Analysis of Spokane CO data

Peter Guttorp

NRCSE

Technical Report Series

NRCSE-TRS No. 002

Analysis of Spokane CO data

Peter Guttorp

National Research Center
for Statistics and the Environment

1. Introduction

This is the second product produced under agreement C9700173 between the author and the Washington State Department of Ecology. In this paper we analyze some data on ambient CO concentrations collected by Ecology (four reference sites) and the Spokane County Air Pollution Control Agency (varying numbers of portable samplers). The monitoring was made to determine whether the four reference sites reasonably represent the CO concentrations in Spokane. The Empire Ford reference site is a particular concern in terms of whether it reflects representative concentrations for downtown Spokane.

An additional question of interest in this report is whether there were locations among those monitored with portable CO samplers that had significantly higher CO concentrations than the four permanent monitoring sites when measured CO concentrations are greater than 7 ppm.

2. Data analysis

In this section I present the basic data analysis, and conclusions relative to the study questions. I look at collocated measurements using the same and different measurement techniques. Some hypotheses needing further study are developed from the data and will be addressed in the next section.

2.1 Data used

My study focuses on portable samplers 1-20 for the winter of 1995-96. I never received data pertaining to the winter of 1996-97, as pointed out in the first report for this contract. The sampling program started in December 1995, but the reference data for 1995 were received too late to be used in many of the analyses. I converted the reference data to 8 hour averages by using the hours 16-24, as done by Dames and Moore (1996).

The locations of the sampling stations were obtained from the map produced by Dames and Moore (1996). I used this map to determine locations (in cm from lower left-hand corner of the mapped area). Portable samplers 10 and 16 were located beyond the boundary of the map, and are therefore only used for comparison of collocated samplers. As mentioned in Dames and Moore (1996), samplers 1, 10 and 19 are primary samplers, and the collocated samplers 17, 16 and 20 are used mainly for comparison purposes.

The data used for the analysis of the portable samplers was “Recorder PPM”. The relationship between this measurement and “Analyzer CO Concentration” was very strong, and better the larger the measurements were. All measurements with “Void?” code 1 were recorded as missing.

The following table contains summary statistics for the portable samplers.

Table 1: Summary statistics for portable samplers. 51 sampling days Dec 7, 95-Feb 20, 96.

Sampler	1	2	3	4	5	6	7	8	9	10
Mean	2.34	3.84	2.75	2.99	2.65	2.34	2.64	2.87	2.85	1.65
SD	1.15	1.90	1.52	1.56	1.29	1.30	1.63	1.64	1.70	1.01
Max	5.9	9.3	6.7	7.2	6.1	5.6	7.3	7.3	7.3	4.3
Missing	9	3	3	6	3	0	3	8	6	3

Sampler	11	12	13	14	15	16	17	18	19	20
Mean	1.90	1.90	2.40	3.36	1.44	1.56	2.35	2.43	3.95	3.97
SD	0.79	1.18	1.15	1.82	0.71	0.94	1.31	0.94	1.91	1.81
Max	4.5	4.6	5.1	8.2	3.7	4.3	6.0	4.9	9.7	9.4
Missing	1	4	4	2	3	4	2	6	5	4

Table 2: Summary statistics for reference sites. 76 sampling days, Dec 7, 95 – Feb 20, 96.

Site	BD Tavern	Empire Ford	Hamilton Street	Spokane Club
Mean	2.34	3.49	3.43	2.49
SD	1.61	2.36	2.29	1.46
Max	6.1	9.6	8.7	5.3
Missing	3	2	2	1

Generally speaking it appears that the means and maxima for the portable stations and for the reference sites are similar in size, particularly for sites that are close together, while the reference sites tend to have somewhat larger standard deviations. .

Standard data analytic techniques, including looking at histograms and boxplots, and the apparent linear relationship between means and variances, indicate that a data transformation may be beneficial in some analyses. A square root transformation was therefore considered for some of the work in section 3, as square root of CO concentrations tend to have a more nearly symmetric distribution.

2.2 Two-way decomposition

A simple and robust way of summarizing temporal and spatial effects is an additive two-way decomposition. Basically, this method looks at the median value in each row (day) and column (site) of the data matrix. The implementation in Splus, using the command twoway, actually implements an additive least absolute value fit to the table. The method easily handles even a fairly large proportion of missing values, and is insensitive to long-tailed distributions. We can think of each observation as composed of a grand effect, a site effect, a day effect, and a noise term. Looking at the largest day effects is a way of identifying high concentration days, while the large site effects corresponds to locations that tend to have high concentrations.

For the portable samplers, having a grand effect of 1.87, the highest concentration days in January-February of 1996 are 2/1 (effect 3.78), 2/12 (3.23), 2/15 (3.18), 2/13 (3.08) and 2/14 (1.98), while the stations with highest effects are 19 and 20 (both 1.65), 2 (1.35) and 7 and 14 (0.55). The corresponding effects for the reference samplers were a grand effect of 2.35, with high values observed on 2/12 (5.22), 2/14 (3.65), 2/13 (3.50), 2/15 (3.36), 2/1 (3.18). Hence the same five highest days are identified with both samplers, albeit in different order. The positive site effects for the reference samplers are for Empire Ford (0.99) and Hamilton Street (0.286). Since sites 19 and 20 are collocated with Empire Ford, and site 2 is the portable sampler closest to Hamilton Street, we see substantial agreement between the two data sets.

Figure 1 (all figures are collected at the end of the report) illustrates the day effects estimated for the January-February data, both for the reference sites (solid line) and the portable samplers (dotted line). Clearly the two effects track each other well. The sizes of the peak effects for the portable samplers is, in effect, pulled down by the number of sites with relatively low values, as compared to the reference sites.

2.3 Comparison of collocated stations

Figure 2 shows the scatter plot of measurements from the collocated portable samplers: 1 and 17, 10 and 16, and 19 and 20. The relationships are strong, with occasional outliers. The following table contains the mean and the SD of differences between these stations. This is an indication of the size of the measurement error.

Table 3: Collocated measurement summaries.

Pair	1 vs. 17	10 vs. 16	19 vs.20	18 vs. Spokane Club	19 vs. Empire Ford
Mean difference	-.055	.096	.023	-.373	.013
SD of difference	.156	.328	.376	.224	.848
Correlation	.993	.951	.982	.976	.909

Figure 3 shows the scatter plots for collocated portable samplers and reference sites: sampler 19 and Empire Ford, sampler 18 and Spokane Club. The means and SDs of differences as well as correlations are given in Table 3. While site 19 does a reasonable job in predicting the Empire Ford measurements, there is some concern that the Spokane Club measurements are uniformly higher than those from sampler 18. Since the data are autocorrelated (see the Discussion section below), once a sampler measurement is substantially below the corresponding reference site measurement, there is a tendency for this to persist. However, another possible explanation is that there is a bias in the sampler measurements compared to the reference site measurements, in spite of the good calibration results. We will assess this possibility in the following section.

3. Assessment of network bias

3.1 The method of kriging

Kriging is a method for prediction of spatial data, developed by a South African mining engineer, and is the foundation for the statistical subfield of geostatistics. A statistically oriented summary of the theory can be found in Cressie (1991). Basically, kriging is a version of least squares prediction, taking into account the spatial dependence of the data. In Splus, kriging is implemented using two functions, `krige` and `predict.krige`.

The kriging method consists of first determining the spatial covariance, or a closely related quantity, the variogram, from observed data at given locations. The choice of variogram is discussed in subsection 3.2. Once a variogram has been estimated, it is possible to predict values at any site using generalized least squares (with the estimated variogram function providing the covariance structure). The computational approach is particular to the geostatistical approach. An important feature of the kriging approach is that it is straightforward to produce prediction standard errors.

In this section kriging is used to predict for each day of observations the two sites (Hamilton Street and BD Tavern) which do not have collocated samplers from the portable sampler network. The predictions for the other two reference sites are identical to the measurements from the collocated sampler, and is therefore uninteresting. If there is a systematic difference between the two types of samplers, this should show up over the whole range of kriging predictions.

As mentioned in the data analysis section of this report, the CO concentration data show a tendency towards skewness, and a square root transformation is expected to improve statistical performance of Gaussian-based methodology. In order to convert the predictions made on the square root scale to predictions on the raw scale, simple formulae are available. Let ξ be the square root scale prediction, and σ_ξ the prediction standard error. Then the raw scale prediction is $\xi^2 + \sigma_\xi^2$, with a prediction standard error of $2\xi\sigma_\xi$ (at least for small values of σ_ξ).

3.2 Variogram estimation

The simplest approach to variogram estimation is to assume that the covariance structure is isotropic, i.e., the same in all directions. Since there is insufficient amounts of data to do a detailed study of this assumption, I will just make this simplifying assumption.

As mentioned above, the analysis will be done using square roots of concentrations. Figure 4 shows pairwise portable sampler covariances as a function of distance (in cm on the Dames and Moore map, which does not have an absolute scale). A fitted exponential covariance function is also shown in the figure. Statistical packages generally deal with missing data for covariance calculations in an inefficient manner: all rows with any missing data are ignored. However, since covariance is a pairwise measure, one should use all pairs of columns without missing data in the estimation of the covariance. In this case both approaches yield similar results.

Once the covariance is estimated, the variogram is obtained by subtracting the covariance from the overall variance, estimated by the average of the diagonal of the covariance matrix. However, in Splus this does not actually need to be done. One just specifies the exponential variogram, and computes the sill (the variance estimate), the nugget (the sill minus the estimated covariance at lag 0), and the range (the value for which correlation goes below 0.1) from the data.

3.3 Kriging results

Figure 5 shows the kriging predictions for BD Tavern. The bars are two prediction standard errors above and below the prediction, and the dots are the observed values. It appears that the surrounding portable samplers are able to predict the BD Tavern values successfully. The predictions for the Hamilton Street reference station in Figure 6, on the other hand, shows a tendency to underpredict on high CO days. In other words, the measurements from nearby samplers (these have the largest influence on the predictions) have trouble finding the highest values. This is not surprising, since kriging is a smoothing method, and we are dealing with a random field which is not very smooth, but rather have substantial spikes at a few locations, while being smooth at most other locations. On low and medium CO days the predictions are accurate. The interpretation is that there is no systematic bias between the two measurement networks.

The basic difference between doing the kriging on the raw scale and on the square root scale lies not in the predictions, but in the standard errors. Very similar conclusions would be drawn even if the square root transformation had not been employed. However, the standard errors produced from the square root scale calculations should be considered more accurate.

4. Assessment of downtown study

The second study was based on locating 17 portable samplers in the downtown region, in order to assess whether or not the Empire Ford site is representative of downtown CO pollution. I do not consider this question well formulated. The CO pollution field appears to be very spiked, and samplers within a block of each other can easily be measuring very different quantities. The data from the portable samplers indicate clearly that the samplers near the Empire Ford site consistently have higher values than the rest of the downtown sites. There is no indication of a systematic bias (within the uncertainty of the data) for the measurements at this site. Table 4 makes this point.

All maxima were obtained at sites within a block of the Empire Ford site, and except for the day with the lowest value (2/27) at one of the samplers (19,20 or 21) at or across the street from the Empire Ford site. In this short study there were no measurements above 7 ppm.

Table 4: Summary of data from downtown study.

Day	2/26	2/27	2/28	2/29	3/1	3/4	3/5	3/6	3/7	3/8
Max site	20	22	21	20	19	21	20&29	21	21	21
Max value	3.8	1.3	2.6	6.1	4.5	2	2.8	4.2	3.5	3.8
Empire Ford	4.0	1.1	2.3	6.9	5.2	1.9	3.1	2.1	2.3	3.0
Hamilton Street	2.7	2.0	3.2	6.7	3.4	2.5	2.2	2.0	2.6	3.0

5. Conclusions and discussion

There is no evidence in the analysis above other than that the Empire Ford site is a well chosen monitoring site. It consistently has the largest or among the largest measurements in both networks. In addition, the Hamilton Street site shows very similar patterns to those of the Empire Ford site. Hence the high concentrations at the Empire Ford site are not due to an inappropriate point source. The detailed downtown study also indicates clearly that the highest downtown concentrations tend to be at the Empire Ford site. In other words, the site is not representative of downtown concentrations in general, but is convincingly representative of large downtown concentrations. There is no indication that any other downtown site has potential for larger values than the Empire Ford site at high CO days.

The daily 8-hour maxima from the reference sites show substantial amounts of autocorrelations. An autoregressive model of order 1 appears to be a suitable fit to the data. The main consequence of the autocorrelation is that it is easy to overestimate the actual sample size. Each autocorrelated observation corresponds to less than one iid observation.

The inherent uncertainty in the kriging predictions is substantial, with approximate 95% confidence intervals of the order of 2 ppm. Hence, the observed potential bias at Spokane Club is much smaller than what could be predicted with these data. In order to improve the precision more samplers (not more observations in time) are needed. I do not think this is a worthwhile investment.

There is a dearth of portable samplers south of the reference sites. This may be a concern, as there appears to be a substantial number of days with southerly winds. I do not know the Spokane geography well enough to assess whether or not this is an important concern.

6. References

Cressie, N. (1991): *Statistics for Spatial Data*. New York: Wiley.

Dames and Moore (1996): *Spokane Metropolitan Area Carbon Monoxide Saturation Study Report.* Draft report, May 13 1996, for Spokane County Air Pollution Control Authority.

Figure legends

Figure 1. Day effects from two-way fits of reference stations (solid line) and portable samplers (dotted line).

Figure 2. Comparison of data for collocated portable samplers.

Figure 3. Comparison of data for portable samplers collocated with reference stations.

Figure 4. Fitted covariance for primary portable samplers (excluding sampler 10) as a function of map distance in cm. The line corresponds to the fitted covariance function. The analysis is done on the square root scale.

Figure 5. Kriging predictions for BD Tavern site. The lines are two kriging standard errors below and above the prediction, while the dots are the observed data.

Figure 6. Kriging predictions for Hamilton Street site. Details as in Figure 5.

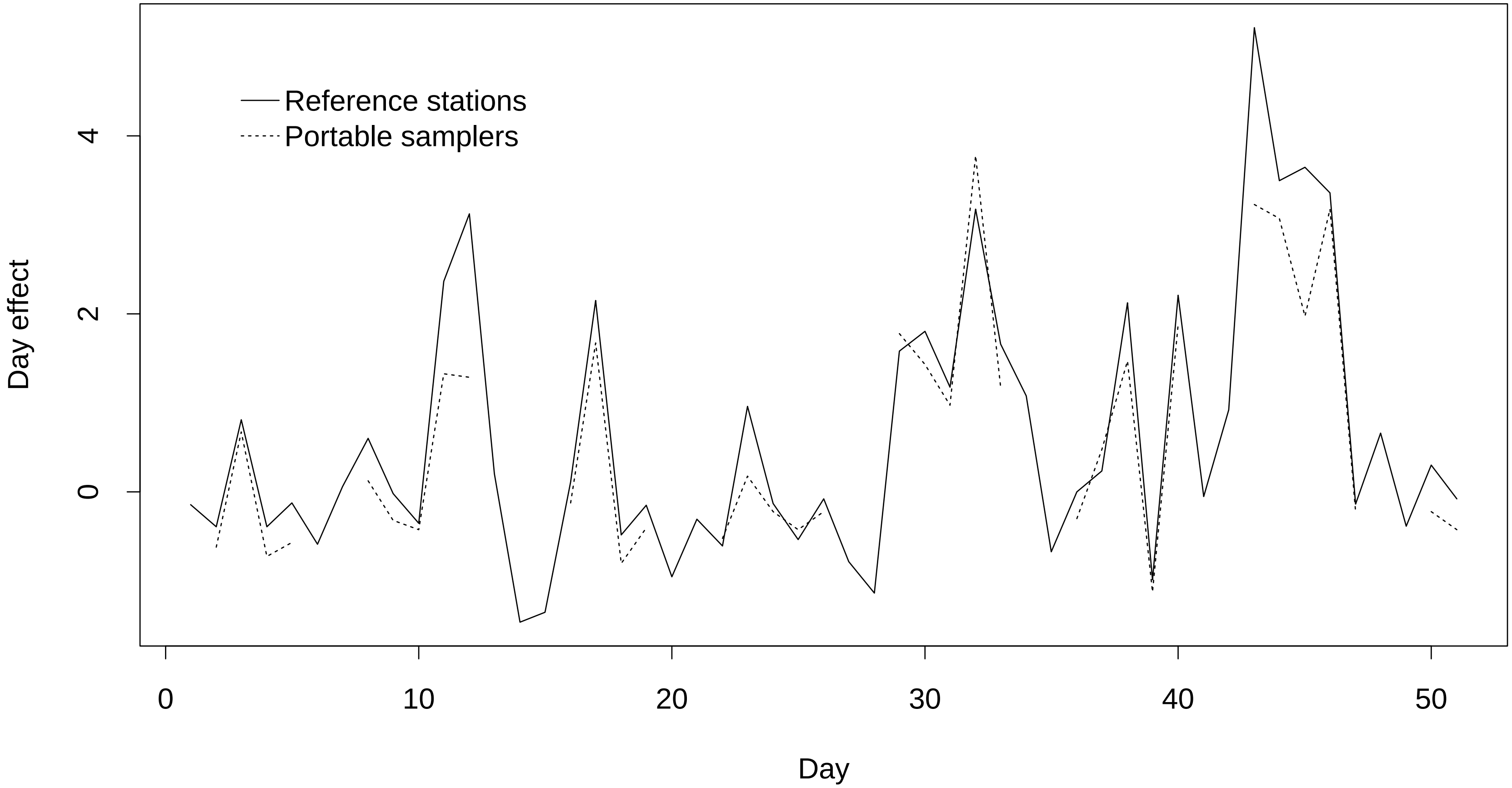

Figure 1: Day effects from two-way fits of reference stations and portable samplers

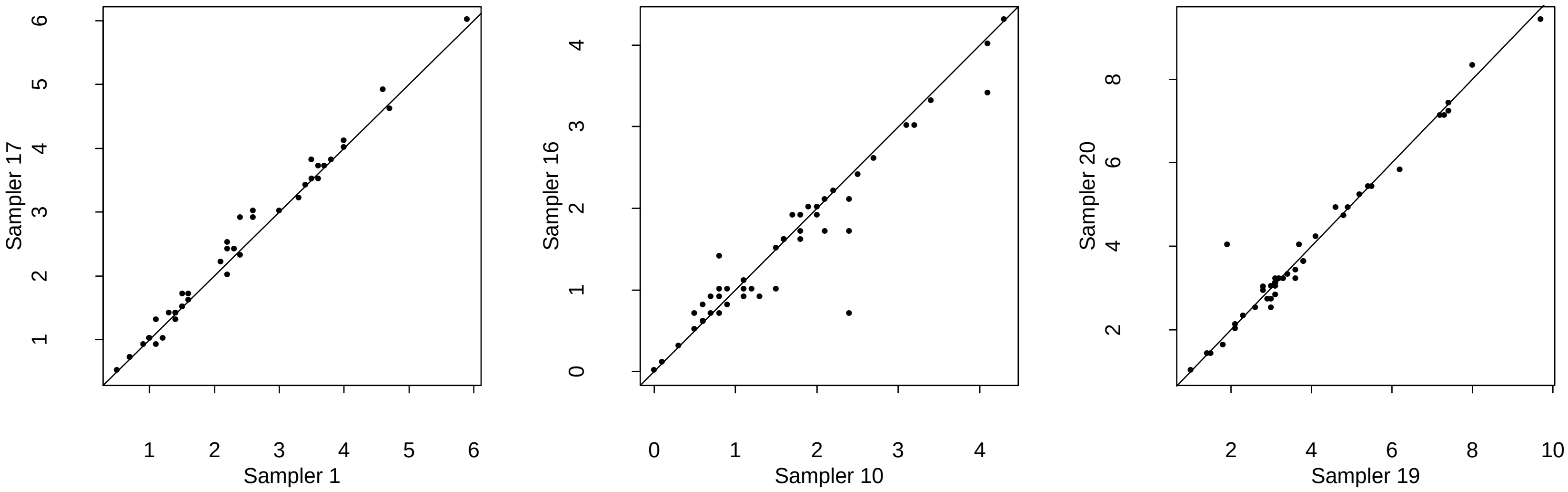

Figure 2: Comparison of data for collocated portable samplers

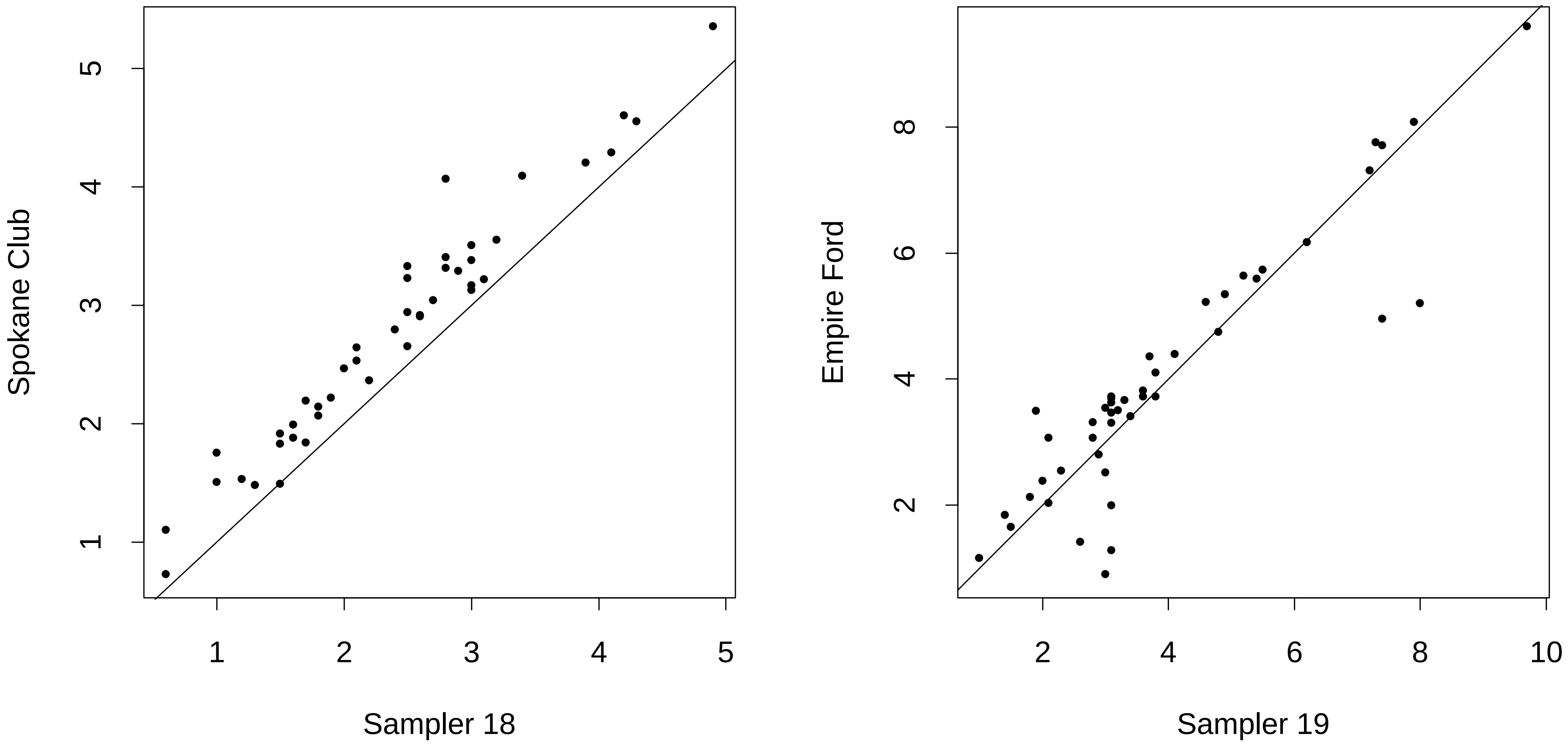

Figure 3: Comparison of data for portable samplers collocated with reference stations

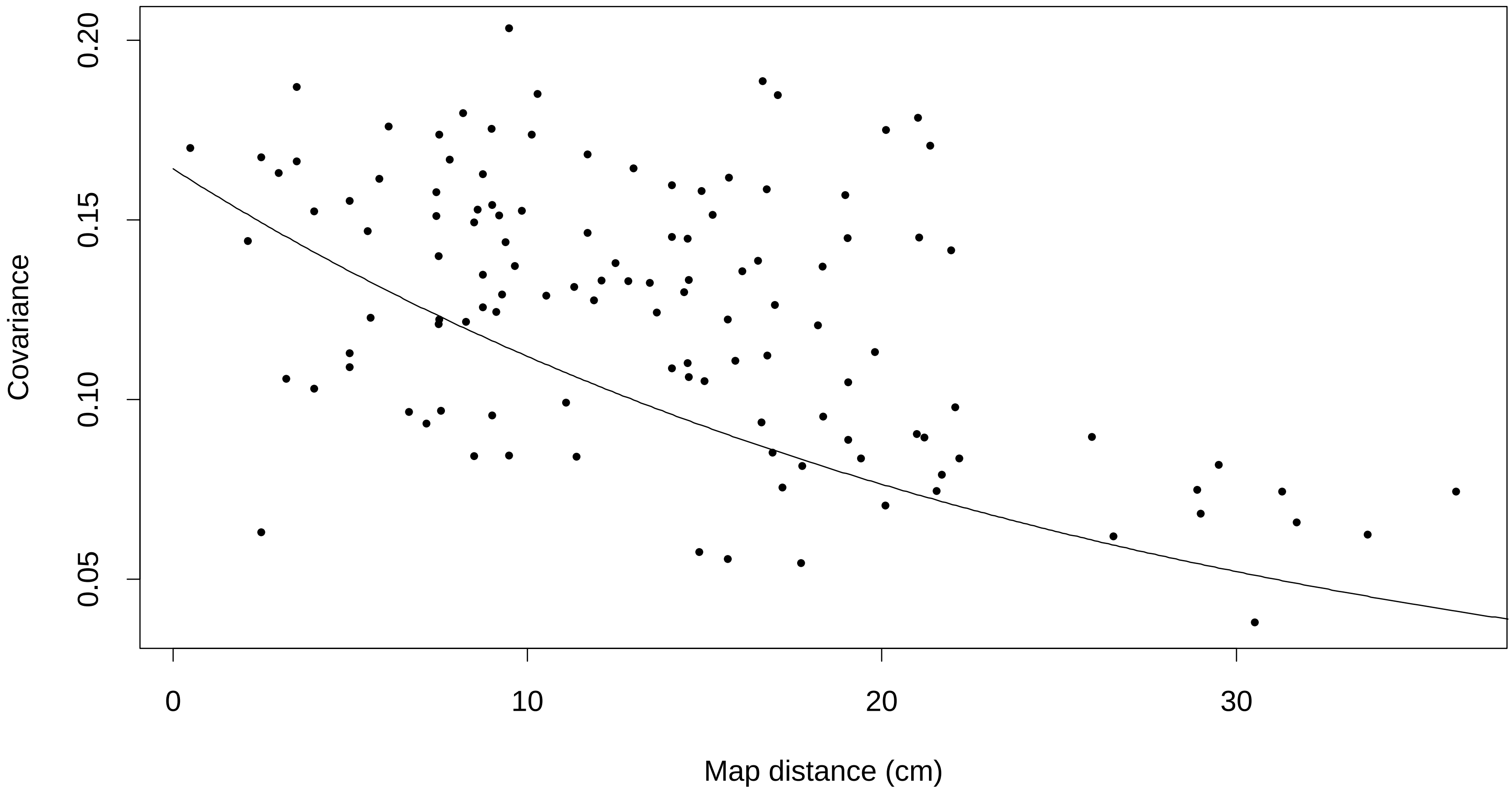

Figure 4: Fitted covariance for primary portable samplers (excluding sampler 10) as a function of map distance in cm. The line corresponds to the fitted covariance function. The analysis is done on the square root scale.

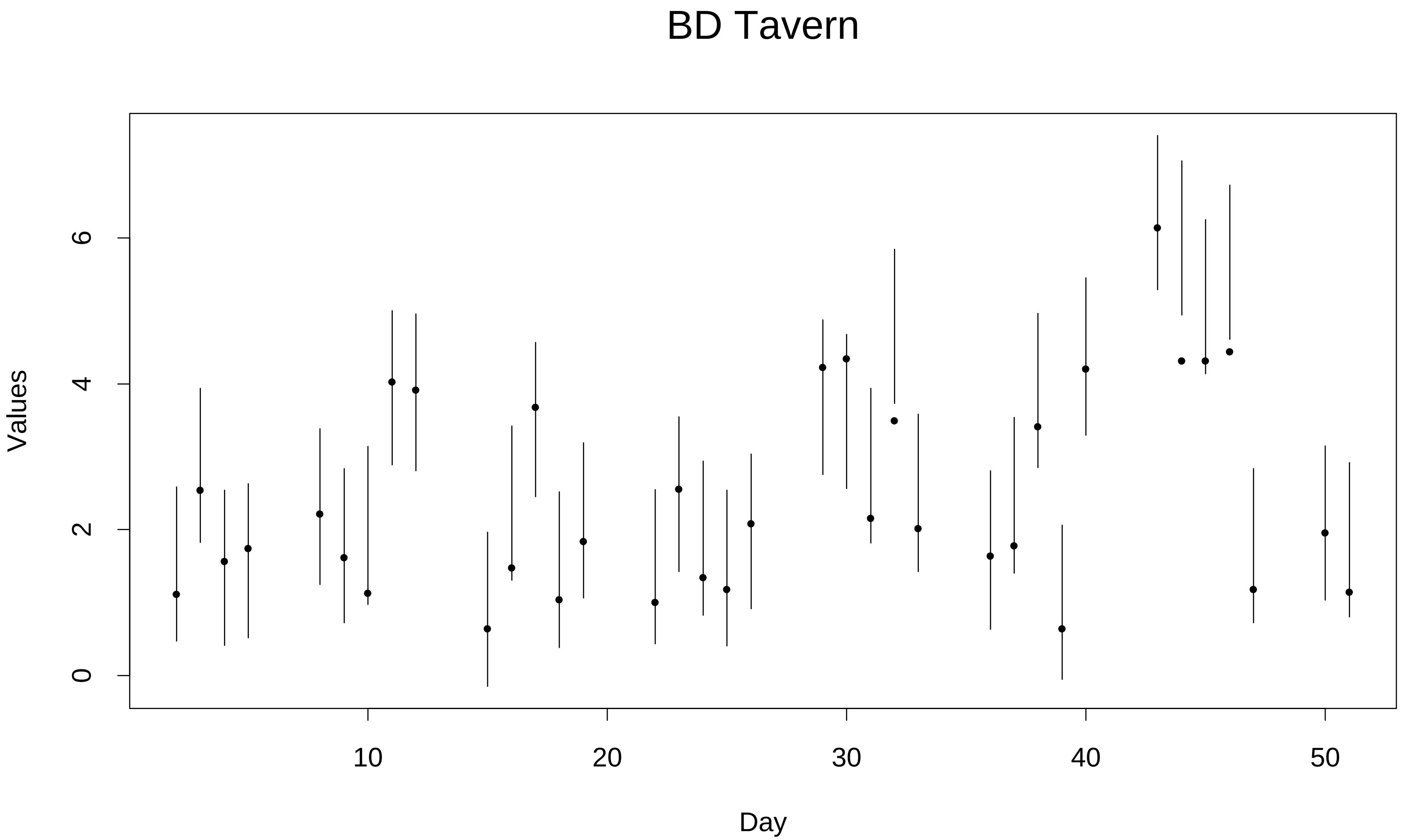

Figure 5: Kriging predictions for the BD Tavern site. The lines are two kriging standard errors below and above the predictions, while the dots are the observed data.

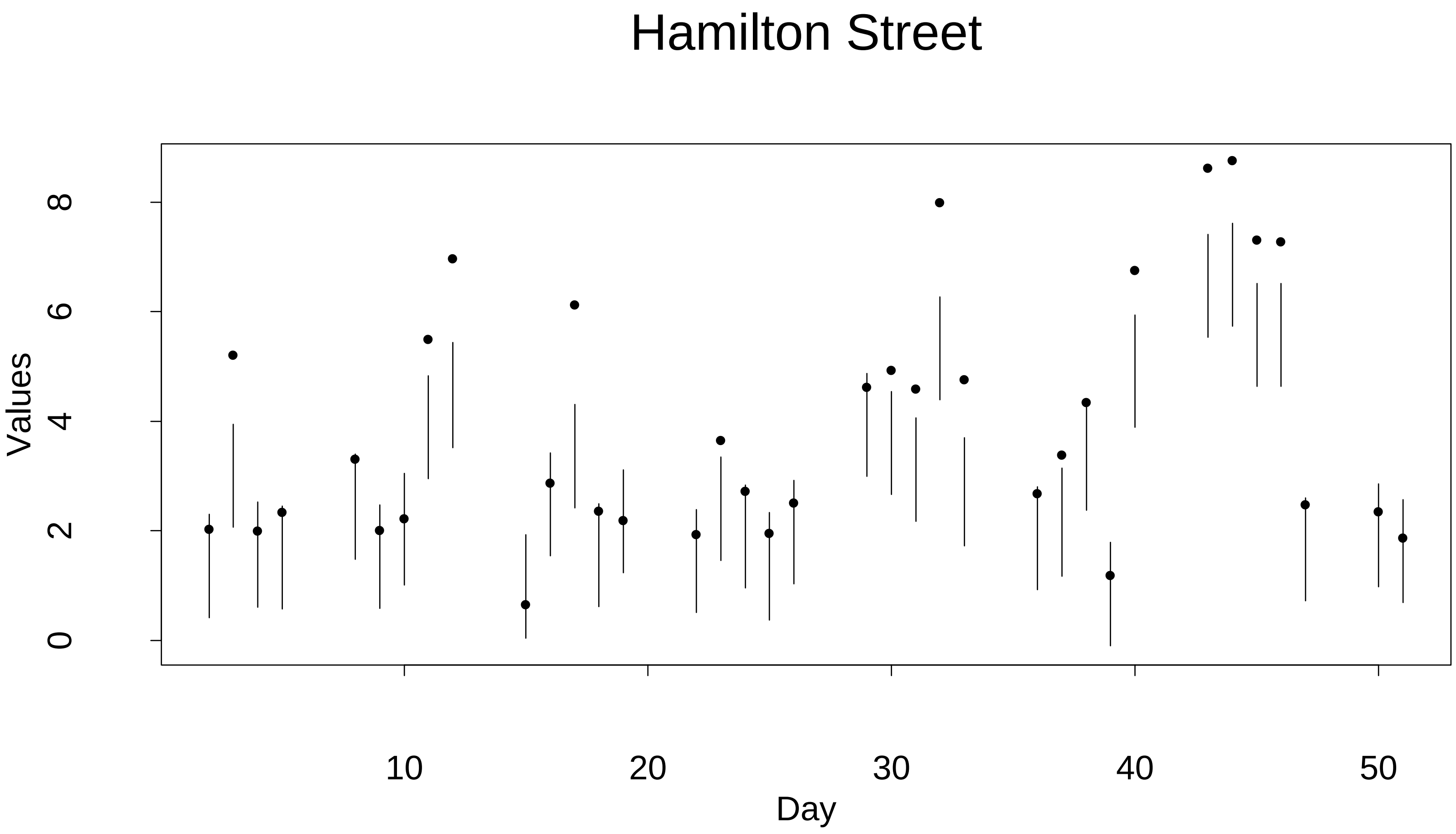

Figure 6: Kriging predictions for Hamilton Street site. Details as in Figure 5