LBL-14235
UC-11 c.2

Lawrence Berkeley Laboratory
UNIVERSITY OF CALIFORNIA

Physics, Computer Science & Mathematics Division

RECEIVED
LAWRENCE
BERKELEY LABORATORY

SEP 30 1982

LIBRARY AND
DOCUMENTS SECTION

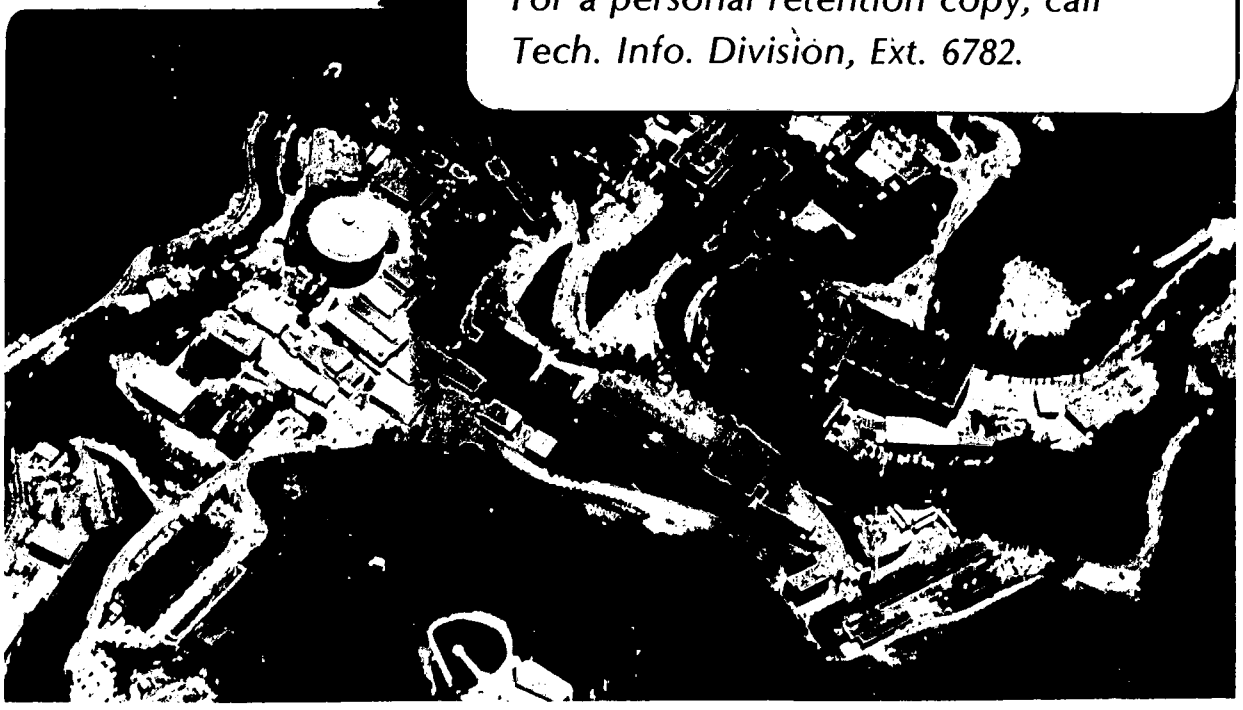
PREDICTING A CONTINUOUS SPATIAL VARIABLE FROM DISCRETE POINT MEASUREMENTS

Laura D. Johnson, D.W. Merrill, and S. Selvin

March 1982

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.



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

LBL-14235 c.2

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-14235

PREDICTING A CONTINUOUS SPATIAL VARIABLE FROM DISCRETE POINT MEASUREMENTS

Laura D. Johnson

D. W. Merrill

S. Selvin

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

March 1982

This work was supported by the Office of Health and Environmental Research, Office of Energy Research, U.S. Department of Energy under Contract number DE-AC03-76SF00098.

Introduction

Estimating an average level of any characteristic for a geographic area based on data collected from specific points in that area is a problem in many fields of study. For example, many atmospheric measurements are made in this way. The problem that we wish to address is the estimation of an average level of pollution for an area based on point data derived from monitoring stations. 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 values and find its average. This has been done in mining fields, for a map of the mineral grade will help plan the mining operation as well as give information on which parcels will have a high enough average grade to make processing economic (Ripley, 1981).

A very simple way to smooth or interpolate is to calculate the value of a surface as a weighted average of the values at the data points. This method is called a moving average. Criteria are imposed so that the surface is smooth at the data points. These weights are chosen as a function of the distance between the data points and the estimation point. The method of estimation that is presented here is a weighted mean (i.e., an arithmetic mean of the logarithm of the geometric mean concentrations) and therefore is an example of a moving average. In this study we have air pollutant monitoring stations. These stations measure pollution a certain percentage of the time and are stationary in that the measurement is always taken at the same latitude and longitude. We have chosen a few different geographic areas to test this method. At each station a log normal distribution is implicitly assumed. A relevant statistical problem is determining the "fitness" of the estimate and is a major part of this paper. For determining the "fitness", we have estimated each station's value from sur-

rounding stations' observed values and compared our estimate with the observed value. The methodology described here is an explicit calculation, not a fit, and so is not affected by ill-defined boundary conditions and other uncertainties.

Data

The Environmental Protection Agency has data for 6625 air quality monitoring stations (active 1974-1976) in the US + territories (Puerto Rico, Guam, V.I.). Ambient air quality was measured at these stations. These pollutants were sampled over one-hour intervals for specific gases like carbon monoxide, or 24-hour intervals for less specific pollutants like total suspended particulate. For sulfur dioxide and nitrogen dioxide, both one-hour and 24-hour sampling intervals were used. The basic air quality file developed by the PAREP (Populations at Risk to Environmental Pollution) project contains latitude and longitude of the 6625 monitoring stations active during 1974-1976, the percent of time active, and the three-year geometric mean concentrations of nine pollutants. This is the file from which data are obtained to investigate interpolation strategies.

Suppose, for example, we want to estimate the level of total suspended particulate for a particular county. We might want to obtain data from all the monitoring stations within the county. However, other stations which are near but outside the county should also be used to estimate the pollution within the county because of their proximity; so we need to include nearby stations. For analysis of long-term epidemiologic effects on a mobile population, highly detailed geographic accuracy is not required. Therefore, the estimate might be a weighted average, where the weight is a function of the distance from the monitoring station to the point of interest.

Estimation

If one is trying to predict the average level of pollution at a particular point over a specified period of time from nearby monitoring stations, there are at least two characteristics which are important. One, already mentioned, is the distance of the monitoring station from the point of estimation. Since the size of the sample is probably correlated with the reliability of the mean value, another important factor is the percentage of time a monitoring station was active. Thus, the weight function used in our estimate is bivariate in these two variables. The weight of station i for predicting at point j is taken to be

$$w_{ij} = p_i e^{-.5 d_{ij}^2 / d_0^2}$$

where p_i is the percent of time station i was active, d_{ij} is the distance of station i to the point j where pollution level needs to be estimated, and d_0 is a scaling parameter from 2 to 20 kilometers. Stations further than five times d_0 were ignored since they have negligible weight. The log mean pollution level (i.e., the logarithm of the geometric mean) at the point of interest j is estimated as

$$e_j = \frac{\sum_{i=1}^{i=n_j} w_{ij} x_i}{\sum_{i=0}^{i=n_j} w_{ij}}$$

where x_i is the log of the observed value of the ith surrounding station and the sum is over all stations close enough to be relevant; thus n_j is the number of "predicting" stations for point j.

In some cases two or more stations have exactly the same latitude and longitude and (in general) different values of pollutant concentration. The 6625 stations have 5777 distinct locations. In our analysis stations at the same location were considered as a single station with a concentration equal to a

geometric mean (weighted by the number of observations) of the individual values. The combining of stations does not affect the calculated values of e_j, but it does affect the loss functions and correlation coefficients defined later in this section.

To choose the "optimum" value of d_0 and investigate the accuracy of these weighted averages, a predicted value for each station was generated from observations from the other nearby stations (excluding one's own value), and compared to the actual value observed at the selected station. By varying d_0, we can generate several estimates for comparison with each observed value. We can then look at correlations and loss functions to pick the "best" d_0 and thus the "best" weighting function for our estimate.

We chose four functions to measure the "goodness of fit". These have been named 1) Correlation Coefficient 2) Loss Function 3) Weighted Correlation Coefficient 4) Weighted Loss Function. The formulas for these are as follows:

Correlation Coefficient:

$$\frac{\sum_{j=1}^{j=n} p_j (x_j - \bar{x})(e_j - \bar{e})}{\sqrt{\sum_{j=1}^{j=n} p_j (x_j - \bar{x})^2 \sum_{j=1}^{j=n} p_j (e_j - \bar{e})^2}}$$

Loss Function:

$$\frac{\sum_{j=1}^{j=n} p_j (x_j - e_j)^2}{\sum_{j=1}^{j=n} p_j}$$

Weighted Correlation Coefficient:

$$\frac{\sum_{j=1}^{j=n} p_j W_j (x_j - \bar{x})(e_j - \bar{e})}{\sqrt{\sum_{j=1}^{j=n} p_j W_j (x_j - \bar{x})^2 \sum_{j=1}^{j=n} p_j W_j (e_j - \bar{e})^2}}$$

Weighted Loss Function:

$$\frac{\sum_{j=1}^{j=n} p_j W_j (x_j - e_j)^2}{\sum_{j=1}^{j=n} p_j W_j}$$

where all the sums are over all stations which have both an observed value x_j and a predicted value e_j. The mean observed value is given by $\bar{x}$ and that for the predicted values is denoted by $\bar{e}$. Observation j and its estimate are denoted by x_j and e_j respectively. The number of stations in the study area predicted is n. The weight for station j, W_j, is

$$W_j = \frac{\sum_{i=1}^{i=n_j} w_{ij}}{\sum_{j=1}^{j=n} \sum_{i=1}^{i=n_j} w_{ij}}$$

The "best" d_0 would be the one which maximizes the correlation, and minimizes the loss function in both the weighted and unweighted case.

Results

We have chosen three areas to experiment with, using this method. The pollutant we have considered is total suspended particulate. The areas are the Detroit Standard Metropolitan Statistical Area (SMSA), the state of California and Los Angeles County. These areas are chosen somewhat arbitrarily, although the results do change with these areas perhaps as a function of size. The maps of the areas and their "relevant" monitoring stations are shown in figures 1, 2, and 3. The size of the circle representing the monitoring station is indicative of the geometric mean pollutant level at that station.

Very different results were obtained in the different areas. The results are summariized in Figures 4,5,6 and Tables 1,2,3. Stations in the Detroit SMSA

(Fig. 4 and Table 1) gave an optimal d_0 of 4 km consistently in all goodness of fit functions and these functions look reasonably smooth. California State (Fig.5 and Table 2) also gave reasonable losses and correlations as a function of d_0, yet unlike Detroit the optimal d_0 was 2 km instead of 4 km. Los Angeles County (Fig. 6 and Table 3) gave the most disconcerting results. The estimate was negatively correlated with the observed value and the optimal d_0 in the correlations is 2 km while in the loss functions it is 20 km. It is also clear that the weighted goodness of fit functions are relatively flat which indicates that these weights make the goodness of fit relatively insensitive to d_0. Figures 4, 5, and 6 and tables 1, 2, and 3 are illustrative of these findings.

Table 1. Detroit SMSA

d_0 (km)	Unweighted Correlation Coefficient	Weighted Correlation Coefficient	Unweighted Loss Function	Weighted Loss Function
2	0.829	0.854	0.0227	0.0236
4	0.849	0.894	0.0139	0.0210
6	0.854	0.870	0.0106	0.0254
8	0.851	0.841	0.0080	0.0307
10	0.832	0.807	0.0072	0.0363
12	0.807	0.775	0.0067	0.0413
14	0.785	0.747	0.0064	0.0455
16	0.766	0.725	0.0061	0.0491
18	0.751	0.708	0.0058	0.0523
20	0.738	0.695	0.0056	0.0551

Table 2. California State

d_0 (km)	Unweighted Correlation Coefficient	Weighted Correlation Coefficient	Unweighted Loss Function	Weighted Loss Function
2	0.744	0.680	0.0500	0.0385
4	0.702	0.671	0.0361	0.0358
6	0.720	0.677	0.0266	0.0370
8	0.716	0.694	0.0196	0.0384
10	0.722	0.705	0.0153	0.0390
12	0.690	0.718	0.0139	0.0391
14	0.699	0.729	0.0110	0.0387
16	0.697	0.739	0.0097	0.0381
18	0.695	0.745	0.0086	0.0376
20	0.694	0.747	0.0075	0.0374

Table 3. Los Angeles County

d_0 (km)	Unweighted Correlation Coefficient	Weighted Correlation Coefficient	Unweighted Loss Function	Weighted Loss Function
2	-0.848	-0.695	0.0964	0.0319
4	-0.788	-0.755	0.0396	0.0654
6	-0.843	-0.829	0.0325	0.0544
8	-0.872	-0.874	0.0199	0.0481
10	-0.888	-0.879	0.0139	0.0438
12	-0.877	-0.861	0.0103	0.0400
14	-0.837	-0.819	0.0078	0.0363
16	-0.761	-0.752	0.0060	0.0332
18	-0.646	-0.659	0.0048	0.0307
20	-0.512	-0.552	0.0039	0.0290

These seemingly contradictory results may be due to sample selection. If we look at the number of stations in our sample for each area, we see a marked difference.

	no. of stations in area.	no. of "relevant" stations.
Detroit	39	198
California	146	210
Los Angeles	13	61

A small number of stations could have a negative effect on the correlation. For example, the extreme case is with only two relevant stations a and b. Their pollutant values might be 100 and 200 for station a and b respectively. If we use station a to predict station b and station b to predict station a, we then would have a perfect negative correlation. The source of negative correlation is indeed an influence when small numbers of stations are used in the estimation procedure. Negative correlations should disappear for larger study areas having discrete clusters of stations with markedly different pollutant levels (i.e., a large "between" cluster variance).

Conclusions

This work is at an early stage. There are many other ways to test this estimate. Because of peculiar effects for small samples, it is probably more meaningful to use larger sample study areas with noticeable clusters of different pollution levels. We also need to look at other pollutants to see how these results might change. Other algorithms, including fitting procedures, should likewise be tested. Tests should be devised for the elimination of spurious data points. Standard deviations should be calculated, which reliably describe the confidence levels of an air quality estimate. This is just the beginning of the search of methods for interpolating point data into a continuous variable, which can apply to many problems in statistical analysis of spatial data.

References

1. Brian D. Ripley, Spatial Statistics, Wiley Publications (1981).

2. S. Selvin, S.T. Sacks, D.W. Merrill and W. Winkelstein, 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 1980.

3. R. Mendelsohn and G. Orcutt, An empirical analysis of air pollution dose-response curves, J. Env. Econ. and Management 6(1979):85-106.

4. B. Efron, Computers and the Theory of Statistics: Thinking the Unthinkable, SIAM Review, 21(4):460-80, October 1979.

5. J.C. Davis and M.J. McCullagh, eds., Display and Analysis of Spatial Data, NATO Advanced Study Institute, Wiley Publications, 1975.

6. M.C. MacCracken and G.D. Sauter, eds, Development of an Air Pollution Model for the San Francisco Bay Area. Final Report to the National Science Foundation. October 1, 1975

OVER 74.4
61.0 - 74.4
51.0 - 60.9
UNDER 51.0
DATA NOT AVAILABLE

GEOMETRIC
MEAN
MICROGRAMS
PER CU METER
1974-76
TOT-SUSP-PART
24-HR

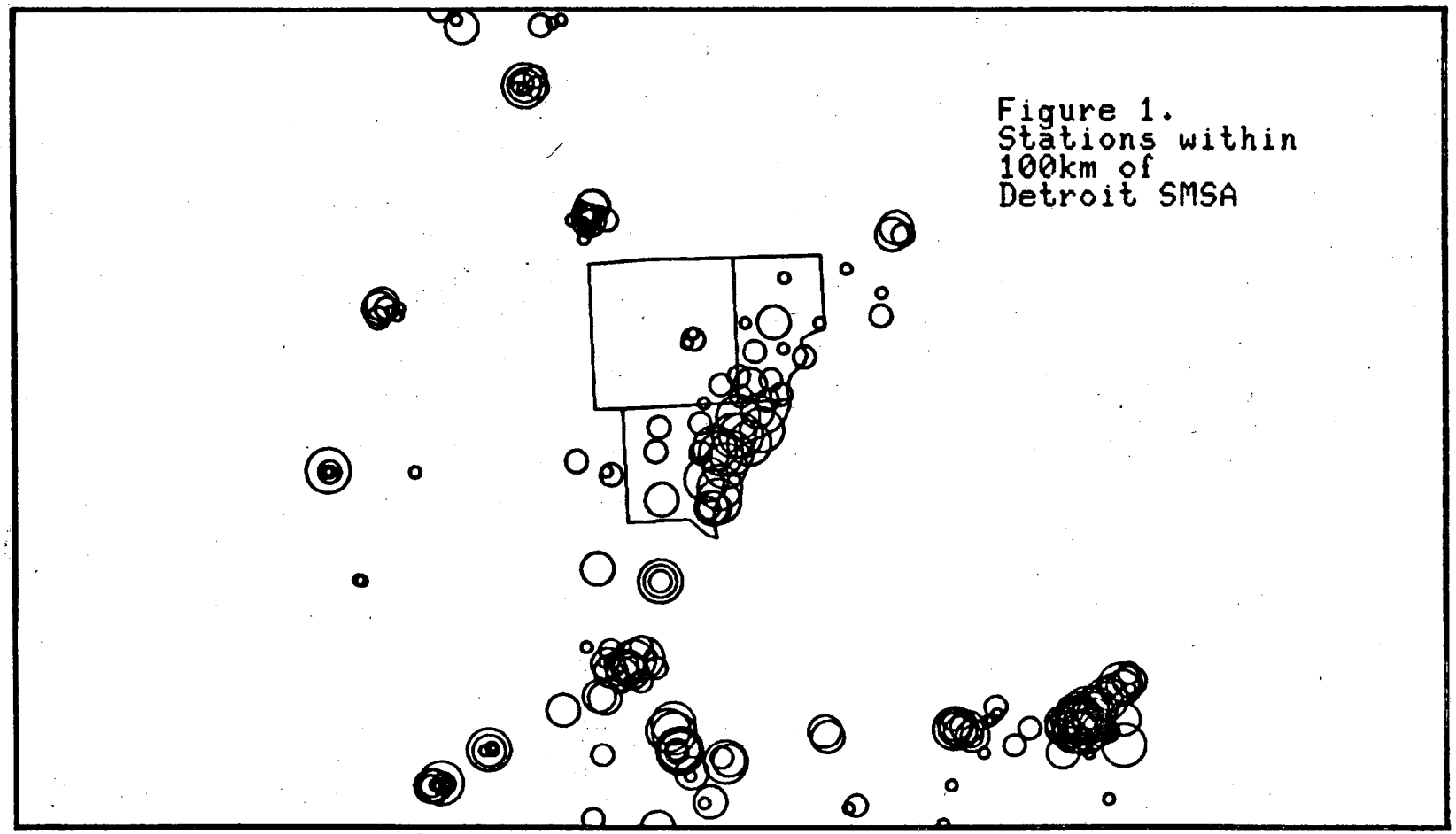


Figure 1.
Stations within
100km of
Detroit SMSA

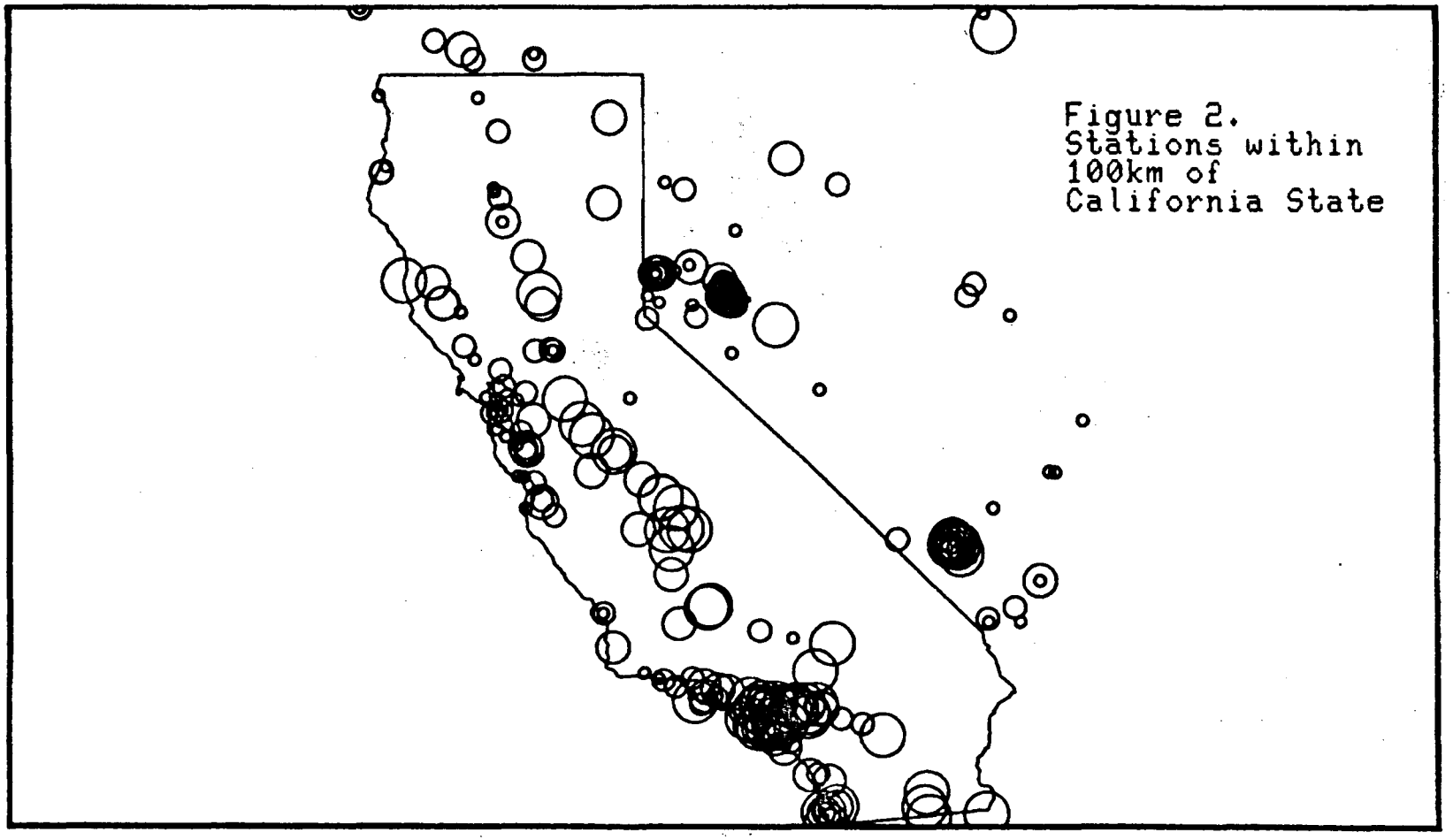


Figure 2.
Stations within
100km of
California State

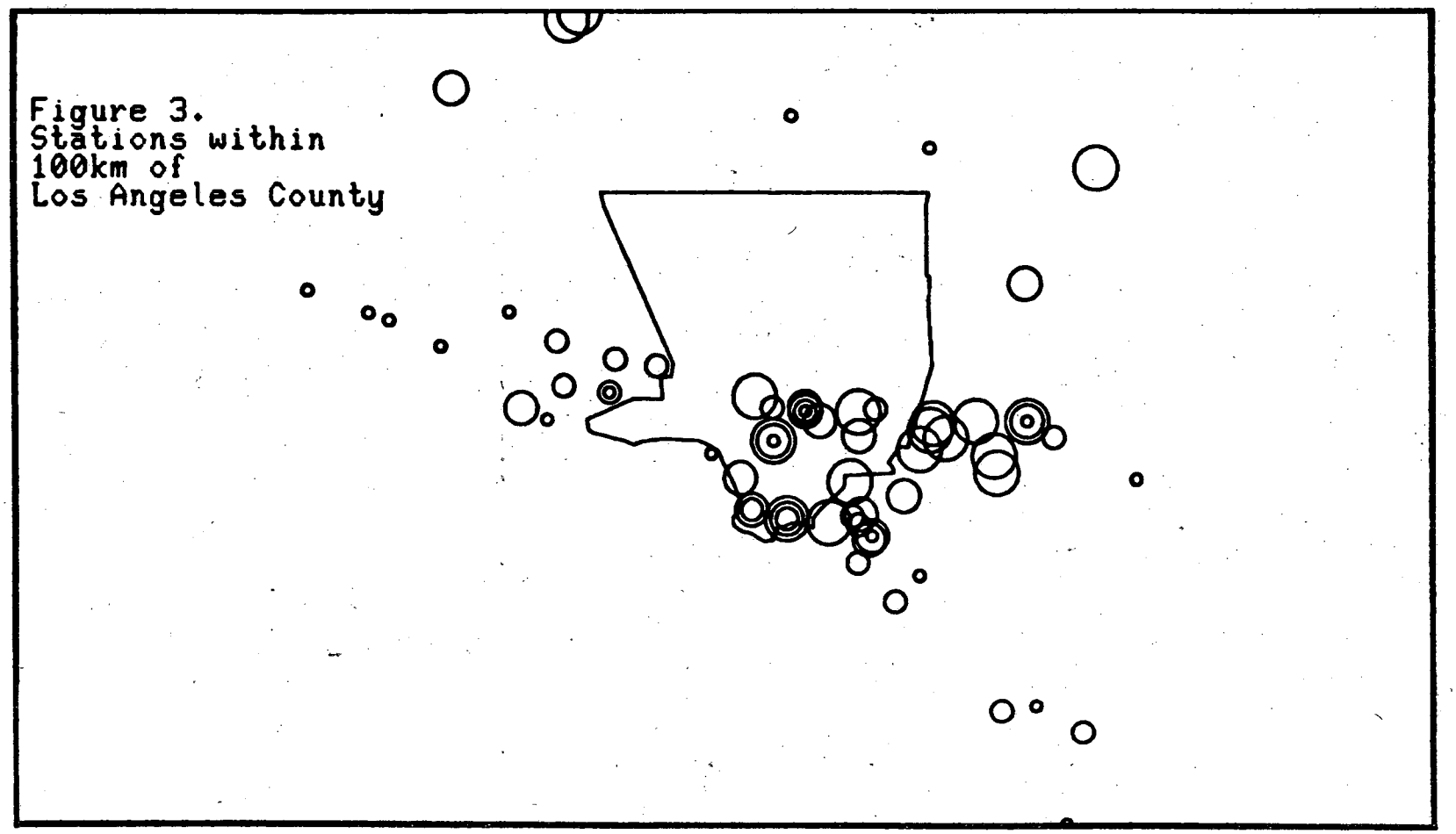


Figure 3.
Stations within
100km of
Los Angeles County

Figure 4. Detroit SMSA

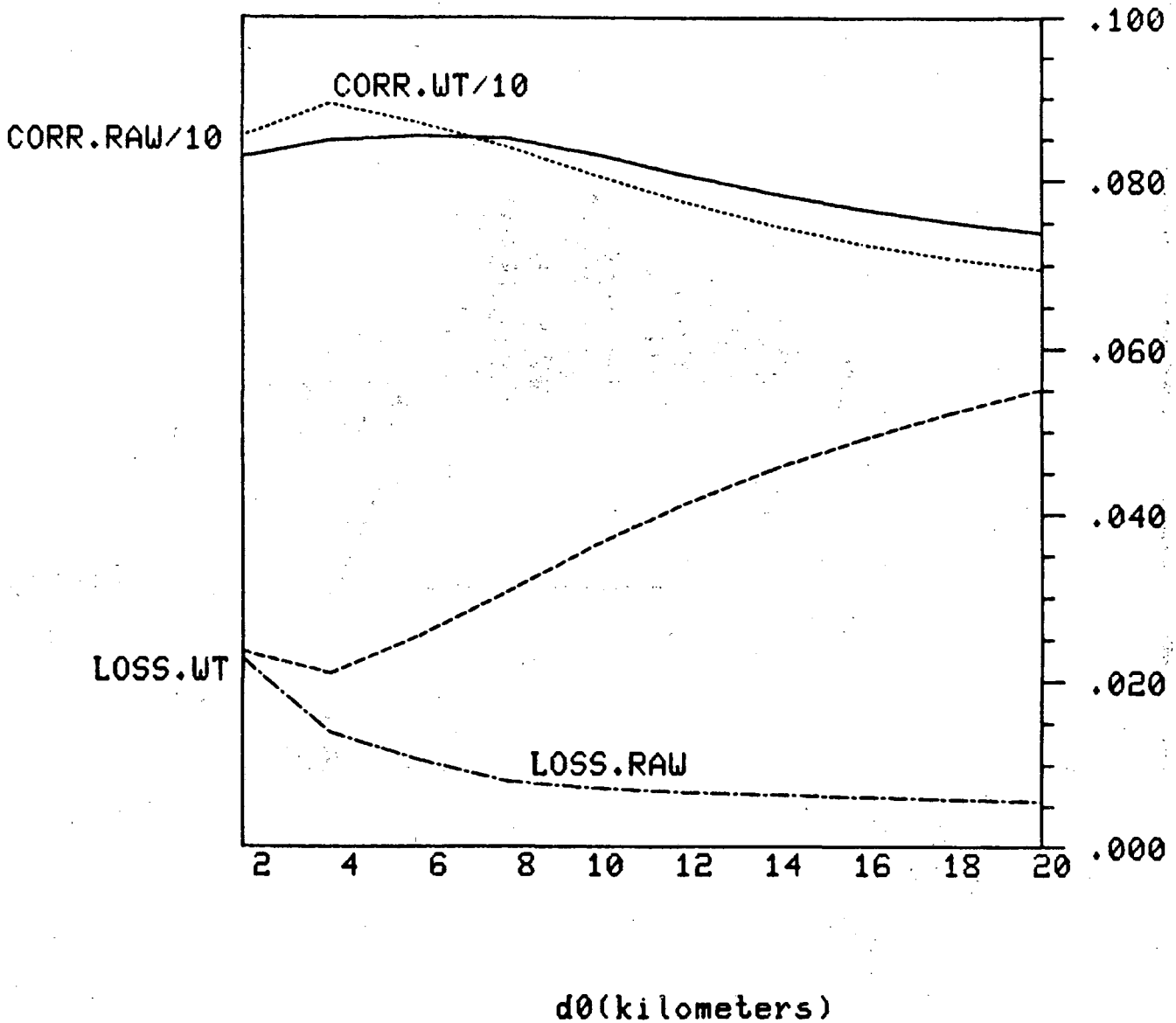


Figure 5. State of California

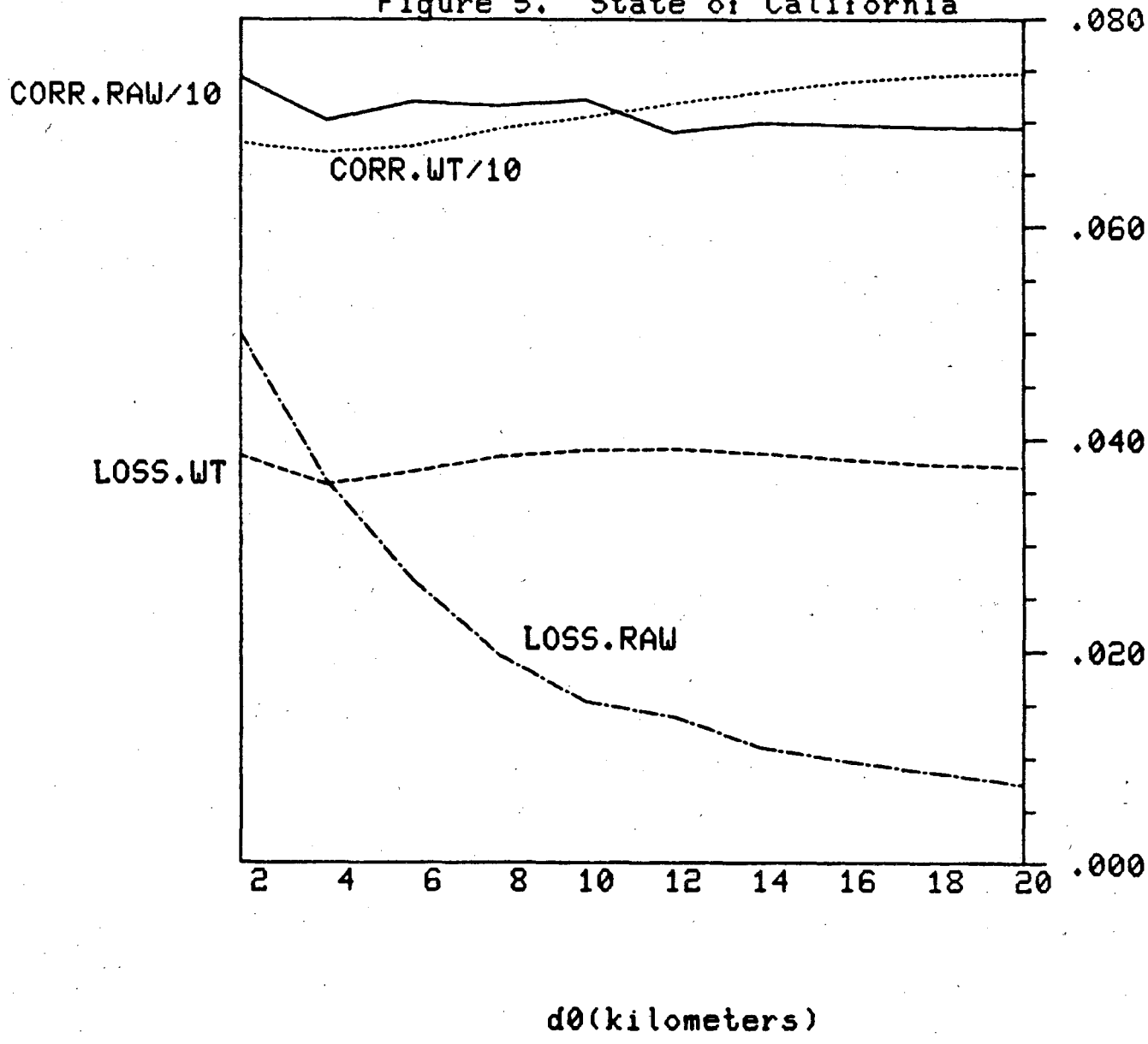
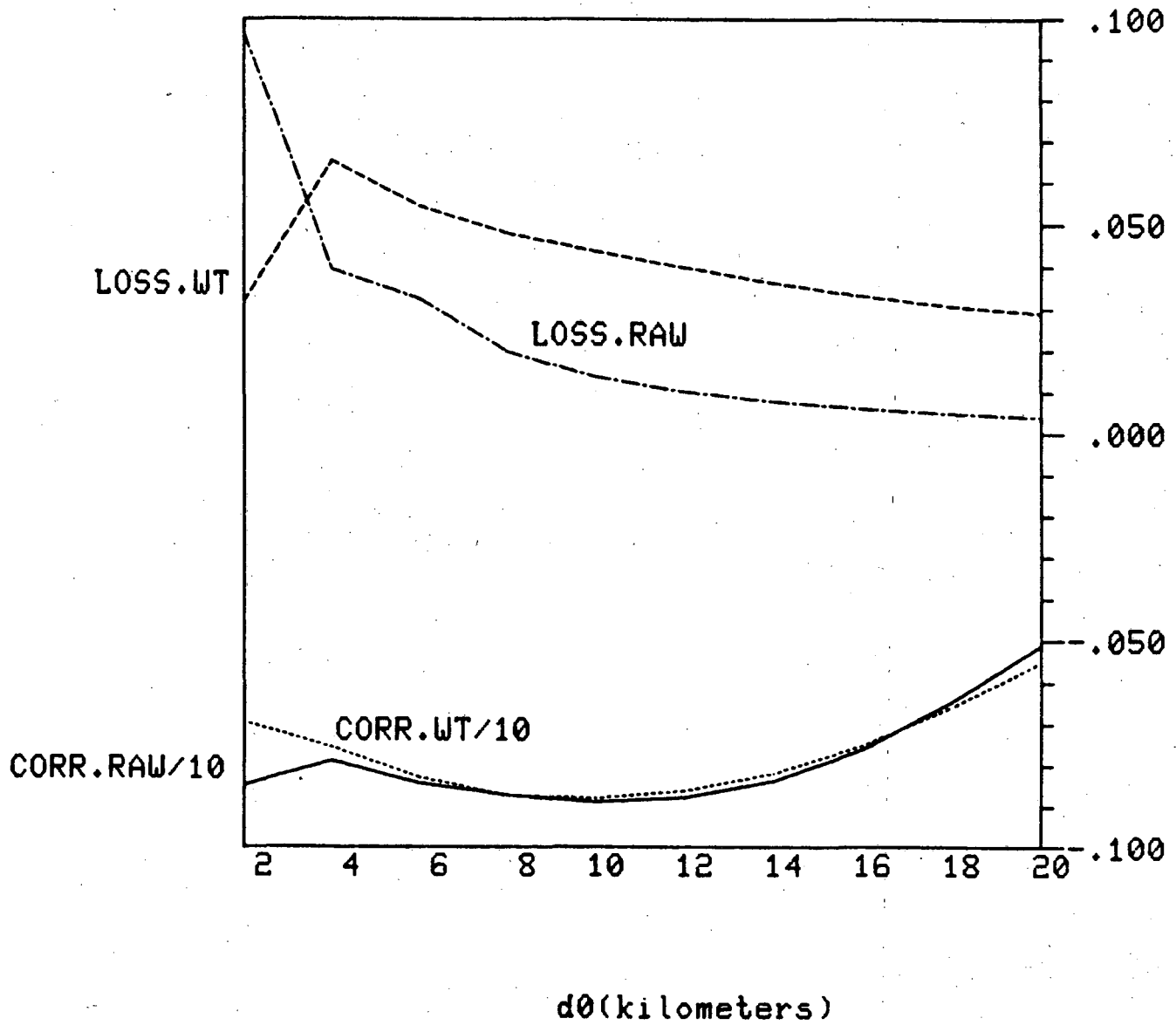


Figure 6. Los Angeles County



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