_{i=1}^{n} p_i (z_i - e_i)^2}{\sum_{i=1}^{n} p_i}$$

(RSE) Ratio of Squared Errors

$$RSE = \frac{1}{nS_z^2} \sum_{i=1}^{n} (z_i - e_i)^2$$

(WRSE) Weighted Ratio of Squared Errors

$$WRSE = \frac{1}{\sum_{i=1}^{n} p_i \; S_{zw}^2} \sum_{i=1}^{n} p_i (z_i - e_i)^2$$

(MAE) Mean Absolute Error, Percent

$$MAE = \frac{1}{n\bar{z}} \sum_{i=1}^{n} \left| z_i - e_i \right| \times 100$$

(WMAE) Weighted Mean Absolute Error, Percent

$$WMAE = \frac{1}{\bar{z} \sum_{i=1}^{n} p_i} \sum_{i=1}^{n} p_i \left| z_i - e_i \right| \times 100$$

(MPE) Mean Percent Error

$$MPE = \sum_{i=1}^{n} \frac{p_i \left| z_i - e_i \right|}{n z_i} \times 100$$

(WMPE) Weighted Mean Percent Error

$$WMPE = \frac{\sum_{i=1}^{n} p_i \frac{\left| z_i - e_i \right|}{z_i}}{\sum_{i=1}^{n} p_i} \times 100$$

These formulas are similar to those used by Breiman (1977) for comparing kernel and Parzen multivariate density estimation techniques. The weights are added because if a station is not very active then the observed value at that station is not as reliable as the more active stations; therefore the reliability of our

estimate is to some extent a function of the station activity as well as the estimation procedure. For these error functions to reflect the adequacy of the estimation procedure and not the station activity they are weighted according to the percentage of time active.

These functions were computed for total suspended particulate data in Ohio. Table 2.1a. shows the values of each function for different values of d_0. Figure 2.4a. presents the same data in graphical form (Values are scaled so that functions appear on the same plot.). If one looks only at the unweighted loss functions, the value of d_0 which minimizes the loss for most of these functions is 5 kilometers. The ratio of squared errors (RSE) is the only one which is not consistent with this choice of an optimal d_0. This probably reflects the small local variability of suspended particulate relative to the global variance. This will be discussed in more detail later in this chapter.

TABLE 2.1a. Unweighted Loss Functions by d_0 (km) for Suspended Particulate in Ohio State.

d_0(km)	LF	RSE	MAE	MPE
1	.07562	*.500	4.9	4.9
2	.06907	.539	4.8	4.9
5	*.06586	.621	*4.8	*4.0
10	.06945	.696	4.9	4.9
20	.08111	.836	5.3	7.4
50	.09337	.963	5.6	5.7

* indicates the minimum value and thus corresponds to the optimum d_0.

In Table 2.1b., the weighted loss functions are computed for the same values of d_0. Figure 2.4b. is a plot of these functions (values are scaled so that functions appear on the same plot.). These functions give a different optimal d_0 than the unweighted measures. All of these weighted measures of error indicate

2 km as an optimal value of d_0 for the weights in the moving average. In the weighted case, we have more consistency amongst loss functions in the choice of d_0 than in the unweighted case.

TABLE 2.1b. Weighted Loss Functions by d_0 (km) for Suspended Particulate in Ohio State.

d_0(km)	WLF	WRSE	WMAE	WMPE
1	.0633	.425	4.5	4.4
2	*.0519	*.404	*4.2	*4.1
5	.0531	.477	4.3	4.2
10	.0565	.533	4.4	4.3
20	.0673	.631	4.7	4.6
50	.0836	.803	5.3	5.2

* indicates the minimum value and thus corresponds to the optimum d_0

Table 2.2a. and 2.2b. are the analogues to Tables 2.1a. and 2.1b. for sulfur dioxide (1-hour and 24-hour data combined). Notice that for all loss functions the sample mean does "better" in terms of squared error than does the moving average model. This could reflect a large local variability in sulfur dioxide relative to the global variance. It also suggests that there is some bias in the cross-validatory estimates of the model. The possibility of bias is explored in Chapter 3. We observe a larger local variablility for sulfur dioxide data than for suspended particulate data. This could arise from measuring problems such as a large measurement error or biased measurements. It could also mean that the moving average model is inherently wrong for modeling sulfur dioxide. Unlike suspended particulate, sulfur dioxide is a specific chemical compound which may diffuse or transform to other compounds in the atmosphere. Hence there might be a high local variability since SO_2 concentration may always be changing, whereas suspended particulate includes many compounds of a certain

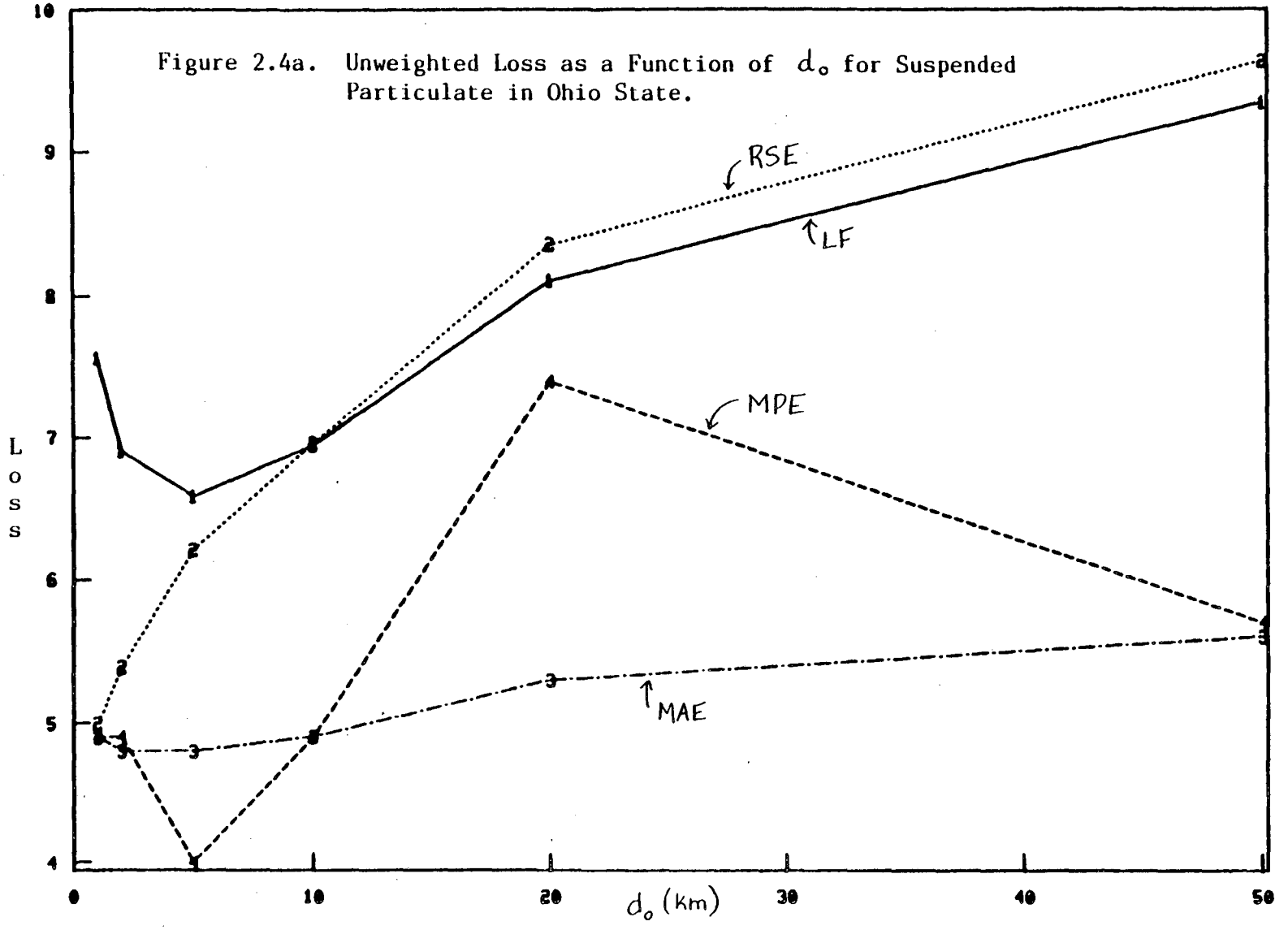


Figure 2.4a. Unweighted Loss as a Function of d_o for Suspended Particulate in Ohio State.

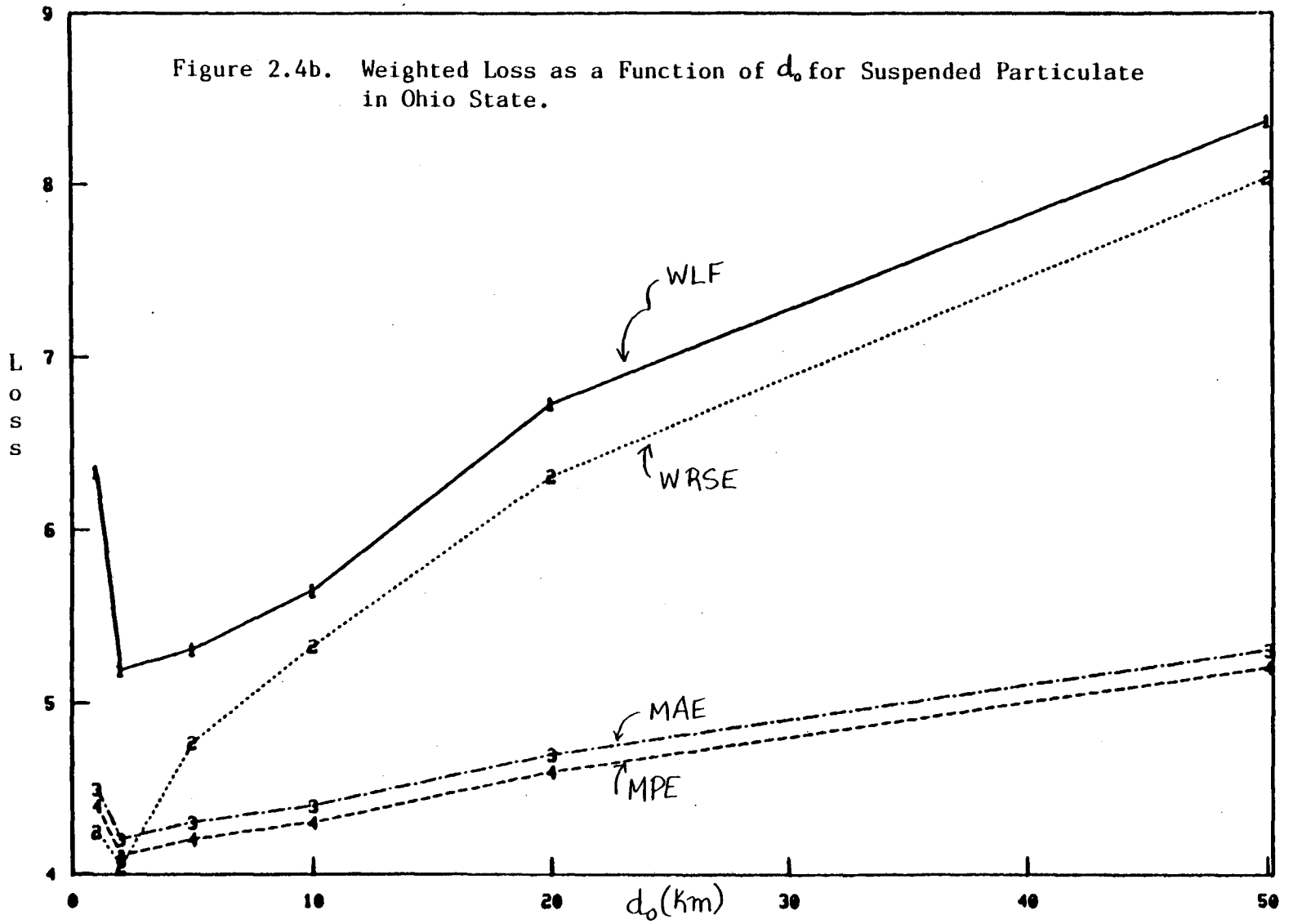


Figure 2.4b. Weighted Loss as a Function of d_o for Suspended Particulate in Ohio State.

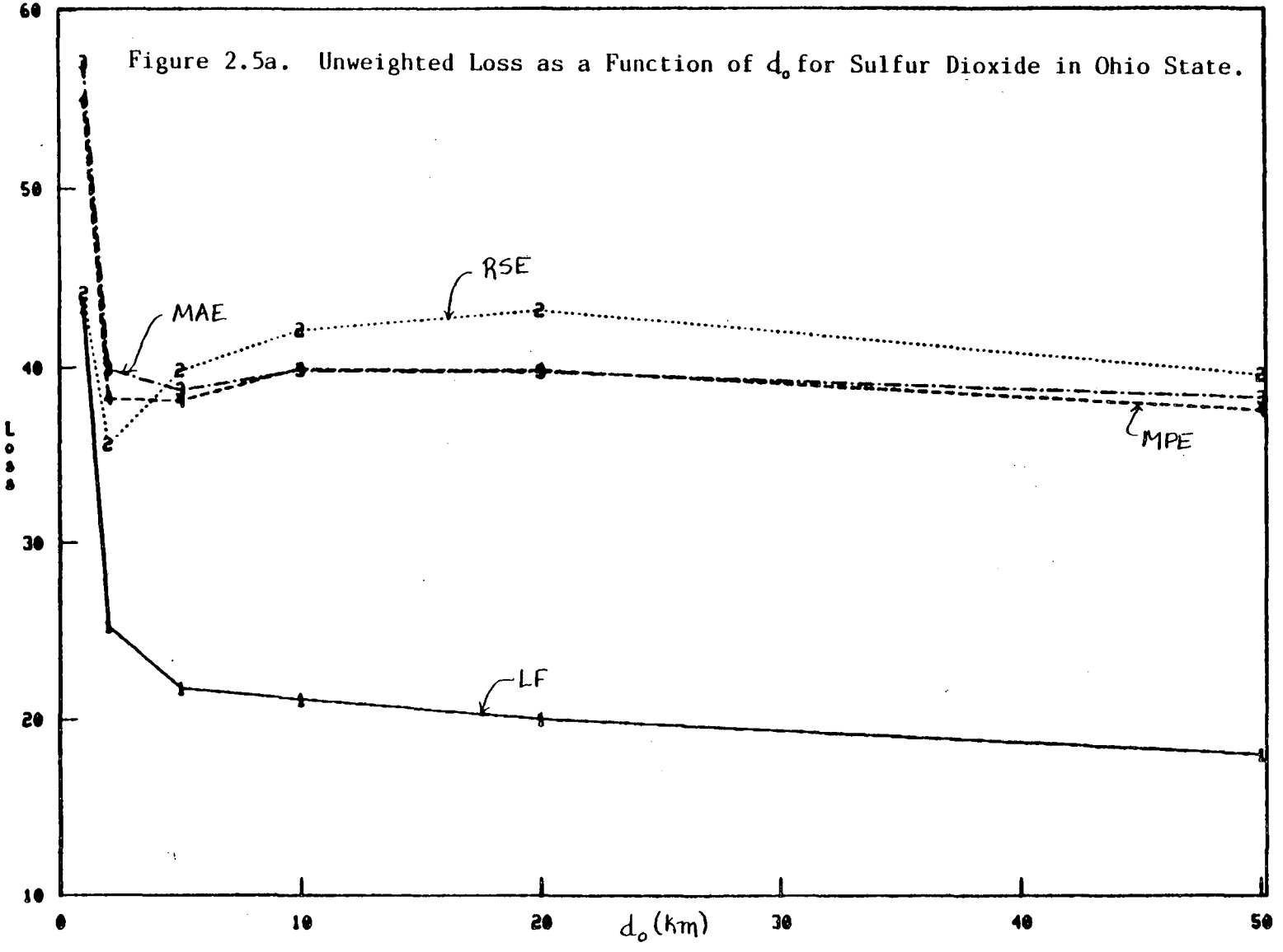


Figure 2.5a. Unweighted Loss as a Function of d_o for Sulfur Dioxide in Ohio State.

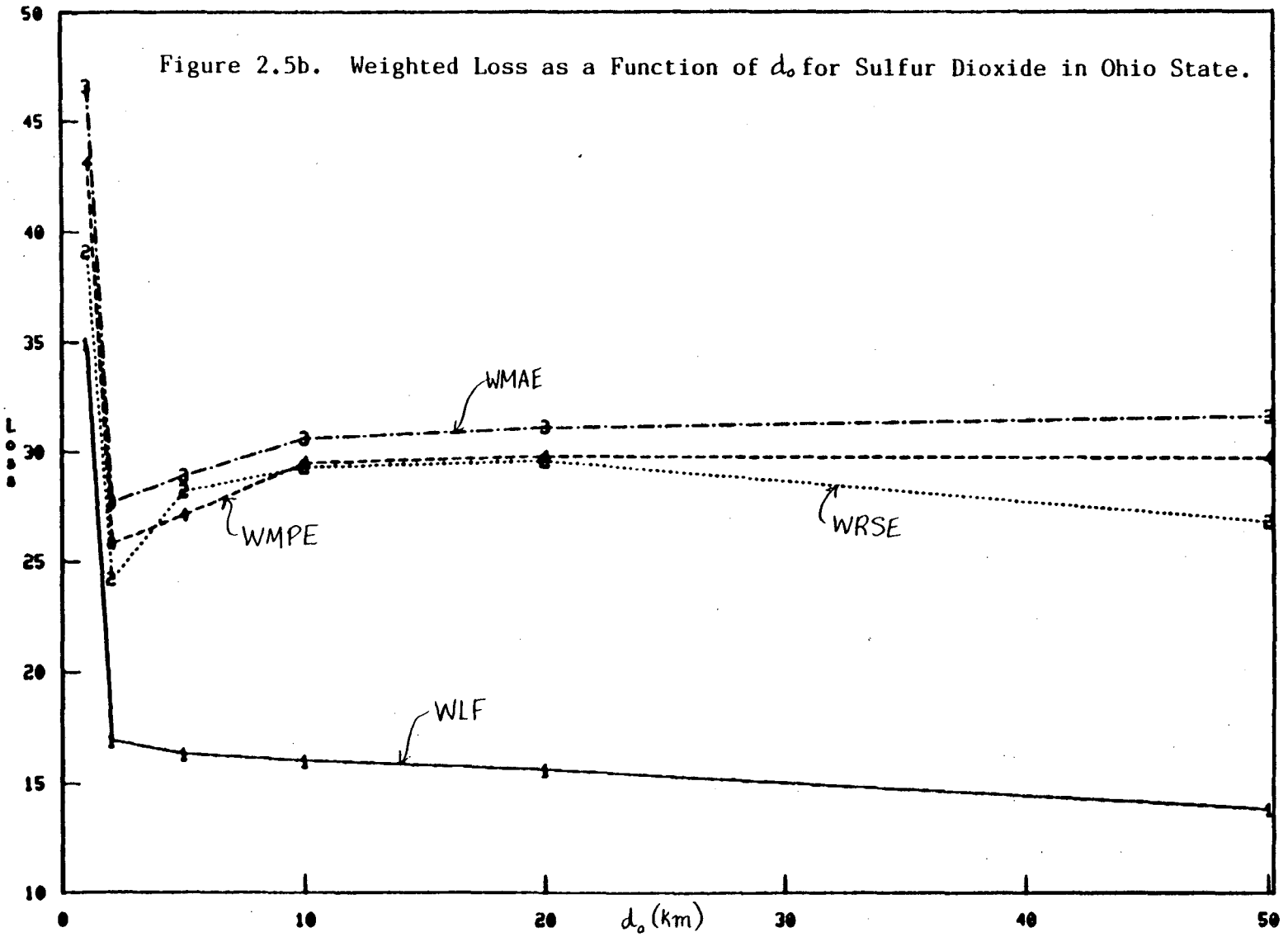


Figure 2.5b. Weighted Loss as a Function of d_o for Sulfur Dioxide in Ohio State.

size (eg. dust) which may not be as apt to change in concentration. Figure 2.5a. and 2.5b. present in graphic form the data of Tables 2.2a. and 2.2b. graphically (Values are scaled so that all 4 functions be presented on the same plot.).

TABLE 2.2a. Unweighted Loss Functions by d_0 for Sulfur Dioxide in Ohio State.

d_0 (km)	LF	RSE	MAE	MPE
1	4.34	4.42	57.0	54.9
2	2.52	*3.56	39.9	38.2
5	2.17	3.98	38.7	38.1
10	2.11	4.21	39.8	39.9
20	2.00	4.32	39.7	39.8
50	*1.80	3.95	*38.2	*37.5

* indicates the minimum value and thus corresponds to the optimum d_0

TABLE 2.2b. Weighted Loss Functions by d_0 for Sulfur Dioxide in Ohio State.

d_0(km)	WLF	WRSE	WMAE	WMPE
1	3.49	3.91	46.6	43.1
2	1.69	*2.42	*27.7	*25.8
5	1.63	2.82	28.9	27.1
10	1.60	2.93	30.6	29.5
20	1.56	2.96	31.1	29.8
50	*1.38	2.68	31.6	29.7

* indicates the minimum value and thus corresponds to the optimum d_0

Measuring Local Variability

Perhaps a more efficient method for finding the optimal d_0 than leaving one out and estimating the moving average for each data point is to find a measure of local variability (LV) as a function of d_0 and find that d_0 which minimizes it. That is, in the moving average model the d_0 chosen by cross-validation should be approximately that d_0, corresponding to a radius of disks drawn around each

point, for which the point values inside the disks are the most homogeneous. In other words, any other choice of d_0 should lead to a disk radius which, on the average, gives more heterogeneity among points within the disks. This section explores a method for estimating this local variability function.

The disk radius is determined by d_0 in the following manner. Recall that the moving average model is a weighted average of the values of the points surrounding the point of estimation. The weights are proportional to

$$w_i = p_i e^{-.5d_i^2/d_0^2}.$$

In the moving average model, points within four times d_0 are used in the average; yet the points that are a distance of four times d_0 give an absolute weight of

$$w_i = e^{-8} = .0003p_i$$

which can be a small relative weight. Since the weights in the model are relative weights, a point which is far away, say 2.9 d_0, can have a weight=1.0 if it is the only point within $3 \times d_0$. Therefore choosing a significant weight as a function of d_0 is an arbitrary choice and may depend on the particular data set and its precision.

Local variability is measured in the following manner. A disk is drawn around every point, defined by the station location, in the entire geographic area of interest. The radius of these disks is chosen as a function of d_0. (See Figure 2.6.) For a large enough radius there will be other points besides the center point contained inside these disks. In addition, each point will be the center of exactly one disk (i.e., for every point there is an associated disk for which it is the center). For each point we can calculate a mean and a variance using the points which lie in its associated disk for a specific radius r. For example, the

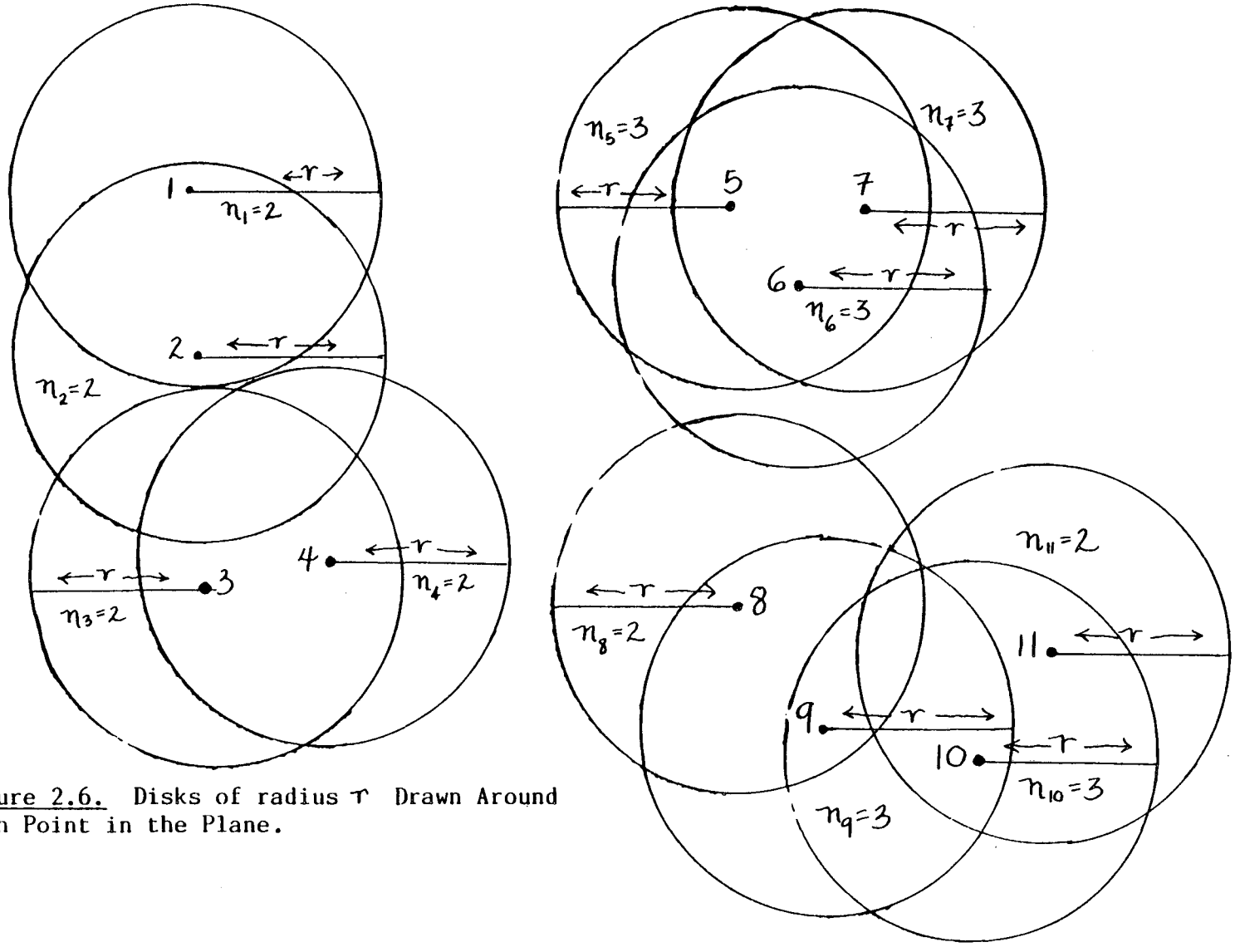


Figure 2.6. Disks of radius r Drawn Around Each Point in the Plane.

disk associated with x_i has the following mean and variance:

Disk Mean

$$\bar{x}_i = \frac{\sum_{j=1}^{k_i} x_j}{k_i}$$

Local Variance for point x_i:

$$\sum_{j=1}^{k_i} \frac{(x_j - \bar{x}_i)^2}{k_i - 1}$$

k_i = the number of points inside the associated disk.

We will call the variance inside each disk the local variance for that point using radius r. To calculate the local variability of the entire region for a particular r, we take a weighted average of the local variances.

Since a point may be in more than one disk we should weight the points accordingly. The precise method by which to weight the local variances for an overall measure of local variability is difficult to decide, so a method is explored which seems to be valid for intuitive reasons. We will examine the two most extreme cases first. Suppose the radii were chosen so small that no point is in more than one disk. We would then have local variances which were undefined since there are no "degrees of freedom". That is, it is impossible to calculate a variance from only one value per disk. If we choose the radius large enough, every point would be in every other point's disk and each disk would have the same mean and variance, namely the mean and variance of the entire region. In this case we would want the weighted average of the so-called local variances to

equal the variance of any one of them and thus there are N-1 degrees of freedom, where N = the number of stations in the region. If we weight each point by $\frac{1}{n_i}$ where n_i = the number of disks containing the point i. (See Figure 2.6.) then the average of the local variances in this case will equal the variance of any one of them, which is what we want.

$$\frac{\sum_{i=1}^{N} \frac{N(x_i-\bar{x})^2}{n_i}}{N-1} = \sum_{i=1}^{N} \frac{(x_i-\bar{x})^2}{N-1}.$$

Now let us examine the cases other than these two extreme examples. We could think of each disk having k_i-1 "degrees of freedom" where k_i is the number of points in each disk. If each disk were independent of every other disk (i.e. no overlapping disks) then this would be valid and the weights would be 1 and local variability could be measured by just the average of the local variance given by each disk. But since there is overlap, we have given each point a weight of $\frac{1}{n_i}$ where n_i= the number of disks containing point i. This way each point still contributes a weight of 1 to the average local variability of the region since each point appears in the sum of squares n_i times. The total weight each point contributes to the total sum of squares is $\sum_{j=1}^{n_i} \frac{1}{n_i} = 1$.

In order to normalize variability we need a denominator which expresses the number of degrees of freedom associated with the local variability. The number of degrees of freedom should take account of both the number of data points in each disk and the amount of overlap. Therefore we have chosen the denominator to be $N-\sum_{i=1}^{N} \frac{1}{n_i}$ which gives the expected answer in a few simple

cases. These cases are given below.

More formally one can express this measure in the following manner.

(LV) Local Variability

$$LV = \frac{\sum_{j=1}^{N}\sum_{i=1}^{k_j}\frac{(x_i-\bar{x}_j)^2}{n_i}}{N-\sum_{i=1}^{N}\frac{1}{n_i}}$$

where

N = the total number of points in the region of interest.

x_i = the value of pollution at a point.

$\bar{x}_j$ = the mean value corresponding to the disk associated with x_i.

Let us examine this formula in light of a few examples to see its plausibilty.

CASE 1. Every point is in only one disk. The disk radius is small so that no variances can be calculated. There are no degrees of freedom. In this case n_i = 1 and thus the denominator of LV is zero. Therefore, as should be expected LV is undefined in this case.

CASE 2. The disk radius is large enough that every point lies in every other point's disk. Then $n_i = N$ and $k_j = N$ and

$$LV = \frac{\sum_{j=1}^{N}\sum_{i=1}^{N}\frac{(x_i-\bar{x}_j)^2}{N}}{N-\sum_{j=1}^{N}\frac{1}{N}} = \frac{\sum_{j=1}^{N}(x_i-\bar{x}_j)^2}{N-1}$$

which is simply the sample variance of the points for the entire region. This is

the variance we would expect in this case since the local variance becomes the global variance with a disk radius this large.

CASE 3. Suppose there are no overlapping disks but there are m discrete clusters of points. In this case each cluster has its own mean and sum of squares associated with it. Each point has the same disk and thus the same mean and sum of squares associated with it as any other point in its cluster. In this case $n_i = k_j$, where k_j is the number of points in the jth cluster, and

$$LV = \frac{\sum_{j=1}^{N}\sum_{i=1}^{k_j}\frac{(x_i-\bar{x}_j)^2}{k_j}}{N-\sum_{i=1}^{N}\frac{1}{k_j}} = \frac{\sum_{j=1}^{m}\sum_{i=1}^{k_j}(x_i-\bar{x}_j)^2}{N-m}.$$

This is similar to the within mean square in one way analysis of variance (Scheffe', 1959).

CASE 4. Suppose each point has only one other point in its disk. So there are a series of two overlapping disks which do not contain any other points. In this case $n_i = 2$ for all i and

$$LV = \frac{\sum_{j=1}^{N}\sum_{i=1}^{2}\frac{1}{2}(x_i-\bar{x}_j)^2}{N-\frac{N}{2}} = \frac{\sum_{j=1}^{N/2}\sum_{i=1}^{2}(x_i-\bar{x}_j)^2}{N-\frac{N}{2}}.$$

This is similar to case 3 with two points in each cluster. This generalises to the case where every point is in k disks which do not overlap, then

$$LV = \frac{\sum_{j=1}^{N/k}\sum_{i=1}^{k}(x_i-\bar{x}_j)^2}{N-\frac{N}{k}}.$$

In other words this is the case, similar to case 3., where there are m clusters with with an equal number of points, k, in each cluster.

Local variability is directly related to the control parameter in the model described earlier. The number of nearby points which are used in the estimate of a particular point is related to the d_0 in this model. If d_0 is large, this implies that the local variability is high and we can give relatively large weight to points far away from the point of interest in its estimation. Likewise, if d_0 is chosen to be small, points far away will be given relatively small weight since the minimum local variability occurs at a smaller disk radius. Points further than $4 \times d_0$ were given zero weight by the two-dimensional moving average. This restriction was arbitrarily chosen because for a particular weighting scheme (i.e. d_0), values beyond $4 \times d_0$ kilometers are assumed to have no influence.

The disk radius where the minimum local variability occurs is similar to d_0 and the $4 \times d_0$ restriction in the two dimensional moving average model. In fact, points within this radius are on the average more similar in concentration than points outside this radius. Therefore we should weight points within this distance relatively high and points beyond the disk radius should be essentially disregarded

Using Ohio as an example, local variability is calculated for various disk radii. The radii were chosen as a function of d_0. For each d_0, two disk sizes were chosen. One was $4 \times d_0$ which is just the same size that the moving average model uses in selecting stations to compute its estimate of a point.[1] Since the weights in the moving average model are distance-dependent, we expect it to do well for disk radii where the local variability is small. Ohio is chosen so that we can com-

[1] Since stations further than about $3d_0$ have negligible weights in the two dimensional moving average model, and since these estimates are suppressed for points further than $3d_0$ from any active station, one might argue that a disk radius of $3d_0$ would have been more appropriate than $4d_0$. Nevertheless, the choice of $3d_0$ or $4d_0$ does not change the conclusions of this section.

pare our choice of disk radius with our cross-validatory choice of d_0 in the previous section. In Figures 2.7a. and 2.7b. the local variability is shown as a function of disk radius for Suspended Particulate and Sulfur Dioxide respectively. In Tables 2.3a. and 2.3b. the actual values which were calculated for Ohio are given. For comparison with the cross-validatory choices of the previous section, we chose disks corresponding to d_0 = 1, 2, 5, 10, 20, and 50 km. This correspondence is arbitrary but is based on the moving average model as it has been used in the PAREP project. The other disk size was chosen so that the absolute weights in the average would never be smaller than .01. The relationship of the absolute weights to the relative weights which are used in the average is not always the same and therefore a disk size which corresponds directly to the weights in the two-dimensional moving average cannot be predicted.

TABLE 2.3a. Local Variability as a function of d_0 for Suspended Particulate in the State of Ohio.

d_0 (km)	disk radius (km)	Local Variability	"df"
1	3.03	.0583	141.80
1	4.00	.0573	173.28
2	4.29	*.0562	181.37
2	8.00	.0597	240.35
5	10.74	.0614	264.68
5	20.00	.0662	326.96
10	21.47	.0671	334.21
10	40.00	.0819	372.48
20	42.94	.0837	375.20
20	80.00	.0927	390.68
50	107.36	.0945	393.21
50	200.00	.0957	396.03
**	500.00	.0972	397.00

* The smallest local variability, corresponding to the optimal d_0.
** d_0 is that value which gives a disk radius large enough to contain all the points in Ohio. Notice that "df"=N-1 since there are 398 active stations for suspended particulate in the region.

The disk radius which gives disks that are more homogeneous than those given by other radii corresponds to a d_0 = 2 km. These results are the same as those given by cross-validation using weighted loss functions (Table 2.1b.) and similar to that obtained using the unweighted loss functions (Table 2.1a.). This result along with the shape of LV suggest that local variability may be another method for choosing the optimal d_0 which is much faster than choice by cross-validation.

TABLE 2.3b. Local Variability as a function of d_0 for Sulfur Dioxide in the State of Ohio.

d_0 (km)	disk radius (km)	Local Variability	"df"
1	3.03	.2129	50.37
1	4.00	.1934	66.70
2	4.29	.1984	71.00
2	8.00	.1930	101.46
5	10.74	*.1887	113.73
5	20.00	.1989	135.56
10	21.47	.2008	137.58
10	40.00	.2526	162.65
20	42.94	.2644	164.31
20	80.00	.3166	177.30
50	107.36	.3468	180.41
50	200.00	.4063	183.10
**	500.00	.5069	184.00

* The smallest local variability, corresponding to the optimal d_0.
** d_0 is that value which gives a disk radius large enough to contain all the points in Ohio. Notice that "df"=N-1 since there are 185 active for sulfur dioxide stations in the region.

For sulfur dioxide, the cross-validatory analysis using weighted loss functions suggests a d_0 = 2 km while the unweighted analysis leans toward a d_0 = 50 km. This implies that sulfur dioxide has a higher local variability than suspended particulate. Local variability analysis shows this is true. Table 2.3b. gives a minimum at a disk radius of 10.74km which corresponds to d_0 = 5 km,

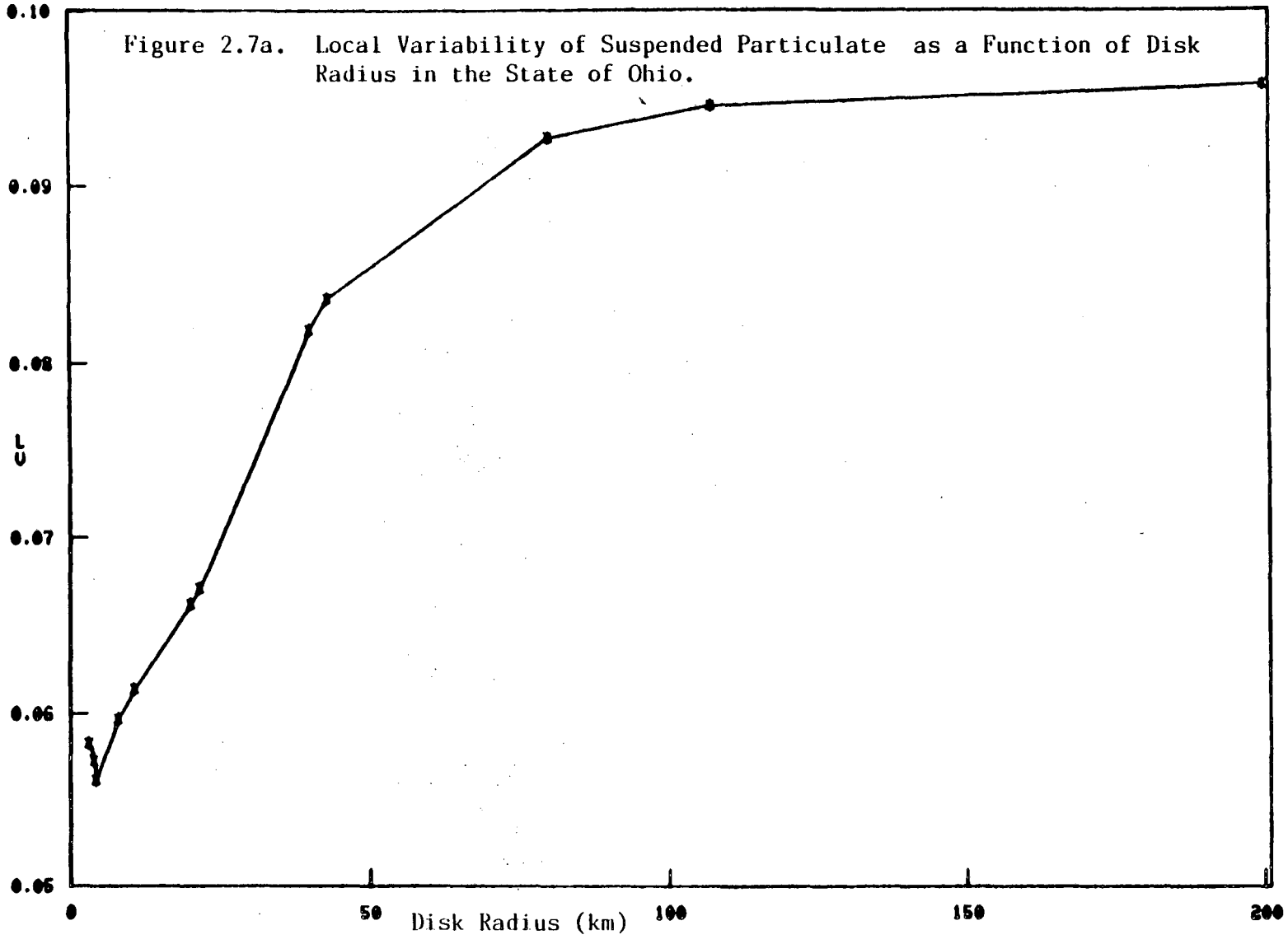


Figure 2.7a. Local Variability of Suspended Particulate as a Function of Disk Radius in the State of Ohio.

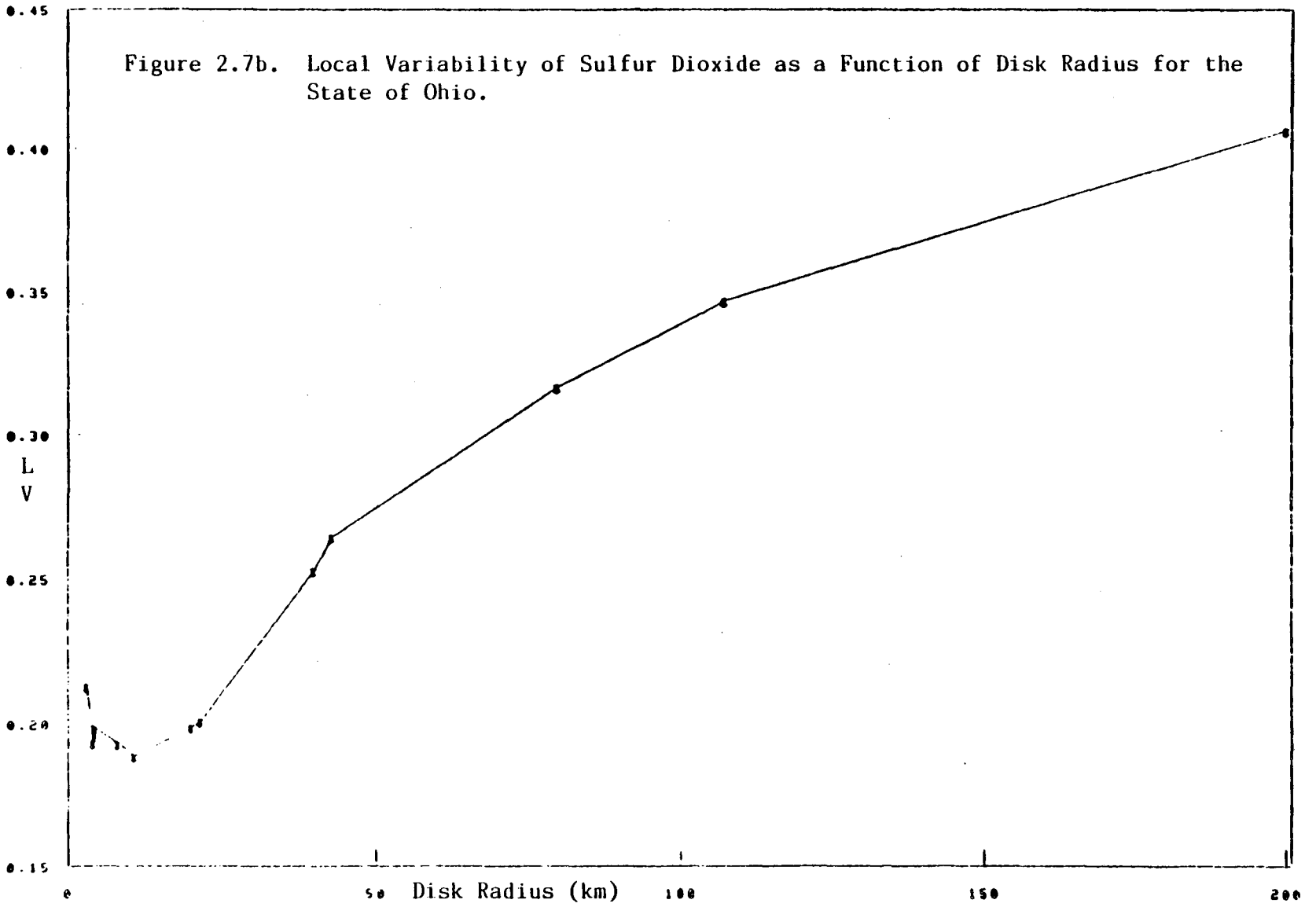


Figure 2.7b. Local Variability of Sulfur Dioxide as a Function of Disk Radius for the State of Ohio.

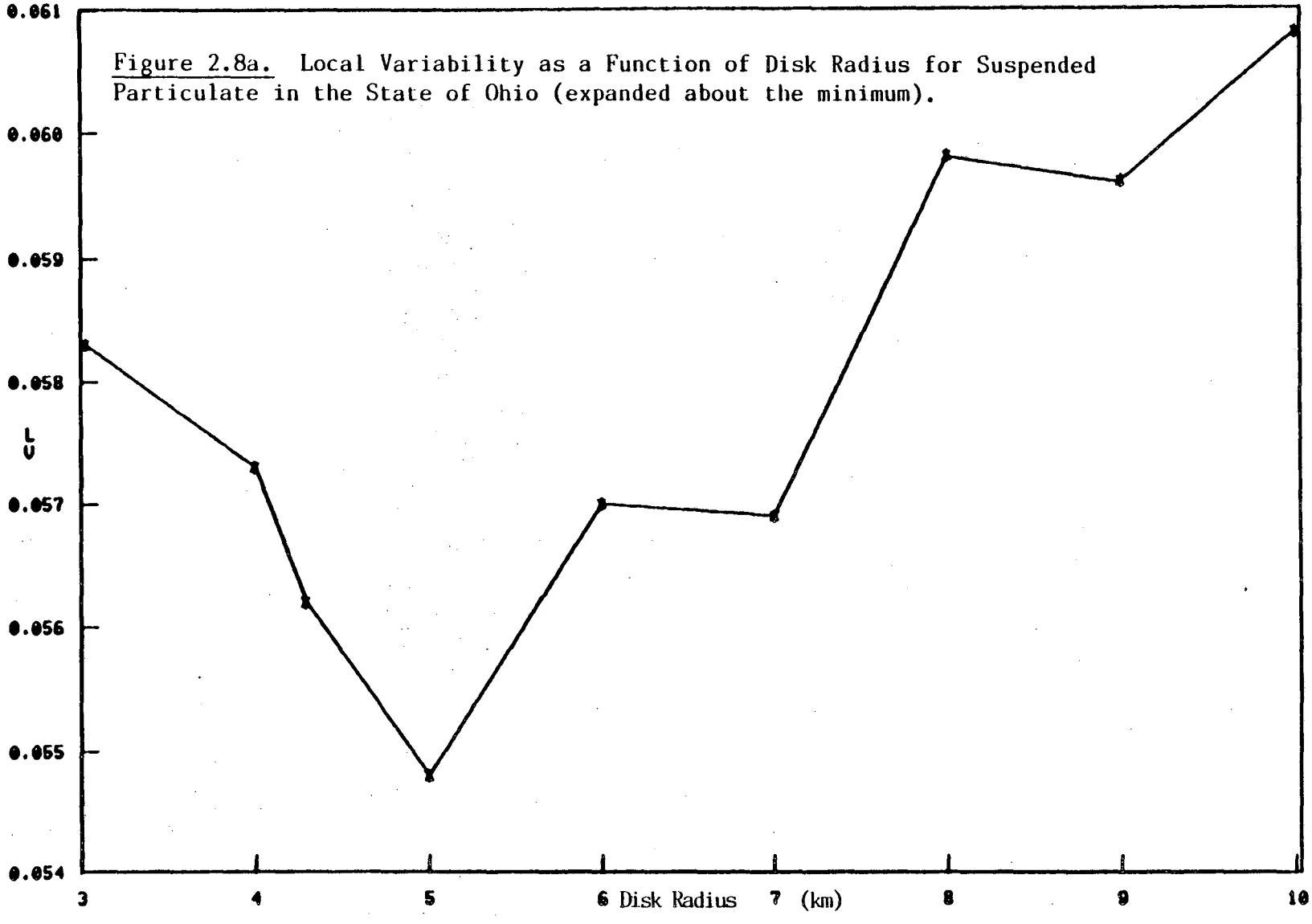


Figure 2.8a. Local Variability as a Function of Disk Radius for Suspended Particulate in the State of Ohio (expanded about the minimum).

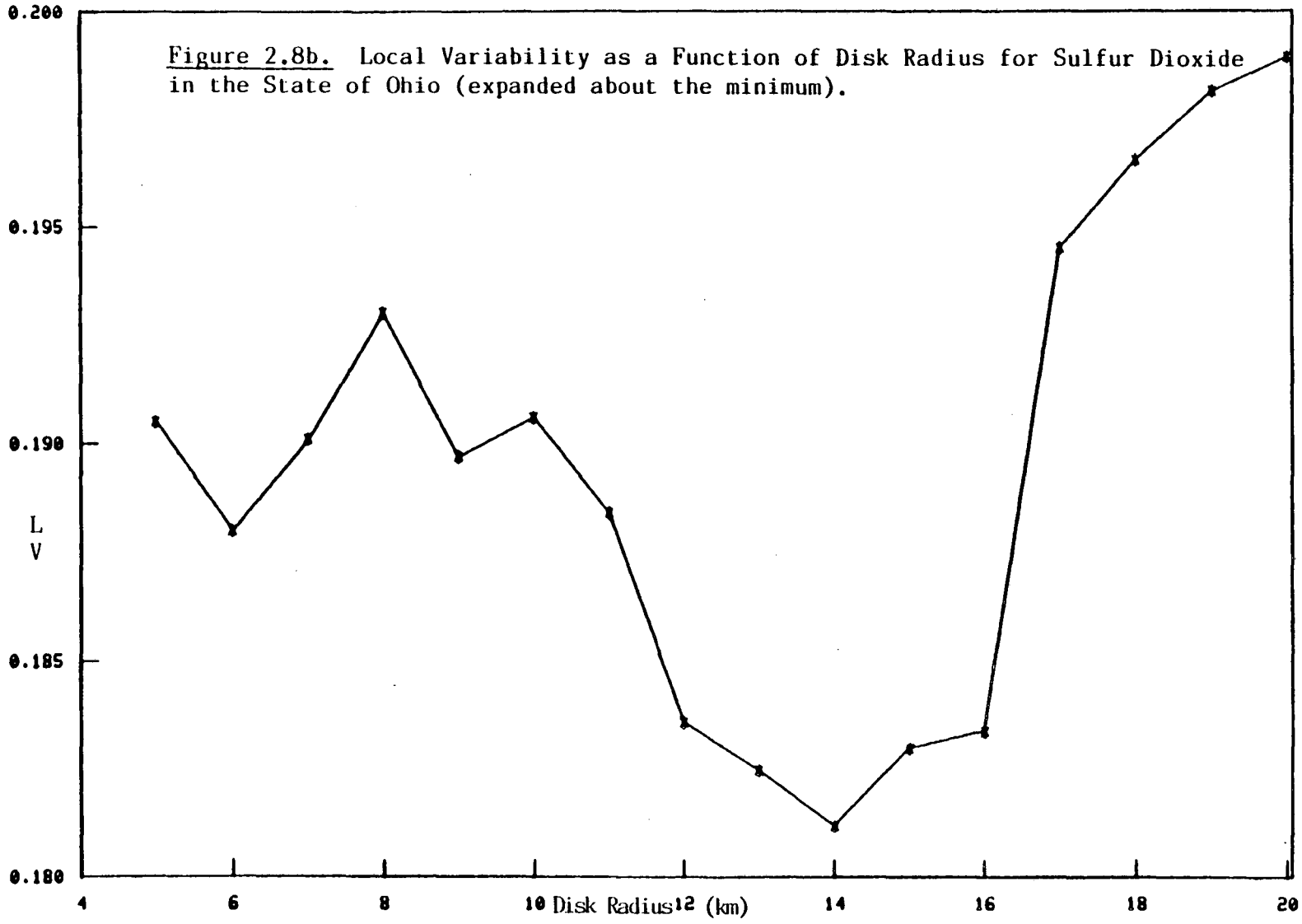


Figure 2.8b. Local Variability as a Function of Disk Radius for Sulfur Dioxide in the State of Ohio (expanded about the minimum).

which is between the two conflicting values in the cross-validatory analysis. There are at least two reasons why sulfur dioxide may have a higher local variability.

In order to determine the actual optimum disk radius, LV was calculated for disk radii increments of one kilometer. In Figure 2.8a. local variability is shown as a function of disk radius between 5 and 10 km for suspended particulate. In Table 2.3a., when local variability is at a minimum the disk radius is equal 4.29 km. Yet, Figure 2.8a. indicates that a radius of 5 km, corresponding to a d_0 of approximately 2 km, gives the minimum local variability for suspended particulate.

In Figure 2.8b. local variability is shown as a function of disk radius betweem 5 and 20 km for sulfur dioxide. This plot shows an expansion around the disk radius equal to 10.74 km giving the minimum shown in Table 2.3b. In the Figure, one can see that beyond 10.74 km, LV decreases further to a minimum of .1812 at disk radius of 14 km; it then increases and continues to increase steadily to the maximum variability (i.e. sample variance for all values in the region). It is also interesting to note the slight instability of the curve at smaller disk radii. This instability is expected since with small k_i in each disk we expect LV to be extremely sensitive to the addition of new points with increments in disk radius. Also, some instability can be expected from other sources of variation, as is seen in Figure 2.8a. Here LV is increasing steadily but also varies from a steady increase.

Consistency in the Shape of Local Variability

In order to further explore our results with Ohio, the same calculations for New York and Florida are shown here. New York and Florida were chosen because these states have the second and third largest number of monitoring stations respectively. For each state, the basic shape of the curve showing local variability as a function of disk radius is the same. Also, the curves showing local variability for sulfur dioxide seem to be similar across all three states. (Tables 2.4a., 2.4b., 2.5a. and 2.5b., along with Figures 2.9a. thru 2.12b.)

TABLE 2.4a. Local Variability as a function of d_0 for Suspended Particulates in New York State.

d_0 (km)	disk radius (km)	Local Variability	"df"
1	3.03	.0484	75.69
1	4.00	.0480	92.47
2	4.29	.0493	98.27
2	8.00	*.0481	132.76
5	10.74	.0516	154.79
5	20.00	.0608	202.26
10	21.47	.0631	208.31
10	40.00	.0808	245.42
20	42.94	.0837	248.50
20	80.00	.0974	265.88
50	107.36	.1033	269.99
50	200.00	.1107	273.30
**	700.00	.1168	275.00

* The smallest local variability, corresponding to the optimal d_0.
** d_0 is that value which gives a disk radius large enough to contain all the points in New York State.

TABLE 2.4b. Local Variability as a function of d_0 for Sulfur Dioxide in New York State.

d_0 (km)	disk radius (km)	Local Variability	"df"
1	3.03	.2023	40.18
1	4.00	.1851	48.51
2	4.29	.1845	51.06
2	8.00	.1615	71.95
5	10.74	*.1614	83.07
5	20.00	.2080	103.33
10	21.47	.2079	104.21
10	40.00	.2588	114.86
20	42.94	.2608	115.95
20	80.00	.3085	124.62
50	107.36	.3537	128.44
50	200.00	.3739	131.49
**	700.00	.3839	133.00

* The smallest local variability, corresponding to the optimal d_0.
** d_0 is that value which gives a disk radius large enough to contain all the points in New York State.

TABLE 2.5a. Local Variability as a function of d_0 for Suspended Particulates in Florida State.

d_0	disk radius (km)	Local Variability	"df"
1	3.03	.0580	48.73
1	4.00	*.0546	69.20
2	4.29	.0563	72.56
2	8.00	.0574	105.35
5	10.74	.0608	122.99
5	20.00	.0654	151.83
10	21.47	.0668	155.72
10	40.00	.0716	176.02
20	42.94	.0716	177.97
20	80.00	.0841	191.81
50	107.36	.0866	194.26
50	200.00	.0918	200.73
**	500.00	.0907	201.00

* The smallest local variability, corresponding to the optimal d_0.
** d_0 is that value which gives a disk radius large enough to contain all the points in Florida State.

TABLE 2.5b. Local Variability as a function of d_0 for Sulfur Dioxide in Florida State.

d_0 (km)	disk radius (km)	Local Variability	"df"
1	3.03	.2365	46.82
1	4.00	.2308	59.74
2	4.29	.2298	62.00
2	8.00	*.2176	81.64
5	10.74	.2325	90.68
5	20.00	.2426	105.83
10	21.47	.2423	107.50
10	40.00	.3138	118.06
20	42.94	.3151	118.75
20	80.00	.3222	127.46
50	107.36	.3452	131.39
50	200.00	.3595	134.17
**	900.00	.4017	184.00

* The smallest local variability, corresponding to the optimal d_0.
** d_0 is that value which gives a disk radius large enough to contain all the points in Florida State.

For all three states, the minimum of local variability for suspended particulate occurs at a disk radius smaller than that for sulfur dioxide. Also each state has a different size disk which gives minimum local variability. Since there are different dispersion and dilution mechanisms in different geographic areas we should expect this inconsistency in choice of disk radius across states for each pollutant. In fact if we look at Figures 2.8b., 2.9b., 2.10b., 2.11b. and 2.12b. where LV are expanded about the minimum we find that the minima seem to occur at the following disk radii:

State	Disk Radius (km) Suspended Particulate	Disk Radius (km) Sulfur Dioxide
Ohio	5	14
New York	1	10
Florida	5	8

To view local variability curves on one graph over the whole range of disk radius (from only one "degree of freedom", i.e. disk radius so small that

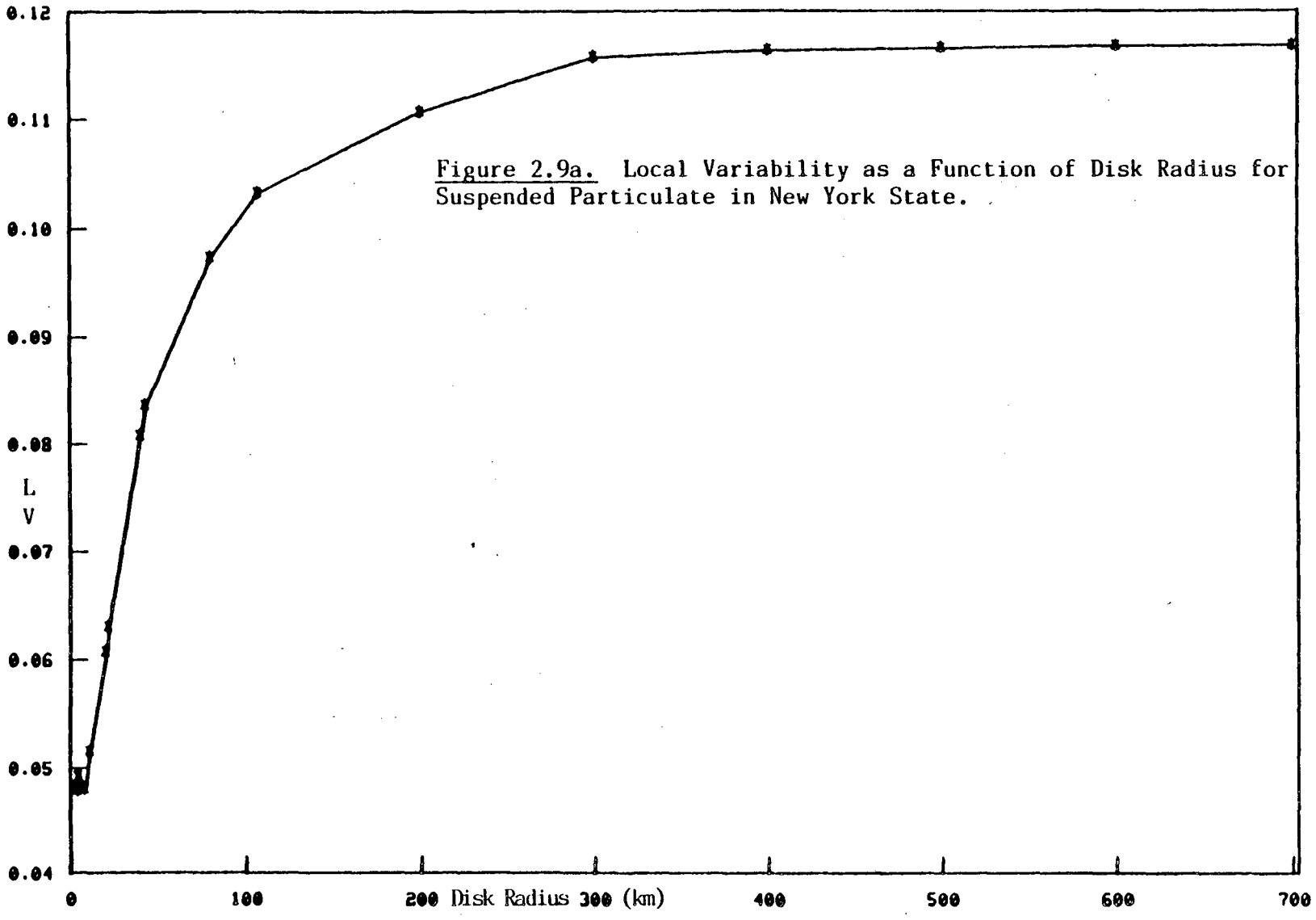


Figure 2.9a. Local Variability as a Function of Disk Radius for Suspended Particulate in New York State.

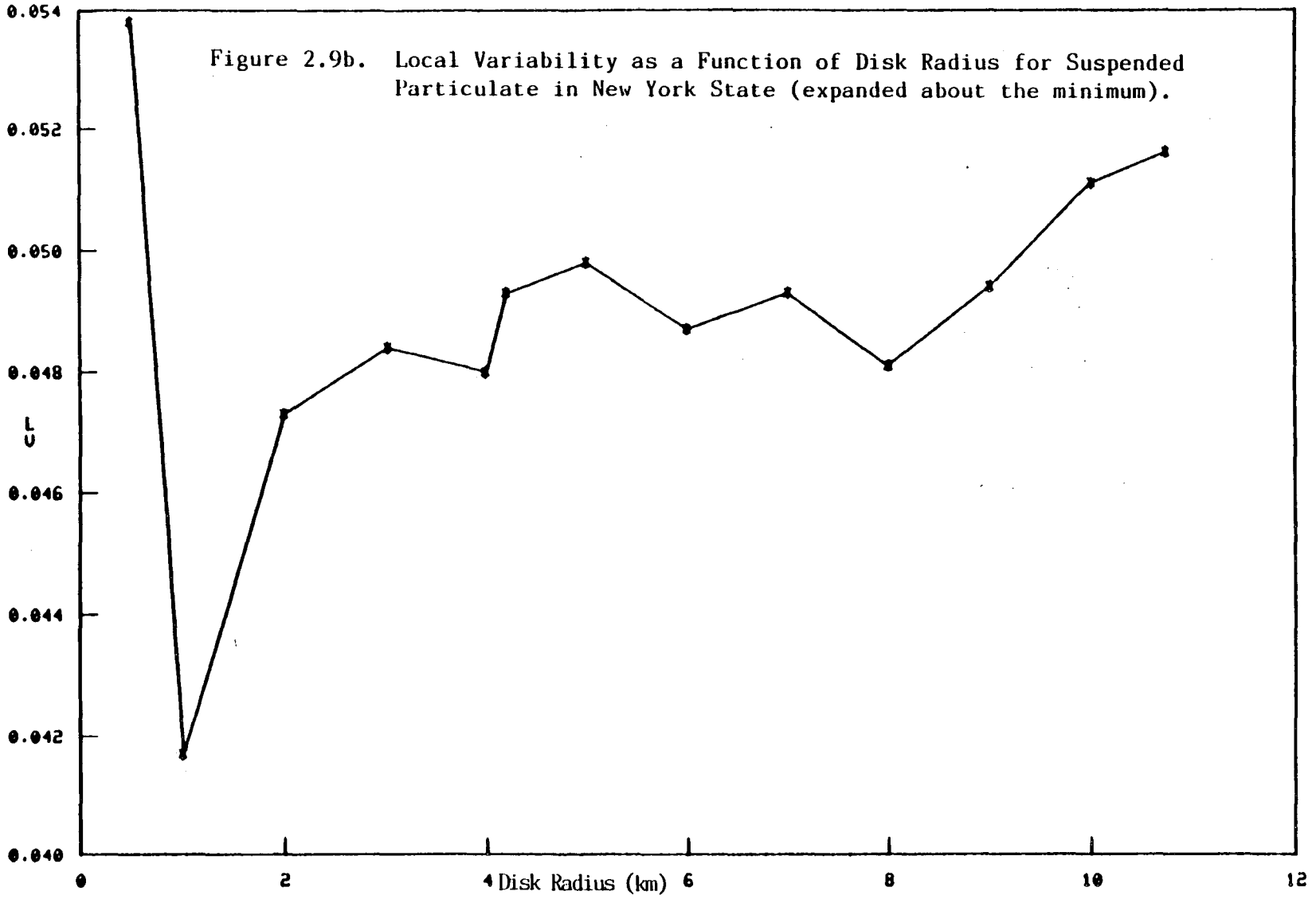


Figure 2.9b. Local Variability as a Function of Disk Radius for Suspended Particulate in New York State (expanded about the minimum).

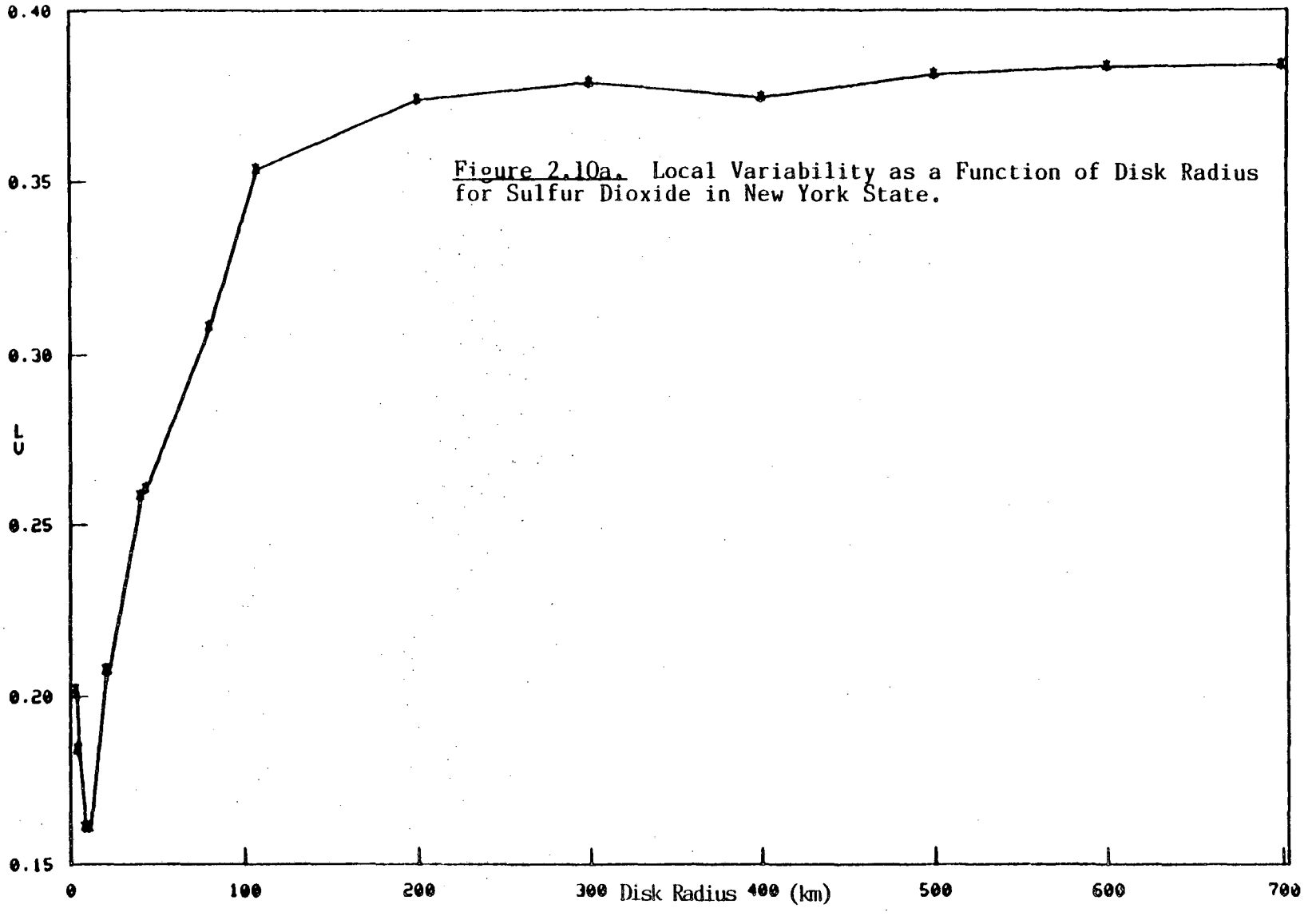


Figure 2.10a. Local Variability as a Function of Disk Radius for Sulfur Dioxide in New York State.

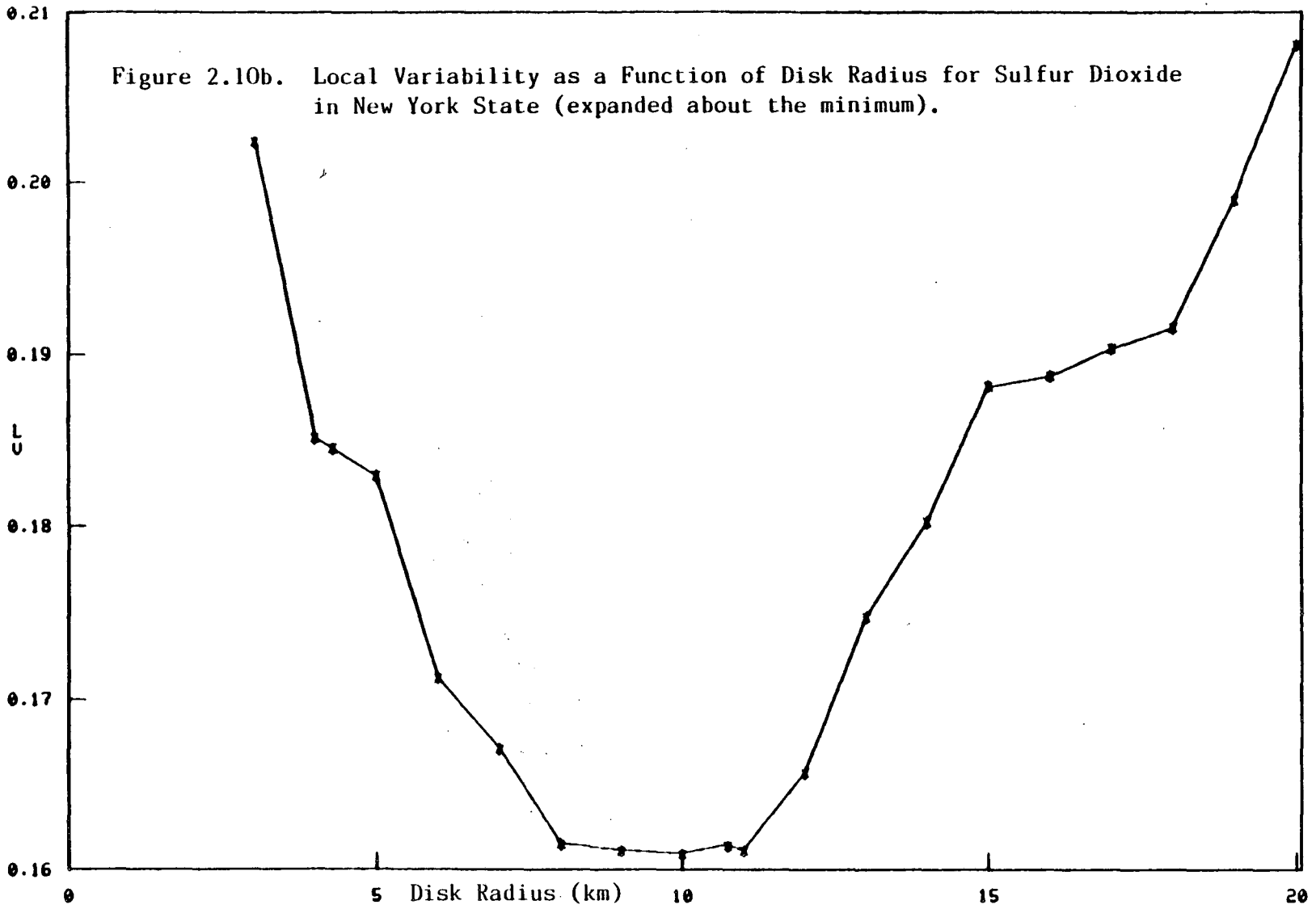


Figure 2.10b. Local Variability as a Function of Disk Radius for Sulfur Dioxide in New York State (expanded about the minimum).

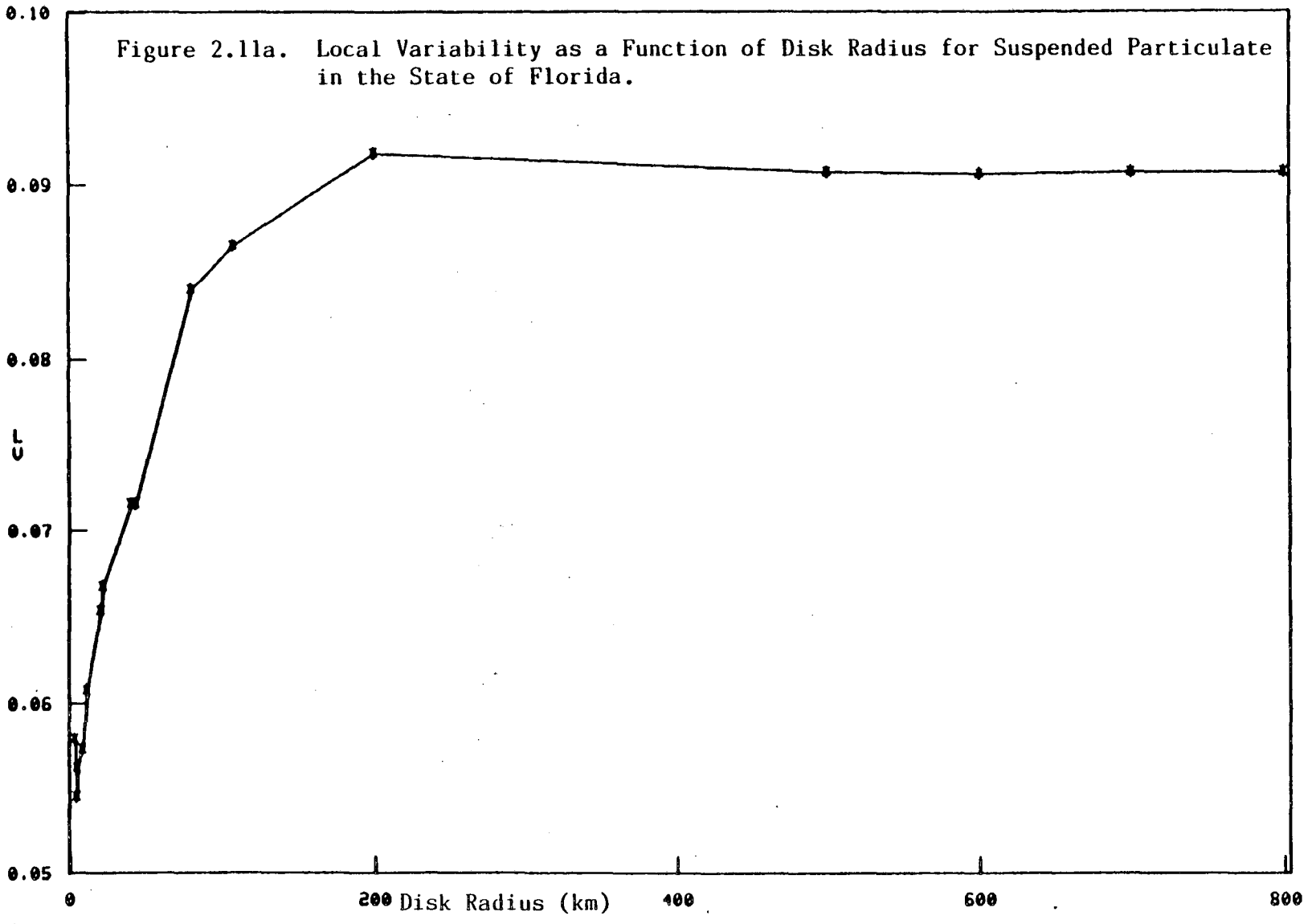


Figure 2.11a. Local Variability as a Function of Disk Radius for Suspended Particulate in the State of Florida.

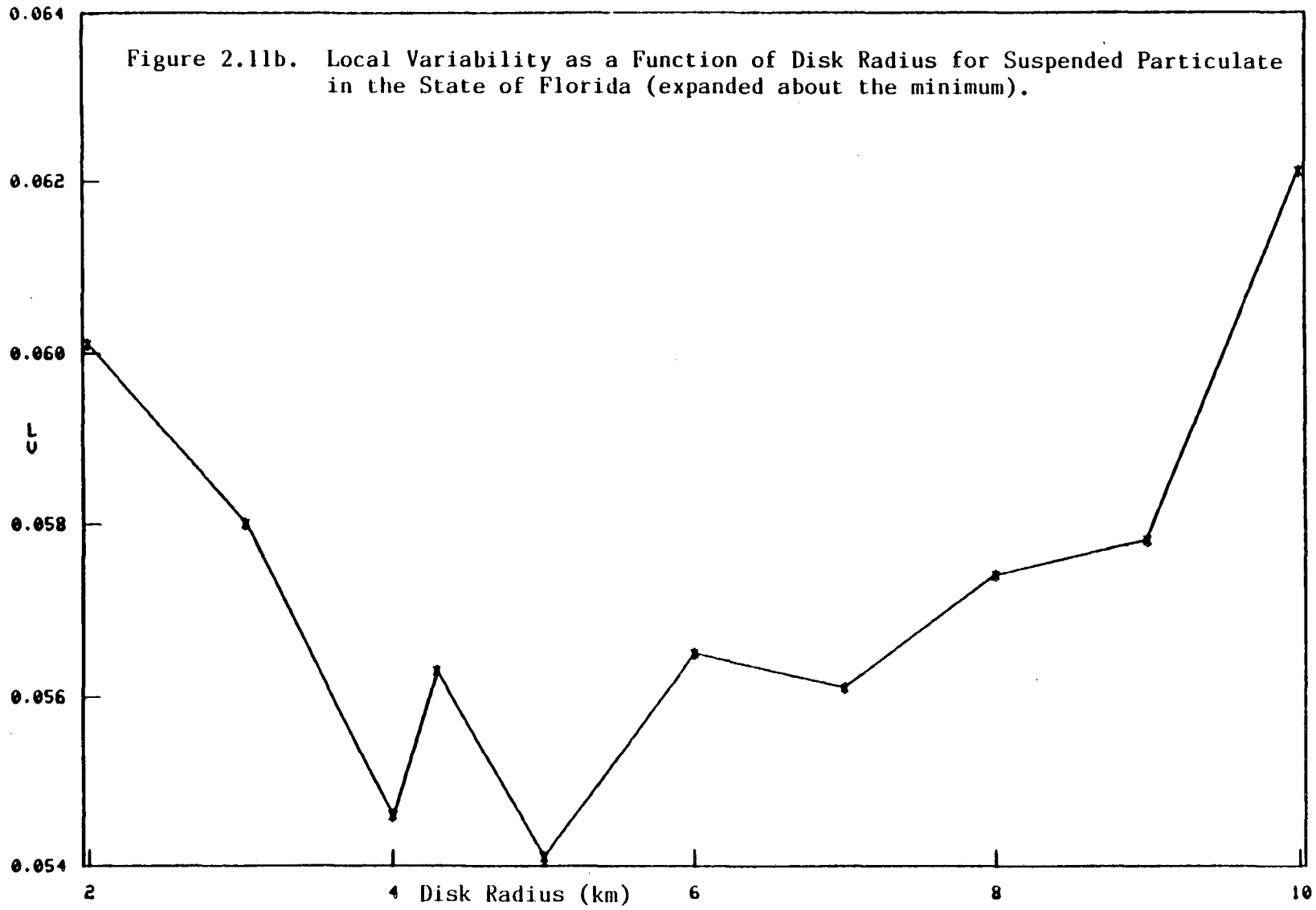
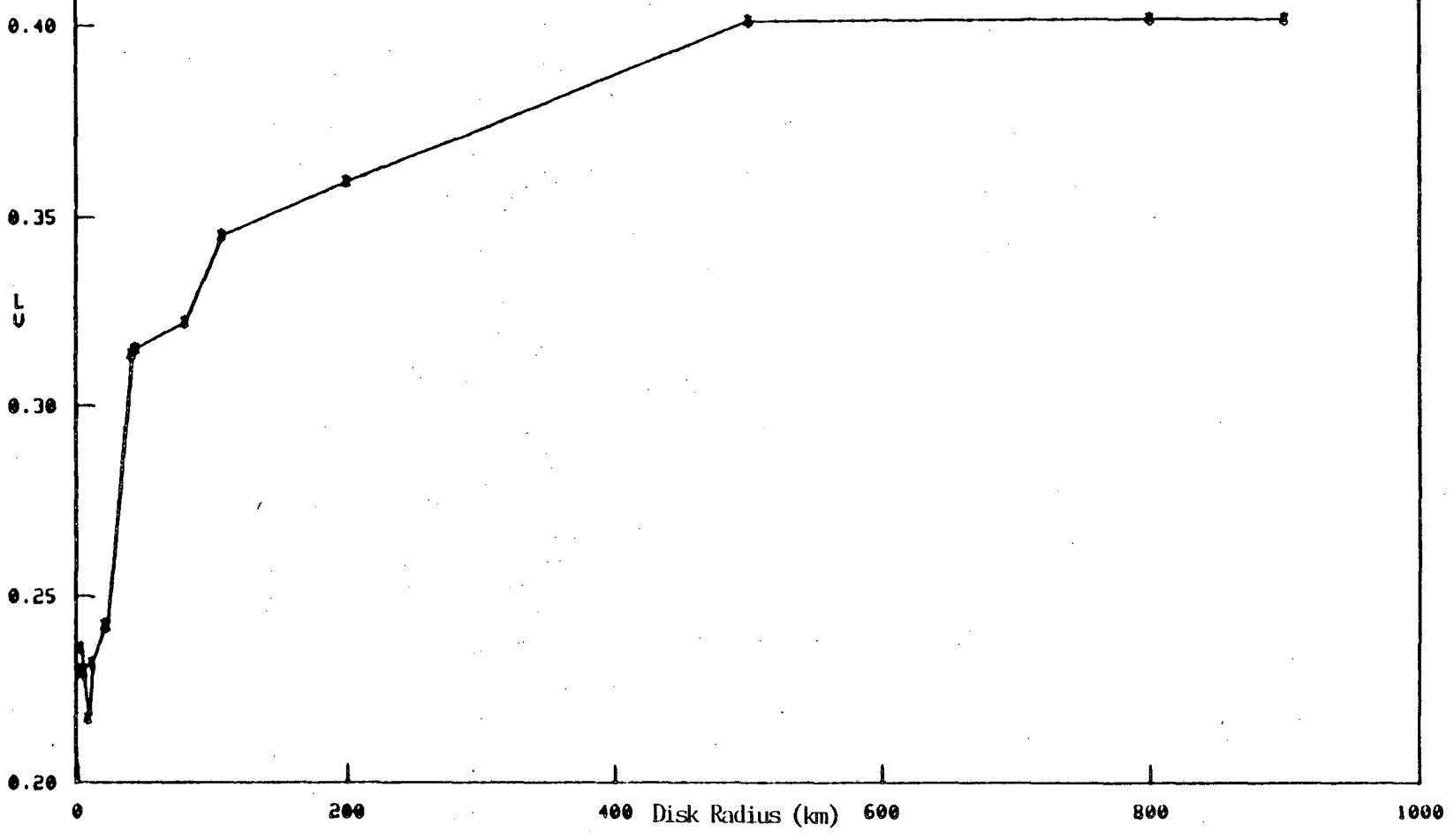


Figure 2.11b. Local Variability as a Function of Disk Radius for Suspended Particulate in the State of Florida (expanded about the minimum).

Figure 2.12a. Local Variability as a Function of Disk Radius for Sulfur Dioxide in the State of Florida.



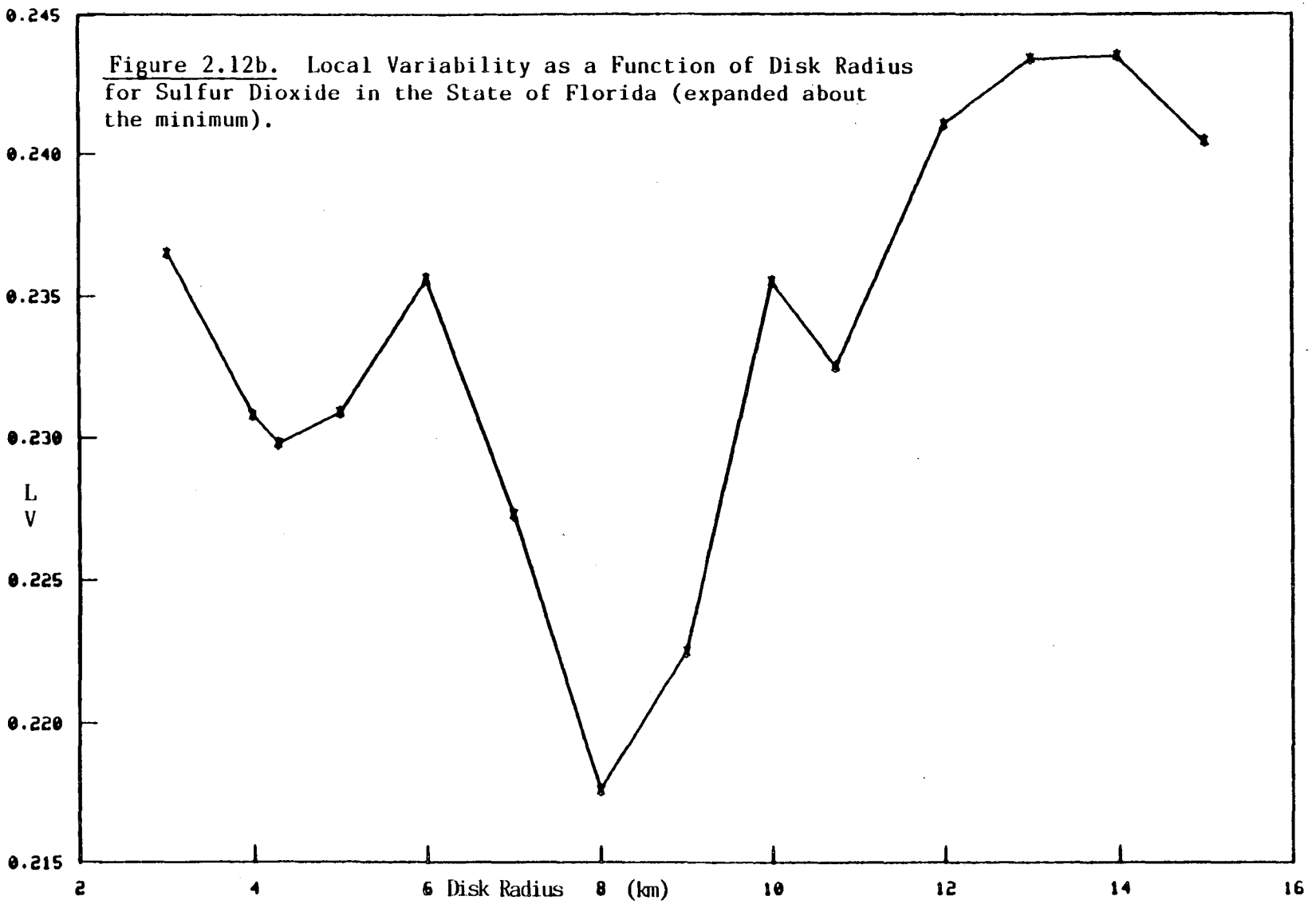


Figure 2.12b. Local Variability as a Function of Disk Radius for Sulfur Dioxide in the State of Florida (expanded about the minimum).

only one disk contains another point all the way to disks of a large enough size to contain the entire region), the logarithm (base e) of the disk radius was plotted against LV. These curves are shown in Figures 2.13a. thru 2.15b.

There are basically three regions on each curve. The first region is called the Undefined Region. This area ranges from disk radius of 0 to a disk radius large enough to have at least one disk containing two points. The size of the undefined region is determined by the particular geographic area one is studying and the geographic distribution of stations in that area. The second region is called the Region of Degeneracy. The Undefined Region is actually included in the Region of Degeneracy. The degenerate region corresponds to disk radii which are too small to have enough points in each disk to estimate variance within each disk. This area is not specifically defined but can be estimated by looking at the "degrees of freedom" and the plots of the logarithm of the disk radius vs. LV. The "degrees of freedom" specify the amount of overlap and therefore how many points lie in each other's corresponding disks. The Region of Degeneracy is designated in each of the plots, yet the line drawn is a dotted one indicating that the boundary of this region is not sharp. The third region is for disk radii greater than those in the degenerate region. These disk radii are large enough to contain enough points so that the variance around a point can actually be measured. It is in this area where the optimum choice exists for disk size.

The choice of these areas is given by the shape of the curves and the degrees of freedom. Except the curve for New York's suspended particulate local variability (Figure 2.14a.), each curve has basically the same shape. The curve climbs initially as the disk radius gets larger. It climbs steeply as the number of degrees of freedom increases. Then at some point it decreases into a

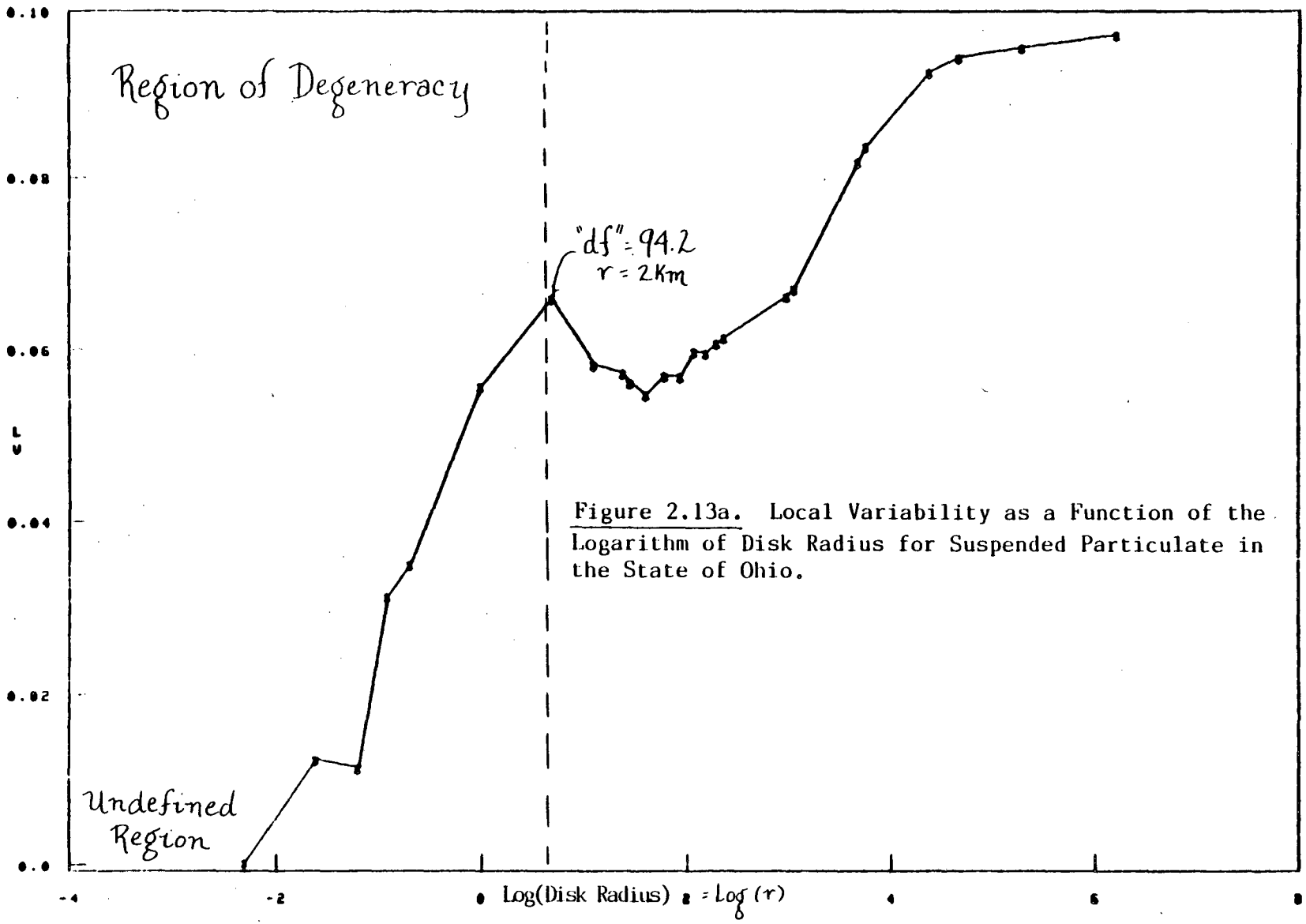


Figure 2.13a. Local Variability as a Function of the Logarithm of Disk Radius for Suspended Particulate in the State of Ohio.

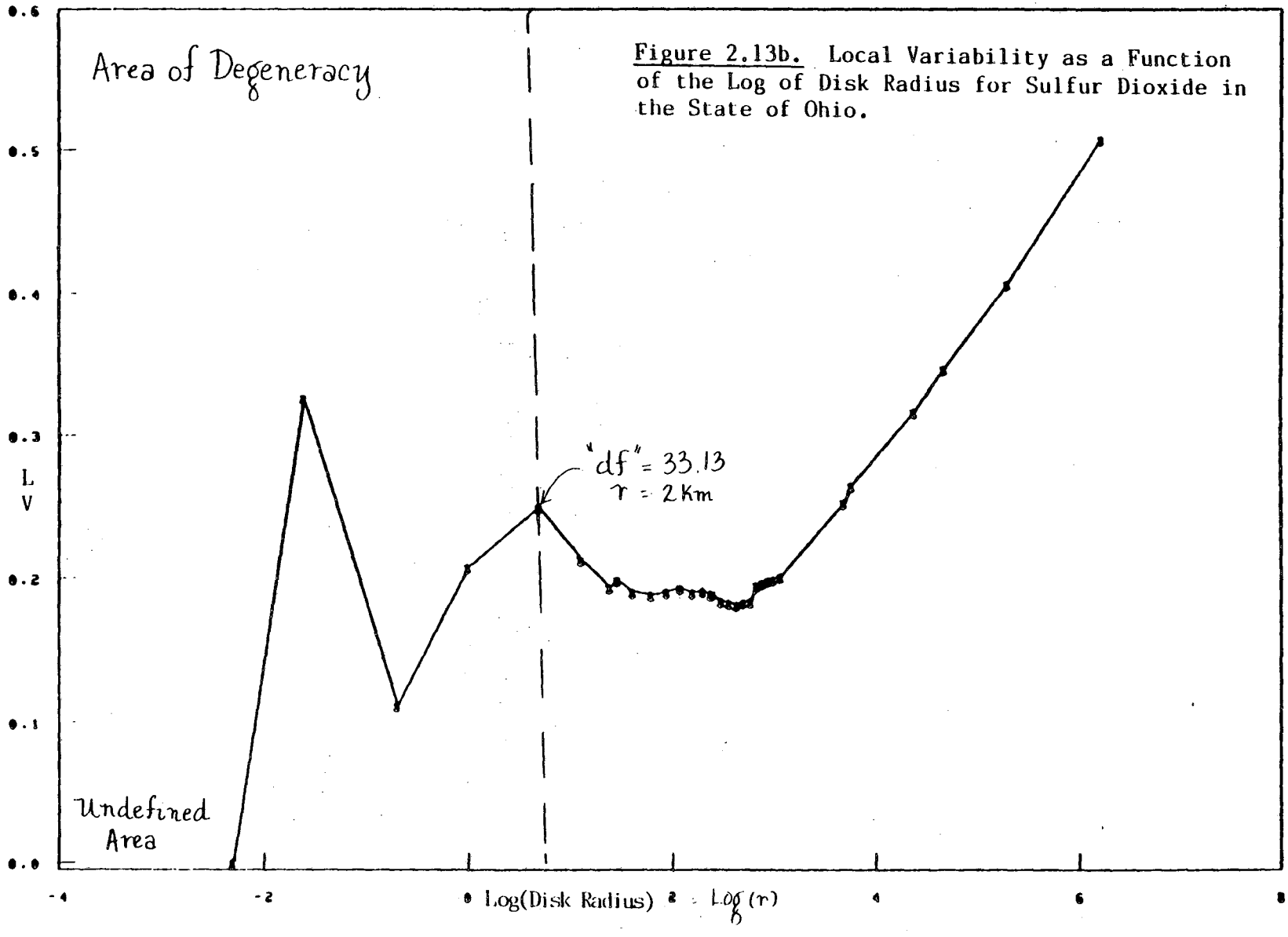


Figure 2.13b. Local Variability as a Function of the Log of Disk Radius for Sulfur Dioxide in the State of Ohio.

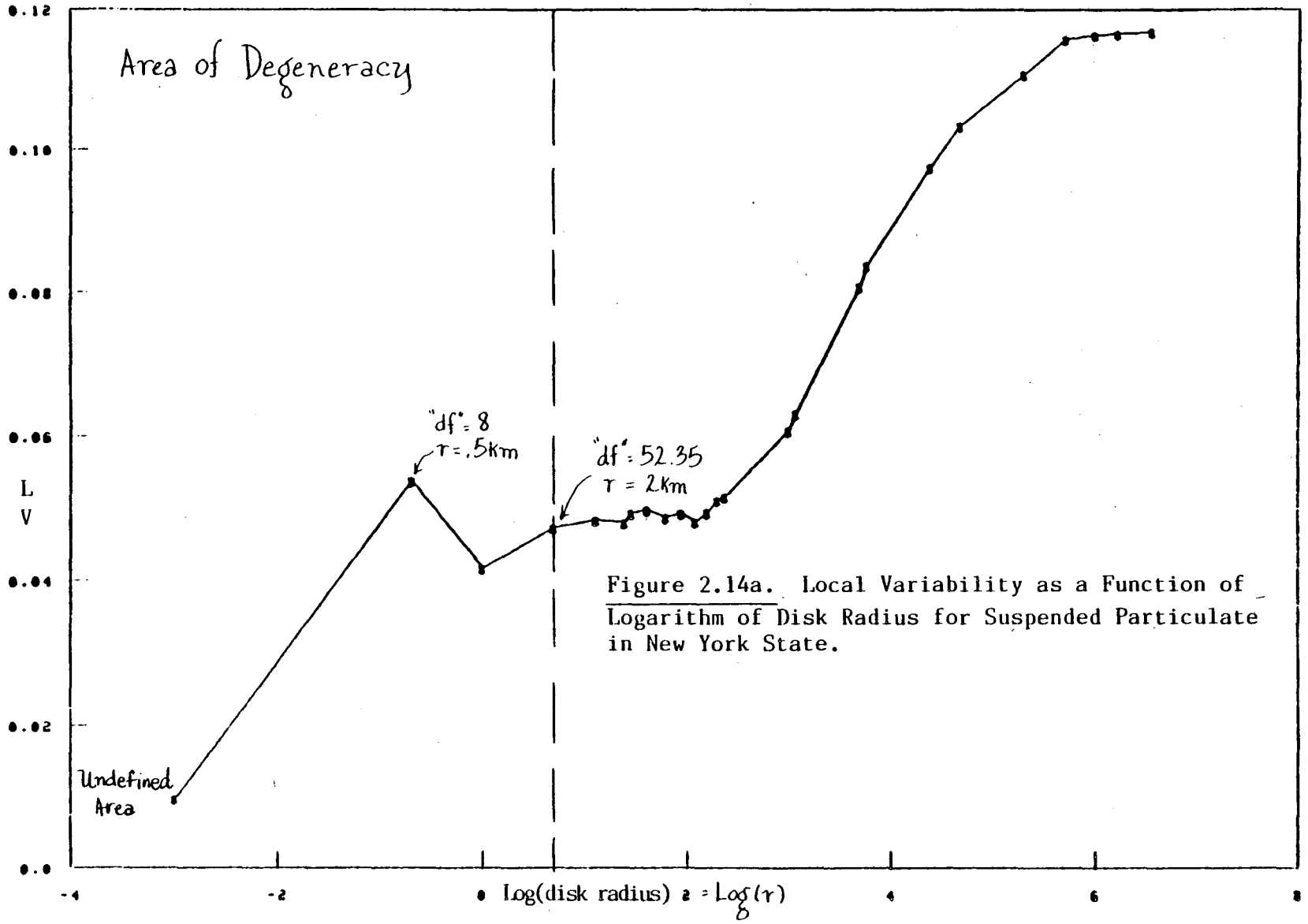


Figure 2.14a. Local Variability as a Function of Logarithm of Disk Radius for Suspended Particulate in New York State.

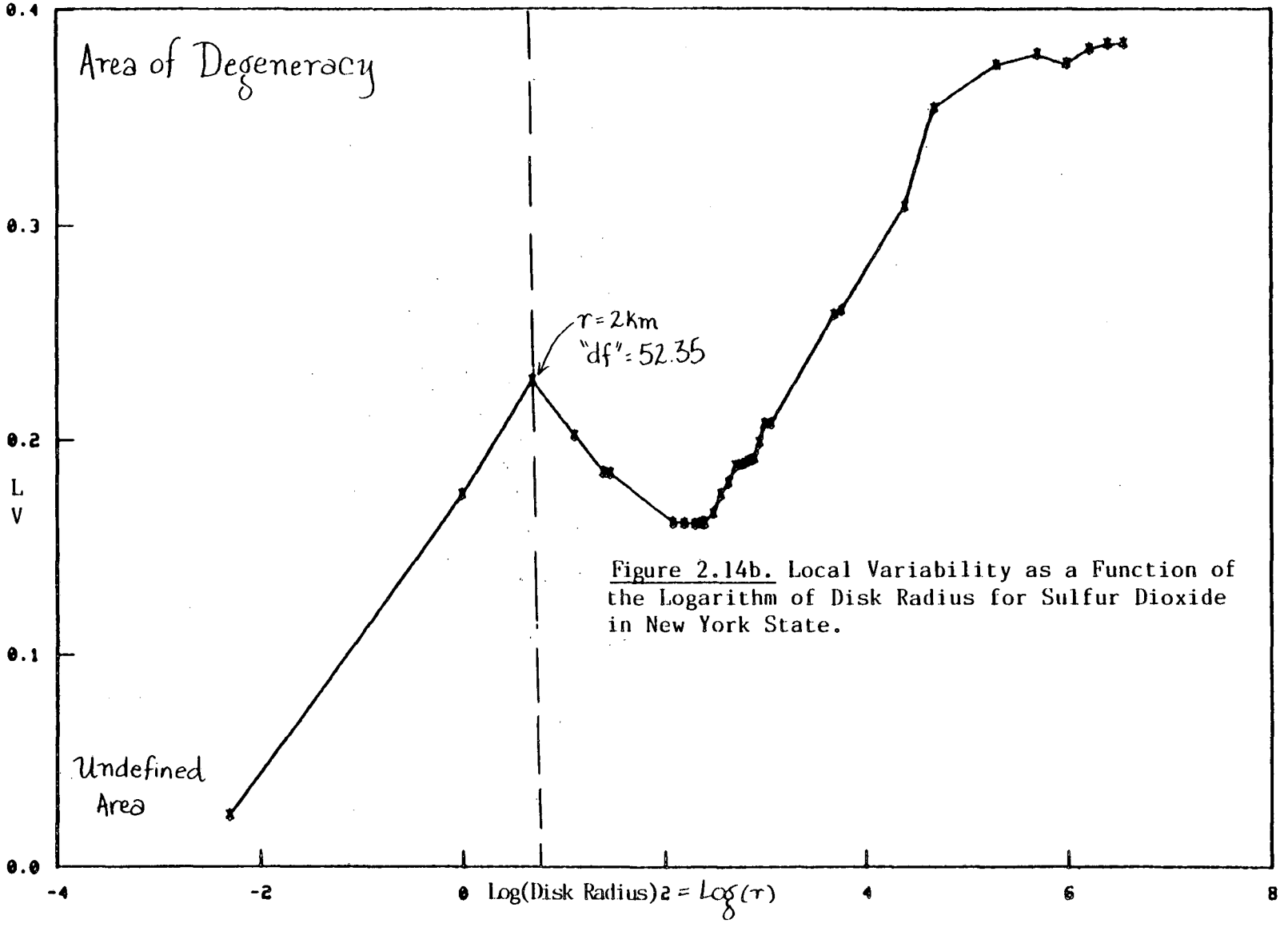


Figure 2.14b. Local Variability as a Function of the Logarithm of Disk Radius for Sulfur Dioxide in New York State.

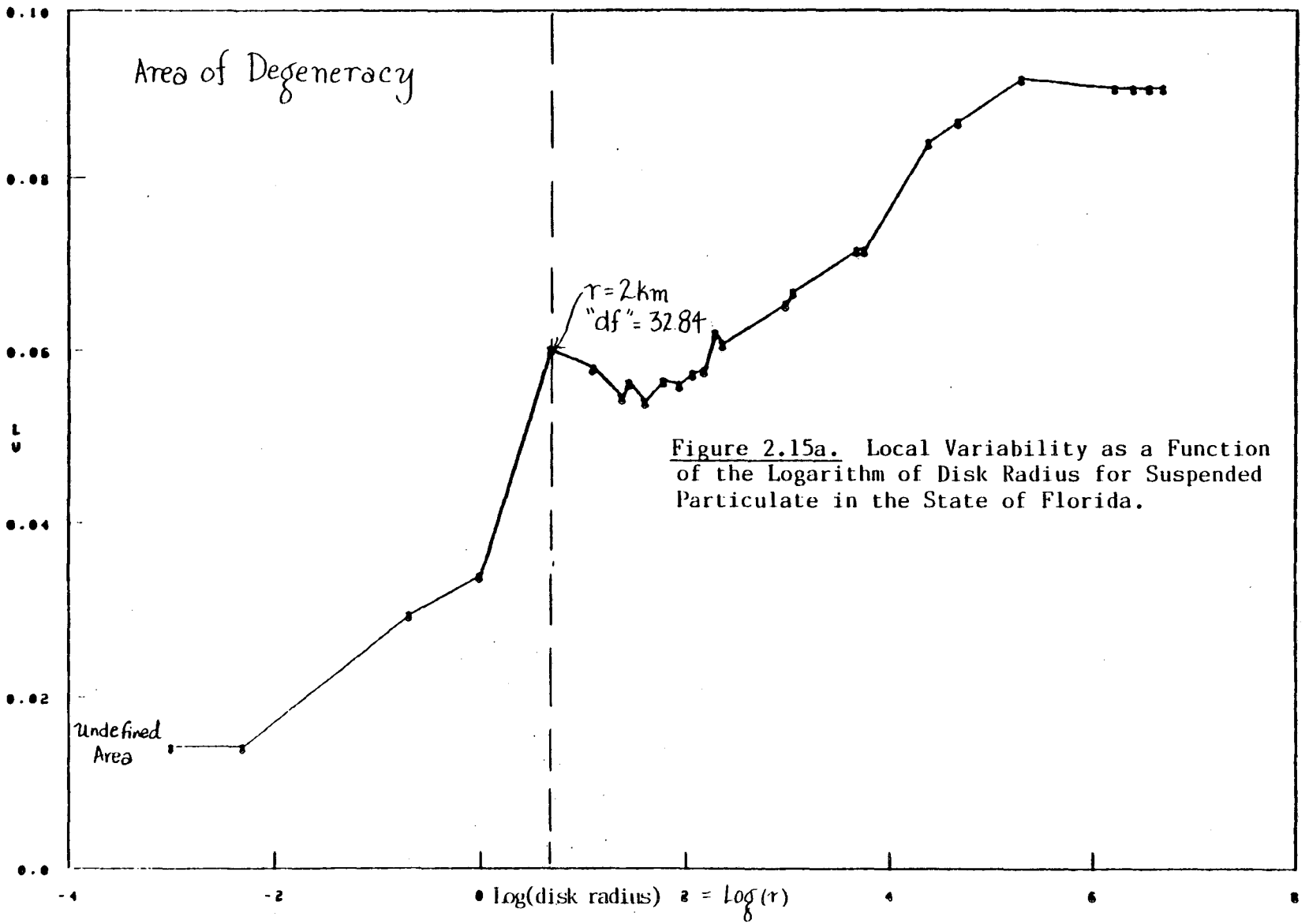


Figure 2.15a. Local Variability as a Function of the Logarithm of Disk Radius for Suspended Particulate in the State of Florida.

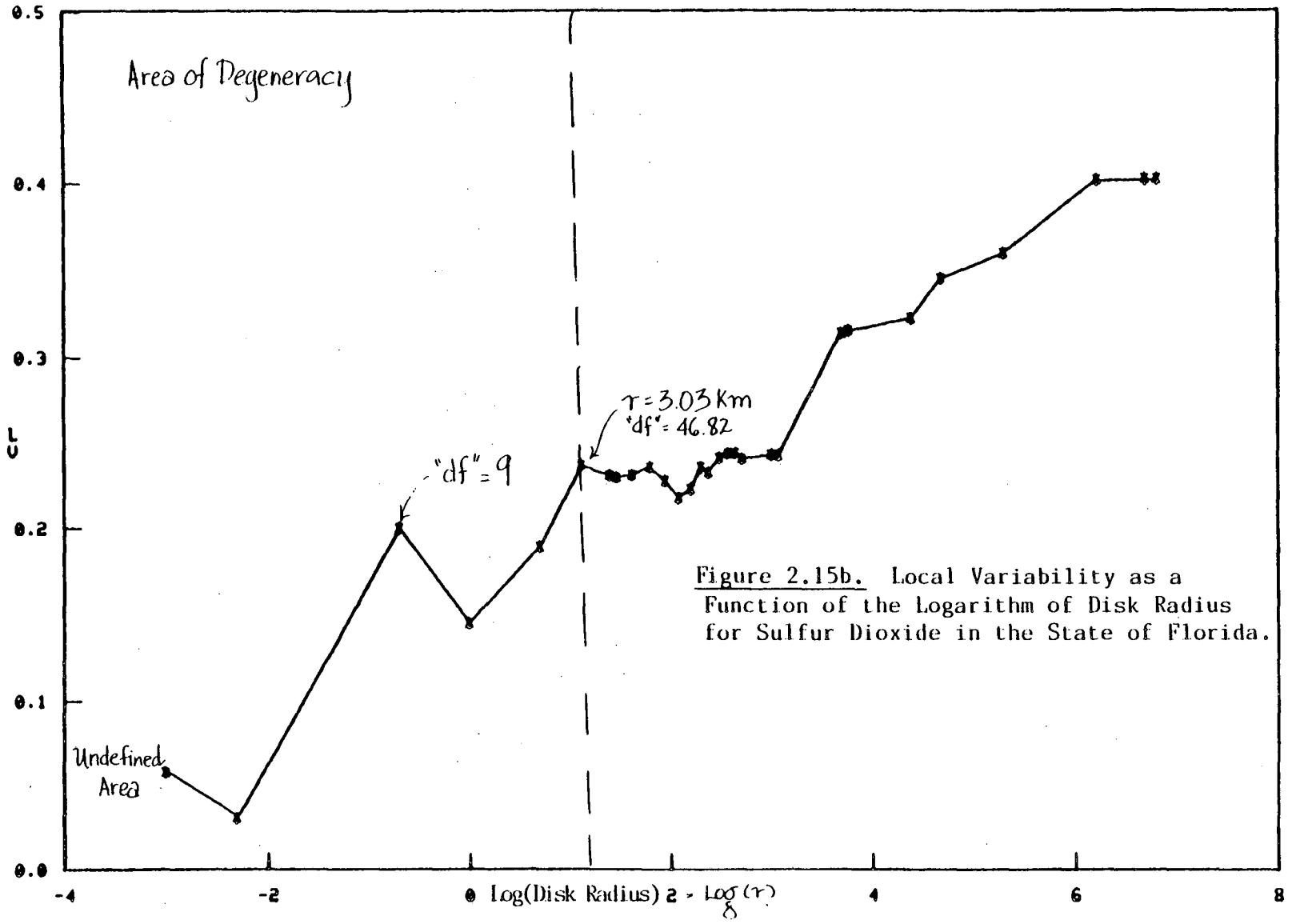


Figure 2.15b. Local Variability as a Function of the Logarithm of Disk Radius for Sulfur Dioxide in the State of Florida.

trough-like area. This is the area where the optimal disk radius is chosen. Then it increases to the maximum, which is the (global) sample variance for the entire state. For the sake of example, plots of the number of degrees of freedom for Ohio state's suspended particulate against disk radius and logarithm of the disk radius are given in Figure 2.16a. and 2.16b. Note that "df" increases quickly when disk radii are small and then levels out as the disks get large enough to include almost every station in the area. Degrees of freedom are a monotonic function of disk radius. The plot of "df" versus the log of the disk radius is an S-shaped curve. The area of degeneracy is chosen as the area where LV is increasing rapidly and stops right before the trough-like area. It increases becauses the number of degrees of freedom are increasing rapidly giving more disks containing more than one station and therefore a variance can be calculated.

It is relatively easy to choose the optimum disk radius if LV has the functional behavior shown in most of these curves. Suppose LV does not increase steadily but bounces around initially as in Figure 2.14a. Using the method described in the previous paragraph, we would choose the minimum which lies in the area of degeneracy, since it is the first trough-like area after a rapid increase. This disk radius is only one kilometer. Knowing something about the spatial distribution of monitoring stations, we see that one kilometer would allow hardly any stations to have another near enough to have a variance associated with it. With this choice, the area of degeneracy stops at a radius of .5 kilometers corresponding to an LV with 8 degrees of freedom. This implies that $\sum_{i=1}^{N} 1/n_i = 268$. Since there are only 276 stations altogether, n_i must be 1 for most stations. Thus at a disk radius of .5 kilometers, there are an

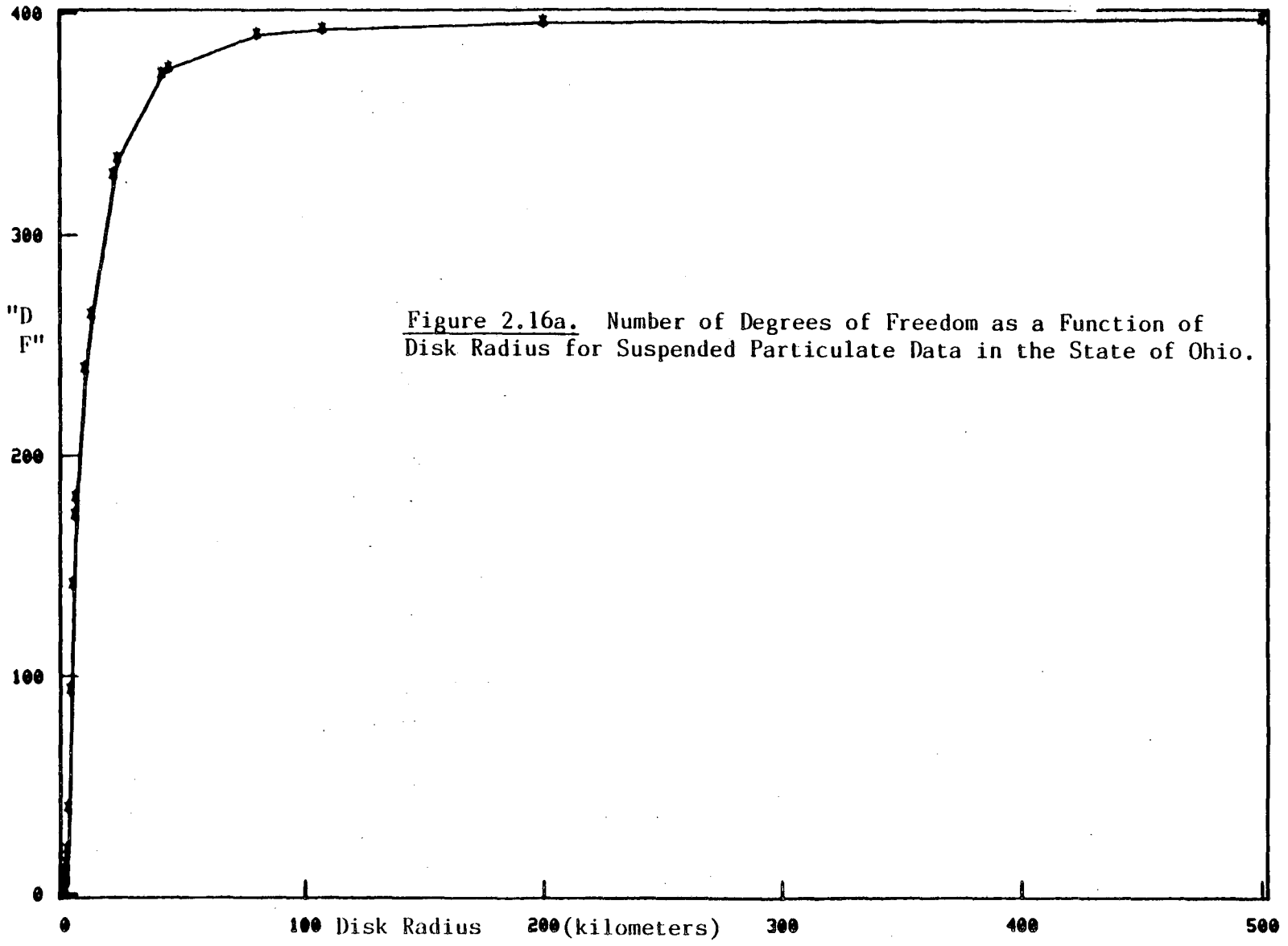


Figure 2.16a. Number of Degrees of Freedom as a Function of Disk Radius for Suspended Particulate Data in the State of Ohio.

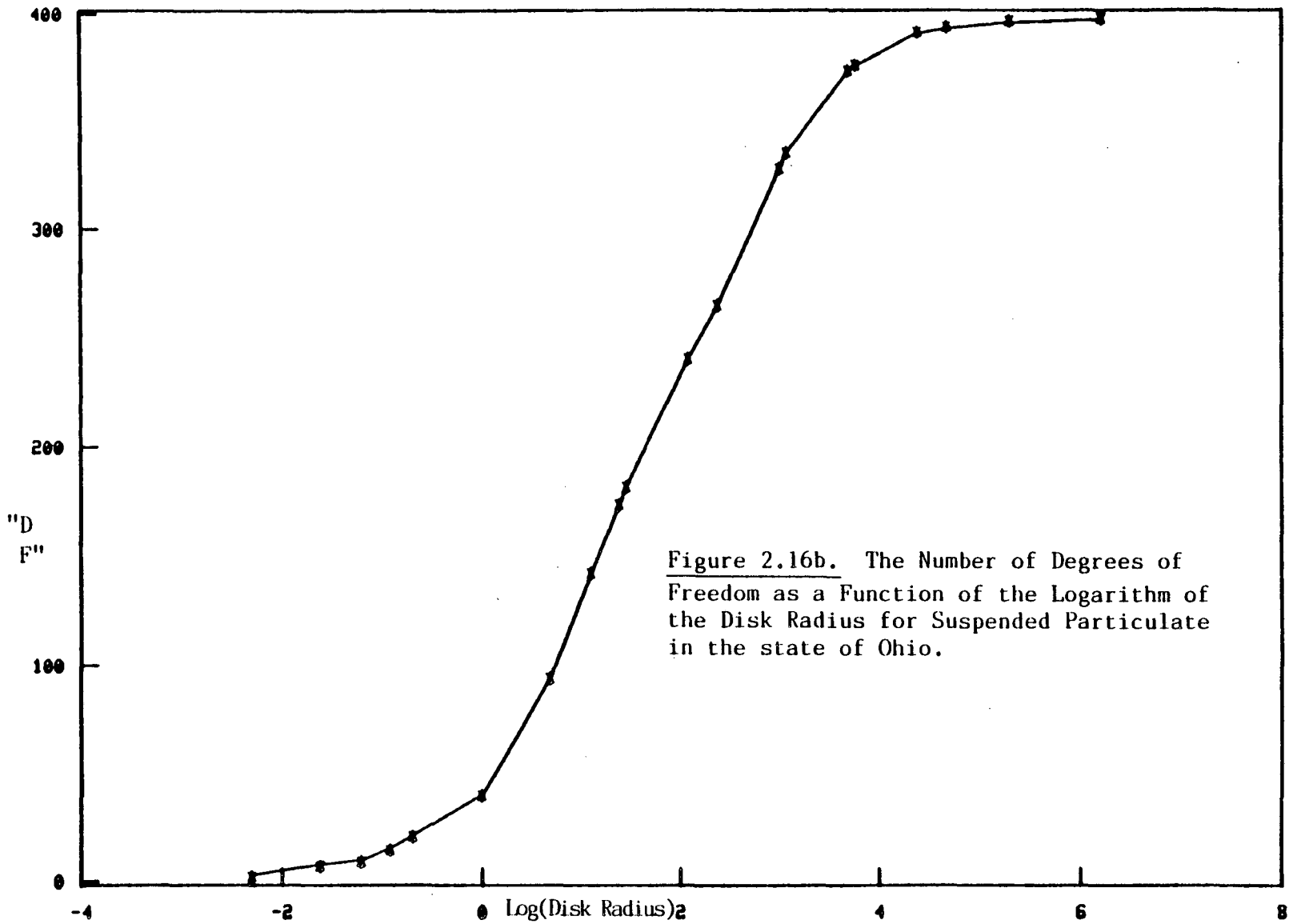


Figure 2.16b. The Number of Degrees of Freedom as a Function of the Logarithm of the Disk Radius for Suspended Particulate in the state of Ohio.

insufficient number of points to calculate local variability and thus LV should be considered degenerate. Therefore, care must be taken in the choice of areas. We must closely examine the number of degrees of freedom to see how reliable LV is at various disk radii.

Perhaps there are two ways one can use local variability to choose the "best" disk radius. One method is to choose the minimum disk radius greater than the degenerate area's upper limit, which is $4 \times d_0$ in the proposed scheme. This method might not be the best since LV is data dependent, causing the curve to fluctuate around the minimum making it difficult to find the true minimum. Another method which is harder to define because it does not involve choosing a particular point, instead involves choosing a region of disk radii for a Region of Homogeneity. In each of the curves of the logarithm of the disk radius versus local variability, after the region of degeneracy there is a trough-like area and then LV starts to increase rapidly. We could call this trough-like area the region of homogeneity. It is better to choose a large disk radius to get more points which will give a a more precise estimate. This argument might suggest a method which chooses the largest possible disk radius before LV starts increasing rapidly. Another method might be to take the disk radius in the middle of the trough. This is analogous to choosing d_0 and $4 \times d_0$ in the moving average. The middle of the trough may be our choice of d_0 and the end of the trough may give us our upper limit. In other words one should not take any points into the estimate which are outside the area of homogeneity. Which point is taken as giving the most homogeneous disk is probably not critical for estimation. Yet it is important, however, to know which disk radii not to use, and those disks can be determined by investigating the relationships between local variability and disk radius.

Minimizing Local Variability as an Alternative to Cross-Validation

The two dimensional moving average model has been used to estimate an average pollution concentration for particular geographic areas such as census tracts and counties. The interest in calculating an average pollution level is one of measuring human exposure. Thus the model is usually calculated for the latitude and longitude equal to that of the population centroids. To determine the best value of the parameter d_0, cross-validation techniques were used (Selvin *et. al.* 1980, Selvin *et. al.* 1981, Johnson *et. al.* 1982). Using the concept of local variability requires less computer time, besides the fact that local variability is of interest in its own right. For example, comparing across states or across pollutants to indicate different dispersion mechanisms can be done by analyzing the local variability function.

For predicting the pollution concentration at the centroid of a geographic area, cross-validation chooses a d_0. This d_0 gives a disk radius which corresponds to a disk for which the centroid is the center. The two-dimensional moving average model is applied to all the points inside this disk for the "best" prediction of the centroid's concentration. Since we have not measured the concentration at the centroid point itself, we cannot use it to choose a disk radius. Yet if we assume that the region of interest has a certain geography associated with it which prescribes a disk radius that is most appropriate, we can then cross-validate every value in the region and choose the d_0 which does the best on the average for this region. We then can use this as the d_0 in our centroid prediction. Another method which is much faster because we do not need to use the model with so many data sets, is to measure local variability as a function of disk radius. We would then use the d_0 which corresponds to the

disk radius which gives us the smallest local variability.

The local variability method is superior to cross-validation for at least two reasons. It is computationally faster; since LV is only a double sum. It is more straightforward; that is, one does not face the problem of which goodness of fit function to look at in the analysis of the residuals in cross-validation. All one needs to do is to select that disk radius which minimizes LV. Another advantage of LV is that it is non-parametric, where in cross-valdation the choice of disk radius (i.e. d_0) depends on the form of the weighting function. Other reasons why it may be superior are given in Chapter 3.

Local variability analysis could probably be used for choosing parameters in other methods of surface estimation. In nearest neighbor regression, LV analysis could give the optimal number of neighbors to include in an estimate. For partition-based methods the size of the partition can be chosen by LV analysis. Local variability can be plotted as a function of the number of nearest neighbors or the size of the partition and thus the minimum local variability could correspond to the optimum number of neighbors or partition size in the same way it does d_0.

CHAPTER THREE

Comparing Two Methods of Surface Estimation

Akima's Method of Bivariate Interpolation

Akima (1978) proposed a method and algorithm for bivariate interpolation and smooth surface fitting for irregularly distributed data points. Since air quality monitoring stations are not distributed regularly, we could perhaps apply the method in this case. Obviously the two-dimensional moving average model is not the only method for estimating a surface from discrete data points. One question that arises is which method is the "best" method for geographic estimation of air quality. Perhaps one method is better for some applications while the other is better for others. Before attempting to answer this question, we must first ask, "What does 'best' mean?". Once we have a definition of "best" then we can find a method for determining the best estimation scheme. In this section, Akima's method of interpolation will be compared to the two-dimensional estimation procedure discussed in the previous chapter.

First, let us outline Akima's general method. The interpolating function is smooth with continuous first order partial derivatives. The xy plane or geographic region of interest is divided into triangular cells using the max-min criterion of Lawson (1972). This is called the Delaunay triangulation which is explained later in this section. Each data point is the vertex of a triangle. Akima's method is based on the following three assumptions. The first assumption is that the value of the function at a point (x,y) in a triangle can be interpolated by a fifth-degree polynomial.

$$z(x,y)=\sum_{j=0}^{5}\sum_{k=0}^{5-j} q_{jk}x^j y^k$$

Thus twenty-one coefficients need to be determined for each triangular region. The second assumption is that the values of the function and its first and second order partial derivatives are given at each vertex of the triangle, which gives eighteen of the twenty-one coefficients needed. The third and final assumption is that the partial derivative of the function differentiated in the direction perpendicular to each side of the triangle is, at most, a polynomial of degree three in the variable measured in the direction of the side of the triangle. The third assumption gives estimates of three additional coefficients for the required number. This method is recommended by Akima, yet air quality data points are so far apart in some cases that estimating first order partial derivatives at the data points would be presumptuous to say the least. If there were information on wind velocity or gas dispersion for these data, estimates of first and second order derivatives is conceivable. Since we do not have any additional data, linear interpolation in the triangles bounded by the data points was used here. The following equation was estimated for each triangle.

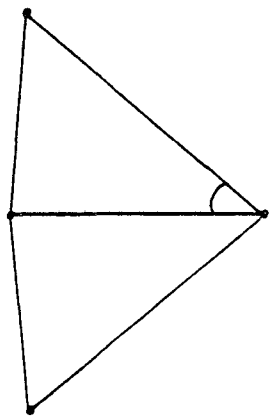
$$Z(\mathbf{x})=q_{00}+q_{10}x+q_{01}y.$$

The three coefficients in linear interpolation are easily defined by the three data points. This technique of surface estimation differs from the two-dimensional moving average model in at least three ways. First of all, Akima's method is defined everywhere within the convex hull of the data and the moving average method will not estimate points which do not have stations within a certain radius. Secondly, linear interpolation estimates a surface which necessarily contains all the data points where the two-dimensional moving average surface does not necessarily go through any of the data points. Lastly, the interpolation

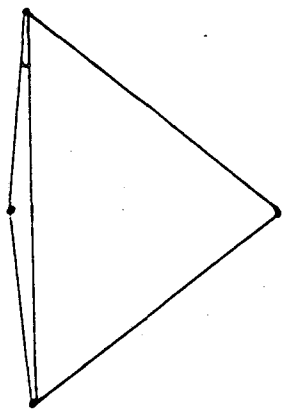
method predicts a surface which is not smooth, that is, on the lines connecting the data points the surface is continuous but not differentiable. Unlike Akima's method, the surface predicted by the moving average is differentiable everywhere it is defined.

It is fairly obvious that a triangulation from several data points is not unique. Different grids or triangulations would of course lead to different piecewise functions for interpolating the data. We used the Delaunay triangulation which can be found using the max-min angle criteria. The following criteria are used. If a set of four points are vertices of a quadrilateral with each interior angle smaller than π, then there are two possible ways to partition it into two triangles (See Figure 3.1). Lawson (1972) prefers the choice that maximizes the minimum interior angle of the two triangles produced. So in this case, we would choose (a) in Figure 3.1. The intuitive reasoning behind this method is that the chosen triangles are as equal in size as possible. Triangles constructed in this manner distribute the data over the area more evenly.

In order to compare Akima's method with the model described earlier, we first triangulate the data points and then fit a plane to the vertices of each triangle. This method clearly does not smooth the data, whereas, where the surface estimated by the two-dimensional moving average is defined, it is smooth. Smooth, in this context, means that at least first order derivatives are defined. The resulting surface of Akima interpolates the data, that is, it passes through all the given points, and is defined everywhere in the convex hull of the data points. It, therefore, is most appropriately applied when precise z-values are given or the errors are negligible. In air quality we know the errors are large and therefore this method is probably not appropriate for that reason alone. However it is useful to consider it as a comparison to other surface-fitting



(a)



(b)

Figure 3.1. The Delaunay triangulation is chosen in (a). It gives the maximum minimum interior angle of the two angles produced.

techniques. Unlike this bivariate linear interpolation (BLI) method, the moving average is (i) smooth where it is defined, (ii) not defined everywhere within the convex hull of the data points, and (iii) does not interpolate the points. First how the definition of "better" needs to be determined and then we can define the superior technique.

Linear Interpolation of Sulfur Dioxide and Suspended Particulate Data in Ohio State

The Statistical Package known as the System S (Becker, 1981) was used to perform linear interpolation on the data from two pollutants, sulfur dioxide and suspended particulate, in Ohio. The estimate of sulfur dioxide was a combination of 1-hour and 24-hour data, defined as described in Chapter 2. The suspended particulate data are taken on a 24-hour basis. Values for 1600 equally spaced points in a 40 by 40 rectangular grid were estimated using linear interpolation. That is, 40 points equidistant over the range of latitude and 40 points equidistant over the range of longitude are chosen. Akima's interpolation is performed for all possible pairs of these chosen latitudes and longitudes. Figure 3.2 and 3.3 show the contour maps derived by linear interpolation for suspended particulate and sulfur dioxide respectively.

In the contour map of suspended particulate most of the higher concentrations occur in metropolitan areas. The three highest levels on the map seem to be in the vicinity of Cleveland, Cincinnati, and Toledo. The area around Steubenville appears as an elevated plateau. Columbus and Findlay also have high levels of suspended particulate concentration. The high plateau effect may

Figure 3.2 Bivariate Linear Interpolation of Suspended Particulate in Ohio State (40 by 40 grid).

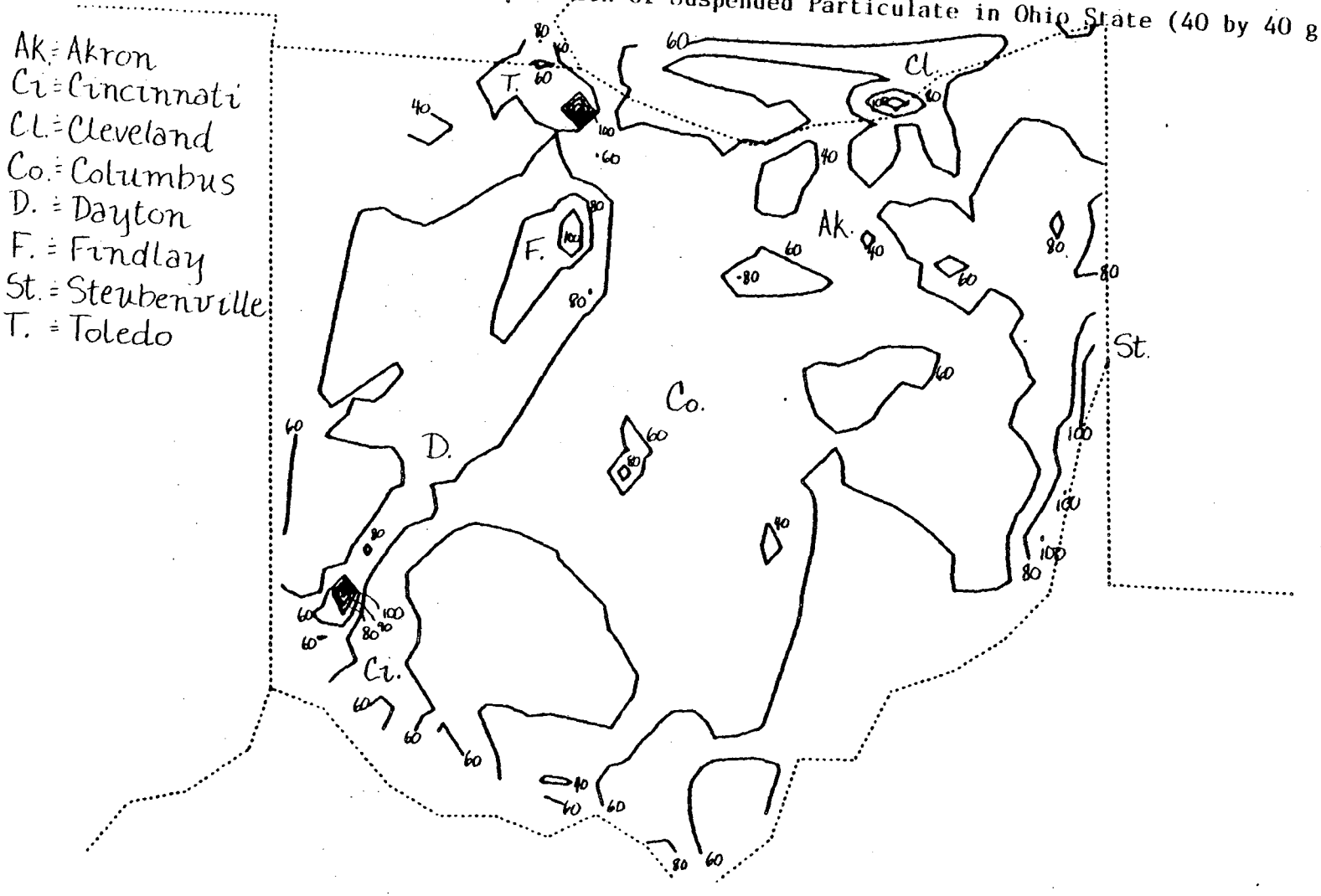
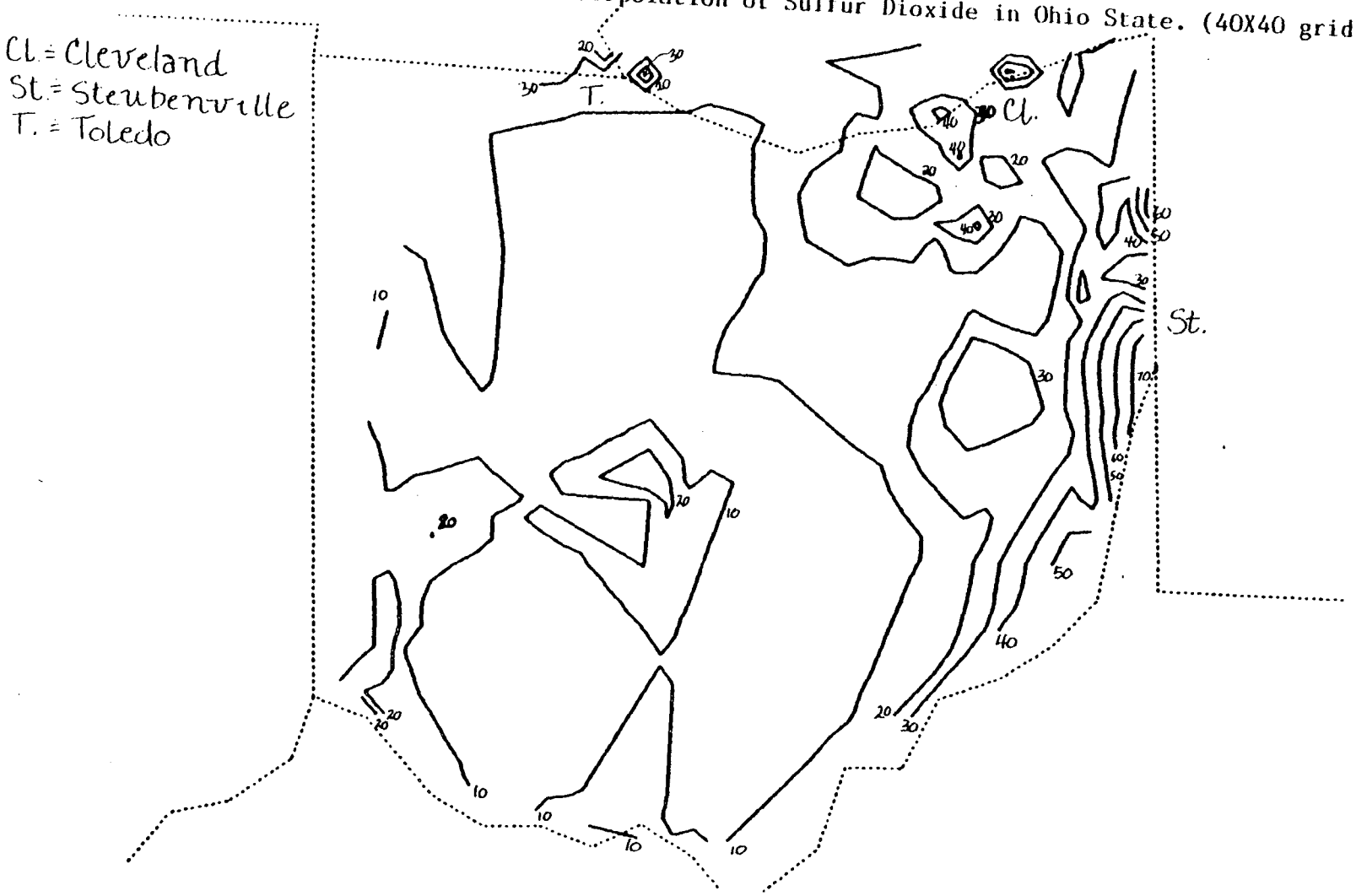


Figure 3.3. Bivariate Linear Interpolation of Sulfur Dioxide in Ohio State. (40X40 grid.)



come from the sparseness of data and the simplification which is made by the triangulation. Also one is likely to see high levels of suspended particulate in agricultural areas and not necessarily metropolitan areas.

Like suspended particulate, sulfur dioxide is high in Toledo, Cleveland and Steubenville as shown by these contour maps. On the other hand, unlike suspended particulate, Cincinnati does not seem to have a very high level of sulfur dioxide. Toledo, Cleveland and Steubenville seem to be the only metropolitan areas of high concentrations of sulfur dioxide.

There are at least three questions which arise in examination of these maps.

1. Are there enough data to make accurate contours using this method?
2. Does the method really apply?
3. Are there enough data to make accurate contours using any method?
4. What is the "variability" of these maps?

We will compare the interpolated contours with contours given by the two-dimensional moving average model. In general this interpolation method shows what we would expect from a contour map of pollution concentrations in Ohio. Usually, high pollutant concentrations in the metropolitan areas and lower concentrations outside these areas are shown. But also, we must remember that it is the metropolitan areas that are most thoroughly monitored and there is some variability associated with these estimated contours. The actual variance of a contour map is not known but its existence has been shown by Diaconis and Efron (1983) using the bootstrap method. The data were resampled and for each sampling an estimated map was drawn from the sampled data. It was found that the estimated contours do change from one sampling to another. Some parts of

the contour maps change more from one sampling to another than other areas. The areas that change more have a higher variance in the contours than those that change less. The map itself could incorporate the variance associated with the contour lines by shading, possibly with color. The areas with a high variance could be colored gray, whereas the areas with a low variance and thus more confidently estimated contours might be shaded a darker color. Contour maps from the two-dimensional moving average estimation procedure will be given in the next section.

Contour Maps Given By The Two-Dimensional Moving Average

For the same 40 by 40 grid that was linearly interpolated in the previous section, a value at each of the 1600 points was estimated using the two-dimensional moving average. The d_0 which was selected by cross-validation and local variability analysis for suspended particulate was 2 km. The d_0 chosen for sulfur dioxide was 5 km using local variability analyses and 2 km using the cross-validation techniques (although not all cross-validatory functions agreed on this choice). The grid was calculated for d_0=1, 2, 5, 10, 20 and 50 km for both sulfur dioxide and suspended particulate. The contour maps from each estimation procedure were drawn. Figures 3.4 thru 3.15 show the contour maps by pollutant and d_0.

Thus, the smaller the parameter d_0, the more difficult it is to estimate a grid point by the two-dimensional moving average because of the window definition. That is, if d_0 is 1 km then only points within 3 km of some active station can be estimated. There are very few points within 3 km of an active sta-

Figure 3.4. Contour Map Produced from the Two-dimensional Moving Average (d_o=1) for Suspended Particulates in Ohio State. (40 by 40 grid.)

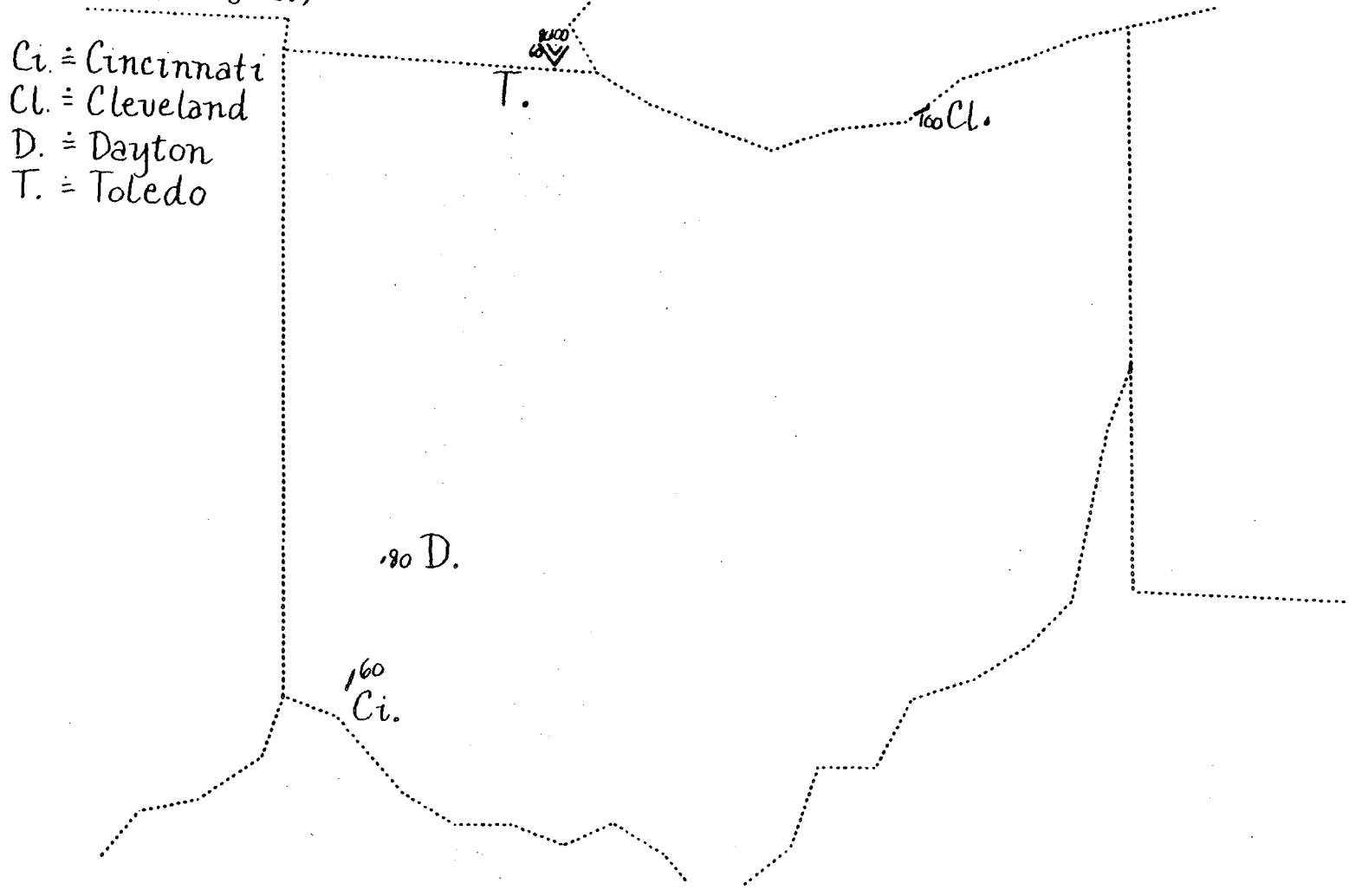


Figure 3.5. Contour Map Produced from the Two-dimensional Moving Average (d_o=2) for Suspended Particulate in Ohio State. (40 by 40 grid.)

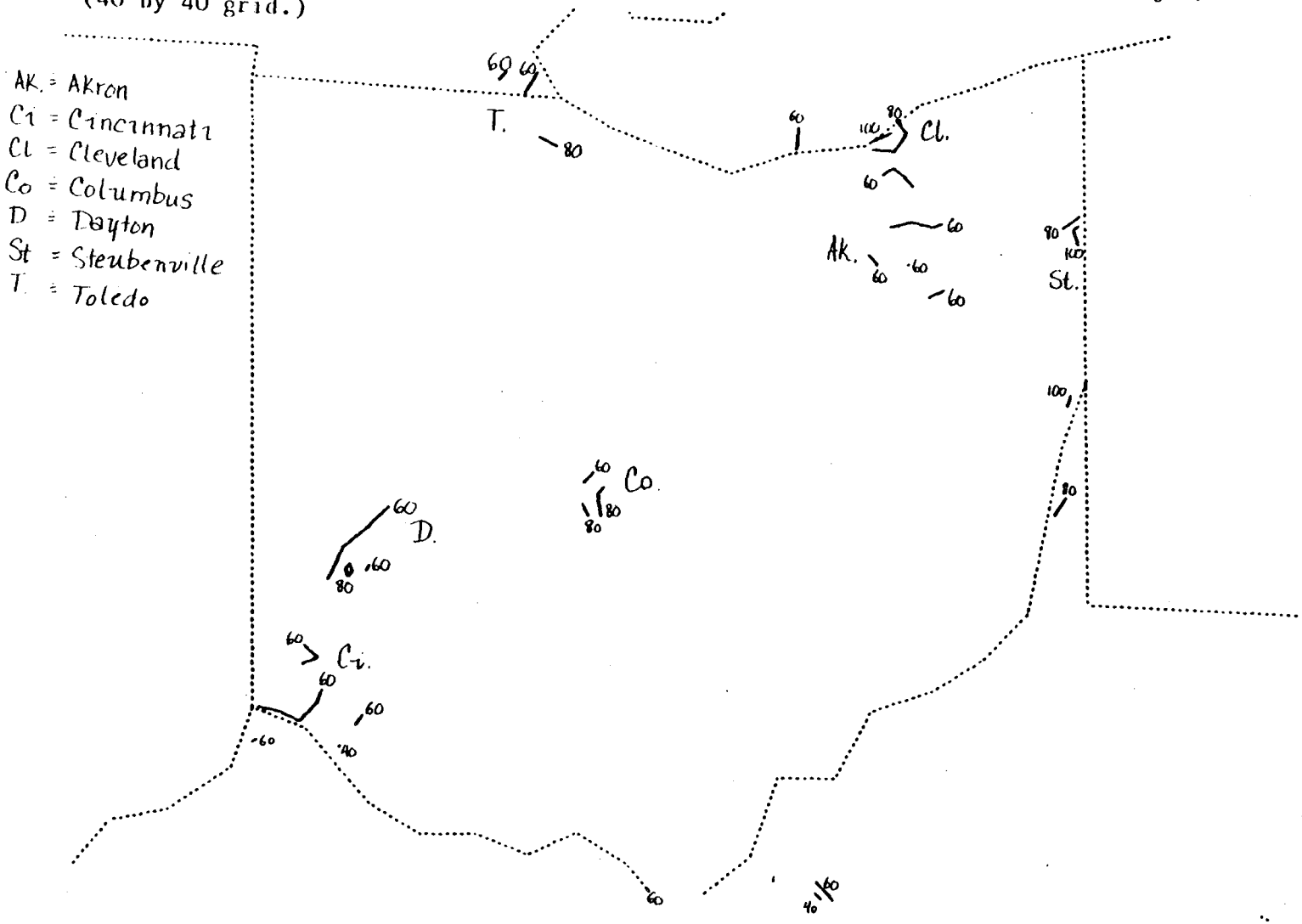


Figure 3.6. Contour Map Produced from the Two-dimensional Moving Average (d_o=5) for Suspended Particulate in Ohio State. (40 by 40 grid.)

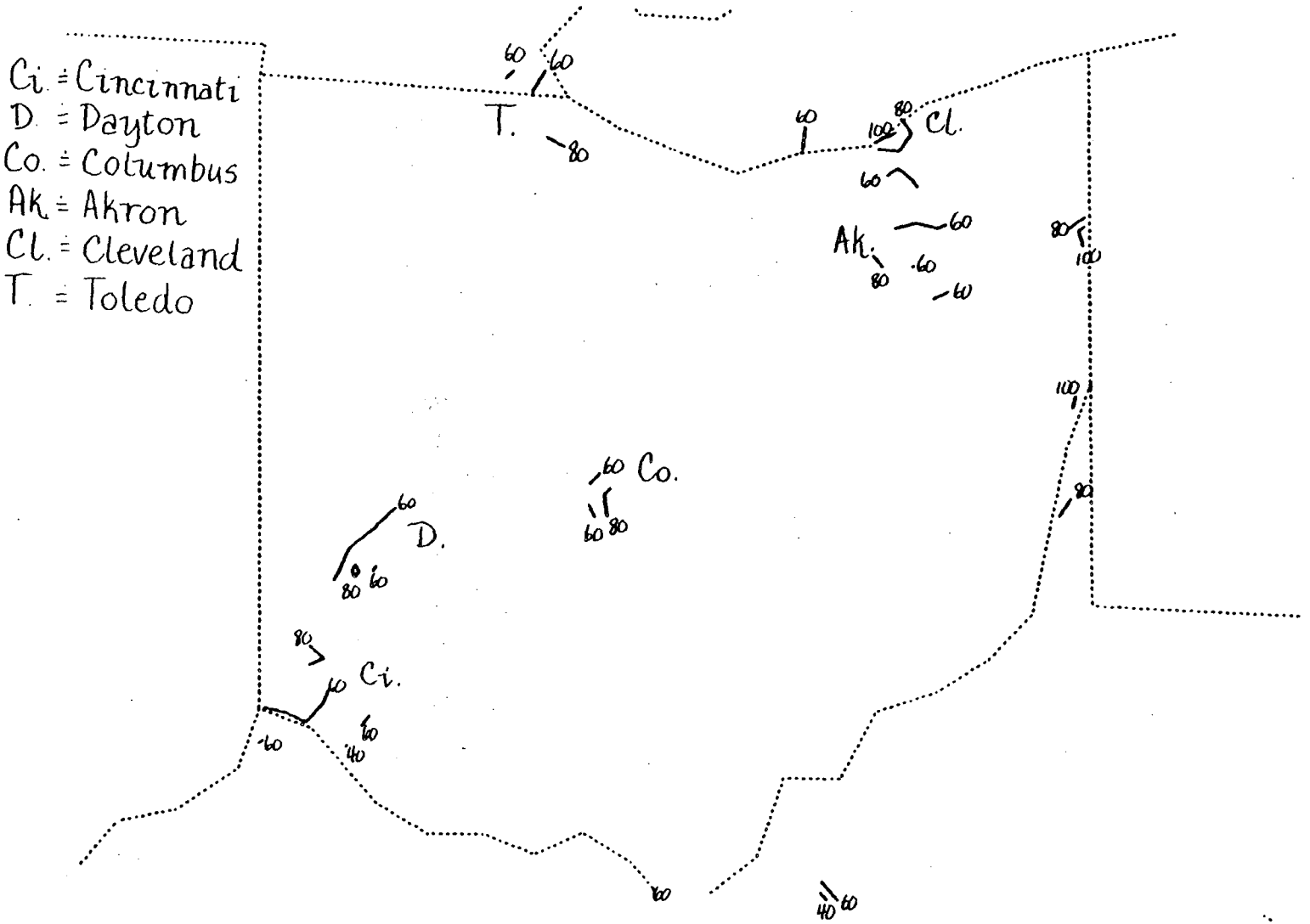


Figure 3.7. Contour Map Produced by the Two-dimensional Moving Average (d_o=10) for Suspended Particulate in Ohio State.
(40 by 40 grid.)

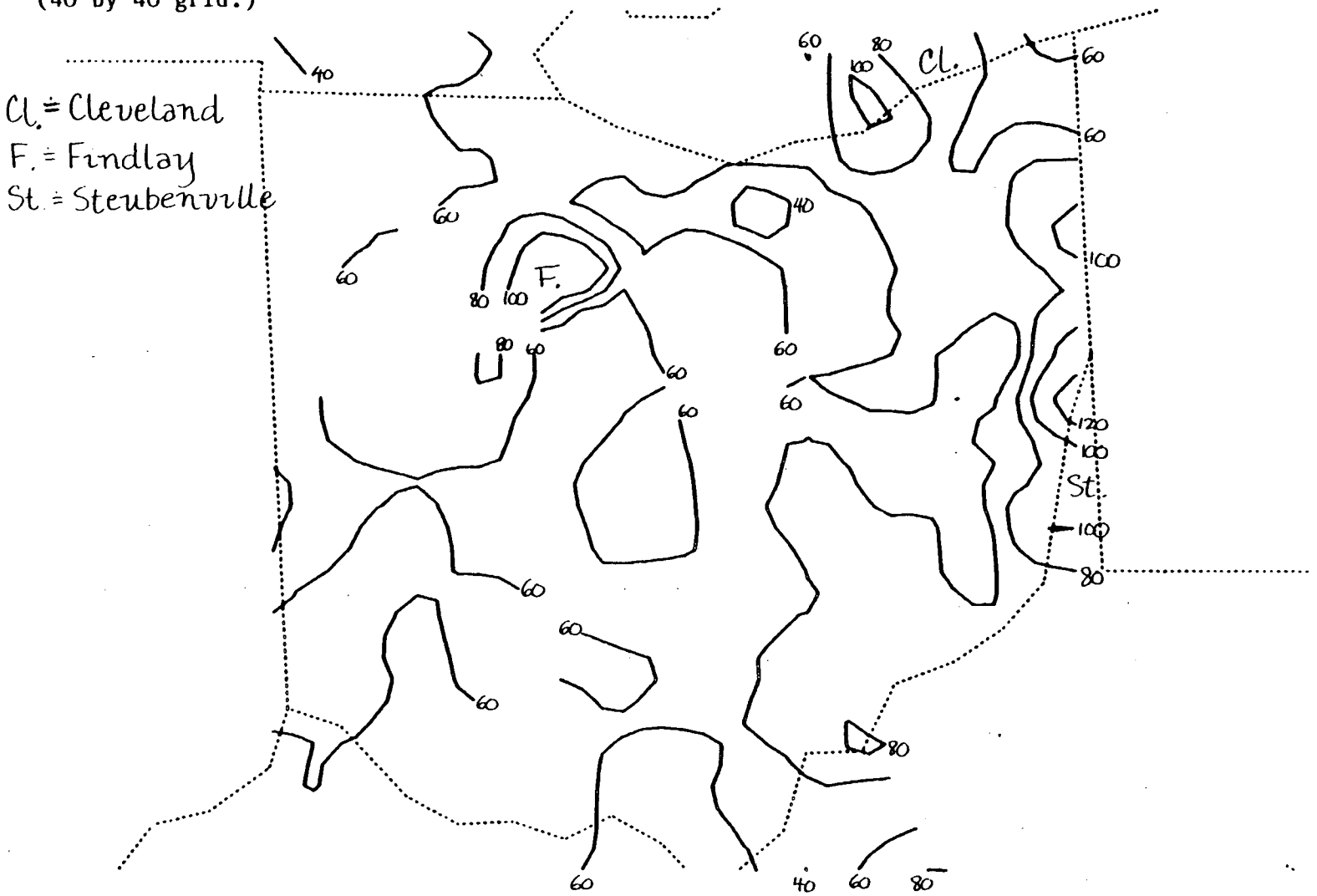


Figure 3.8. Contour Map Produced by the Two-dimensional Moving Average for Suspended Particulate in Ohio State. (d_o=20) (40 by 40 grid.)

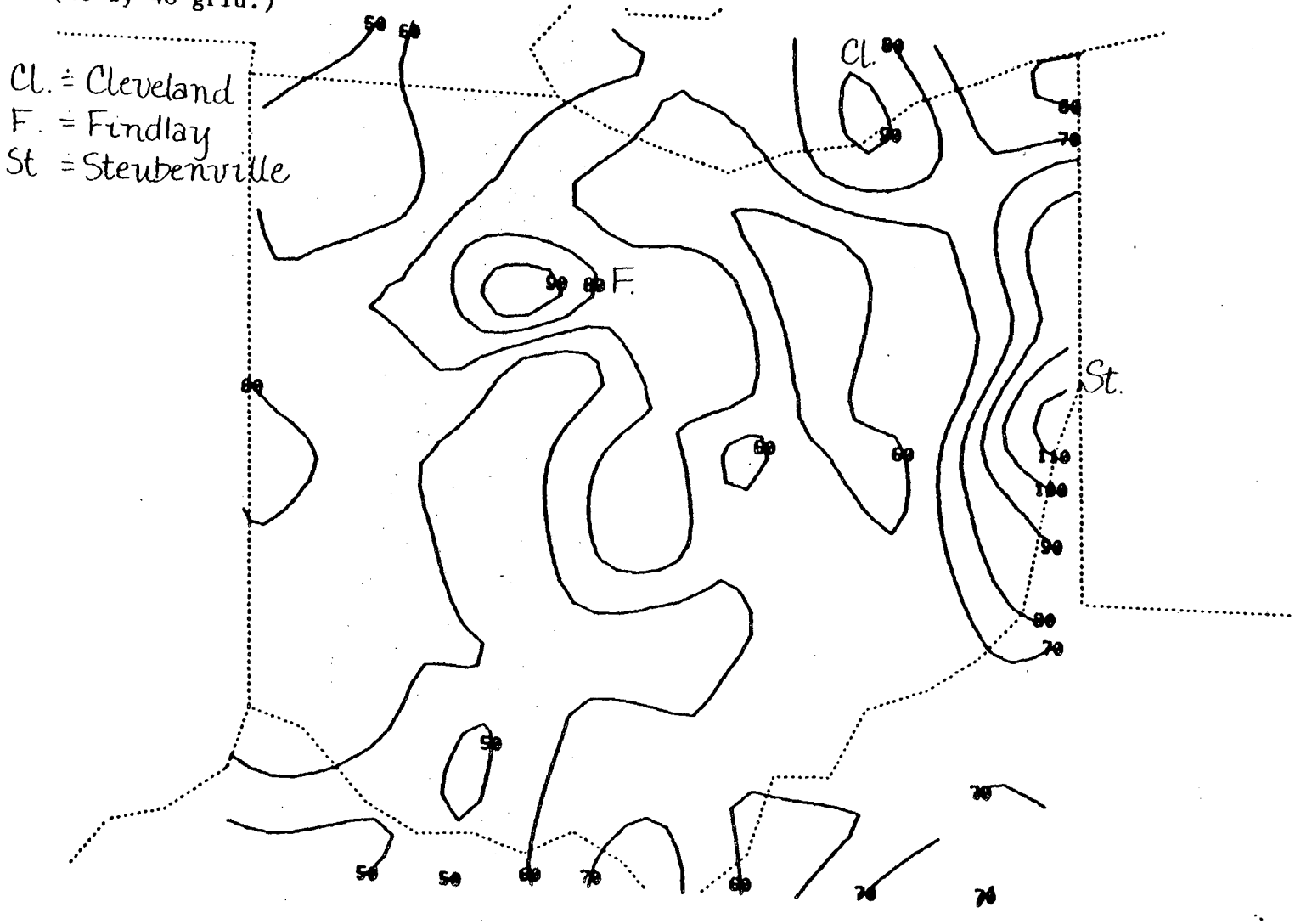


Figure 3.9. Contour Map Produced by the Two-dimensional Moving Average (d_o=50) for Suspended Particulate in Ohio State. (40 by 40 grid.)

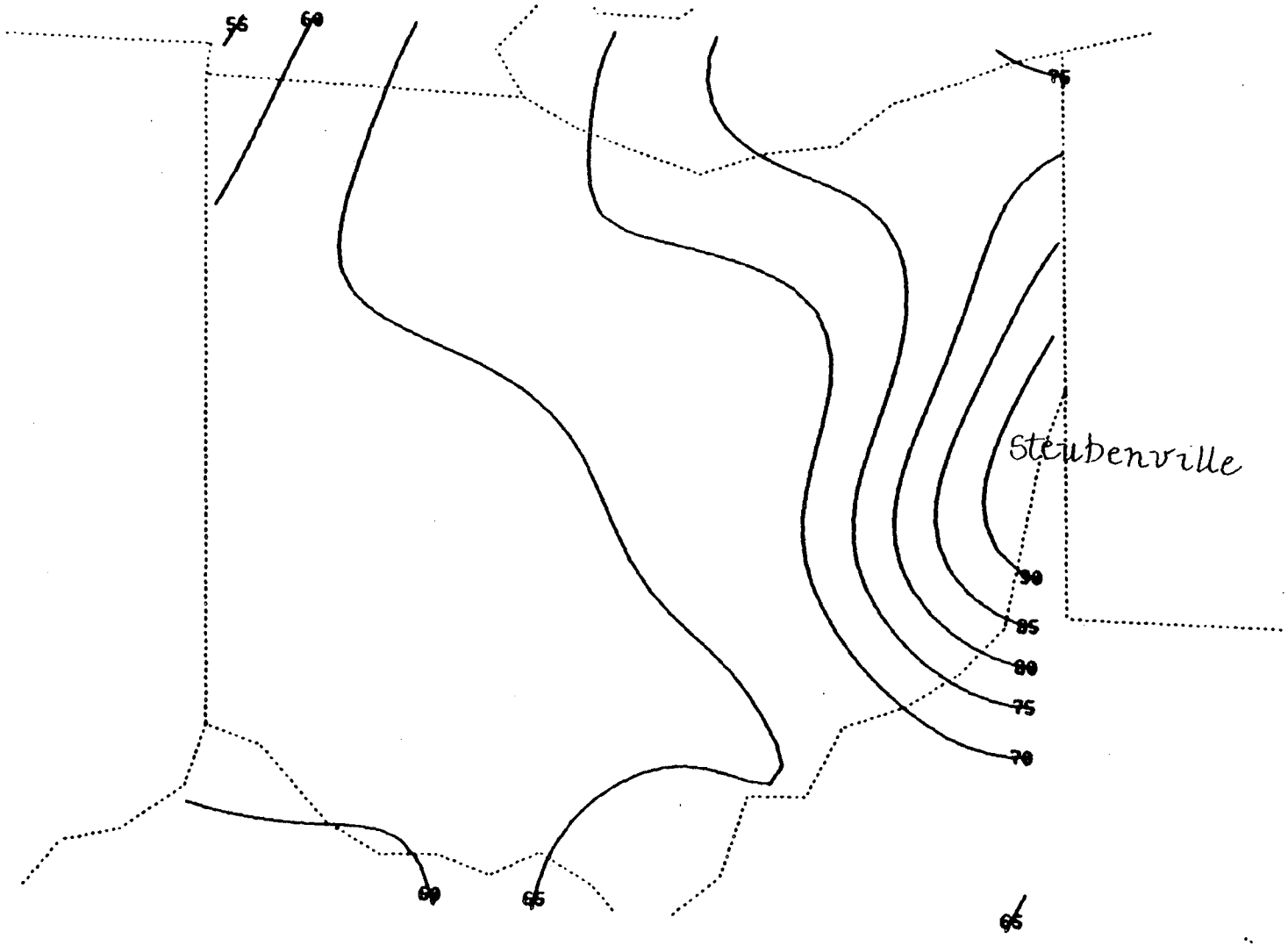


Figure 3.9a. Active Monitoring Stations Measuring Suspended Particulate in the State of Ohio, 1974-76.

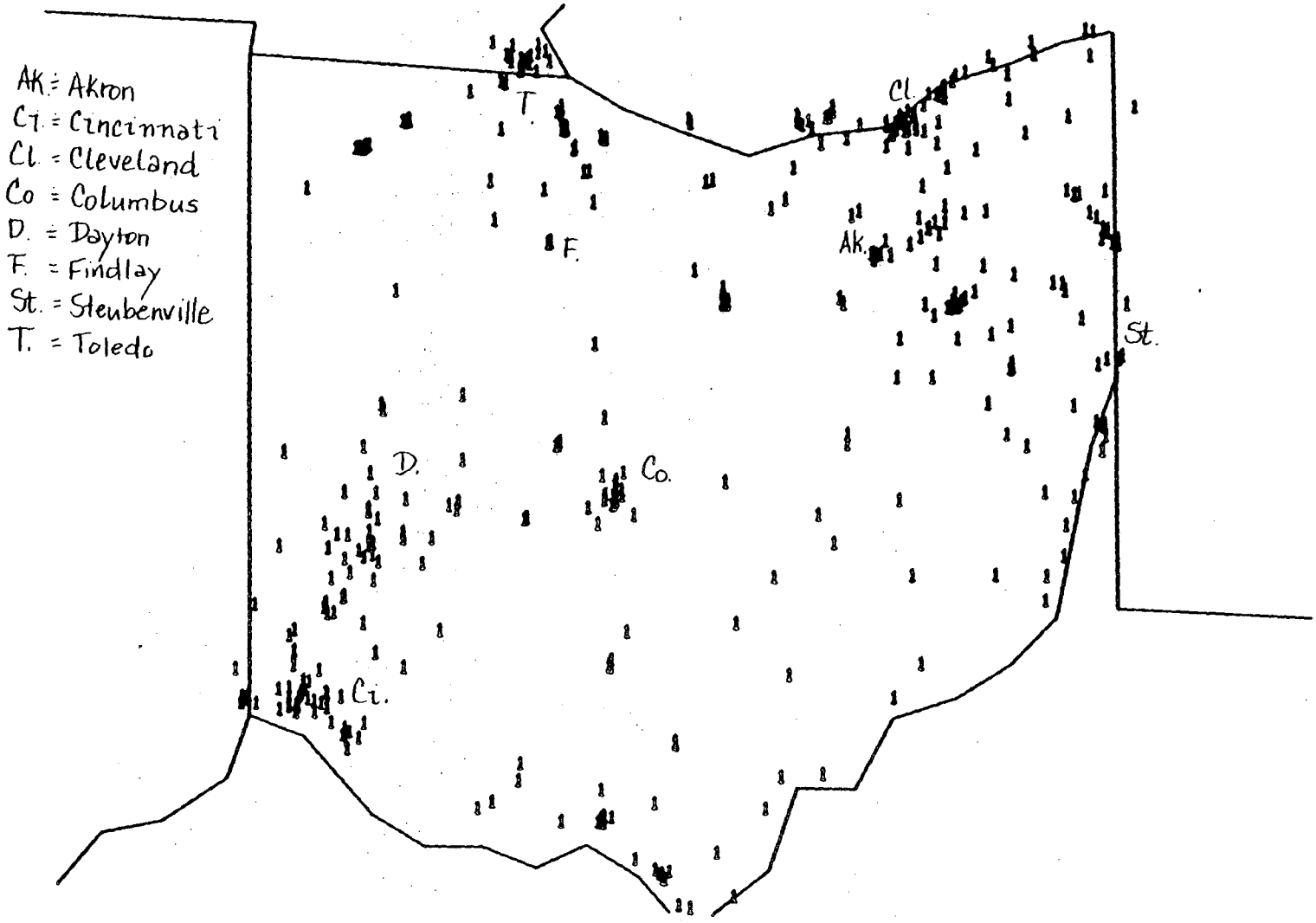


Figure 3.10. Contour Map Produced by the Two-dimensional Moving Average (d_o=1) for Sulfur Dioxide in Ohio State.
(40 by 40 grid.)

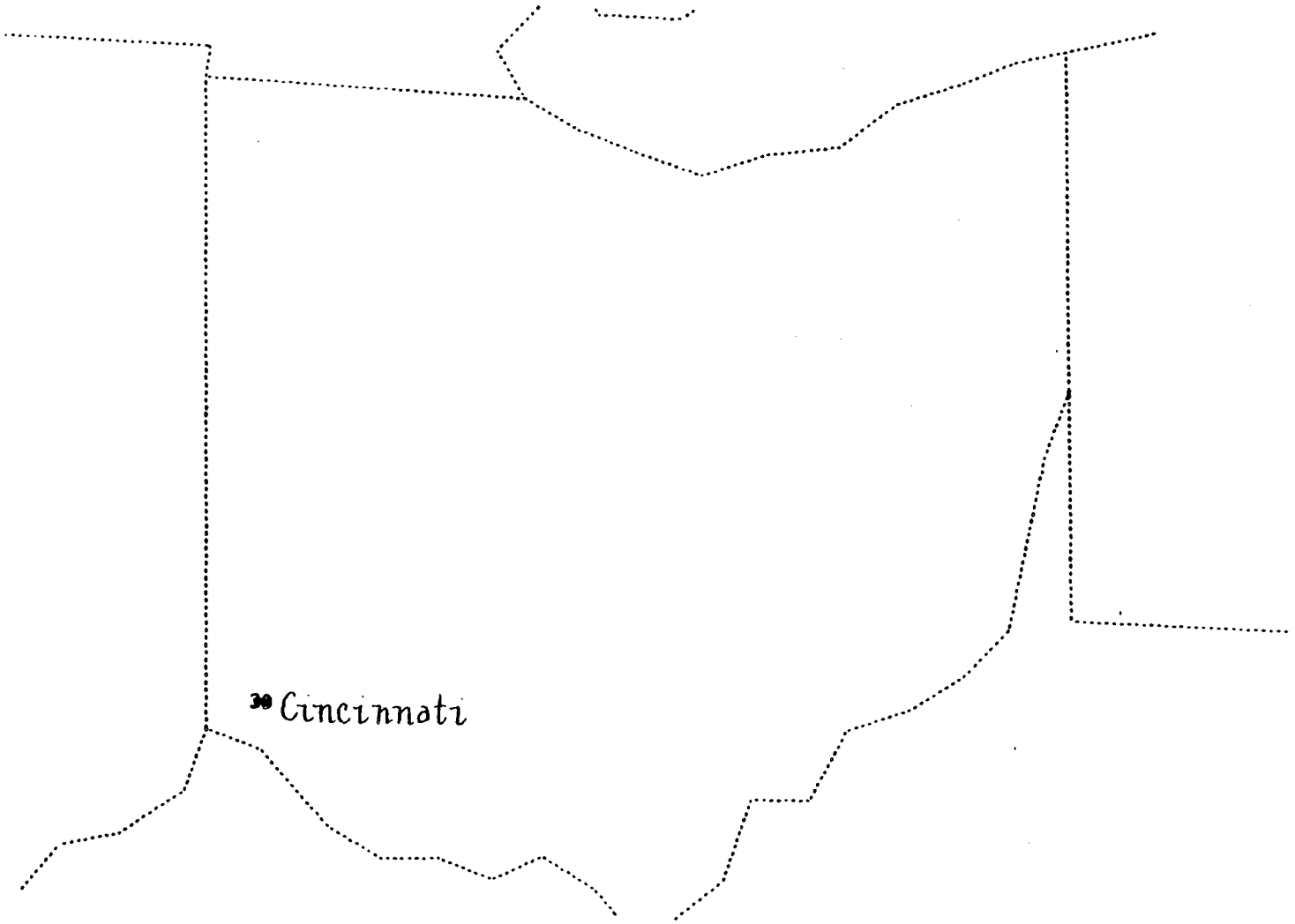


Figure 3.11. Contour Map Produced from the Two-dimensional Moving Average (d_o=2) for Sulfur Dioxide in Ohio State.
(40 by 40 grid.)

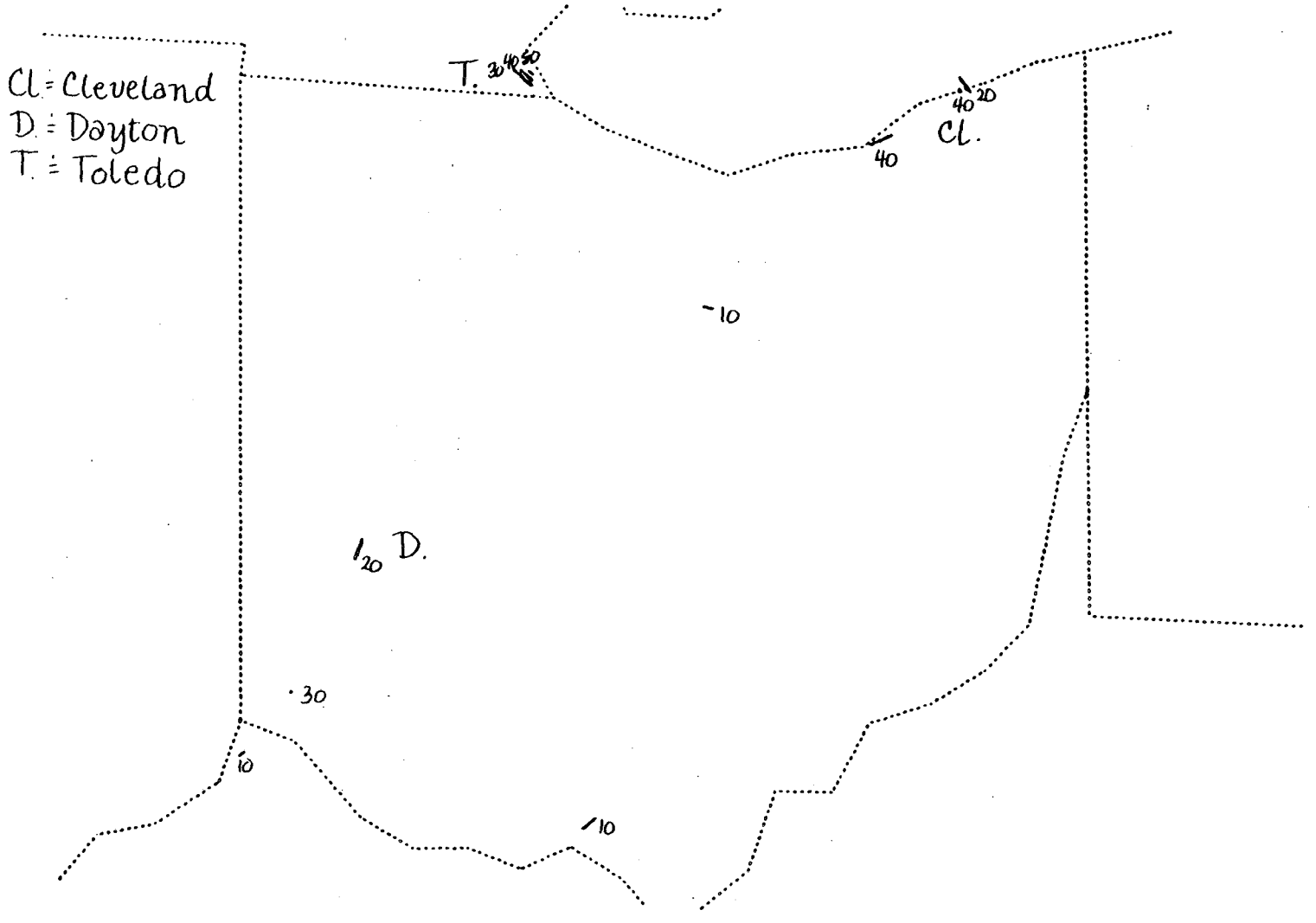


Figure 3.12. Contour Map Produced from the Two-dimensional Moving Average (d_o=5) for Sulfur Dioxide in Ohio State.
(40 by 40 grid.)

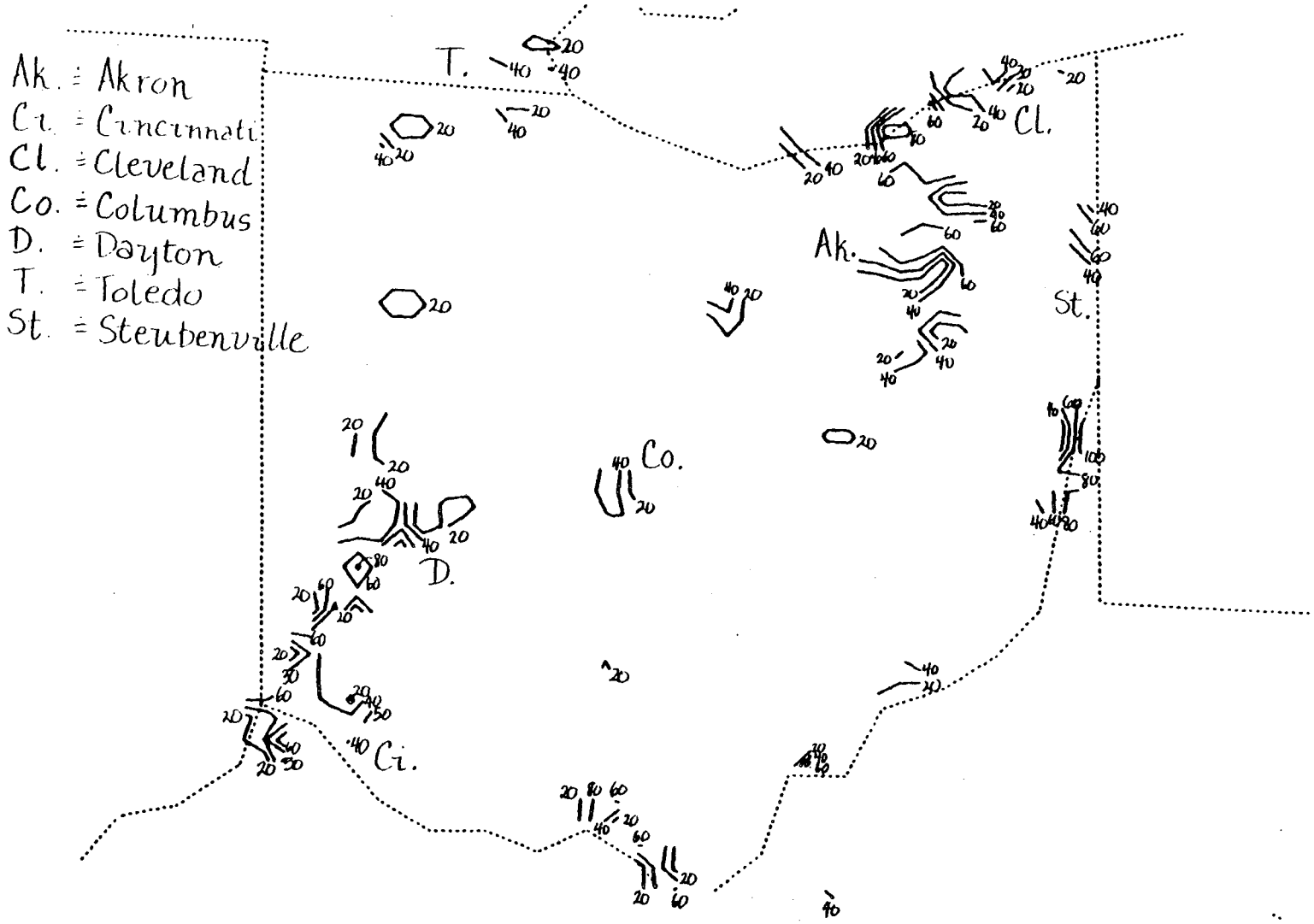


Figure 3.13. Contour Map Produced from the Two-dimensional Moving Average (d_o=10) for Sulfur Dioxide in Ohio State. (40 by 40 grid.)

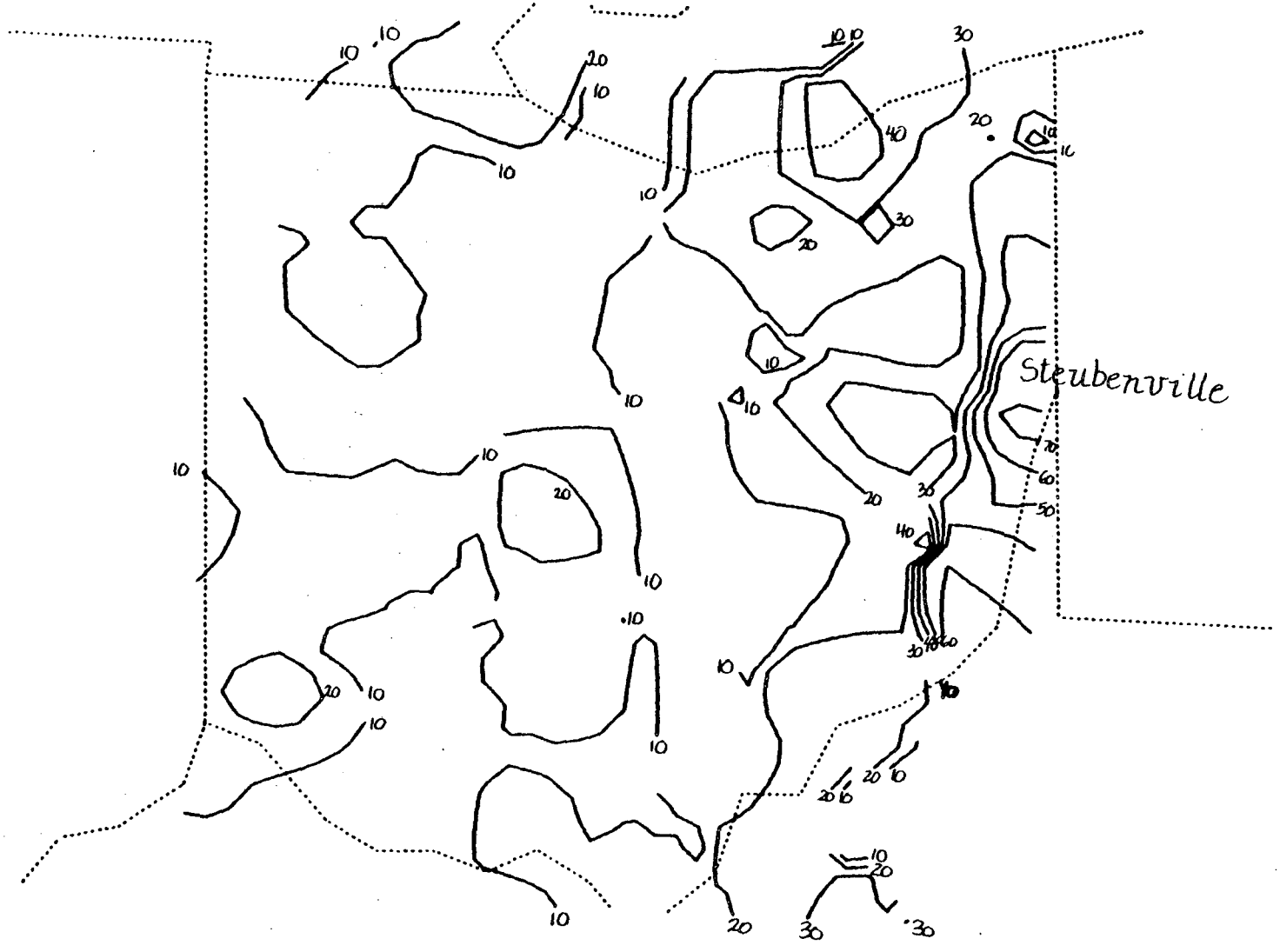


Figure 3.14. Contour Map Produced from the Two-dimensional Moving Average (d_o=20) for Sulfur Dioxide in Ohio State.
(40 by 40 grid.)

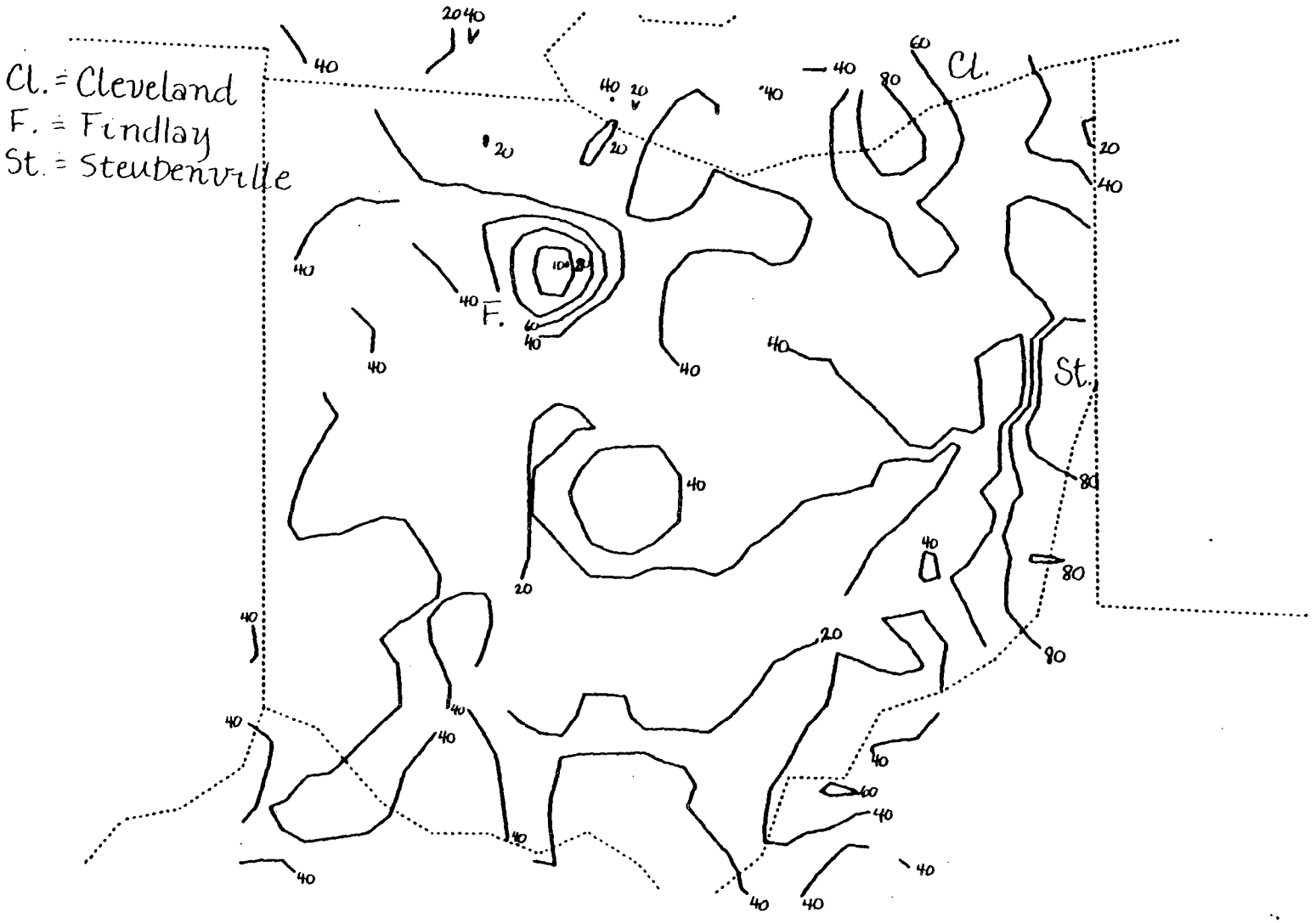


Figure 3.15. Contour Map Produced by the Two-dimensional Moving Average (d_o=50) for Sulfur Dioxide in Ohio.
(40 by 40 grid.)

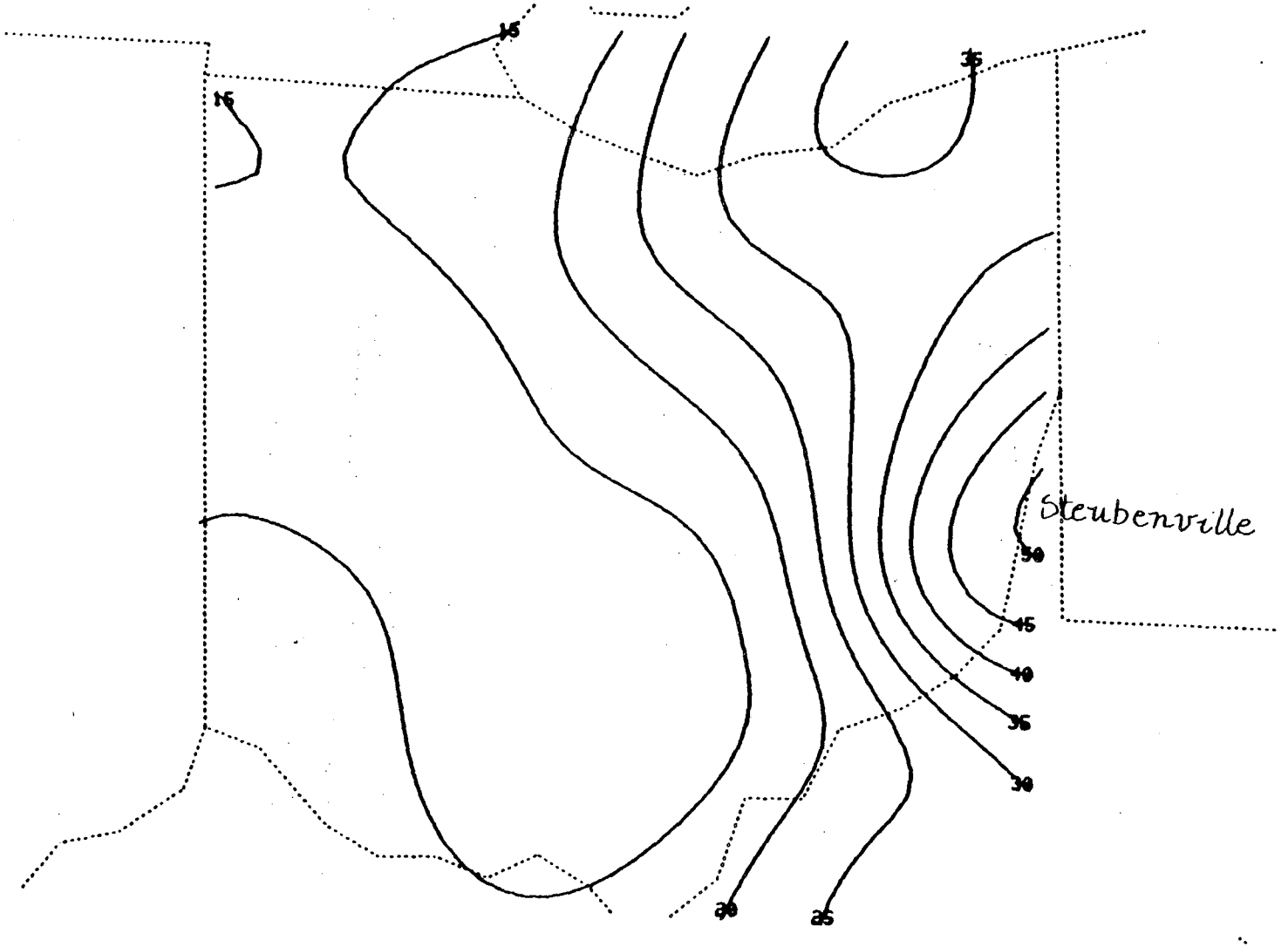
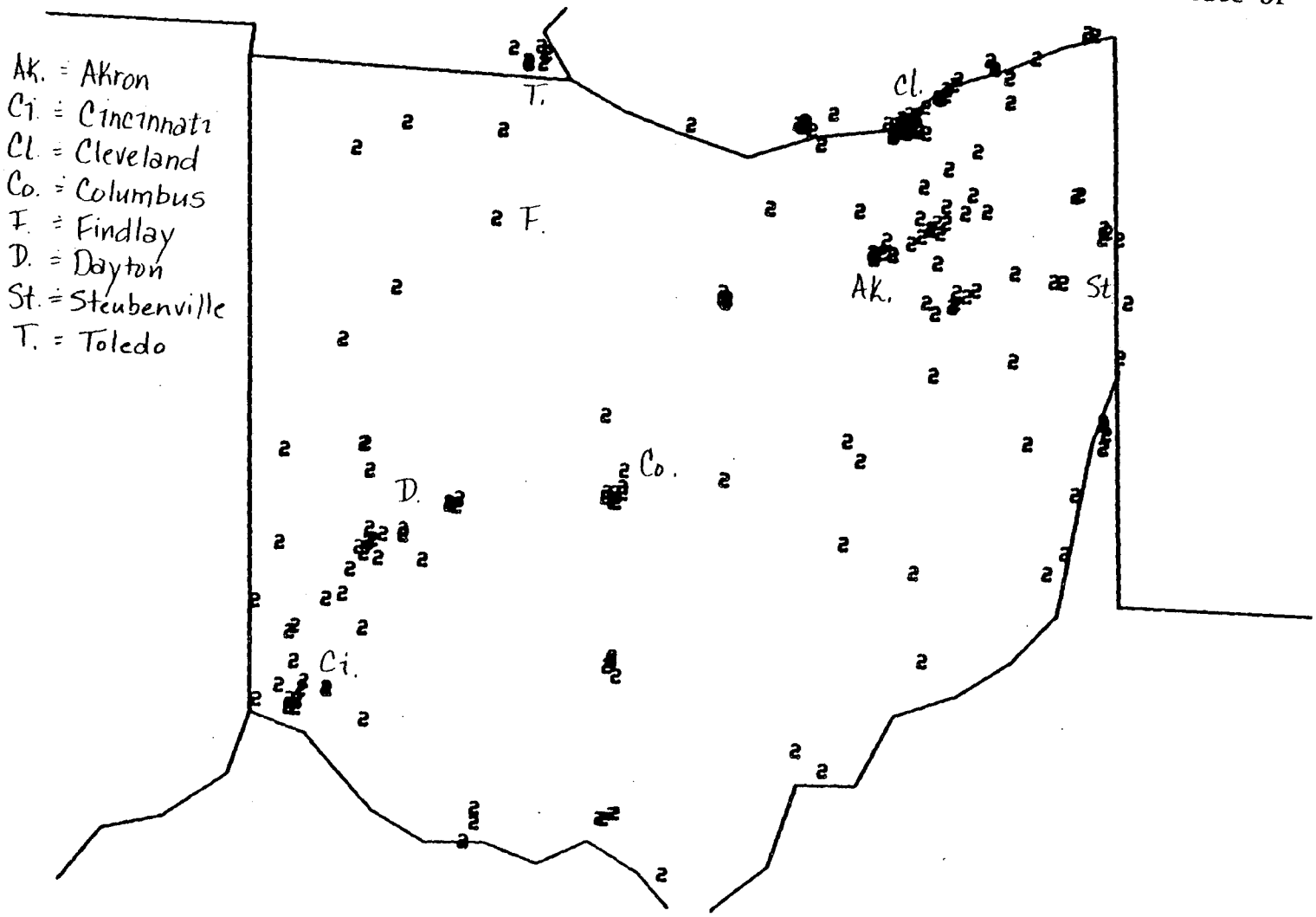


Figure 3.15a. Active Monitoring Stations Measuring Sulfur Dioxide in the State of Ohio, 1974-76.



tion. In fact only 56 points out of 1600 are estimated by the moving average with d_0 = 1 km for sulfur dioxide. Using the contouring algorithm described in System S, these points are not close enough to each other to estimate any contour lines. Only 96 out of 1600 grid points were estimated for suspended particulate with d_0 = 1 km. As d_0 increases the contours get flatter because more stations are being taken into the estimate.

In chapter 2, both cross-validation and local variability analysis show that for suspended particulate a d_0 equal to 2 km is preferred. For these contour maps, a d_0 = 2 km does does not give a very informative contour map. Several inferences could be made from this observation. One is that these data are too sparse to derive contour maps from them. Another might be that the window definition is too stringent. For d_0 = 1, 2 or 5 km, the contours are not estimable from a 40 by 40 grid whereas for d_0 = 20 or 50 km the contours are too flat to distinguish any particularly high or low concentrations. Therefore, it seems that a d_0 = 10 km gives the most information in a contour map. Sulfur dioxide contour estimation by the two-dimensional moving average model behaves very similarly. The value of d_0 which gives the most informative contour maps is around 10 or 20 km. This d_0 is greater than that which is chosen for suspended particulate probably because there are not as many active stations for sulfur dioxide and therefore we may need a larger d_0 to estimate a grid point. These maps produced by the moving average with the optimal d_0 indicate areas where there is no information. Thus if we use a larger d_0, the contours produced will have more variability.

It is helpful to compare these contour maps with the spatial distribution of stations in Ohio. Figures 3.9a and 3.15a show the active monitoring stations on the map for suspended particulate and sulfur dioxide. The variance of

the contour lines is increased for regions of no data.

Comparing Contour Maps Obtained by Two Methods of Estimation

If we want to compare the two different methods of contour map estimation that we have described, first we must choose a d_0 for weights in the moving average so that we can compare one contour map from the two-dimensional moving average estimation procedure to one from bivariate linear interpolation. Suppose we choose the d_0 which was determined by local variability analysis in the previous chapter. Since this analysis does not give an exact choice, we select a d_0 which is near the minimum. We will get the most information from a d_0 which is large enough to include a sizeable number of stations in the estimate but not so large as to substantially increase the local variability function. In chapter 2, this is described as the end of the "trough-like" region. In order to estimate a grid point by the two-dimensional moving average model there must be a station within $3 \times d_0$. The larger the d_0, the more grid points there are that are estimable with any confidence by this model. Using more stations in the estimate also improves precision. Thus d_0 should be taken as large as possible before the local variability increases rapidly.

If we look at Figure 2.13a to choose the proper d_0 for suspended particulate in Ohio, we see that the largest point we can choose before local variability increases quickly towards the maximum is at a disk radius of about 21 km which corresponds to a d_0 of approximately 5 km (Table 2.4a., chapter 2). This correspondence is achieved by dividing the disk radius by 4. Since this d_0 is not large enough for this model to adequately predict contours (See Figure 3.6),

perhaps the next best map to use for comparison of the moving average predicted contours of suspended particulate with those estimated by Akima's method is that estimated with d_0 = 10 km. This d_0 corresponds to a total disk radius of 40 km or log of radius of 3.7 which does not give a particularly high local variability (See Figure 2.13a.). Yet if we increase d_0 to obtain nice contours it is done at the cost of increasing the variability of the contours.

We choose d_0 for sulfur dioxide in the same manner. In chapter 2, Figure 2.13b, the largest disk radius we can choose before local variability approaches its maximum value is also about 21 kilometers. Again, this corresponds to a d_0 of about 5 kilometers. Unfortunately, this d_0 is also not large enough to predict a full contour map using this 40 by 40 grid (see Figure 3.12.) In the same way, the next best map to use for comparison is that predicted by the moving weighted average with d_0 = 10 kilometers. This d_0 also does not give an excessively high value for local variability.

It is evident that the two-dimensional moving average model as it is defined here is not particularly useful for generating contour maps. That part of this model which chooses d_0 seems to select a disk radius which is not large enough to predict with any confidence the grid points of a particular size grid. Even if bivariate linear interpolation is not the intuitive method to use for these data, it is able to estimate all the points within the convex hull of the data points. Perhaps in using this estimation procedure for contouring, we need to choose the grid size after we choose the d_0. For if we have a grid with a higher density we might be able to have stations within the proper disk radius so that this model can estimate with confidence a greater number of grid points. In Figure 3.16a, we have an area in the grid which is not estimable with any reliability by this moving average model because the stations are too far from the grid

Figure 16a. Grid points are too far from Stations to be 'estimated'.

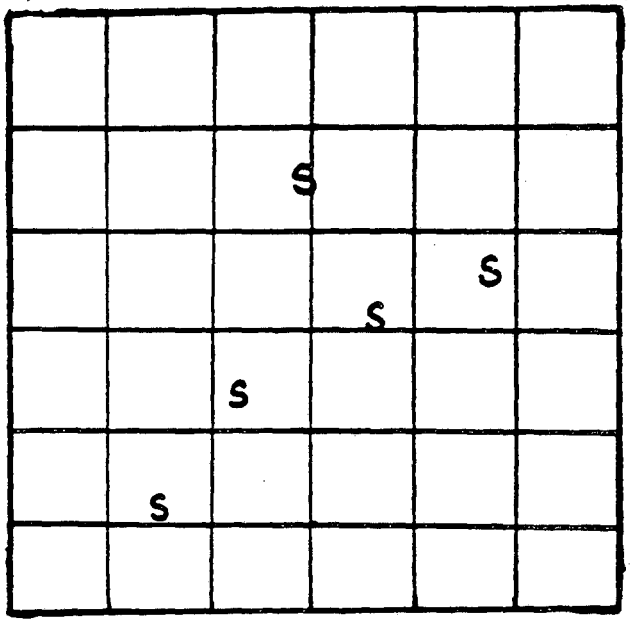
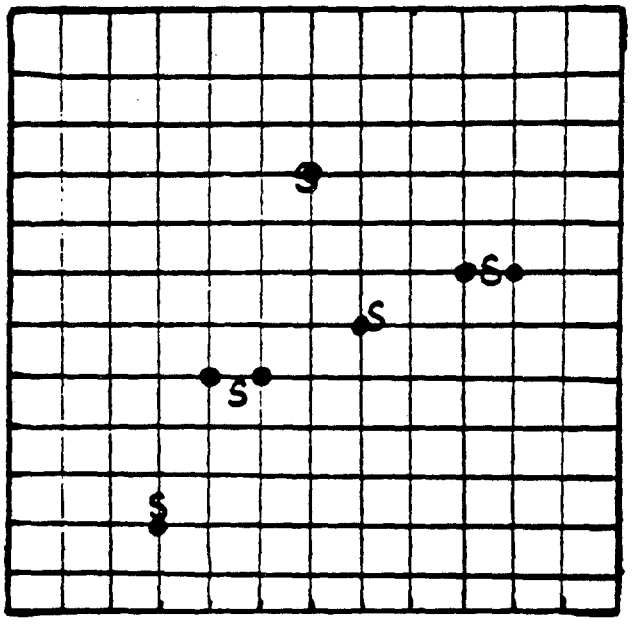


Figure 16b. By doubling the grid density in 16a. some grid points are now 'estimable'.



points. Yet, if we double the grid density as in Figure 3.16b, at least some points in this same area are now reliably estimable since the grid points are now closer to the stations.

The state of Ohio is contained within a circle with a radius equal to 500 kilometers. A 40 by 40 grid gives grid points, at most, about 12.5 kilometers apart. Because of the irregular spatial distribution of the data points the two-dimensional moving average model cannot estimate all the points in a 40 by 40 grid. Let us examine how the estimated contour reliability changes if we double the grid point density. In other words we will choose an 80 by 80 grid which will give grid points, at most, 6.25 kilometers apart.

Figures 3.17 and 3.18 show bivariate linear interpolated contours using an 80 by 80 rectangular grid. If we compare these Figures with Figures 3.2 and 3.3 respectively, where the contours were estimated by bivariate linear interpolation using a 40 by 40 grid, we see that the grid density does not change the general shape of the map very much. As expected, the 80 by 80 grid estimation does give a little more detail in the contour lines and this is the only difference in the contours produced from these two grid densities. These contours are produced by the algorithm "contour" from system S (Becker, 1981).

If we use an 80 by 80 grid to predict contours from the two-dimensional estimation procedure, rather than from bivariate linear interpolation, we will see, for small d_0, very different contour lines from those obtained with a 40 by 40 grid. For large d_0, that is d_0 equal to 20 or 50 km, the contour maps obtained from an 80 by 80 grid are exactly those that were obtained from a 40 by 40 grid. This is shown by comparing Figure 3.9 with Figure 3.19 and Figure 3.8 with Figure 3.20. Yet for smaller d_0 the maps do change drastically with the grid density. For d_0 = 10 km, contours estimated by a 40 by 40 grid (Figure 3.7) are not

Figure 3.17. Contour Map of Suspended Particulate in Ohio Obtained from Bivariate Linear Interpolation Using an 80 by 80 grid.

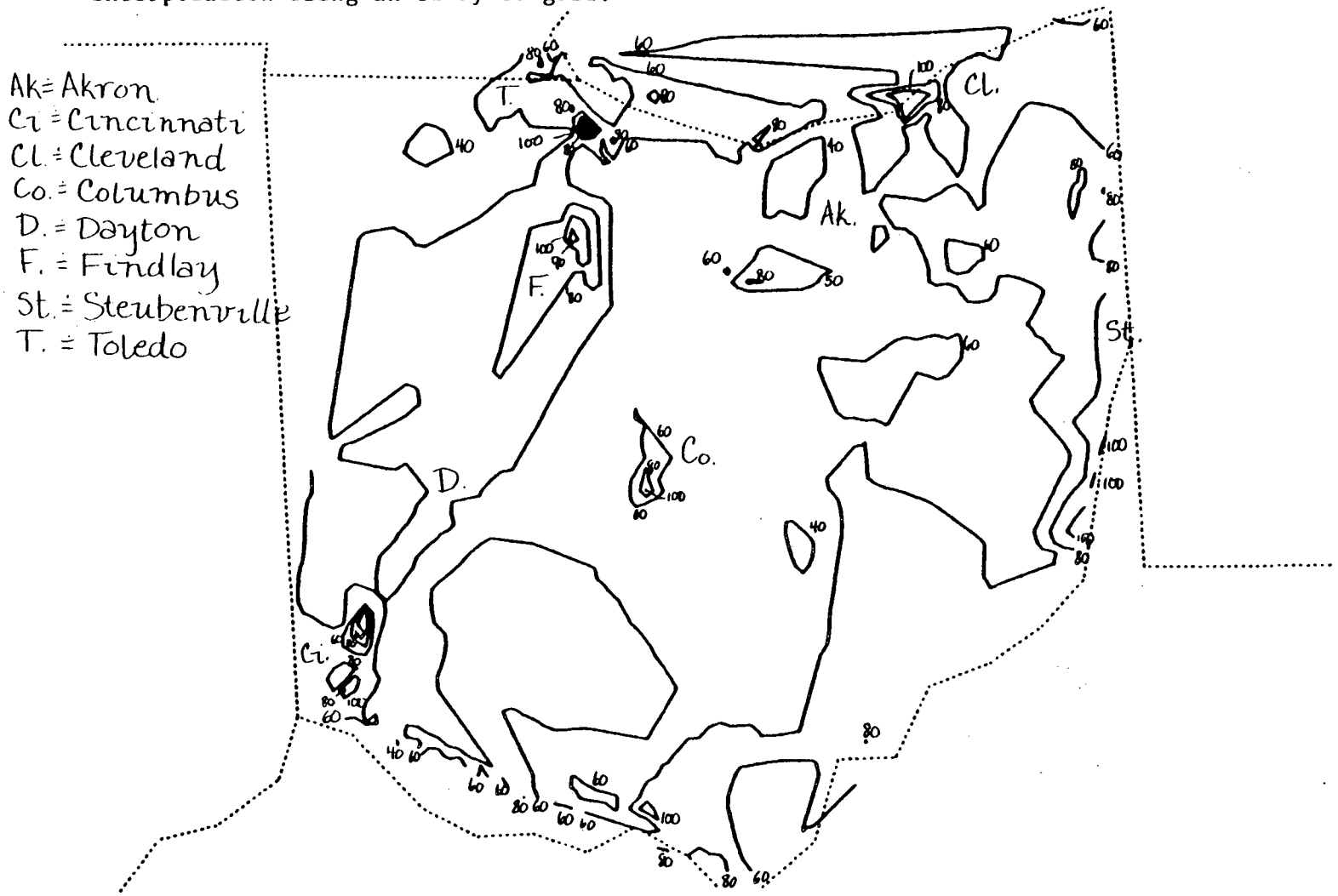


Figure 3.18. Contour Map of Sulfur Dioxide in Ohio State Obtained by Bivariate Linear Interpolation Using an 80 by 80 Grid.

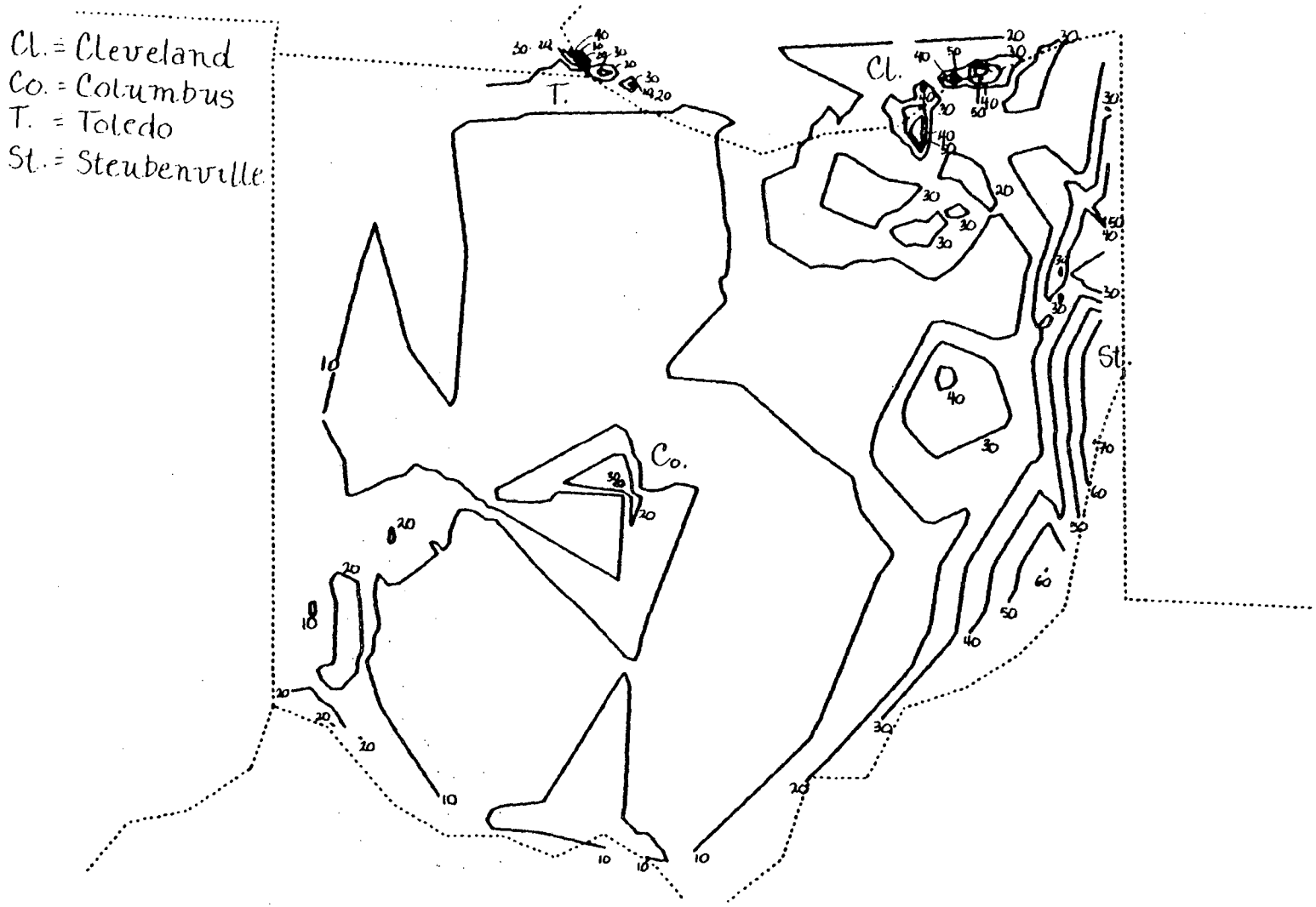


Figure 3.19. Contour Map of Suspended Particulate in Ohio State Using the Two-dimensional Moving Average with d_0=50 km for an 80 by 80 Grid.

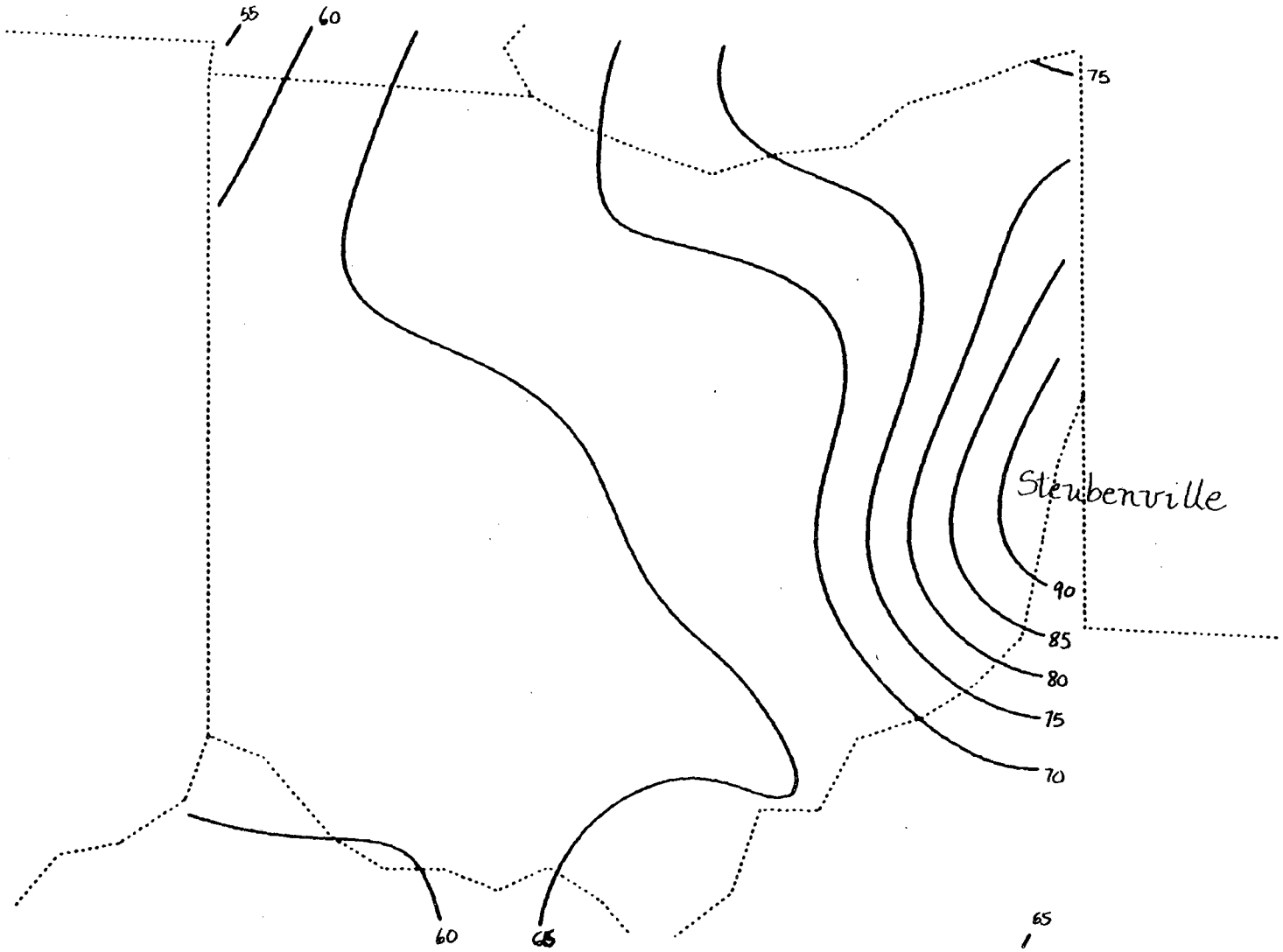


Figure 3.20. Contour Map of Suspended Particulate in Ohio State Using the Two-dimensional Moving Average with d_o=20 km for an 80 by 80 Grid.

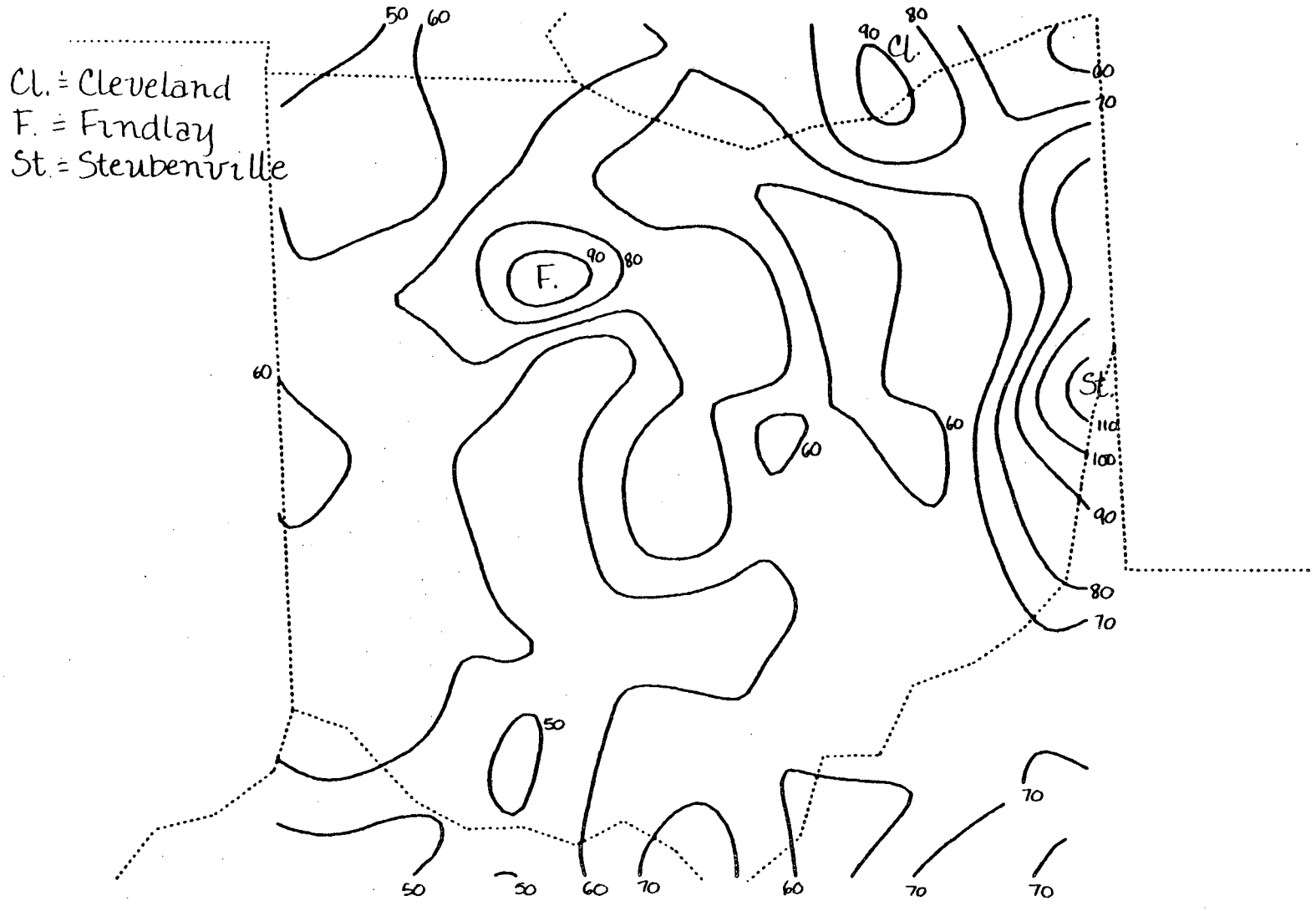


Figure 3.21. Contour Map of Suspended Particulate in Ohio Using the Two-dimensional Moving Average (d_o=10) for an 80 by 80 Grid.

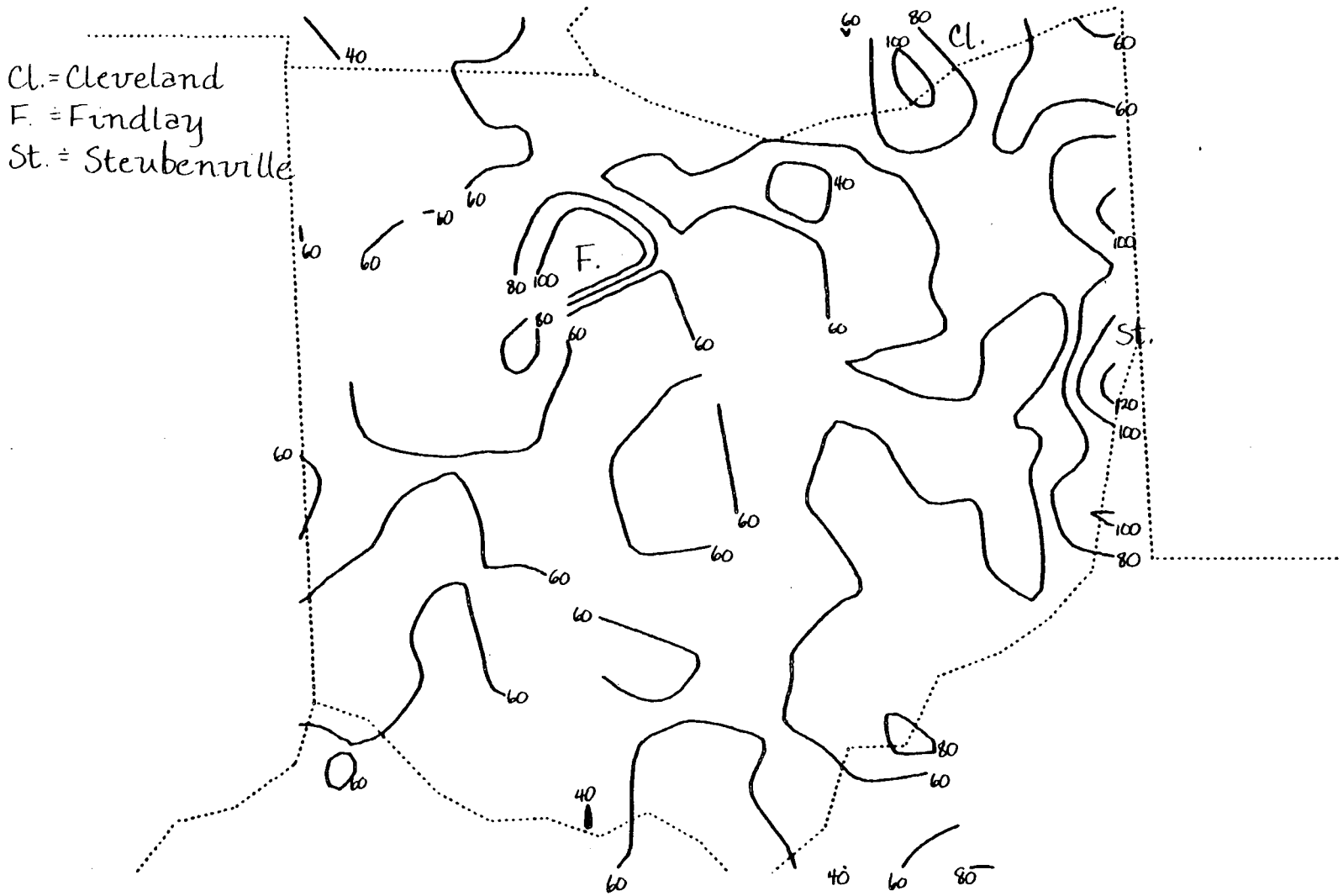


Figure 3.22. Contour Map of Suspended Particulate in Ohio State Using the Two-dimensional Moving Average (d_o=5 km) for an 80 by 80 Grid.

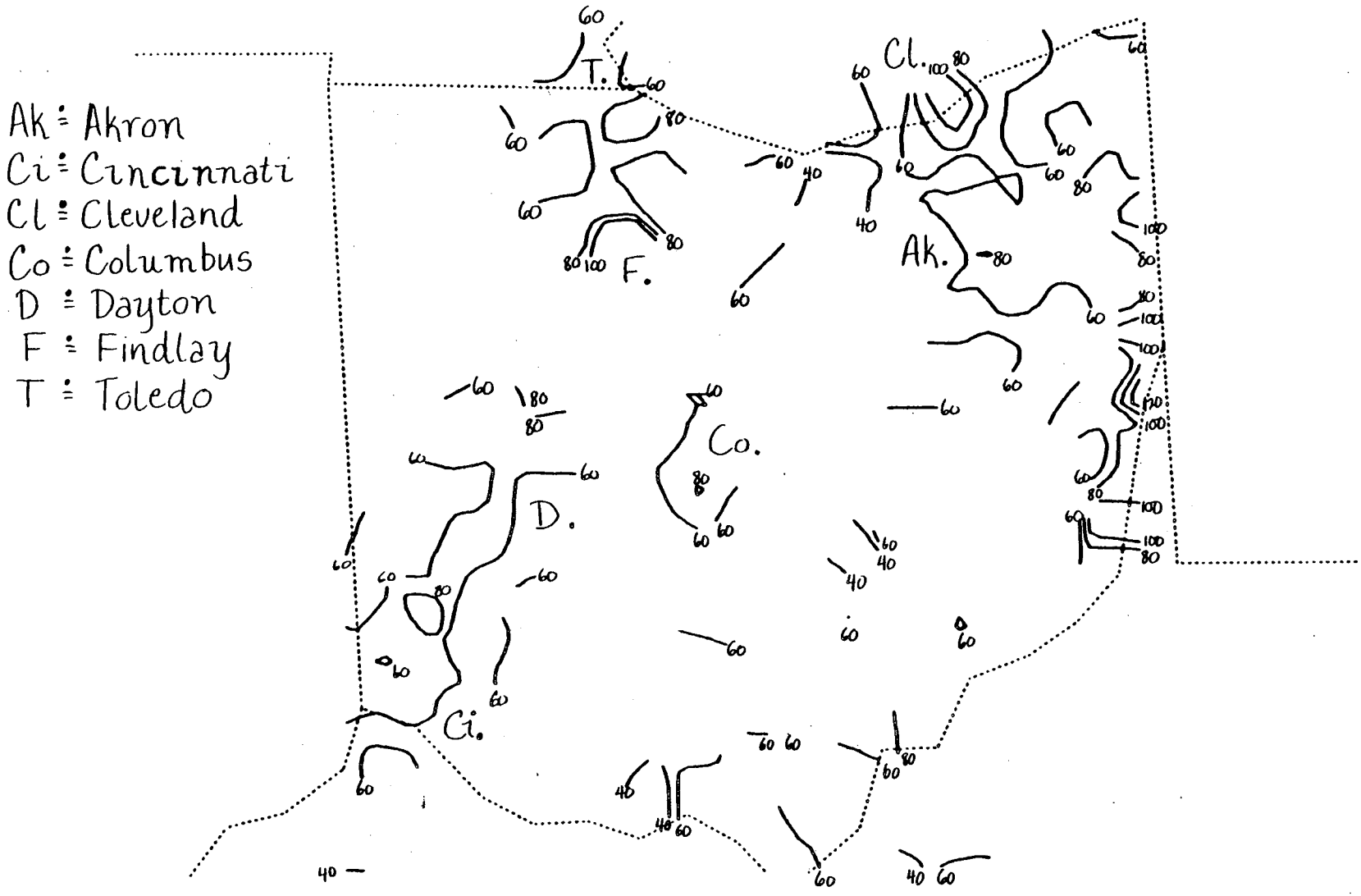
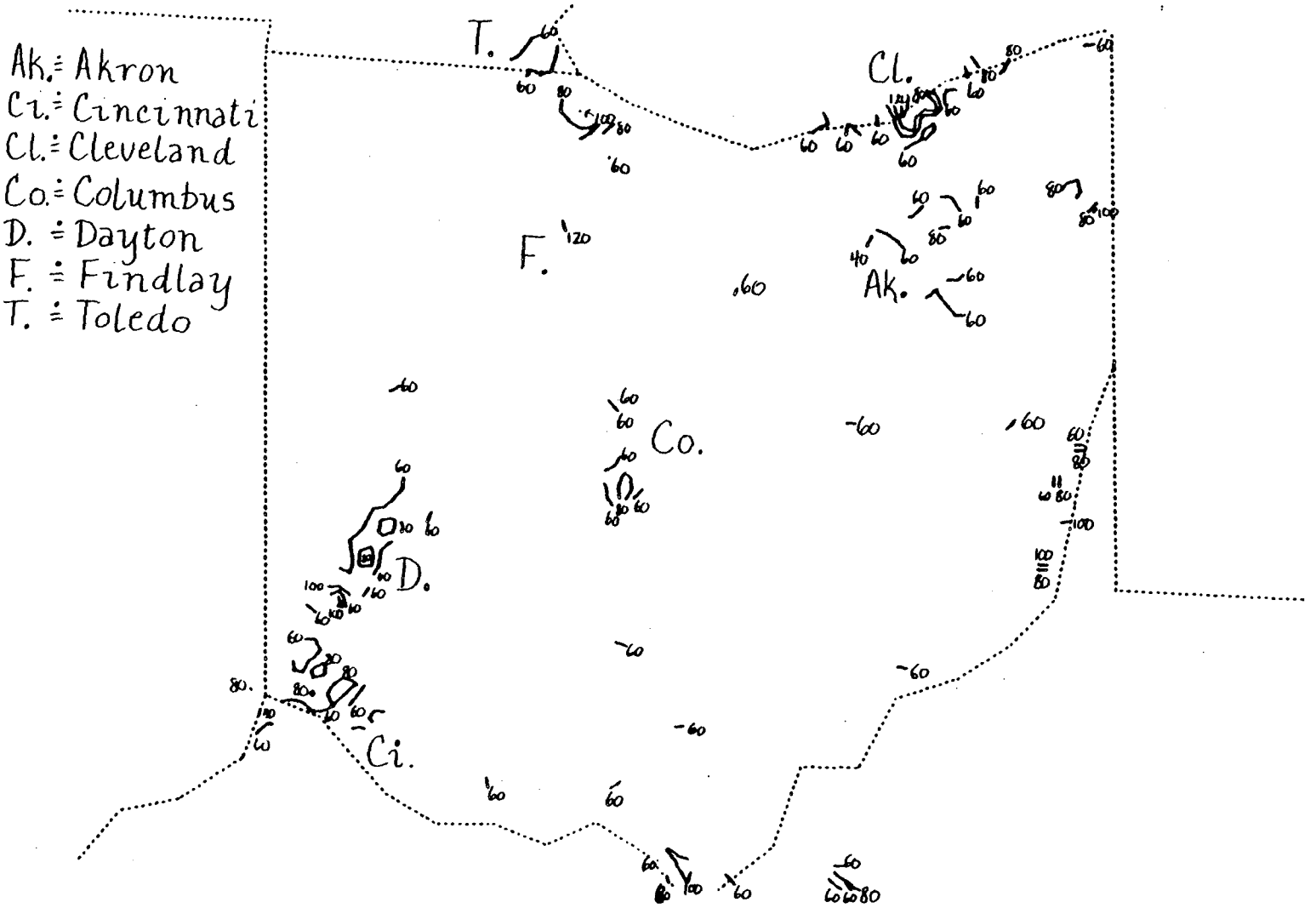


Figure 3.23. Contour Map of Suspended Particulate in Ohio Obtained from the Two-dimensional Moving Average (d_0=2 km) for an 80 by 80 Grid.



extremely different from those estimated for an 80 by 80 grid (Figure 3.21) using the same model. The only difference is that in the 80 by 80 grid estimation with d_0 = 10, contour lines are extended further and appear smoother than in the 40 by 40 grid. In Figure 3.22, contours estimated using this model with d_0 equal to 5 km and an 80 by 80 grid, more grid points are estimable than in the same estimation with a 40 by 40 grid (Figure 3.6). Comparing Figure 3.5 with 3.23 where d_0 = 2 km, more estimable points with the 80 by 80 grid are shown. Thus in using this model to produce contour lines, the estimability of these lines is dependent on the grid density used.

This is not a fundamental difference between the moving average method and bivariate linear interpolation. In order to generate reasonable contour lines, the grid must be dense enough to track detailed variability in the predictive function. In the moving average the amount of detail depends upon d_0, whereas, in bivariate linear interpolation, it does not.

Although more grid points are estimable in an 80 by 80 grid, the contour maps with d_0 equal to 2 or 5 km are still not very informative. Looking at the contour map with a grid density of 6400 points for the two-dimensional moving average model with d_0 = 5 km, which was chosen by LV (local variability) analysis, we see that many points are estimated but the contour lines are not complete (Figure 3.22). This is a expected when using this model for contour analysis. For small d_0, this moving average is a discontinuous surface because of the way it is conditioned. A point cannot be estimated if there are no active monitoring stations within $3 \times d_0$. Now, if d_0 is large enough or if we had a higher station density this situation will never arise. But for small d_0 and the station density as it is, inestimability is likely to arise. In the ignorance of topographical and meteorological factors, the $3 \times d_0$ restriction in this model is sensible in

the sense that it prevents the two-dimensional moving average from producing meaningless estimates.

Comparison of Point Prediction from Two Methods of Geographic Estimation

In the previous section, it was concluded that for estimating contours, bivariate linear interpolation does a more complete job possibly at the cost of contour variability than the two-dimensional moving average estimation procedure. In this section a comparison of these two methods of surface estimation in terms of their ability to correctly predict values of specific points is attempted. Cross-validation is done for each method. That is, using each station's location and omitting the observed value at that station, a concentration is predicted by the other n-1 stations. Bivariate linear interpolation and the two-dimensional moving average model are used to obtain two estimates of the concentration. Assuming that the observed values are true values, we obtain an assessment of each method in estimating the true concentrations.

Comparing techniques by cross-validation when one method involves cross-validation as an inherent part requires an inordinate number of calculations and is beyond the scope of this dissertation. As Stone (1974) points out in his article on cross-validatory choice and assessment of statistical predictions, this comparison would require a "two-deep" cross-validation analysis. For example, let us consider the two-dimensional moving average estimation procedure that is presented here. If we were to make the two-dimensional moving average model estimate of each station's concentration, we should omit that station and,

then taking the n-1 points which remain, run the model. As a part of this procedure we need to choose a value of d_0, by cross-validating the remaining n-1 points. In other words we would be using n-2 points to predict each of the n-1 points in our choice of d_0.

It is difficult to clearly see whether the choice of d_0 would change in the moving average from station to station in the "two-deep" cross-validatory analysis. Since local variability analysis is given as an alternative to the cross-validatory choice, let us examine how this function may change in a technique assessment situation. One question which arises is that of the need to calculate a local variability function for each station from n-1 stations leaving the station of interest out. This needs to be done because we are comparing two methods, one of which requires LV (local variability) analysis. Leaving one station out probably does not change the shape or the trough area of the LV curve drastically Since it is these aspects of the curve which are important in choosing d_0 we should not have to recompute the LV curve for each station. Figure 3.24 shows five different LV curves for Ohio. Each curve corresponds to leaving a different station out. These stations were chosen at random from a random-digit table (Hoel, Port and Stone, 1971). The stations are arranged in alphabetical order by location and these random numbers were used to choose the five stations by using them as the order of the stations. The numbers chosen were 039, 385, 175, 326, and 001. This was done so that a random number of stations were chosen rather than just the first five. The value for the concentration for these numbers is 56.9897, 67.9664, 45.8159, 167.003, 61.6569 micrograms per cubic meter. The range of concentration in micrograms per cubic meter for Ohio State is from 26.1539 to 194.6105. The five numbers chosen seem to be a fair sampling from this range. They are not all in the center.

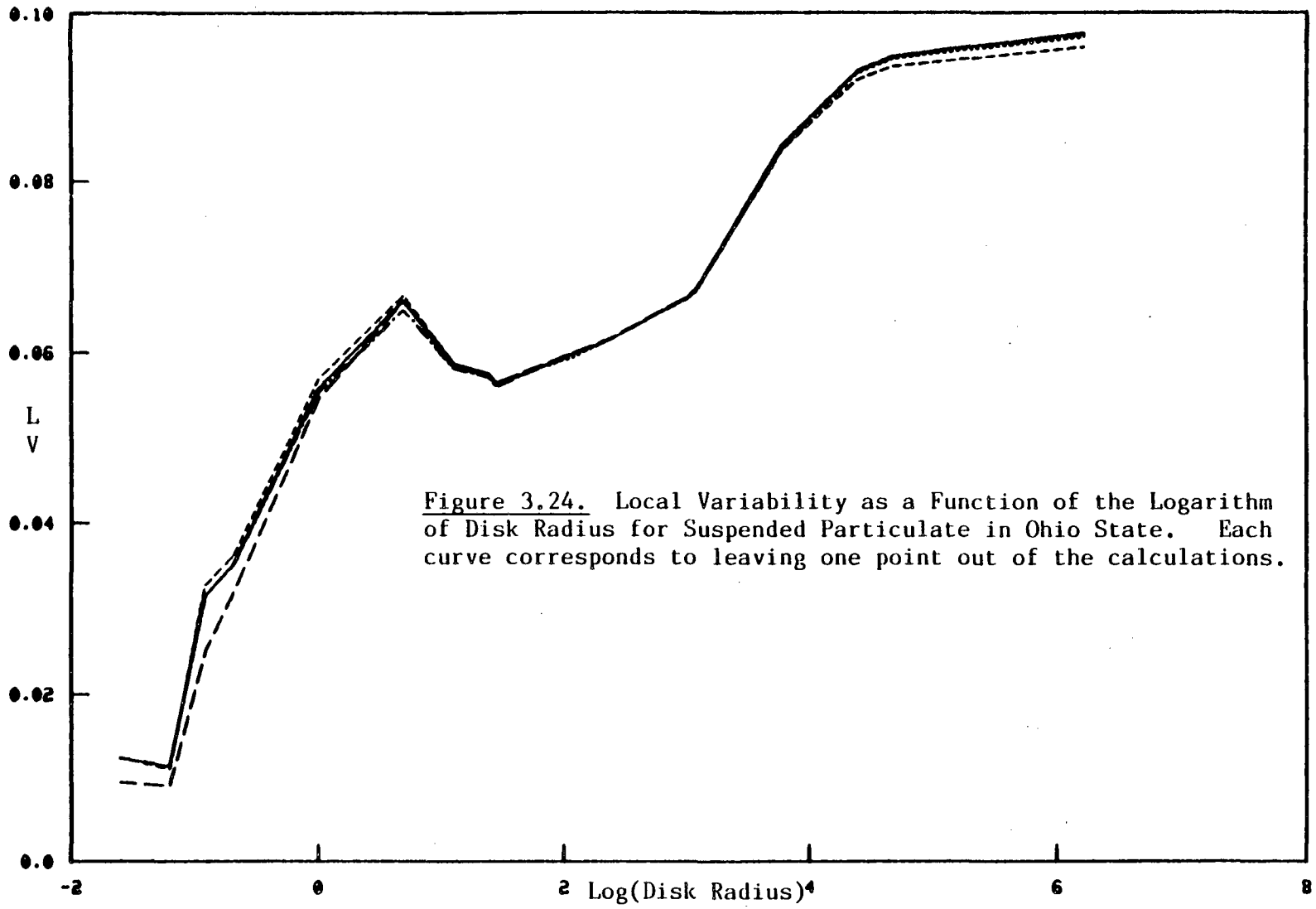


Figure 3.24. Local Variability as a Function of the Logarithm of Disk Radius for Suspended Particulate in Ohio State. Each curve corresponds to leaving one point out of the calculations.

Figure 3.24 shows that the shapes of the LV curves are very much the same. All these curves have the same area of degeneracy and the same trough-like area as the curve for LV vs. logarithm of the disk radius for all the stations. This function is shown in Chapter 2, Figure 2.13a. Therefore for each of the five stations we would choose the same area of degeneracy and the same d_0 based on the same minimum disk radius. In light of this example, it seems clear that the shape of the LV function for a particular area is fairly insensitive to the omission of a station. Therefore for comparing the two techniques, the minimum disk radius given by the original LV curve with no stations omitted was used for the estimation of each station's concentration. This leave-one-out method is a way to understand the variability of the local variability function and thus the variability associated with the chosen d_0. This is very similar to the Jackknife method for estimating the variance of an estimate. Bootstrapping may be another method for estimating the variance of the LV curve. Instead of leaving one out, the bootstrap would resample the data with replacement and find the LV curve for each sample.

Different points that are left-out in the above analysis affect the LV curve in various ways. The disk radius at which the left-out point would be included depends on the distance to the nearest station. It is at this disk radius that the LV curve is affected by the left out point. The relative size of the concentration of the left out point affects the size of the change in the local variability.

Using a d_0 of 2 kilometers for suspended particulate in Ohio state, we have calculated the two-dimensional moving average model for each data point, leaving the point itself out of the calculation. Bivariate linear interpolation of the same data points was also done. For comparing adequacy of estimation, the

four functions described in chapter 2 were used. These functions are weighted by the percentage of time active for predicted station (See Chapter 2). To remind the reader of their construction they are given below.

(WLF) Weighted Loss Function

$$WLF = \frac{\sum_{i=1}^{n} p_i (z_i - e_i)^2}{\sum_{i=1}^{n} p_i}$$

(RWSE) Ratio of the Weighted Squared Errors

$$RWSE = \frac{1}{\sum_{i=1}^{n} p_i \; S_{zw}^2} \sum_{i=1}^{n} p_i (z_i - e_i)^2$$

(WMAE) Weighted Mean Absolute Error, Percent

$$WMAE = \frac{1}{\bar{z} \sum_{i=1}^{n} p_i} \sum_{i=1}^{n} p_i \left| z_i - e_i \right| \times 100$$

(WMPE) Weighted Mean Percent Error

$$WMPE = \frac{\sum_{i=1}^{n} p_i \frac{\left| z_i - e_i \right|}{z_i}}{\sum_{i=1}^{n} p_i} \times 100$$

Table 3.1 and 3.2 show the results from the cross-validation of each method for suspended particulate and sulfur dioxide respectively. For suspended particulate stations, 302 out of 398 were estimable for cross-validating the moving average model with d_0 = 2. This means that 302 stations had at least one station

within 6 km. Similarly, the number of stations estimable by the moving average model with d_0 = 5 for sulfur dioxide was 155 out of 185.

TABLE 3.1. Loss Functions for Cross-Validation of the Moving Average Model (d_0 = 2 km) and Bivariate Linear Interpolation for Suspended Particulate in Ohio State.

Function	The Model	BLI
WLF	.0519	.0537
RWSE	.404	.515
WMAE	4.2%	4.03%
WMPE	4.1%	3.99%

TABLE 3.2. Loss Functions for Cross-Validation of the Moving Average Model (d_0 = 5 km) and Bivariate Linear Interpolation (BLI) for Sulfur Dioxide in Ohio State.

Function	The Model	BLI
WLF	1.63	.2959
RWSE	2.82	.566
WMAE	28.9%	12.0%
WMPE	27.1%	12.0%

For suspended particulate in Ohio State, the squared error loss and the ratio of weighted squared errors are smaller for the two-dimensional moving average model than they are for Akima's method. The mean absolute error and the mean percent error are very close although those for BLI are slightly smaller. The greatest difference seems to be in the ratio of weighted squared errors. In the next section the plots show that for suspended particulate the estimates produced by both methods are unbiased. Thus the difference in RWSE for the two methods shows that the moving average model explains about 10 percent more of the variance than does prediction by BLI. Therefore the two-dimensional moving average estimation procedure is possibly a "better" method to use than BLI for point prediction of suspended particulate in Ohio.

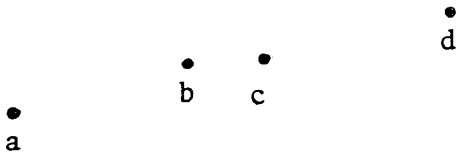
In Table 3.2, the results are shown for sulfur dioxide point prediction. For all loss functions, BLI is far superior to the two-dimensional moving average estimation procedure. Squared error loss is reduced by one fifth. Absolute and percent errors are cut in half by using BLI instead of this moving average model. The ratio of weighted squared errors is larger than 1.0 in the moving average model. In the next section it is shown that the cross-validation moving average estimates are biased for sulfur dioxide. That is, they consistently underestimate the observed value. The bias is estimated at about -1.06 km. The bias was subtracted from each estimate and the loss functions were recalculated. Table 3.3 gives the results of the moving average estimate of sulfur dioxide with biased removed compared to BLI. There is no apparent bias in bivariate linear interpolation of sulfur dioxide. Table 3.3 shows that bivariate linear interpolation still prevails over the two-dimensional moving average model for sulfur dioxide point estimation.

TABLE 3.2a. Loss Functions for Cross-Validation of the Moving Average Model, (d_0 = 5 km), Minus the Bias and Bivariate Linear Interpolation (BLI) for Sulfur Dioxide in Ohio State.

Function	The Model	BLI
WLF	.951	.2959
RWSE	1.86	.566
WMAE	25.5%	12.0%
WMPE	23.6%	12.0%

The Occurrence of Bias In Cross-Validation Estimates

Bias can occur in cross-validation estimates using a moving average method. Let us consider the following four points as monitoring stations.



We will cross-validate using a two-dimensional moving average similar to the model described. Suppose that point a has a very high concentration of 50 $\mu g / m^3$, and points b, c and d reported low concentrations, all equal to 1 μ / m^3. Now let us predict each point by cross-validation.

1. Prediction of concentration at station a, leaving the observed value at a out of the calculation, involves points b an c since the window around a contains a, b and c. The weights applied are .6 and .4 to b and c respectively. Point b has a slightly higher weight than point c because it is slightly closer to point a. Then,

Cross-Validation Estimate of a = .6(1) + .4(1) = 1 $\mu g / m^3$.

Therefore,

Residual of Cross-Validation Estimate = 1 - 50 = -49 $\mu g / m^3$.

The residual is defined as the estimate minus the observed.

2. Predicting the concentration at point b, leaving b's own observed value out involves a , c and d. Point a has weight equal .01 since it is relatively far away. Point c has weight .5 and point d has weight .49. Thus,

Cross-Validaion Estimate of b = .01(50) + .5(1) + .49(1) = 1.04 $\mu g / m^3$.

And,

Residual of Cross-Validation Estimate = 1.04 - 1.00 = .04 $\mu g / m^3$.

3. Predicting point c by cross-validation involves points a , b and d. The weight assigned to point a for estimating c is .008. Point b is given weight equal to .492 and point d is given weight .5. Therefore,

Cross-Validation Estimate of c = .008(50) + .492(1) +.5(1) = 1.03$\mu g / m^3$.

And,

Residual of Cross-Validation Estimate of c = 1.03 - 1.00 = .03 $\mu g / m^3$.

4. Prediction of point d is rather easy since it involves points b and c. Both these points have concentrations = 1 $\mu g / m^3$. Therefore the estimate of d's concentration by cross-validation is 1 which has a residual of zero.

The sum of the residuals in this example is -48.3 which gives an average of -12.1, similar to what is shown with the cross-validation estimates of sulfur dioxide is large and negative (Figure 3.28 and 3.31). This example exaggerates the effect that is seen in Ohio. The average bias for Sulfur Dioxide in Ohio is estimated to be -1.06. Also the simple average of these four points does better in this example than the cross-validation estimates. For example, the mean value is 13.25, which gives the following four residuals: -36.75, 12.25, 12.25, and

12.25. Naturally, the sum of the residuals is zero since the mean is unbiased. The squared error loss for the simple average is 1800, whereas for the weighting scheme with cross-validation, the squared error is about 2401. Subtracting the bias, as calculated, from squared error, we obtain 2257 for the variance of the cross-validatory estimates. This is still larger than using the sample mean of all the data. The crucial difference between the sample mean's squared error and the cross-validation squared error is the fact that the sample mean contains all the data while cross-validation leaves the point out that we are trying to estimate. This plays a large part in creating the bias that we see here. This probably happens because of an occasional high concentration reported in an isolated area or where most concentrations are low. The data are sparse so it is likely that an occasional high value will not be near many other points. Perhaps this does not happen with BLI because points are weighted equally, no matter how far they are from the estimated point. This is an interesting phenomena that occurs when cross-validating some estimation procedures with limited data sets. It is also interesting to note that the average of all four points does better in terms of squared error loss in this example.

The "Better" Method

Obviously, it is impossible to recommend a method of surface estimation which works "well" in all circumstances. Therefore it is imperative that we be very cautious in evaluating a technique before employing it. This chapter has shown methods for the evaluation of a technique in comparison with another method. We have compared the two-dimensional moving average estimation

procedure and Akima's bivariate linear interpolation in terms of their ability to estimate contours and predict values at specific points. Although Akima's method of bivariate linear interpolation seems questionable because it may give too much weight to outliers with the predicted surface going through every data point, it seems to do almost as "well" and sometimes better than this moving average model.

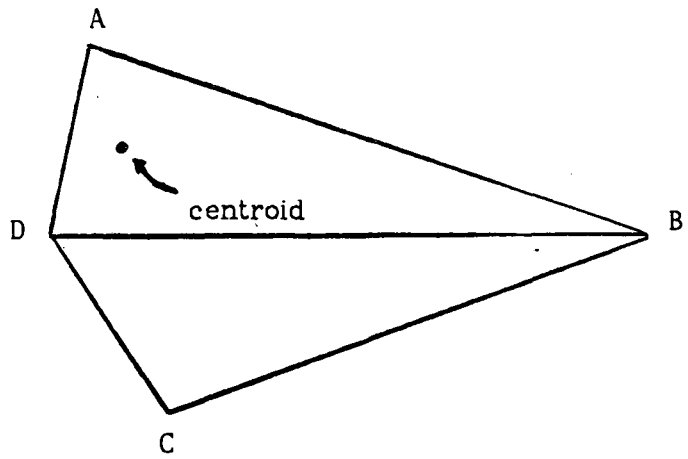
Regarding contour map estimation, because of the missing regions in the surface estimated by the two-dimensional moving average, contour lines are incompletely predicted. Although Akima's surface is not differentiable everywhere, it is continuous within the convex hull of the data points and therefore the estimated contour lines using this method are complete. Confusion with the moving average model lies in the way d_0 is chosen. It is chosen on the basis of point prediction and the best d_0 for predicting points may not be the best d_0 for estimating contour lines. The best d_0 for contouring relies heavily on the sparseness of the data and the window, which in our construction of the moving average model is $4 \times d_0$. From this analysis one might immediately conclude that, for contouring, a method which estimates a continuous surface like Akima's, is "better" than something censored like the model presented in the previous chapter. Statistically speaking, contour lines drawn in areas of no information are very unreliable and there is question as to whether they are worth considering as valid estimates. It is then the comparison of these two techniques which is useful. From the contour maps of both methods we can get a feeling for the reliability of these estimated contour lines. We can look at the Akima map to get a feeling for the high and low spots. Then by examining the map estimated by the two-dimensional moving average we can see that if we

truly believe in the $4 \times d_0$ [1] cut-off, that some of the Akima contour lines are drawn right through areas of no data. The moving average model model is "better" than bivariate linear interpolation in the sense that it shows the areas where there is little or no information.

Point prediction is really what the two-dimensional moving average estimation procedure was designed to do. It was proposed by Selvin, *et. al.* (1981) as a method for estimating the air quality of one point, the population centroid of a county or any geographic area. This estimate was used as a value for ecological regression analysis of air quality on mortality. If we were to use Akima's method to estimate this point it would be similar to a nearest neighbor approach but not exactly. We would need to draw the Delaunay triangulation of several points in the area of interest, find the triangle containing the population centroid, then estimate a plane using the vertices of that triangle, and find the value of the population centroid in that plane. This is almost like taking the three points nearest to the centroid except the Delaunay Triangulation may not yield the triangle with the three nearest points in all cases. For example, in the following Figure, the 3 closes points are ACD; yet the triangulation uses ABD.

[1] Since stations farther than about $3 \times d_0$ have negligible weights in the model, and since model estimates are suppressed for points further than $3 \times d_0$ from any active station, one might argue that the window is actually of $3 \times d_0$. Nevertheless, the choice of $3 \times d_0$ or $4 \times d_0$ does not change the conclusions of this section.

Figure 3.25. The Vertices of the Triangle Containing the Centroid are not the Closest Vertices to the Centroid.



It may be that for point prediction, taking the three nearest point rather than the Delaunay triangle does better although these are probably the same in most cases. At any rate, in comparing the moving average with Akima's method for point prediction, there were very different results for the two pollutants used. For suspended particulate, the two-dimensional moving average does a little "better" than the interpolation technique since the ratio of the weighted squared error is smaller. The values of the loss functions for the moving average model are consistently and substantially smaller than those for BLI. Yet for sulfur dioxide, interpolation does "better" than the moving average procedure for point prediction because of the consistently and substantially smaller values of the loss functions for BLI.

We have compared the two methods using several summary measurements including loss functions, percent variance, absolute error and absolute relative error, all weighted by the percentage of time active. These are the measures that were used to determine, on the average, which method is "better". Emphasis should be given to the fact that these measures are "on the

average" and often we are interested in predicting only one specific point. Figures 3.26 thru 3.33 show the cross-validation estimates compared with the observed values on a point by point basis. If we contrast these plots by method for sulfur dioxide (Figures 3.28 and 3.29), it is clear that bivariate linear interpolation predicts points "well" much more consistently than the two dimensional moving average model; while with suspended particulate, the scatterplots, Figure 3.26, for the moving average, and 3.27, for Bivariate Linear Interpolation, of estimate versus observed values are not very different. All these plots have a definite positive trend except the plot of the moving average estimate versus the observed for sulfur dioxide concentrations in Figure 3.28. This plot has a slight positive trend but also shows consistent underestimation of sulfur dioxide concentration by this method. The average bias was estimated to be about -1.06 km. Occurrence of bias when cross-validating a limited data set like this one is explained in the previous section.

The residual is defined as the estimate minus the observed value. The residual is plotted as a function of the observed for all methods with each pollutant in Figures 3.30 thru 3.33. All the residual plots, except the moving average for sulfur dioxide are roughly centered at zero. This is because of the bias which has occurred in the moving average estimate of sulfur dioxide. Bias has not occurred for any of the other method-pollutant combinations. Bias probably does not occur with bivariate linear interpolation for sulfur dioxide because the stations are weighted equally. These plots confirm and show, in more detail, what the summary measures in the previous section suggest.

The summary measures of fit are sensitive to outliers in the data for we are comparing our estimates to observed values. These observed values may not be close to the true value because of statistical variation or simply because

they are outliers. Therefore we must be cautious in the interpretation of these kind of analyses and the conclusions which are drawn. Drawing scatterplots like those shown in Figures 3.26 thru 3.33 are useful for examining the estimation procedure in more detail than just one summary statistic.

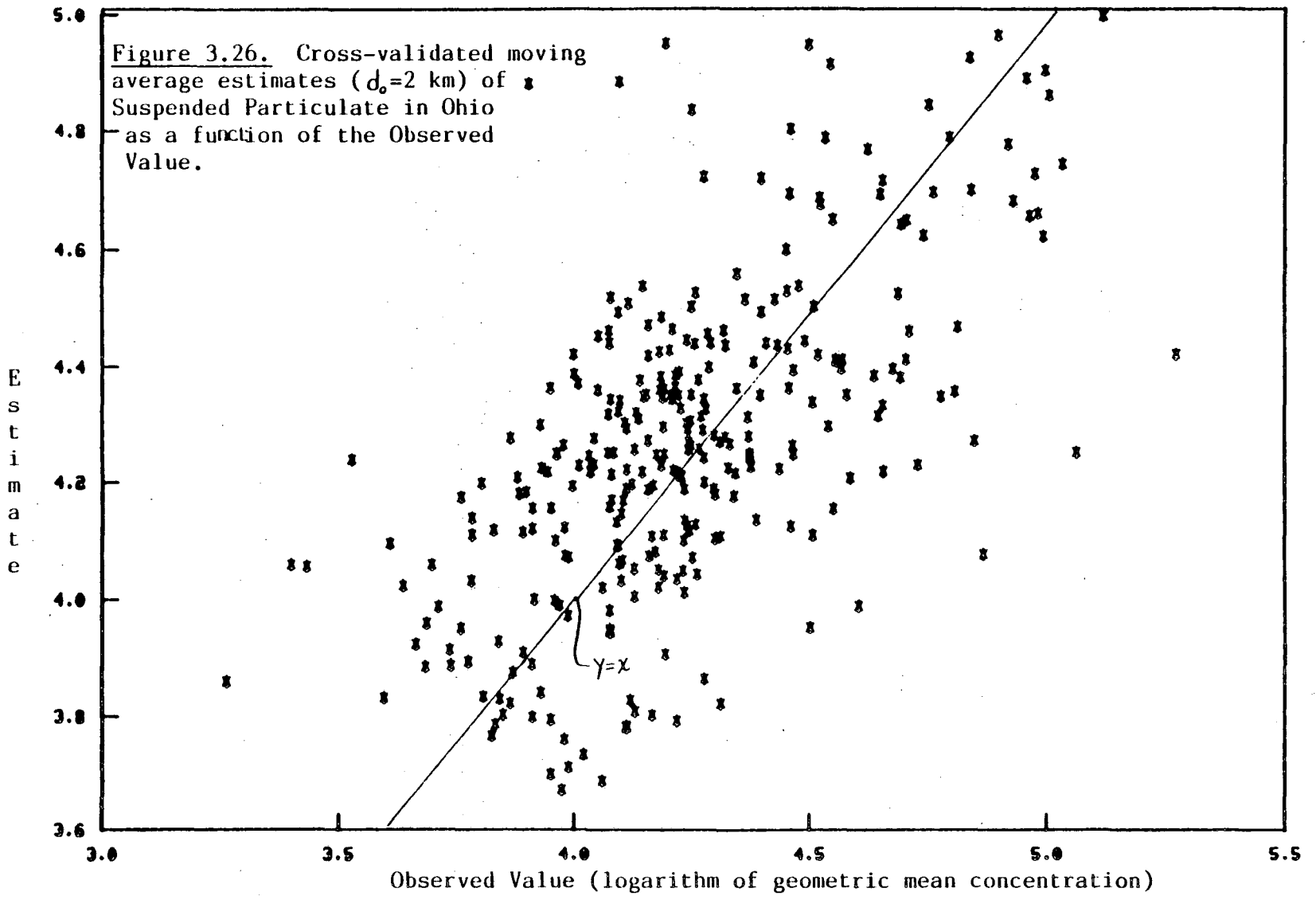


Figure 3.26. Cross-validated moving average estimates (d_o=2 km) of Suspended Particulate in Ohio as a function of the Observed Value.

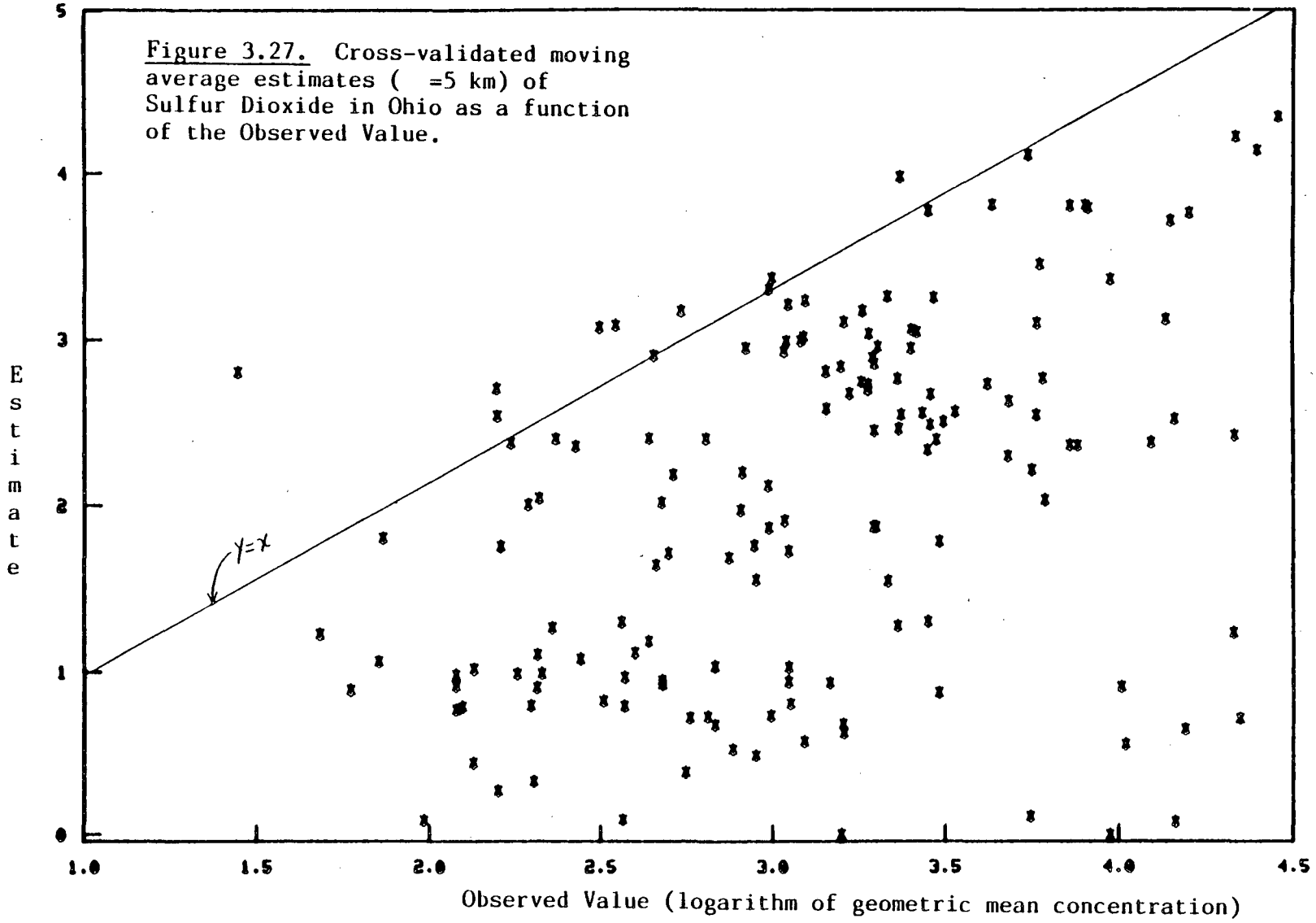


Figure 3.27. Cross-validated moving average estimates (=5 km) of Sulfur Dioxide in Ohio as a function of the Observed Value.

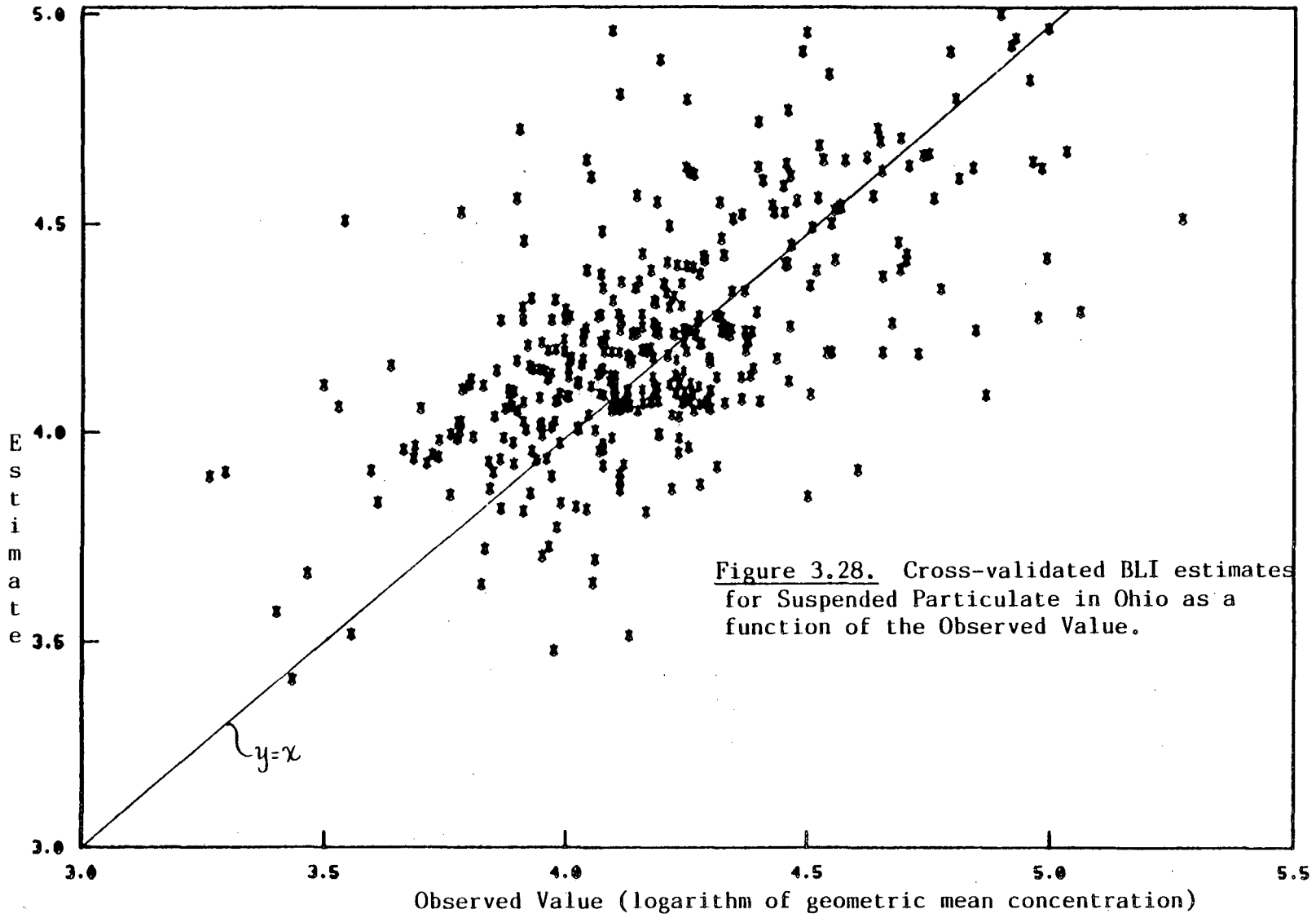


Figure 3.28. Cross-validated BLI estimates for Suspended Particulate in Ohio as a function of the Observed Value.

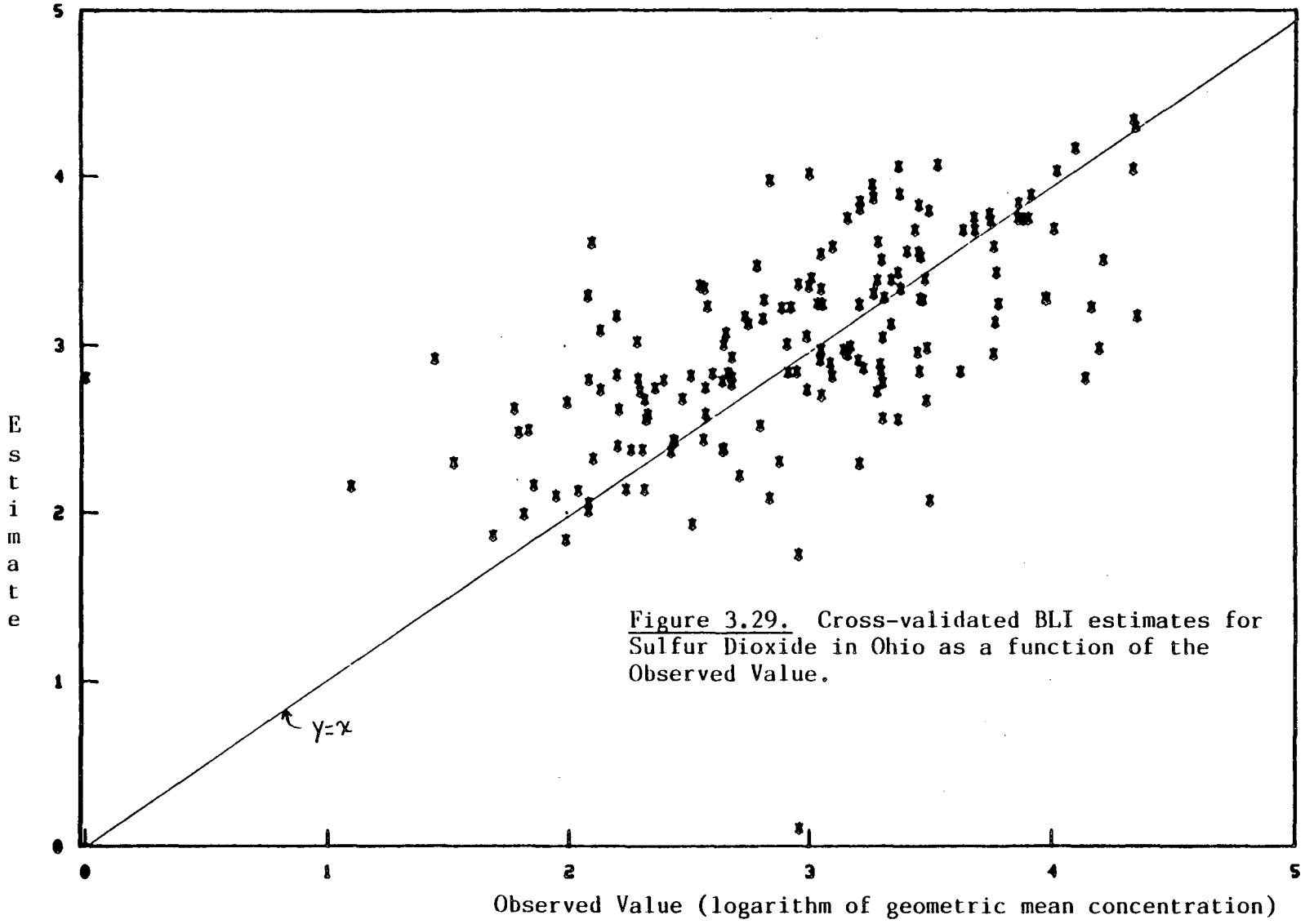


Figure 3.29. Cross-validated BLI estimates for Sulfur Dioxide in Ohio as a function of the Observed Value.

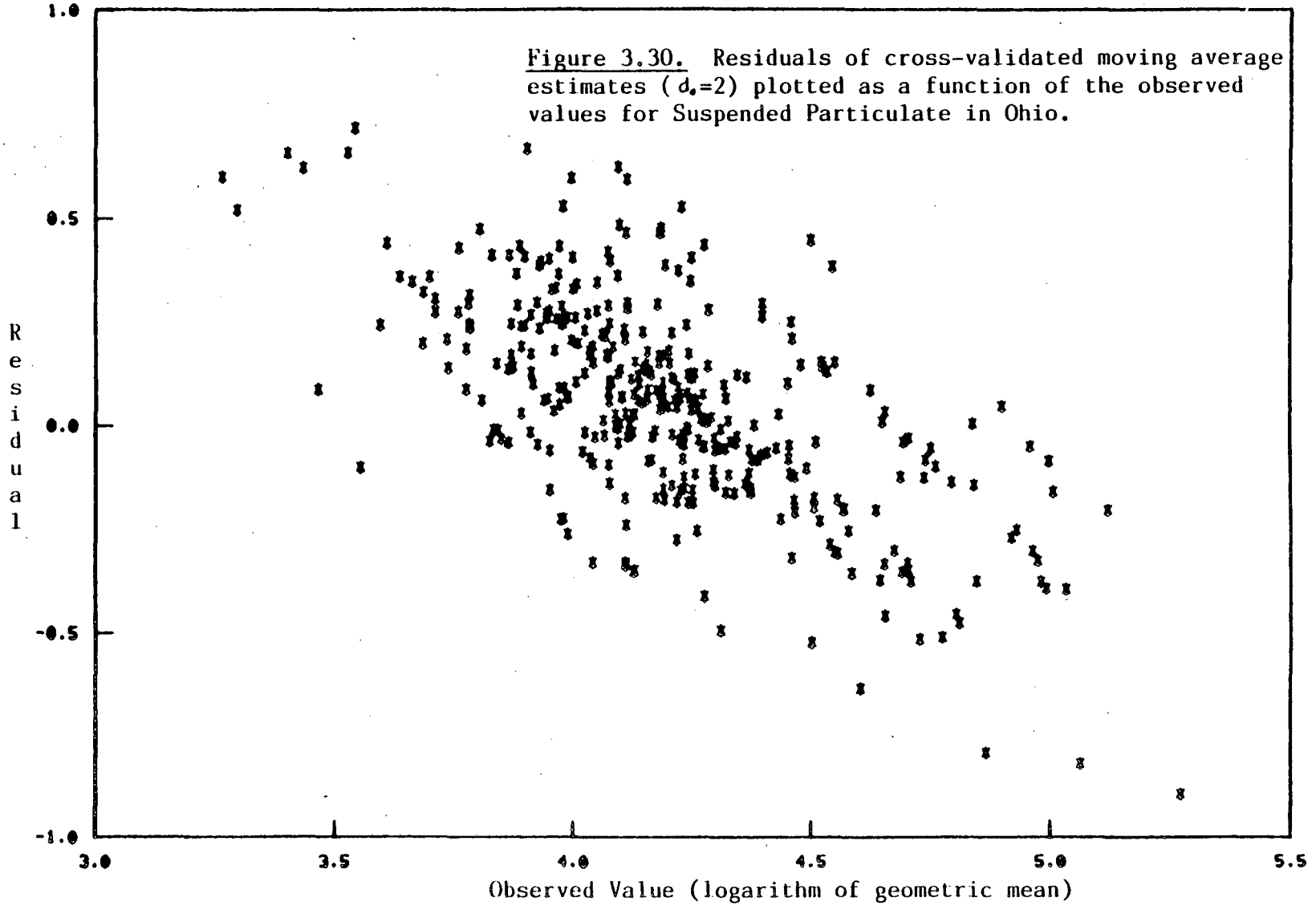


Figure 3.30. Residuals of cross-validated moving average estimates (d_e=2) plotted as a function of the observed values for Suspended Particulate in Ohio.

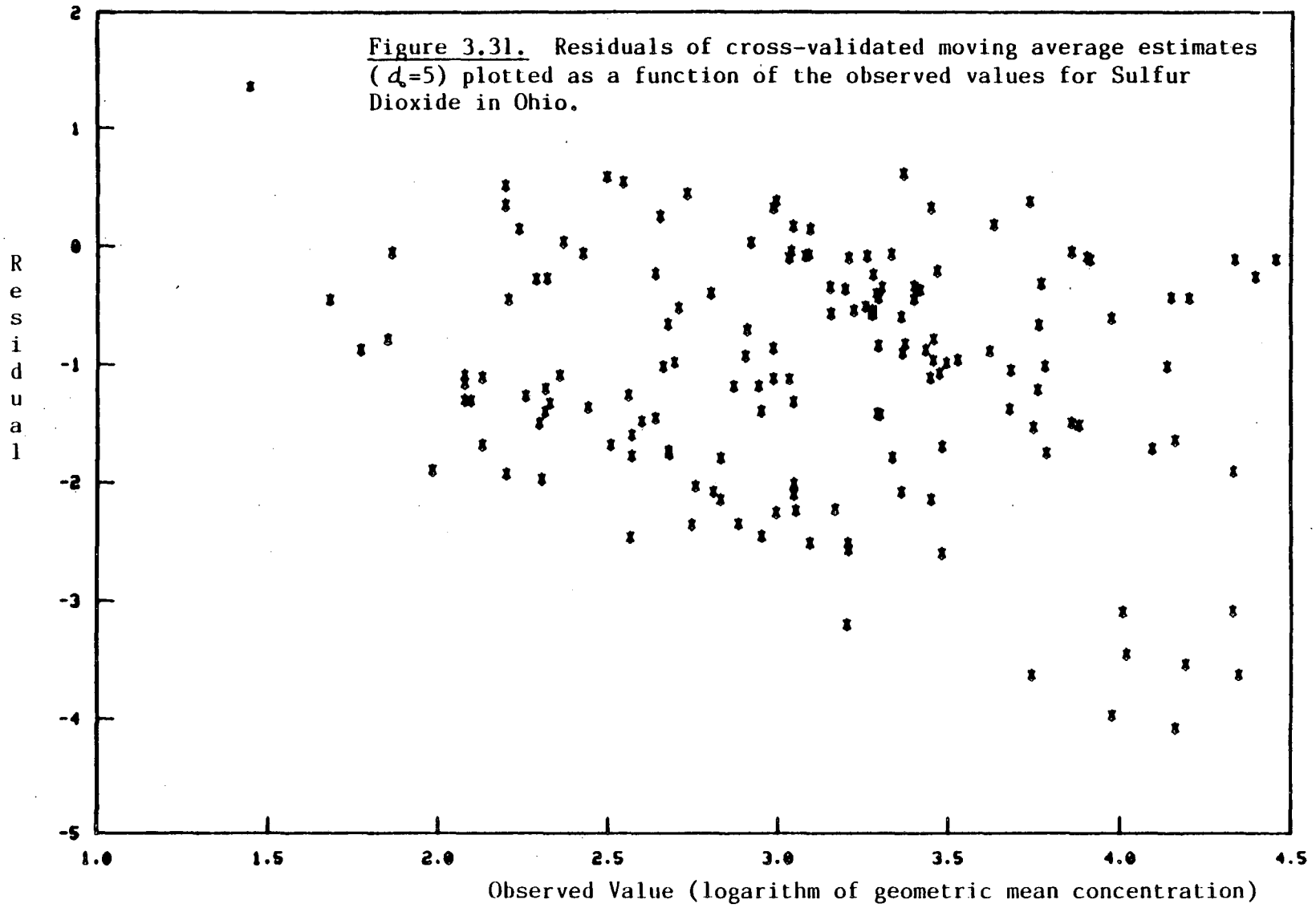


Figure 3.31. Residuals of cross-validated moving average estimates (d_0=5) plotted as a function of the observed values for Sulfur Dioxide in Ohio.

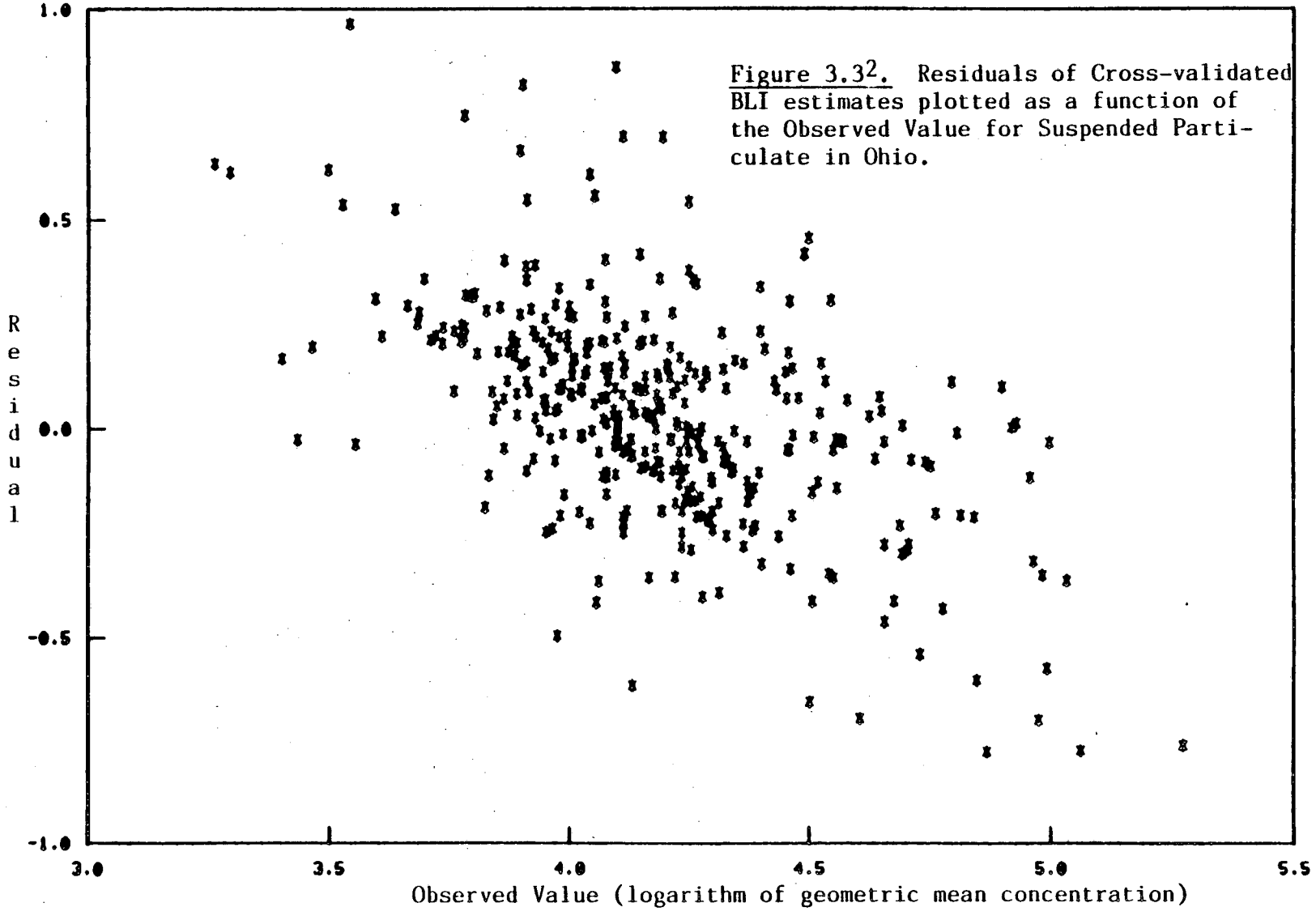


Figure 3.3[2]. Residuals of Cross-validated BLI estimates plotted as a function of the Observed Value for Suspended Particulate in Ohio.

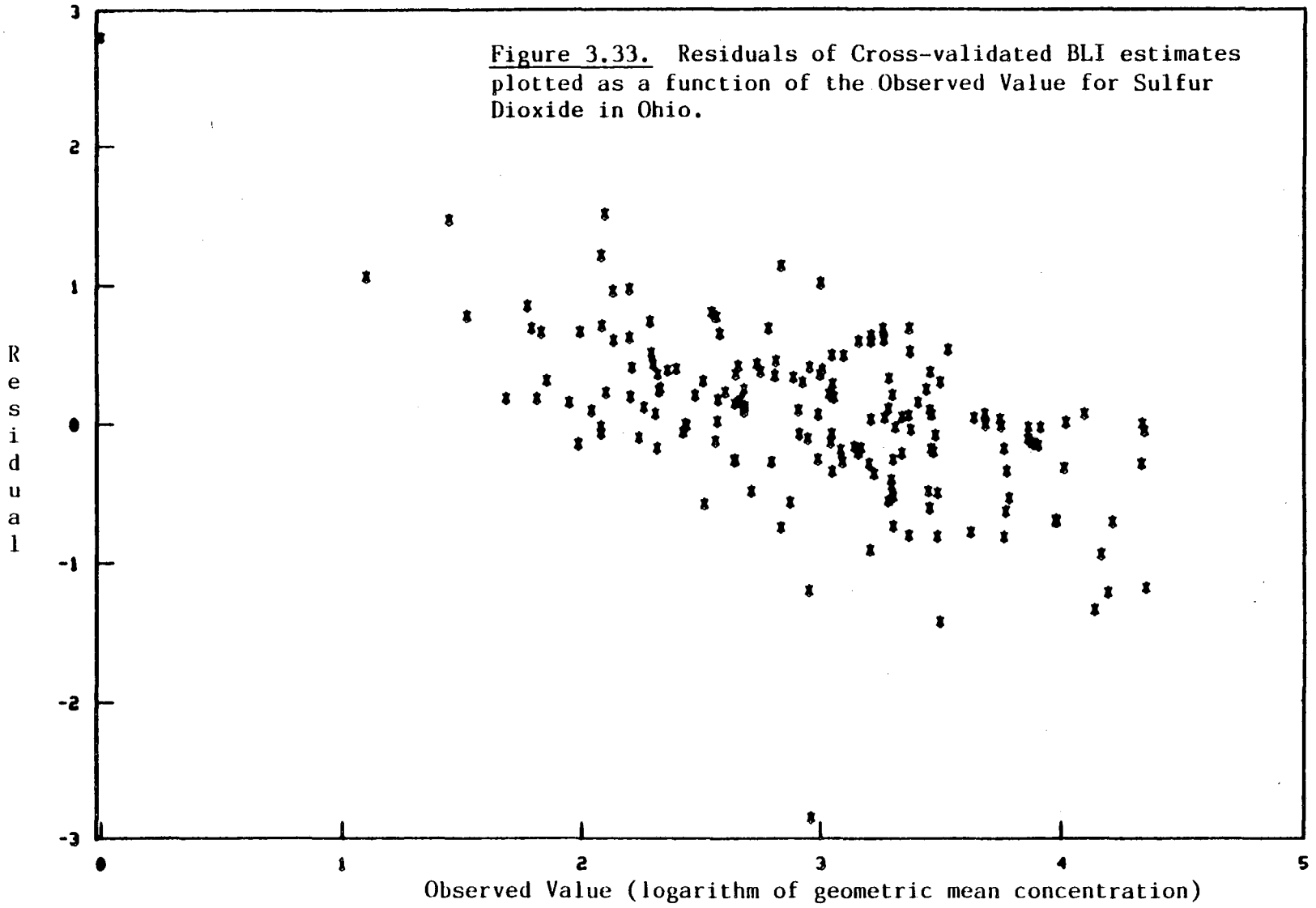


Figure 3.33. Residuals of Cross-validated BLI estimates plotted as a function of the Observed Value for Sulfur Dioxide in Ohio.

CHAPTER FOUR

Conclusions and Further Applications

Summary

In this document, 1974-76 air quality data for two pollutants from the EPA SAROAD data bank for two pollutants, have been described. Several summary statistics have been calculated for all pollutants in the whole United States to obtain a sense of the completeness of the data set. Limitations of these data have been discussed. The data have been summarized in such a way so as to facilitate analysis. Two pollutants and one state have been chosen to test the analytic techniques; in principle the same techniques can apply to any pollutant and any region if data were complete enough. The chosen pollutants and state have the highest average monitoring density. Limitations of statistical methods commonly used are discussed.

A two-dimensional moving average model is chosen for estimating pollutant concentrations from these data. This model is a weighted average with the weights being an exponential function of the squared distance of the data from the point of estimation. Distant data points, which have negligible weight, are ignored. Properties of this model are discussed; its drawbacks and advantages are pointed out. The functional form of the weights has a smoothing parameter which controls the relative weighting of nearby and distant data points. The parameter is chosen by cross-validation. Several functions are used to select the 'best' value of the parameter. Different functions yield different optimum values. Local variability as a concept is explored and a statistic which measures local variability is created. The smoothing parameter, d_0, is discussed

as a function of local variability. The optimal smoothing parameter is suggested as being equivalent to that which gives the minimum local variability. Local variability is discussed as an alternative to cross-validation for choosing the parameter in the moving average model. Local variability and cross-validation analyses are compared for sulfur dioxide and suspended particulate data in Ohio State. Consistency in the shape of the local variability function is shown in three states, Ohio, New York and Florida. The actual value of the optimal d_0 varies across these different geographic regions. The optimum value of the smoothing parameter is found to be larger for sulfur dioxide than for suspended particulates in each state.

Assessment of the moving average model is made by comparing it to a more commonly used method. Akima's method of bivariate linear interpolation is used for comparison and is explained. Linear interpolation is performed for Ohio using suspended particulate and sulfur dioxide data. Contour maps of these data are drawn using both methods. The model is used to estimate contour maps with several distinct values of the smoothing parameter covering a wide range. The dependency of the model's ability to estimate contour lines on the grid density is shown. These methods are compared in terms of their ability to estimate contour maps and to predict specified points. Local variability is used to make these comparisons by cross-validation rather than a "two-deep" cross-validatory analysis. Residuals are plotted and summary statistic calculations are discussed for choosing the 'best' method. It is found that the two-dimensional moving average model does as well and sometimes better than Akima's method of interpolation for suspended particulate prediction. For the sulfur dioxide data, Akima's method appears to make more accurate and less biased predictions than the moving average. The occurrence of bias in cross-

validating geographically sparse data using the moving average method is shown.

Further Research

Simulations of these data from actual distributions may be very helpful in determining what values a parameter like d_0 should have, how densely scattered the data points should be in order to perform a valid analysis, and possibly investigating the distribution theory of local variability. There are several distributions one could assume in order to simulate air quality data. One possibility is that each city's industrial centroid is like a point source of air pollution. In this case one might assume a bivariate normal distribution centered on the industrial portion of the city. One could vary the bivariate normal distribution to yield many different simulations. There are many other possibilities, for example treating each industry, commuting freeway, and other emitting source separately, treating the city as a sum of bivariate normal distributions. The possibilities are endless and probably none of them adequately represent the true situation.

The author believes that one of the most original and statistically promising results of this thesis is the concept local variability. The statistical theory of the measure we call local variability, that is defined in Chapter 2 may prove very useful for choosing parameters such as d_0. The statistical properties of local variability might lend themselves to the calculation of a confidence interval associated with the chosen disk radius. Local variability has been proposed as an alternative to cross-validation for this problem; it could be explored

as a possible alternative in other problems which involve cross-validation. It could also be applied to time series or trend analysis in one dimension, especially in cases where the data are irregular.

Boot-strapping, a new inventive method for estimating the distribution of a statistic (Diaconis, 1983), could be applied to the local variability function. Through the bootstrap, a confidence interval on the chosen disk radius could be estimated. That is, with each bootstrap sample the minimum local variability may change slightly giving different optimal disk radii, one from each bootstrap sample. There would then be a distribution of optimal disk radii, one from each bootstrap sample, from which a confidence interval could be estimated. The bootstrap could also be used for estimating variability associated with the contour maps and the predicted points. For predicting points, jack-knifing may also be helpful for examining the distribution of the estimate. The bootstrap, jack-knife and local variability are all data driven methods. These techniques do not require any distributional assumptions, therefore they are very useful for interpreting these data.

Time trends in the air quality data could be analyzed. Some ten years of data of reasonable quality are now available through the Environmental Protection Agency. The time trend analyses could be very helpful in understanding these data which are three-year averages, a single point in time. A time-geographic series could also result in some new data analytic techniques.

This thesis analyzes some of the problems associated with geographic data and creates a concept applicable to other analyses. Many problems remain to be solved. The variability of estimates is an area of great interest and is still a wide open field for study. Although not addressed here, the inclusion of other information like meteorologic and topologic factors certainly lends itself to

understanding geographic air quality data. From a statistical standpoint, the concept of local variability could be applied to many data analytic problems. Through this dissertation there is hope that others will find interesting ideas for their own research.

References

Akima, Hiroshi (1978) *A Method of Bivariate Interpolation and Smooth Surface Fitting for Irregularly Distributed Data Points*. ACM Transactions on Mathematical Software, Vol. 4, no. 2, June 1978.

Akima, Hiroshi (1978) *Algorithm 526: Bivariate Interpolation and Smooth Surface Fitting for Irregularly Distributed Data Points*. ACM Transactions on Mathematical Software Vol. 2, no. 1, pp. 160-164.

Becker, Richard A. and Chambers John M. (1981) *S a language and system for data analysis*. Bell Laboratories, January.

Bengtsson, B.E. and Nordbeck, S. (1964) *Construction of isarithms and isarithmic maps by computers*. BIT, Vol. 4, pp. 87-105.

Breiman, Leo, Meisel, W. and Purcell, E. (1977) *Variable Kernel Estimates of Multivariate Densities*. Technometrics, Vol. 19, no. 2.

David, M. *Geostatisticsl Ore Reserve Estimation*. Elsevier, Amsterdam.

Davis, J.C. and Cullagh, M.J. (Editors, 1975) *Display and Analysis of Spatial Data*. Wiley, New York.

Delfiner, P. (1976) *Linear Estimation of nonstationary spatial phenomena. Advanced Geostatistics in Mining Industry*. M. Guarascio, C.J. Huijbregts, and M. David, Eds. Reidel, Dordrecht. pp. 49-68.

Diaconis, Persi and Efron, Brad (1983) *Computer-intensive methods in Statistics*. Scientific American, Vol. 248, no. 5, pp. 116.

C.A. Eades, *Chart - A Graphic Display Analysis System.* Lawrence Berkeley Lab Report, Unpublished LBL Report.

EPA (1976) *AEROS, Aerometric and Emissions Reporting System Information.*

R. Faith and Sheshinski, R. (1979) *Misspecification of trend in spatial random function interpolation with application to oxidant mapping.* SIMS Technical Report No. 28, September.

Grant, F. (1957) *A problem in the analysis of geophysical data.* Geophysics, Vol. 22, pp. 309-344.

Green, P.J. and Sibson, R. (1978) *Computing Dirichlet Tessellations in the Plane.* Computer Journal, Vol. 21, pp. 168-173.

Hatten, Dorothy (1978) *Interpretation, Interpolation and Adjustment of Oceanographic Fields.* Norfolk State College, Dwight Dierterle and Arthur Tingle, Research Advisors.

Hoel, Port and Stone (1971) *Introduction to Statistical Theory.* Houghton Miflin, pp. 220-221.

Y. Horie and D. Bicker (1979) *Final Report for Development of Data Base for Human Exposure to Air Pollution in the South Coast Air Basin.* Contract A7-163-30, October.

Johnson, Laura D. (1982) *1974-1976 Air Quality by Location.* SEEDIS data document LBID-357-CD, January.

Johnson, Laura D., Merrill, D.W., and Selvin, S. (1982). *Predicting a continuous*

spatial variable from discrete point measurements. LBL Report LBL-14235, March.

Journel, A.G. and Huijbregts, C.J. (1978) *Mining Geostatistics*. Academic Press, London.

Kendall, M.G. (1976) *Time Series*. Second Edition, Hafner Press.

Landau, E. (1971) *The Nashville Air Pollution Study. Sulfur Dioxide and Bronchial Asthma - A Multivariate Analysis*. Int. J. Env Studies Vol. 2 pp. 41-45.

Lawson, C.L. (1972) *Generation of a Triangular Grid with application to contour plotting*. Section 914, Tech. Mem. No. 299, California Institute of Technology.

Lehmann, E.L. (1975) *Nonparametrics, Statistical Methods based on Ranks*. Holden-Day, Inc., San Francisco.

MacCracken M.C. and Sauter G.D., Editors, (1975) *Development of an Air Pollution Model for the San Francisco Bay Area*. Lawrence Livermore Laboratory Report UCRL-51920, Volumes 1 and 2, October.

Matheron, G. (1965) *Les Variables Regionalisees et leur Estimation*. Masson, Paris.

McLain D.H. (1976) *Two dimensional interpolation from random data*. Computer Journal, Vol. 19 pp. 178-181; 384.

R. Mendelsohn and Orcutt G. (1979) *An Empirical analysis of air pollution dose-response curves*. J. Env. Management Vol. 6 pp. 85-106.

Merrill, Deane W. (1982) *Problems in Spatial Data Analysis*. LBL Report LBL-

14047, February.

D.W. Merrill,Jr., Sacks, S.T., Selvin, S., Hollowell, C.D., and Winkelstein, Jr., W. (1978) *Populations-At-Risk-To-Air Pollution (PARAP): Data Base Description and Prototype Analysis*. LBL Report UCID-8039, August.

Merrill, D. (1981) *1974-1976 Air Quality for Individual Monitoring Stations*. SEEDIS data document LBID-357-AZ January.

W. R. Ott (1980) *Models of Human Exposure to Air Pollution*. SIMS Technical Report No. 32, July.

Powell, M.J. and Sabin, M.A. (1977) *Piecewise quadratic approximations on triangles*. ACM Transactions on Mathematical Software, Vol. 3, pp. 316-325.

Brian D. Ripley (1981) *Spatial Statistics*. Wiley, October.

Sager, Thomas W. (1976) *Relating Spatial Distributions of Pollutants to* Health Effects. SIMS Technical Report No. 1.

Scheffe', Henry (1959) *Analysis of Variance*. Wiley Publications in Statistics, pp. 57-58.

Selvin, S., Merrill, D., Kwok, L., and Sacks, S. (1981) *Ecologic Regression Analysis and the Study of the Influence of Air Quality on Mortality*. LBL Report LBL-12217, September.

Selvin, S., Sacks, S.T., Merrill, D.W., and Winkelstein, W. (1980) *The relationship between cancer incidence and two pollutants (total suspended particulate and carbon monoxide) for the San Francisco Bay Area*. LBL Report LBL-10847, June.

Sibson, R. (1980) *The Dirichlet Tesselation as an aid in data analysis.* Scandinavian Journal of Statistics, Vol. 7, pp. 14-20.

Sibson, R. (1980) *A vector identity for the Dirichlet Tessellation.* Mathematical Proceedings of the Cambridge Philosophical Society, Vol. 87 pp. 151-155.

Sibson, R. (1978) *Locally equiangular triangulations.* Computer Journal, Vol. 21, pp. 243-245.

Snedecor and Cochran (1967) *Statistical Methods.* Iowa State University Press.

Wahba, G. and Wold, S. (1975) *A Completely Automatic French Curve: Fitting Spline Functions by Cross Validation.* Communications in Statistics, Vol. 4, no. 1, pp. 1-17.

Whittemore, Alice S. and Korn, Edward L. (1980) *Asthma and Air Pollution in the Los Angeles Area.* SIMS Technical Report 35, March.

Winkelstein, W., Levin, Lynn, Johnson, Kathryn (1982) *RE: "Health Effects of Particulate Pollution: Reappraising the Evidence" Letters to the Editor.* AJE, Vol. 115, no. 3, pp. 471-475.

This report was done with support from the Department of Energy. Any conclusions or opinions expressed in this report represent solely those of the author(s) and not necessarily those of The Regents of the University of California, the Lawrence Berkeley Laboratory or the Department of Energy.

Reference to a company or product name does not imply approval or recommendation of the product by the University of California or the U.S. Department of Energy to the exclusion of others that may be suitable.

TECHNICAL INFORMATION DEPARTMENT
LAWRENCE BERKELEY LABORATORY
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720