Ecol Res (2012) 27: 1–13
DOI 10.1007/s11284-011-0863-y

ORIGINAL ARTICLE

Jung-Joon Park · Key-Il Shin · Joon-Ho Lee
Sung Eun Lee · Woo-Kyun Lee · Kijong Cho

Detecting and cleaning outliers for robust estimation of variogram models in insect count data

Received: 12 February 2011 / Accepted: 13 June 2011 / Published online: 7 July 2011

Abstract Outlier detection and cleaning procedures were evaluated to estimate mathematical restricted variogram models with discrete insect population count data. Because variogram modeling is significantly affected by outliers, methods to detect and clean outliers from data sets are critical for proper variogram modeling. In this study, we examined spatial data in the form of discrete measurements of insect counts on a rectangular grid. Two well-known insect pest population data were analyzed; one data set was the western flower thrips, *Frankliniella occidentalis* (Pergande) on greenhouse cucumbers and the other was the greenhouse whitefly, *Trialeurodes vaporariorum* (Westwood) on greenhouse cherry tomatoes. A spatial additive outlier model was constructed to detect outliers in both the isolated and patchy spatial distributions of outliers, and the outliers were cleaned with the neighboring median cleaner. To analyze the effect of outliers, we compared the relative nugget effects of data cleaned of outliers and data still containing outliers after transformation. In addition, the correlation coefficients between the actual and predicted values were compared using the leave-one-out cross-validation method with data cleaned of outliers and non-cleaned data after unbiased back transformation. The outlier detection and cleaning procedure improved geostatistical analysis, particularly by reducing the nugget effect, which greatly impacts the prediction variance of kriging. Consequently, the outlier detection and cleaning procedures used here improved the results of geostatistical analysis with highly skewed and extremely fluctuating data, such as insect counts.

Keywords Variogram models · Spatial additive model · Outlier cleaner · Western flower thrips · Greenhouse whitefly · Box–Cox transformation

J.-J. Park
Institute of Life Science and Natural Resources,
Korea University, Anam-dong, Sungbuk-ku,
Seoul 136-701, Korea

K.-I. Shin
Department of Statistics, Hankuk University
of Foreign Studies, Yongin 449-791, Korea

J.-H. Lee
Entomology Program,
Department of Agricultural Biotechnology,
Seoul National University, Seoul 151-921, Korea

S. E. Lee
Nanotoxtech Inc, #906, Gyeonggi Technopark,
Ansan 426-901, Korea

W.-K. Lee · K. Cho (✉)
Division of Environmental Science
and Ecological Engineering, Korea University,
1-5ka Anam-dong, Sungbuk-ku, Seoul 136-701, Korea
E-mail: kjcho@korea.ac.kr
Tel.: +82-2-32903064
Fax: +82-2-9251970

Introduction

The spatial dependency of insect pests should be taken into consideration when ecologically based pest management programs are implemented because most insect pests are observed at specific points in their habitats and appear to be concentrated within a given spatial distribution (Southwood 1978; Binns and Nyrop 1992; Park et al. 2004). To implement strategies to control insects, it is often necessary to characterize whether the insect species is equally abundant throughout the entire sampling space and whether its abundance at sampling locations is positively associated with the occurrence of certain spatially aggregated environmental conditions (Nansen et al. 2003). Geostatistics is a useful tool for interpreting spatial patterns and the dependency of organisms. Variogram, which is widely used in practical geostatistics, is the average degree of similarity between sample values as a function of their separation distances (Isaaks and Srivastava 1989; Rossi et al. 1992; Cressie 1993; Kitanidis 1997). If the two values are close together, their difference will typically be small, and as the values get farther apart, their differences become larger as well as the variance of the difference.

For variogram modeling, the most important step is appropriately calculating the experimental variogram from the field spatial data (Olea 2007). Once the appropriate calculation has been performed, the best variogram model can be selected without the need for any special or additional analysis. The accuracy and precision of kriged predictions can depend on these fitted variogram models, which in turn may be significantly influenced by the initial method used to estimate the variogram (Olea 2006).

Srivastava and Parker (1989) reported that the variogram often erratically and unreliably characterizes the spatial dependency, especially when the data are highly skewed or when the data are aggregated, which is likely to be the case in many insect count data sets. Geostatistics involves the fitting of spatially continuous models to spatially discrete data (Diggle et al. 2010). The variogram model may also provide an incomplete description of the spatial pattern because the variogram is strongly affected by small-scale or local mean and variance differences (Rossi et al. 1992).

Some transformation methods have been developed to reduce the effect of skewed and biased data for geostatistical analyses. Bardossy and Kundzewicz (1990) reported that a power transformation will change highly skewed raw data so that they more closely resemble the normal distribution, leading to more robust experimental variograms. It has often been reported that environmental variables are lognormal (Sichel 1952; Miesch and Riley 1961) or positively skewed (Zhang et al. 1995; Zhang and Selinus 1998) and that data transformation is necessary to normalize such data sets. Box–Cox data transformations have been widely accepted as effective data transformation methods, conferring normality on skewed data (Box and Cox 1964). Park et al. (2004) evaluated mathematical restricted variogram models using greenhouse whitefly data after conducting various data transformations.

Data transformation of raw data tends to improve approximation to normality by reducing the values of skewness; however, data transformation does not seem to solve the problems caused by outliers (Kerry and Oliver 2007). In statistics, the term outliers refer to abnormally large or small values sometimes seen on histograms. The problem when outliers are present is that the mean and the variance are disproportionately affected by a few extreme values. These outliers result in the inappropriate calculation of experimental variograms. Variogram noise resulting from outliers can even completely mask the spatial structure and produce a pure nugget effect (Liebhold et al. 1993).

Statistical parametric methods that assume a known underlying distribution of the observations (e.g., Hawkins 1980; Rousseeuw and Leory 1987; Barnett and Lewis 1994) or are based on statistical estimates of unknown distribution parameters (Hadi 1992) are frequently used to detect outliers. These methods flag outliers as those observations that deviate from the model assumptions. Another simple way to detect outliers in spatial data sets is based on visualization, namely, illustrating the distribution data difference in a figure and visually identifying the points in particular portions of the figure that are outliers included in frequency graphs, such as histograms and box plots (Goovaerts 1997; Zar 1999). One potential problem with these methods is that they use simple arithmetic averages to estimate the overall behavior of a set of neighbors, and they do not consider the impact of spatial relationships on the neighborhood comparison. The *h*-scattergram is useful for observing outliers or spurious values and can be part of a clean-up process. However, the decision to remove aberrant data should not be made based on a single lag distance, since the same datum may not yield outlier pairs for the other lags or in other directions (Goovaerts 1997). In addition, the *h*-scattergram method identifies outliers in a subjective manner.

The other methods used to identify outliers depend on performing statistical tests to discover local inconsistency. Breunig et al. (2000) suggested that, in some situations, local outliers are more important than global outliers. They proposed the concept of a local outlier factor, which defines how isolated an object is with respect to its surrounding neighborhood rather than to the whole data set. In addition, Mugglestone et al. (2000) proposed the neighboring median method to detect and clean outliers in which each outlier is replaced by the median of its four nearest neighbors on the lattice.

In this paper, we evaluated a novel method to detect and clean outliers for geostatistical analysis of insect count data. Two species of insect populations, the western flower thrips, *Frankliniella occidentalis* (Pergande), and the greenhouse whitefly, *Trialeurodes vaporariorum* (Westwood), were counted in cucumber and cherry tomato greenhouses, respectively. Both species are very serious pests to many greenhouse crops, including cucumbers and cherry tomatoes (Lewis 1973; Mound and Halsey 1978), and they are resistant to many insecticides (Immaraju et al. 1992; Sanderson and Roush 1992). We also compared and validated the resulting geostatistical parameters using the leave-one-out cross-validation method.

Materials and methods

Study plot and sampling

Thrips population

Frankliniella occidentalis populations were monitored in two commercial cucumber greenhouses (TGA and TGB) located near Jeju City (33°29′55″N, 126°29′55″E) on Jeju Island, Korea, from May to July of 1996. Three-week-old greenhouse-grown cucumber (*Cucumis sativus* L.) plants were transplanted on March 5 in each greenhouse. Both greenhouses were 24 × 45 m and surrounded

by potato fields. All greenhouses were managed using standard recommended production practices, including the use of fertilizers and pesticides at the grower's discretion. The surveyed greenhouses consisted of 12–15 beds on approximately 0.8-m centers, and plants were grown using the modified vertical cordon training system (Cho et al. 1998). Horizontal support wires were positioned directly over the row of plants at a height of 1.8–2.0 m. Initially, each plant was trained vertically along and around a plastic twine support and fastened with plastic snap-on clips. As the plant reached the top supporting wire, the grower removed the clips and released the reserved twine, which moved the plant approximately 0.3 m closer to the ground so that the lower section of the stem was on the ground. Therefore, the youngest leaf was always at the top of the canopy.

Cucumber leaves were visually inspected weekly to measure temporal and spatial changes in the density of the thrips population. To determine the temporal abundance of thrips, at least 30 plants were selected randomly, and the numbers of adult and immature thrips from the seventh leaf from the growing tip were counted. Those leaves were fully matured, non-senesced leaves and were located approximately 1.0 m above ground level.

To analyze the spatial distribution patterns of thrips, an approximate quadrat of permanent sampling positions was established for each greenhouse. The sampling array for each greenhouse consisted of 72 visual counts in TGA and 55 in TGB, which were laid out in 12×6 quadrat patterns in TGA and 11×5 quadrat patterns in TGB. The distance between the plant stations was approximately m in the down-row and approximately 5 m in the cross-row. Thus, each quadrat covered approximately 15 m^2 and contained approximately 50 plants. One cucumber plant located at the center of the quadrat was selected, and the numbers of immature thrips were counted on each sampling date. TGA and TGB were surveyed six and eight times, respectively, throughout the growing season.

Whitefly population

Two commercial cherry tomato (*Lycopersicon esculentum*) greenhouses (WGA and WGB) in Buyeo (36°15′55″N, 126°53′56″E), Chungcheongnam-do, Korea, were selected for sampling of *T. vaporariorum* larvae during the growing season of 1998. Each greenhouse was about 3,000 m^2 in size. In each greenhouse, cherry tomatoes were transplanted in early August and early November. Plants were spaced ≈0.3 m apart in a single row on beds (height 0.1 m, width 0.5 m) of soils covered with black polyethylene mulch. The greenhouses surveyed consisted of 12–15 beds on approximately 0.8-m centers, and plants were grown using the ventral cordon system (Papadopoulos 1991). Horizontal support wires were positioned directly over the row of plants at a height of 1.8–2.0 m. Initially, each plant was trained vertically along and around the plastic support twines and tied with plastic snap-on clips. As the plant reached the top supporting wire, the clips were released, allowing the plant to be lowered by approximately 0.3 m such that the lower section of the stem lay on the ground. The lowest foliage was removed to promote flower and fruit production, while the youngest foliage was at the top of the canopy.

Late larvae (3rd and 4th instars) of greenhouse whitefly were counted visually on a seven-leaflet leaf near the middle of each cherry tomato plant that was sampled (approx. 0.8–1.2 m above ground). Approximate grids of 40 permanent sampling locations were established in each greenhouse; each grid measured approximately 35 m^2 and contained approximately 180 plants. The distance between sampling location was approximately 7 m within the rows and approximately 5 m across the rows.

Geostatistical modeling

Data transformations

As in conventional statistics, a normal distribution for the variable under study is desirable in linear geostatistics (Clark and Harper 2000). Although normality may not be strictly required, serious deviations from normality, such as very high skewness and the presence of outliers, can impair the variogram structure and modeling (McGrath et al. 2004). It has often been reported that environmental variables are lognormal (Sichel 1952; Miesch and Riley 1961) or positively skewed (Zhang et al. 1995; Zhang and Selinus 1998), and data transformation is necessary to normalize such data sets.

In our study, the Box–Cox transformation was used to make the data more normally distributed and with less skewness (Box and Cox 1964). The Box–Cox transformation is given by:

$$x' = \begin{cases} x^{\lambda} & \text{if } \lambda \neq 0 \\ \ln(x) & \text{if } \lambda = 0 \end{cases}, \quad (1)$$

where x' is the transformed value, and x is the value to be transformed. For a given data set, the parameter λ is estimated based on the assumption that the transformed values are normally distributed. When $\lambda = 0$, the transformation becomes the logarithmic transformation. Bartlett (1936) proposed the transformation of data as $x' = (x + 0.5)$. This transformation is preferred on theoretical grounds and is especially preferable when some of the observed values are small numbers, particularly for data sets that contain a high number of zero counts. Park et al. (2004) demonstrated that natural logarithmic ($\lambda = 0$) and fourth root transformations ($\lambda = 1/4$) with the actual value of $+0.5$ were better than any other transformations for aggregated insect pest populations. Thus, we applied these two data transformations to the raw data sets before the outlier analysis was conducted.

Outlier detection and data cleaning

Outlier detection and cleaning steps used in our analysis were based on the spatial additive outlier model (Eq. 2), which is a generalization of models by Hawkins and Cressie (1984) as the basis for a robust form of kriging and by Martin and Yohai (1986, 1991) to study outliers in time series:

$$Y_{u,v} = X_{u,v} + Z_{u,v}v_{u,v}, \tag{2}$$

where $Y_{u,v}$, a 'possible outlier included' part, is an observation at the point of (u, v) on the $U \times V$ lattice $(u = 1, \ldots, U;\ v = 1, \ldots, V)$, $X_{u,v}$ is an 'outlier not included-core' part, $Z_{u,v}$ is an assignment process that takes the values 0 and 1, and $v_{u,v}$ is a 'possible outlier part' that is superimposed on X. The v is typically taken to be large compared to the variance of X with mean zero (Nirel et al. 1998; Mugglestone et al. 2000).

The spatial outlier cleaner was calculated as follows:

$$\psi(y_{u,v}; M, g^0, g^1) = \begin{cases} y_{u,v}, & \text{if } \left|y_{u,v} - g^0(y_{u,v})\right| \le M \\ g^1(y_{u,v}), & \text{otherwise} \end{cases}. \tag{3}$$

A g^0 is the median of the observed values. The M value is called the 'turning' constant, and it controls the rate of replacement (Mugglestone et al. 2000). The optimal M value can be derived by finding the minimum of $[\mathrm{var}\{\Psi(M)\} - 1]^2$ (Nirel et al. 1998). An observation at the point of (u, v), $y_{u,v}$, was determined to be an outlier if $\left|y_{u,v} - g^0(y_{u,v})\right| > M$. The data cleaner replaces each outlier by $g^1(y_{u,v})$, leaving the rest of the observation unchanged. A function of g^1 is obtained by using simple median smoothing, replacing each outlier by the median of its four nearest neighbors on the sampling value. The detailed computation of the M value and outlier detection and cleaning protocols can be found in Mugglestone et al. (2000); however, it is not easy to apply other systems for the detail method in given mathematical references, especially calculation of the M value. Thus, we provided the script code for computing the M value using R language (Ihaka and Gentleman 1996) (see Appendix) and calculated M values for outlier detection and cleaning in each sample week and data transformation separately.

Estimation of mathematical restricted variogram parameters

Experimental variograms were calculated both before and after the outliers were cleaned from the data sets. Directionality was not included into the variogram analysis because the data sets used in this study had an insufficient number of paired observations (< 30) within a given direction (Nansen et al. 2003). Therefore, isotropy was assumed and omnidirectional variograms were used for all the data sets throughout this paper.

The experimental variograms were calculated from the data counted, according to the following equation:

$$\hat{\gamma}(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [z(x_i) - z(x_i + h)]^2, \tag{4}$$

where $\hat{\gamma}(h)$ is the estimated experimental variogram value for lag distance h, $N(h)$ is the number of pairs of points separated by h, and $z(x_i)$ and $z(x_i + h)$ are the two data points separated by h. Experimental variograms before and after the outliers were cleaned from the data set were estimated for each greenhouse sampling week. Experimental variograms were calculated using a lag distance of 3.0 m with a tolerance ± 2.0 m for the thrips population and 5.0 m with a tolerance ± 3.0 m for the whitefly population. The most common choice for the lag tolerance is one-half the lag distance between two neighboring classes. This results in an increase in the number of data pairs that can be used in the variogram calculation (Isaaks and Srivastava 1989). In geostatistical analysis, at least 30 data pairs per lag distance are required to adequately estimate the variance (Isaaks and Srivastava 1989), and the maximum lag distance for all variograms should be at least half the shortest dimension of the sampling space (Nansen et al. 2003). In this study, the area surveyed in each cucumber greenhouse for the thrips population and in each cherry tomato greenhouse for the whitefly population was 24×45 m and 44×62 m, respectively, which demonstrated that the variograms could not ideally account for lag distances > 12.0 m for the thrips population and > 22.0 m for the whitefly population, respectively.

The experimental variograms were modeled using the following three mathematical restricted variogram models (Cressie 1993).

The spherical model

$$\gamma(h) = \begin{cases} C_0 + \left(\frac{3h}{2\alpha} - \frac{1h^3}{2\alpha^3}\right)\sigma^2 & \text{for } 0 \le h \le \alpha \\ C_0 + \sigma^2 & \text{for } h \ge \alpha \end{cases}, \tag{5}$$

where C_0 is the nugget variance, $\sigma^2(\sigma^2 > 0)$ is the structural variance, and α ($\alpha > 0$) is the effective range.

The exponential model

$$\gamma(h) = C_0 + \sigma^2 \left[1 - \exp\left(-\frac{h}{l}\right)\right], \tag{6}$$

where C_0 is the nugget variance, $\sigma^2(\sigma^2 > 0)$ is the structural variance, and $l(l > 0)$ is the length parameter, and the effective range is $\alpha \approx 3l$

The Gaussian model

$$\gamma(h) = C_0 + \sigma^2 \left[1 - \exp\left(-\frac{h^2}{l^2}\right)\right], \tag{7}$$

where C_0 is the nugget variance, $\sigma^2(\sigma^2 > 0)$ and $l(l > 0)$ are the structural variance and length parameter, respectively. The effective range α is $\alpha \approx 7l/4$ (Srivastava and Parker 1989; Cressie 1993).

For all three of the mathematical restricted variogram models, $C_0 + \sigma^2$ is commonly referred to as the sill. The range is defined as the distance at which the data are no longer autocorrelated. Lower ranges indicate that the data are correlated only with data in close proximity, and high ranges indicate that the data are correlated over much larger distances. A mathematical restricted variogram model was fitted to an experimental variogram using optimization techniques, in the form of a nonlinear weighted least squares regression, and the three variogram models were evaluated based on the weighted sum of square residuals $(Q * (\theta))$ (Cressie 1985).

Validations with leave-one-out method

In order to properly compare the results of the mathematical restricted variogram parameters, comparison should be made with the same data treatment, because the means and variances of each of the treated data sets are on different scales (Zar 1999). In our study, the performance of the outliner detection and cleaning system was assessed using a leave-one-out cross-validation procedure (Cressie 1993; Kohavi 1995). The leave-one-out cross-validation is a commonly applied method in geostatistics because no reserved data are required for data validation. This approach works by first deleting one data point at a given position and then performing kriging with the remaining sample values to estimate the value at the location of the deleted sample. This procedure is repeated for all of the data points presented in a data set. For the cross-validation results, the final estimates are calculated with the back-transformed value obtained by the leave-one-out procedure. Back-transformation of $\exp(x)$ and x^4 is used for logarithmic and fourth root transformation, respectively.

To validate outlier cleaning using the leave-one-out method, four sampling weeks were selected: one from each greenhouse. Thrips counts from the fifth sample week at TGA and the fourth sample week at TGB were selected; for the whitefly data sets, the ninth sample week at WGA and the second at WGB were selected. The reason for selecting these data sets is that they showed maximum and minimum mean densities in each insect pest population (see Fig. 1).

The original data were then compared with the corresponding leave-one-out estimates for all sample points. Estimation errors were evaluated using the root mean squared error (RMSE), and correlation between the estimates and original data was also investigated. The correlation coefficients between the outlier-cleaned and outlier-included data were compared using Fisher's Z transformation test (Zar 1999). If the relationship between the cross-validation results and the actual values was highly positive, we can conclude that the estimated parameters properly explained the spatial dependency.

Statistical tools

The S-Plus program with SPATIALSTATS module (ver. 6.0, release 2) and SAS PROC NLIN were used to calculate experimental variograms, to estimate the

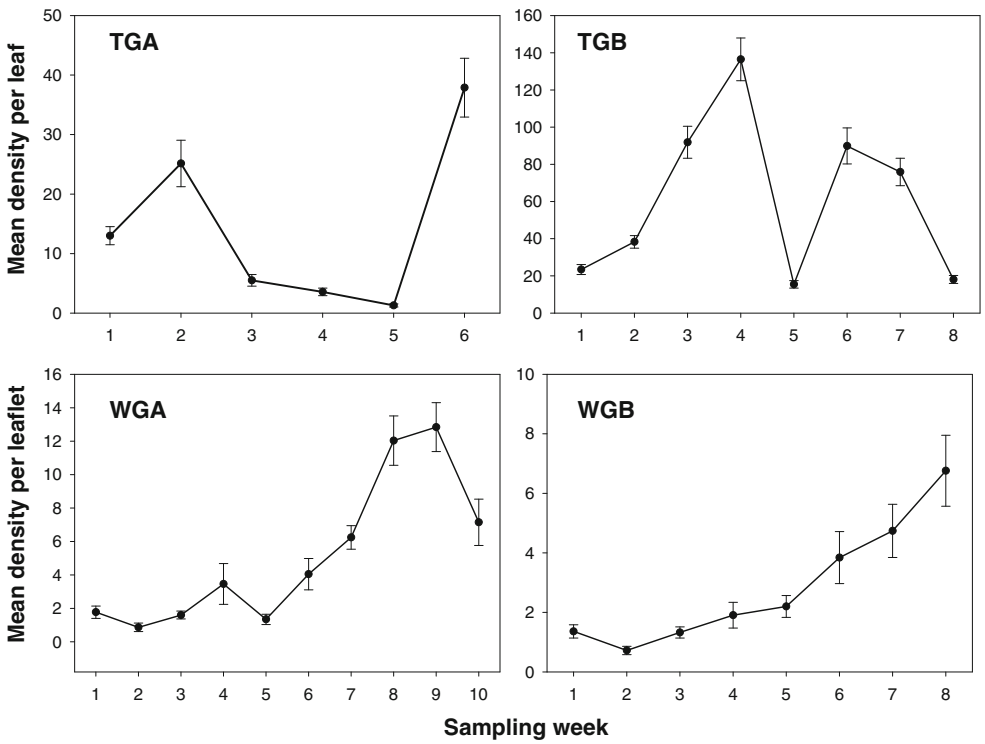

Fig. 1 Mean density of thrips larvae in the two cucumber greenhouses on Cheju Island (*TGA*, *TGB*; *upper panels*) and whitefly larvae in the two cherry-tomato greenhouses in Buyeo (*WGA*, *WGB*; *bottom panels*)

weighted least squares of the mathematical restricted variogram model parameters, and to validate the leave-one-out method (SAS Institute 1996; Kaluzny et al. 1998; Insightful Corporation 2001).

Results

Data description and outlier cleaning

Mean densities of thrips and whiteflies fluctuated between greenhouses and sampling times (Fig. 1). At TGA and TGB, the mean densities of thrips varied from 1.3 to 37.9 and from 15.5 to 136.4 per leaflet, respectively. At WGA and WGB, the mean of whiteflies varied from 0.9 to 12.8 and from 0.7 to 6.8 per leaflet, respectively. The fluctuations in the densities of the thrips were more intense than those of the whitefly densities.

To evaluate the raw data sets, the variance-to-mean ratio (VMR) and the percentage of the zero counts in the data set from each greenhouse and sample week were calculated (Table 1). The VMR provides a convenient measure of the degree of over-dispersion or aggregation of insect populations in the fields. All population count data were highly aggregated (VMR $\gg 1$), which occurs when there are more differences between low and high counts.

More than 20% of the zero counts were observed in nine of 32 count cases, namely, the third, fourth, and fifth sample weeks in TGA, the second, fourth, fifth, and sixth sample weeks in WGA, and the second and fourth sample weeks in WGB greenhouse (Table 1). Park et al. (2004) reported that no data transformations for analyzing the geostatistics were satisfactory for correcting data sets when the empty grids (zero counts) in the sample data consisted of $>20\%$ of the total samples.

Estimation of mathematical restricted variogram models

The outlier cleaning procedure was applied to the transformed data sets with $\ln(x+0.5)$ and $(x+0.5)^{1/4}$. The proportions of the data points identified as outliers are listed in Table 1. These varied between sample weeks and greenhouses, but all the data sets contained at least one outlier. Of all data sets, 50% (16 out of 32 cases) contained $>10\%$ outliers from among the total data points.

The spatial autocorrelation structure of the data sets was explored by estimating experimental variograms ($\hat{\gamma}(h)$) from pooled data sets by insect and greenhouse. By pooling the data sets over time, data from different sampling dates were treated as replicates. This process allowed for more precise estimates of spatial model parameters, especially for the small-scale components (Cressie 1993). Thus, experimental variograms could be fitted to the mathematical restricted variogram models with a robust and precise estimation of model parameters (Table 2).

Calculated nuggets and sills from each data transformation produced different scales; therefore, the relative nugget effect (RNE) was calculated in order to compare each data treatment and mathematical restricted variogram model (Tables 2, 3). The RNE, which is the ratio of the nuggets (C_0) to the sills ($C_0 + \sigma^2$), can be used to evaluate sampling error and fine-scale spatial effects (Isaaks and Srivastava 1989). In all cases of insect count data sets, the RNE for the outliers cleaned was smaller than that when the outliers were included, regardless of the data transformations. In the thrips count data sets, the RNE varied from 0.17 to 0.55 (median 0.24) when outliers were cleaned, but varied from 0.34 to 0.98 (median 0.69) when the outliers were included in the data sets. Similar differences in the RNE were observed between whitefly count data sets.

Table 1 Variance-to-mean ratio, zero count samples, and outliers detected at each sampling week from each greenhouse

Greenhouse	Data description	Sample week									
		1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
TGA	VMR	12.9	43.7	12.5	8.0	3.6	46.4				
	Zero count (%)	6.9	0.0	25.0	27.8	55.6	4.2				
	Outlier (%)	4.2	11.1	6.9	6.9	12.5	12.5				
TGB	VMR	16.7	16.3	44.2	53.4	14.7	57.8	39.7	13.8		
	Zero count (%)	3.6	1.8	0.0	0.0	5.5	0.0	0.0	7.3		
	Outlier (%)	1.8	1.8	3.6	1.8	7.3	10.9	7.3	16.4		
WGA	VMR	3.0	3.0	1.4	17.3	2.8	8.7	3.2	7.2	6.6	10.7
	Zero count (%)	15.0	42.5	12.5	25.0	32.5	20.0	2.5	0.0	0.0	2.5
	Outlier (%)	10.0	22.5	10.0	25.0	10.0	17.5	5.0	5.0	2.5	17.5
WGB	VMR	1.4	1.1	1.1	3.9	2.4	8.0	6.7	8.4		
	Zero count (%)	10.0	32.5	12.5	20.0	10.0	5.0	12.5	7.5		
	Outlier (%)	15.0	12.5	2.5	15.0	10.0	7.5	7.5	20.0		

TGA, TGB, Two commercial cucumber (*Cucumis sativus* L.) greenhouses in which *Frankliniella occidentalis* (thrips) populations were monitored during the growing season of 1996; WGA, WGB, Two commercial cherry tomato (*Lycopersicon esculentum*) greenhouses selected for sampling of *T. vaporariorum* (whitefly) larvae during the growing season of 1998; VMR, variance-to-mean ratio
Numbers of data points in each sample week are 72, 55, 40, and 40 for TGA, TGB, WGA and WGB, respectively

Table 2 Estimated parameters of three mathematical restricted variogram models with outliers cleaned and included *Frankliniella occidentalis* count data

Greenhouse	Transformation	Outlier	Model	Nugget (C_0)	Sill ($C_0 + \sigma^2$)	RNE[a]	Range	$Q^*(\theta)$[b]
TGA	$\ln(x + 0.5)$	Cleaned	Spherical	0.28	1.12	0.25	24.59	0.503
			Exponential	0.25	1.08	0.23	14.83	0.669
			Gaussian	0.31	0.94	0.33	12.53	0.510
		Included	Spherical	1.04	1.29	0.81	21.38	0.972
			Exponential	1.07	1.75	0.61	11.29	0.993
			Gaussian	1.15	1.42	0.81	16.30	0.921
	$(x + 0.5)^{1/4}$	Cleaned	Spherical	0.05	0.29	0.17	26.40	0.024
			Exponential	0.09	0.47	0.19	21.96	0.042
			Gaussian	0.09	0.25	0.36	13.50	0.086
		Included	Spherical	0.22	0.32	0.69	28.43	0.892
			Exponential	0.20	0.58	0.34	29.20	1.095
			Gaussian	0.27	0.28	0.98	14.45	0.753
TGB	$\ln(x + 0.5)$	Cleaned	Spherical	0.29	1.35	0.21	24.62	0.659
			Exponential	0.25	1.49	0.17	15.23	0.755
			Gaussian	0.30	1.27	0.24	10.56	0.815
		Included	Spherical	1.15	1.75	0.66	20.00	1.518
			Exponential	1.41	1.89	0.75	12.89	1.929
			Gaussian	1.63	1.84	0.89	16.35	1.855
	$(x + 0.5)^{1/4}$	Cleaned	Spherical	0.15	0.34	0.44	18.00	0.224
			Exponential	0.12	0.29	0.41	17.66	0.142
			Gaussian	0.18	0.32	0.55	19.34	0.186
		Included	Spherical	0.25	0.47	0.53	19.17	0.992
			Exponential	0.20	0.44	0.45	11.20	0.895
			Gaussian	0.27	0.45	0.60	19.73	0.653

[a]Relative nugget effect $= C_0/(C_0 + \sigma^2)$
[b]Weighted sum of square residuals

Table 3 Estimated parameters of three mathematical restricted variogram models with outliers cleaned and included in the *Trialeurodes vaporariorum* count data

Greenhouse	Transformation	Outlier	Model	Nugget (C_0)	Sill ($C_0 + \sigma^2$)	RNE	Range	$Q^*(\theta)$
WGA	$\ln(x + 0.5)$	Cleaned	Spherical	0.21	0.64	0.33	8.96	0.080
			Exponential	0.21	0.68	0.32	8.52	0.071
			Gaussian	0.28	0.67	0.42	8.24	0.075
		Included	Spherical	0.40	0.58	0.69	8.97	0.210
			Exponential	0.49	0.64	0.77	10.47	0.244
			Gaussian	0.45	0.66	0.68	8.23	0.237
	$(x + 0.5)^{1/4}$	Cleaned	Spherical	0.02	0.05	0.40	8.99	0.049
			Exponential	0.03	0.07	0.44	2.68	0.004
			Gaussian	0.02	0.08	0.25	4.35	0.096
		Included	Spherical	0.04	0.06	0.60	8.99	0.147
			Exponential	0.04	0.09	0.42	2.74	0.109
			Gaussian	0.04	0.07	0.55	10.99	0.125
WGB	$\ln(x + 0.5)$	Cleaned	Spherical	0.16	0.51	0.31	10.47	0.087
			Exponential	0.17	0.52	0.33	3.88	0.098
			Gaussian	0.27	0.41	0.66	6.53	0.085
		Included	Spherical	0.21	0.52	0.41	12.22	0.285
			Exponential	0.21	0.56	0.38	4.59	0.363
			Gaussian	0.35	0.57	0.61	12.62	0.226
	$(x + 0.5)^{1/4}$	Cleaned	Spherical	0.01	0.14	0.07	10.17	0.090
			Exponential	0.01	0.11	0.10	9.88	0.098
			Gaussian	0.03	0.12	0.26	10.23	0.083
		Included	Spherical	0.07	0.13	0.54	15.78	0.128
			Exponential	0.06	0.14	0.43	9.20	0.133
			Gaussian	0.09	0.13	0.69	13.66	0.109

The differences in the effective ranges from the outlier-cleaned data were slightly smaller than those from the outlier-included data sets; the range varied between 10.56 and 26.40 m (median 17.66 m) and 11.20–29.20 m (median 16.35 m) in thrips count data sets, respectively, and from 2.68 to 10.47 m (median 8.52 m) and 2.74 to 15.78 m (median 9.2 m) for whitefly count data sets (Tables 2, 3).

The weighted sum of square residuals $(Q*(\theta))$ was selected as a criterion for relevant spatial predictor in the fit of the theoretical variogram models. Based on the comparison of the mean $Q*(\theta)$ within the same insect species and data transformations, the outlier cleaned data sets always had lower $Q*(\theta)$ values than the data sets containing the outliers (Tables 2, 3), indicating that all of the variogram models tested in this study were very sensitive to outliers and that the outlier cleaning process provided better estimations of theoretical variogram model parameters from experimental variograms.

Leave-one-out validation

The leave-one-out cross-validation estimation errors are summarized by RMSE, which ranged from 0.51 to 1.30 in the outlier-included data sets and from 0.11 to 0.97 in the outlier-cleaned data sets (Table 4). The RMSE in the outlier-cleaned data sets was always smaller than that of the outlier-included data sets, suggesting the variogram models with outliner-cleaned data fit the data better than those containing outliers.

The relationship between the predicted values and observed data was investigated using a Pearson's correlation efficient (r) (Table 4). In all cases, significantly higher r values at $p = 0.05$ were achieved from the outlier-cleaned data sets, regardless of greenhouse and data transformation. The r values from the outlier-cleaned data ranged from 0.57 to 0.78 for the thrips count data sets and from 0.67 to 0.79 for the whitefly count data sets, respectively. A major factor in these apparently low correlations in the outlier-included data sets was the very high nugget effect present in the data; the estimated RNE from the outlier-included data sets was always greater than those from the outlier-cleaned data sets (Tables 2, 3).

Discussion

Geostatistics allows ecologists to organize and summarize data and thus make meaningful inferences about the dynamics of target organisms in space (Rossi et al. 1992; Kitanidis 1997; Tilman et al. 1997). Ecologists and applied entomologists have recently begun to use two geostatistical techniques for describing the spatial distribution of insect populations (Schotzko and O'Keeffe 1989; Liebhold et al. 1993; Midgarden et al. 1993; Cho et al. 2001; Kim et al. 2001; Wright et al. 2002). These two approaches are the variogram, which is a way to model spatial dependency, and kriging, which provides estimates of unrecorded locations (Isaaks and Srivastava 1989; Rossi et al. 1992; Cressie 1993; Kitanidis 1997; Brandhorst-Hubbard et al. 2001). Both geostatistical techniques should be evaluated using appropriate experimental variograms from field spatial data (Olea 2006, 2007).

Unfortunately, skewed population distributions and outliers in data sets prevent appropriate experimental

Table 4 Root mean square error and correlation coefficient between the original and estimate data by leave-one-out cross validation

Greenhouse[a]	Transformation	Model	Root mean square error		Correlation coefficient (r)	
			Outlier included	Outlier cleaned	Outlier included	Outlier cleaned
TGA	$\ln(x + 0.5)$	Spherical	1.14	0.94	0.21	0.65*
		Exponential	1.15	0.94	0.21	0.61*
		Gaussian	1.13	0.97	0.18	0.60*
	$(x + 0.5)^{1/4}$	Spherical	0.48	0.39	0.20	0.67*
		Exponential	0.48	0.39	0.24	0.76*
		Gaussian	0.51	0.43	0.23	0.78*
TGB	$\ln(x + 0.5)$	Spherical	0.76	0.40	0.21	0.69*
		Exponential	0.77	0.41	0.14	0.57*
		Gaussian	0.80	0.42	0.23	0.60*
	$(x + 0.5)^{1/4}$	Spherical	0.52	0.31	0.11	0.65*
		Exponential	0.51	0.30	0.34	0.70*
		Gaussian	0.50	0.30	0.33	0.71*
WGA	$\ln(x + 0.5)$	Spherical	1.30	0.79	0.19	0.67*
		Exponential	1.30	0.76	0.17	0.68*
		Gaussian	1.29	0.79	0.16	0.67*
	$(x + 0.5)^{1/4}$	Spherical	0.78	0.20	0.20	0.77*
		Exponential	0.78	0.20	0.22	0.79*
		Gaussian	0.79	0.20	0.19	0.76*
WGB	$\ln(x + 0.5)$	Spherical	0.68	0.42	0.31	0.71*
		Exponential	0.57	0.49	0.28	0.77*
		Gaussian	0.65	0.41	0.23	0.79*
	$(x + 0.5)^{1/4}$	Spherical	0.57	0.14	0.21	0.77*
		Exponential	0.53	0.13	0.21	0.78*
		Gaussian	0.61	0.11	0.13	0.77*

* Significant correlation coefficient between outlier-cleaned and -included data set by Fisher's Z transformation test

[a]Samples were selected from the 5th sample week at TGA, the 4th sample week at TGB, the 9th sample week at WGA, and the 2nd sample week at WGB

variogram models from being developed (Isaaks and Srivastava 1989; Cressie 1993; Kitanidis 1997). The most important part of any analysis of any data is the identification of outliers. When the analysis is concerned with second moments (such as variances with autocorrelations or related measures, such as a variogram), outliers can have a particularly dramatic effect and have long been recognized as a potential source of serious problems. Outliers in a data set can make the variogram display erratic behavior; whereas data transformation can dampen the difference between extreme values (Gringarten and Deutsch 2001). Park et al. (2004) examined the typical structure of samples from aggregated populations and found that many samples contained few or no individuals of a particular species, whereas some samples contained an extremely high number of individuals. They also reported that no data transformations in the analysis of geostatistics could correct the data sets when the zero counts in the sample data were $>20\%$. In our study, all data sets were clumped, and nine of the 32 examined cases contained $>20\%$ zero counts. For these data sets, an outlier cleaning method may help in obtaining appropriate data sets for geostatistical analysis (Mugglestone et al. 2000; Olea 2006).

Existing work on outliers in spatial processes deals mainly with variogram estimation and kriging procedures (Nirel et al. 1998). Up to now, a distinction has not clearly been established between 'isolated' and 'patchy' outliers arising from purely random and spatially structured patterns of contamination, respectively. The effects of additive outliers can be dramatic: estimates of auto-covariance function and auto-regressive moving-average parameters can be severely biased and inefficient (Guttman and Tiao 1978; Martin and Yohai 1986). In our study, an outlier cleaning method from Mugglestone et al. (2000), which is a model-based 'datacleaner', was developed for spatial lattice data. This approach defines observations that are associated with spatial regions, where the regions can be regularly (as in a grid) or irregularly spaced with varying distances between the region's centroids (Birkhoff 1967). Our data set was classified as spatial lattice data because each data point was separated in a rectangular shape, and species were sampled at the same spatial location in each greenhouse. In addition, one data point represented the population density in the rectangular space of the sampling position in the greenhouse. Outliers were then detected and cleaned by replacing the outlier value with the median value of the four closest neighbors (Mugglestone et al. 2000). In this case, the outlier cleaning procedure operated locally rather than systematically, which is appropriate for small systems, such as the greenhouse environment. The outlier detection and cleaning procedures were successful in replacing outliers in sound manner and showed better performance in the small-scale greenhouse system.

In this study, the outlier cleaning procedure was shown to improve geostatistical analysis and reduce the RNEs of extremely fluctuating and aggregated insect population data. A higher nugget effect relative to the sill in the outlier-included data set indicated a poor spatial continuity (Olea 2006). In addition, a higher nugget effect makes the prediction values more of a simple average of the available data. Furthermore, an increase in the nugget effect indicates a lack of spatial correlation (Isaaks and Srivastava 1989). Also, several previous studies have reported that the large nugget effect is not desirable in terms of prediction variances (Rossi et al. 1992; Kitanidis 1997). Thus, the outlier cleaning process developed in this study allows for more robust model estimation.

The outlier cleaning procedure also enabled hidden spatial dependence patterns to be identified within the greenhouse. Masking and swamping effects are very well-known problems that can arise when detecting outliers. Masking occurs when an outlier is not detected because of the presence of other outliers, i.e., an outlier is being masked by others. Swamping occurs when an observation is incorrectly identified as an outlier due to the effect of other outliers (Ben-Gal 2005). For instance, when a contour map of thrips in TGB at the 7th sampling week was drawn (Fig. 2), two large spatial points

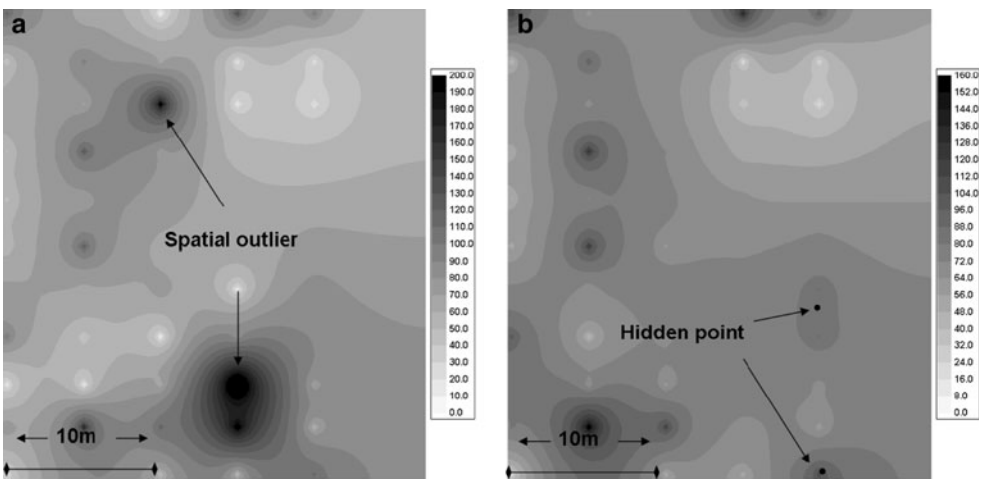

Fig. 2 Contour maps of thrips larvae distribution including outliers (**a**) and not including outliers (**b**) in TGB during the 7th sample week. Note that hidden spatial patches appeared after the outlier cleaning procedure was applied

(i.e., extreme high-density values) were shown to significantly affect the whole sampling plot and spatial dependency (Fig. 2a). However, after applying the outlier cleaning procedure, some hidden spatial points appeared (Fig. 2b). Diggle et al. (2010) reported that unintentional hot spots can mislead geostatistical inferences. In this case, the hidden spatial points were blocked by two outliers, which would lead to a misinterpretation of the spatial dependency. The zone containing the outliers and the other high values should be separated from the rest and treated separately. Kerry and Oliver (2007) reported that aggregated outliers had different effects on the variogram shape from those that were randomly located and these effects also depended on whether the outliers were aggregated near to the edge or to the center of the field. Thus, the outlier cleaning procedures improve the interpretation of the spatial dependency through geostatistical analysis by removing data with extreme values and a large number of zero values. This study clearly demonstrates that outlier detection and cleaning procedures are needed before the spatial dependency of insect population data sets in a local greenhouse environment can be analyzed.

Insect pest populations vary spatially and temporally from a field to a larger region scale. As all insect populations change over time, temporal aspects should be also considered. Moreover, spatial and temporal relationships exist among spatial objects at various scales. There are several space–time models based upon physical process and geographical information systems in wide open areas (Hristopulos and Christakos 2001; Christakos et al. 2002, 2005), but the detection and cleaning of temporal outliers or spatio–temporal outliers have been seldom discussed in ecological studies of insect populations. Future studies should focus on the effect of the outliers of the spatial and temporal scales for a better understanding of the population dynamics and, ultimately, for more efficient pest management practices.

Acknowledgments The authors thank the two cucumber growers who graciously allowed us to use their greenhouses for this work. This work was supported in part by a grant (H0487401) from the KOSEF and Korea University Special Research Grant to K Cho, and by the research fund of Hankuk University of Foreign Studies to K-I Shin.

Appendix

R language program script for computing M value (from detail computation method of Mugglestone et al. 2000).

```
Mvalue <- read.table("c:\\data\\The input data.txt[1]")



y <- Mvalue$V5[2]
y.mc <- y- median(y)
s.n <- median(abs(y.mc))*1.483
mean.y.mc <- mean(abs(y.mc))
mean.y.mc.sq <- (mean((y.mc)^2) )
w.2 <- 1
a1 <- mean.y.mc*sqrt(3.141592/(2*s.n^2 ))
a2 <- mean.y.mc.sq/(s.n^2)
gamma <- (2*a1 - a1^2 -1)/(2*a1 - a2 - 1)
k.2<- ((a2 + gamma - 1)/gamma ) - 1
tau.2 <- 0.36


rel.bias <- function(gamma, tau2, kappa2, omega, m)
{
tmp1 <- gamma*(tau2-kappa2-1)*pnorm(-m/sqrt(omega^2+kappa2))
tmp2 <- (1-gamma)*(tau2-1)*pnorm(-m/omega)
ans <- abs(2*(tmp1+tmp2)+gamma*kappa2)
return(ans)
}
m.s <- optimize(rel.bias, c(-3,3), gamma=gamma, tau2=tau.2, kappa2=k.2,
omega=1, maximum=F)$minimum
m.0 <- 3.08 + 0.63 * m.s - 0.33 * m.s^2 - 11.47 * gamma + 29.23 * gamma^2 +
0.02 * sqrt(k.2)
M <- m.0 * s.n


M
```

[1]"the input data.txt" is in the "data" folder of the "C drive of your computer" as a simple ASCII (text) file, such as may be exported from a spreadsheet or word processor, or even created by hand using Windows Notepad, or in the input panel itself.

This is the structure of the input files:

First, put the *m* sampling weeks with *n* records of your data into a file in the following form:

```
x coordinate 1<tab>y coordinate 1<tab>sample 1 count 1<tab>sample 2 count 1<tab>.....<tab>sample m count 1
x coordinate 2<tab>y coordinate 2<tab>sample 1 count 2<tab>sample 2 count 2<tab>.....<tab>sample m count 2
x coordinate 3<tab>y coordinate 3<tab>sample 1 count 3<tab>sample 2 count 3<tab>.....<tab>sample m count 3
.....
x coordinate n<tab>y coordinate n<tab>sample 1 count n<tab>sample 2 count n<tab>.....<tab>sample m count n
```

where the x and y coordinates for each pair (x_k, y_k; $k = 1, \ldots, n$), and the count (c_k, $k = 1, \ldots, n$) at the m th sampling week (1st week, $\ldots$, mth week), would be read in, on the same line, as either integers and real numbers. Each column should be separated by tap.

Note that there are no headers or column titles in the data file.

[2]After *R* reads the data file, they automatically put internal name as $V_1, V_2, \ldots, V_{m+2}$ for each column in data. V_1 and V_2 always refer 'x coordinate' and 'y coordinate', respectively. Thus, the count data column starts from V_3.

References

Bardossy A, Kundzewicz ZW (1990) Geostatistical method for detection of outliers in groundwater quality spatial fields. J Hydrol 115:343–359

Barnett V, Lewis T (1994) Outliers in statistical data. Wiley, New York

Bartlett MS (1936) The square root transformation in analysis of variance. Suppl J Stat Soc 3:22–28

Ben-Gal I (2005) Outlier detection. In: Maimon O, Rockach L (eds) Data mining and knowledge discovery handbook: a complete guide for practitioners and researchers. Kluwer, Dordrecht, pp 1–16

Binns MR, Nyrop JP (1992) Sampling insect population for the purpose of IPM decision making. Annu Rev Entomol 37:427–453

Birkhoff G (1967) Lattice theory, 3rd edn. American Mathematical Society, Providence

Box GEP, Cox DR (1964) An analysis of transformations. J R Stat Soc B 26:211–243

Brandhorst-Hubbard JL, Flanders KL, Mankin RW, Guertal EA, Crocker RL (2001) Mapping of soil insect infestations sampled by excavation and acoustic methods. J Econ Entomol 94:1452–1458

Breunig MM, Kriegel H-P, Ng RT, Sander J (2000) LOF: identifying density-based local outliers. ACM SIGMOD Rec 29:93–104

Cho K, Kang SH, Lee JO (1998) Spatial distribution of thrips in greenhouse cucumber and development of a fixed-precision sampling plan for estimating population density. J Asia Pac Entomol 1:163–170

Cho K, Lee JH, Park JJ, Kim JK, Uhm KB (2001) Analysis of spatial pattern of *Frankliniella occidentalis* (Thysanoptera: Thripidae) on greenhouse cucumbers using dispersion index and spatial autocorrelation. Appl Entomol Zool 36:25–32

Christakos G, Bogaerts P, Serre M (2002) Temporal geographical information systems: advanced functions for field-based applications. Springer, New York

Christakos G, Olea RA, Serre ML, Yu H-L, Wang L-L (2005) Interdisciplinary public health reasoning and epidemic modeling: the case of Black Death. Springer, New York

Clark I, Harper WV (2000) Practical geostatistics 2000. Ecosse North America LTD, Columbus

Cressie N (1985) Fitting variogram models by weighted least squares. Math Geol 17:563–586

Cressie N (1993) Statistics for spatial data. Wiley, New York

Diggle PJ, Menezes R, Su TL (2010) Geostatistical inference under preferential sampling. J R Stat Soc C Appl Stat 59:191–232

Goovaerts P (1997) Geostatistics for natural resources evaluation. Oxford University Press, New York

Gringarten E, Deutsch CV (2001) Teacher's aide: variogram interpretation and modeling. Math Geol 33:507–534

Guttman I, Tiao GC (1978) Effect of correlation on the estimation of a mean in the presence of spurious observations. Can J Stat 6:229–247

Hadi AS (1992) Identifying multiple outliers in multivariate data. J R Stat Soc Ser B 54:761–771

Hawkins DM (1980) Identification of outliers. Chapman and Hall, New York

Hawkins DM, Cressie N (1984) Robust kriging—a proposal. Math Geol 16:3–18

Hristopulos DT, Christakos G (2001) Practical calculation of non-Gaussian multivariate moments in spatiotemporal Bayesian maximum entropy analysis. Math Geol 22:543–568

Ihaka R, Gentleman RR (1996) R: a language for data analysis and graphics. J Comput Graph Stat 5:299–314

Immaraju JA, Paine TD, Bethke JA, Robb KL, Newman JP (1992) Western flower thrips (Thysanoptera: Thripidae) resistance to insecticides in coastal California greenhouses. J Econ Entomol 85:9–14

Insightful Corporation (2001) S-PLUS 6 for Windows user's guide. Insightful Corporation, Seattle

Isaaks EH, Srivastava RM (1989) Applied geostatistics. Oxford University Press, New York

Kaluzny SP, Vega SC, Cardoso TP, Shelly AA (1998) S + Spatial statistics: user's manual for Windows and Unix. Springer, New York

Kerry R, Oliver MA (2007) Determining the effect of asymmetric data on the variogram. II. Outliers. Comput Geosci 33:1233–1260

Kim J-K, Park J-J, Park H, Cho K (2001) Unbiased estimation of greenhouse whitefly, *Trialeurodes vaporariorum*, mean density using yellow sticky trap in cherry tomato greenhouses. Entomol Exp Appl 100:235–243

Kitanidis PK (1997) Introduction to geostatistics: applications to hydrogeology. Cambridge University Press, New York

Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. IJCAI 2:1137–1143

Lewis T (1973) Thrips: their biology, ecology and economic importance. Academic Press, New York

Liebhold AM, Rossi RE, Kemp WP (1993) Geostatistics and geographical information system in applied insect ecology. Annu Rev Entomol 38:303–327

Martin RD, Yohai VJ (1986) Influence curves for time series. Ann Stat 11:1608–1630

Martin RD, Yohai VJ (1991) Bias robust estimation of autoregression parameters. In: Stahel W, Weisberg S (eds) Directions in robust statistics part 1. Springer, Berlin, pp 233–246

McGrath D, Zhang C, Carton OT (2004) Geostatistical analyses and hazard assessment on soil lead in Silvermines area, Ireland. Environ Pollut 127:239–248

Midgarden DG, Youngman RR, Fleuscher SJ (1993) Spatial analysis of counts of western corn rootworm (Coleoptera: Chrysomelidae) adults on yellow sticky traps in corn: geostatistics and dispersion indices. Environ Entomol 22:1124–1133

Miesch AT, Riley LB (1961) Basic statistical methods used in geochemical investigation of Colorado Plateau uranium deposits. HIMMP Trans (Mining) 220:247–251

Mound LA, Halsey SH (1978) Whitefly of the world. British Museum, Chichester

Mugglestone MA, Barnett V, Nirel R, Murray DA (2000) Modelling and analyzing outliers in spatial lattice data. Math Comput Model 32:1–10

Nansen C, Campbell JF, Phillips TW, Mullen MA (2003) The impact of spatial structure on the accuracy of contour maps of small data sets. J Econ Entomol 96:1617–1625

Nirel R, Mugglestone MA, Barnett V (1998) Outlier-robust spectral estimation for spatial lattice processes. Commun Stat Theory Methods 27:3095–3111

Olea RA (2006) A six-step practical approach to semivariogram modeling. Stoch Environ Res Risk Assess 20:307–318

Olea RA (2007) Declustering of clustered preferential sampling for histogram and semivariogram inference. Math Geol 39:453–467

Papadopoulos AP (1991) Growing greenhouse tomatoes in soil and in soilless media. Agriculture Canada Publication no 1865/E. Agriculture Canada, Ottawa

Park J-J, Shin K, Cho K (2004) Evaluation of data transformations and validation of a spatial model for spatial dependency of *Trialeurodes vaporariorum* populations in a cherry tomato greenhouse. J Asia Pac Entomol 7:289–295

Rossi RE, Mulla DJ, Journel AG, Franz EH (1992) Geostatistical tools for modeling and interpreting ecological spatial dependence. Ecol Monogr 62:277–314

Rousseeuw P, Leory A (1987) Robust regression and outlier detection. Wiley Series in probability and statistics. Wiley, New York

Sanderson JP, Roush RT (1992) Monitoring insecticide resistance in greenhouse whitefly (Homoptera: Aleyrodidae) with yellow sticky cards. J Econ Entomol 85:634–641

SAS Institute (1996) SAS user's guide. SAS Institute, Cary

Schotzko DJ, O'Keeffe LE (1989) Geostatistical description of the spatial distribution of *Lygus hesperus* (Heteroptera: Miridae) in lentils. J Econ Entomol 82:1277–1288

Sichel HS (1952) New methods in the statistical evaluation of mine sampling data. Lond Inst Min Metall Trans 61:261–288

Southwood TRE (1978) Ecological methods, 2nd edn. Chapman and Hall, London

Srivastava RM, Parker HM (1989) Robust measures of spatial continuity. In: Armstrong M (ed) Geostatistics, vol 1. Kluwer, Dordrecht, pp 295–308

Tilman D, Lehman CL, Kareiva P (1997) Population dynamics in spatial habitats. In: Tilman D, Kareiva P (eds) Spatial ecology. Princeton University Press, Princeton, pp 3–20

Wright RJ, Devries TA, Young LJ, Jarvi KJ, Seymour RC (2002) Geostatistical analysis of the small-scale distribution of European corn borer (Lepidoptera: Crambidae) larvae and damage in whorl stage corn. Environ Entomol 31:160–167

Zar JH (1999) Biostatistical analysis, 4th edn. Prentice Hall International, Upper Saddle River

Zhang CS, Selinus O (1998) Statistics and GIS in environmental geochemistry-some problems and solutions. J Geochem Explor 64:339–354

Zhang CS, Zhang S, Zhang LC, Wang LJ (1995) Background contents of heavy metals in sediments of the Changjiang River system and their calculation methods. J Environ Sci 7:422–429