Title
PREDICTING A CONTINUOUS SPATIAL VARIABLE FROM DISCRETE POINT MEASUREMENTS

Permalink
https://escholarship.org/uc/item/1bt1v18j

Authors
Johnson, L.D.
Merrill, D.W.
Selvin, S.

Publication Date
1982-03-01
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.
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.
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.
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 su-
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

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

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

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

where $x_i$ is the log of the observed value of the $i$th 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:**

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

**Loss Function:**

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

**Weighted Correlation Coefficient:**

\[
\frac{\sum_{j=1}^{i=n} p_j W_j (x_j - \bar{x})(e_j - \bar{e})}{\sqrt{\sum_{j=1}^{i=n} p_j W_j (x_j - \bar{x})^2 \sum_{j=1}^{i=n} p_j W_j (e_j - \bar{e})^2}}
\]
Weighted Loss Function:
\[ \sum_{j=1}^{n} p_j W_j (x_j - e_j)^2 \]
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
\[ W_j = \frac{\sum_{i=1}^{n_j} w_{ij}}{\sum_{i=1}^{n_j} \sum_{j=1}^{n_y} w_{ij}} \]

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

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

Table 3. Los Angeles County

<table>
<thead>
<tr>
<th>$d_0$ (km)</th>
<th>Unweighted Correlation Coefficient</th>
<th>Weighted Correlation Coefficient</th>
<th>Unweighted Loss Function</th>
<th>Weighted Loss Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>-0.848</td>
<td>-0.695</td>
<td>0.0964</td>
<td>0.0319</td>
</tr>
<tr>
<td>4</td>
<td>-0.788</td>
<td>-0.755</td>
<td>0.0396</td>
<td>0.0654</td>
</tr>
<tr>
<td>6</td>
<td>-0.843</td>
<td>-0.829</td>
<td>0.0325</td>
<td>0.0544</td>
</tr>
<tr>
<td>8</td>
<td>-0.872</td>
<td>-0.874</td>
<td>0.0199</td>
<td>0.0481</td>
</tr>
<tr>
<td>10</td>
<td>-0.886</td>
<td>-0.879</td>
<td>0.0139</td>
<td>0.0438</td>
</tr>
<tr>
<td>12</td>
<td>-0.877</td>
<td>-0.861</td>
<td>0.0103</td>
<td>0.0400</td>
</tr>
<tr>
<td>14</td>
<td>-0.837</td>
<td>-0.819</td>
<td>0.0078</td>
<td>0.0363</td>
</tr>
<tr>
<td>16</td>
<td>-0.761</td>
<td>-0.752</td>
<td>0.0060</td>
<td>0.0332</td>
</tr>
<tr>
<td>18</td>
<td>-0.646</td>
<td>-0.659</td>
<td>0.0048</td>
<td>0.0307</td>
</tr>
<tr>
<td>20</td>
<td>-0.512</td>
<td>-0.552</td>
<td>0.0039</td>
<td>0.0290</td>
</tr>
</tbody>
</table>

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.
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


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
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

![Graph showing CORR.RAW/10, CORR.WT/10, LOSS.WT, and LOSS.RAW against d0 (kilometers).]
Figure 5. State of California

- CORR.RAW/10
- CORR.WT/10
- LOSS.WT
- LOSS.RAW

dθ (kilometers)
Figure 6. Los Angeles County

- LOSS.WT
- LOSS.RAW
- CORR.WT/10
- CORR.RAW/10

`dθ(kilometers)`
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.