APPENDIX G

SEMIVARIANCE ANALYSIS

> "Everything is related to everything else, but near things are more related than distant things."

Tobler's first law of geography

> "One event may cause other nearby events, but events far off or in the distant past can have significant influence to events being observed in the present."

David Lord

1. INTRODUCTION

Semivariance methods deal with autocorrelated spatial data. These methods were first applied to geological problems, hence they are often referred to as geostatistics. As reviewed by Madden et al. (2), these methods have been applied to plant disease spatial analyses since 1988. Sometimes, semivariance is spelled as a hyphenated word, semi-variance. For internet searches, it is advisable to include all three terms (semi-variance, semivariance and geostatistics) as keywords for maximum results.

Semivariance analysis provides spatial statistics of a particular quantifiable variable located within a designated space. This is different from the spatial point pattern (SPP) analysis which provides statistics based on the points' locations relative to each other. Both are tools of spatial statistical analysis, and the valid use of either of these tools depends on the application.

The 2002 published article by Gottwald et al provides the procedure and results of the semivariance analysis (1). There are no other presentations. The article can be difficult to follow in places as both semivariance and SPP analysis are discussed together.

Tobler's first law of geography seems relevant to semivariance analysis, "Everything is related to everything else, but near things are more related than distant things." This is, of course, a philosophical idea, not a scientific law. Also, plant disease epidemiology is not geography, so it seems more relevant to discuss events rather than things. I added an additional piece of philosophy regarding events. Observational studies provide only limited purview of events within a defined window in time and space, thus can lead to distorted analyses and erroneous conclusions.

A list of eight published articles related to plant disease spatial analysis using semivariance methods is provided at the end of the chapter. These articles were published from 1989 to 2014, and all are available to the public, free of charge, from the APS website.

The semivariance analysis is typically applied to a spatial data set with values of a well quantifiable continuous variable, with values z_i located at Cartesian coordinate locations *(x_i, y_i)*. It is convenient to express the set of a random field variable *Z(**u**)* where ***u*** refers to a spatial location. The basis for the

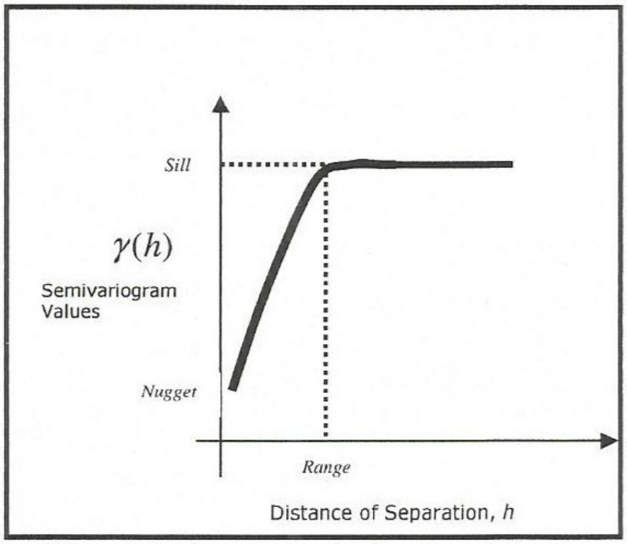
A variogram (or semivariogram) is used to summarize the relationship at each lag distance (distance between points, which can be individuals or grid cells centers). Sample gamma values $\hat{\gamma}(h)$ can be calculated for lag distance *h* is defined as:

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

,

where *N(h)* is the number of pairs of points separated by *h*, $z(x_i)$ is the data value for the point x_i, and $z(x_i+h)$ is the data value at cells separated from x_i by the lag distance *h* in the chosen direction. The calculated gamma values at lag distances (h) are also referred to as empirical or estimated semivariances. The plot of these values verses lag distances as shown in Figure 1 may be referred to as the experimental variogram.

Figure 1: Variogram with nugget and sill shown on y-axis and range on x-axis



The nugget value is the semivariance value at a separation distance of zero, and theoretically must be equal to zero. However, to improve curve fitting, the nugget value is one of the parameters adjusted in curve fitting of a model. From the nugget value to the sill point, theoretical variograms show increasing values of semivariance values corresponding to an increasing dissimilarity in the compared Z values at higher separation distances. The range is the distance beyond which the semivariance values are no longer spatially autocorrelated. The sill value is equal to the variance of the data.

Transitional models are can be fitted to calculated semivariance values. Common transitional models are Gaussian, spherical and exponential. They have similar shapes in the nugget to sill curve and it may be difficult to know which model best fits the calculated semivariances. The intent of semivariance analysis is generally to evaluate whether plant characteristics such as disease incidence are clustered or random. If incidence is clustered, "neighbors" are more likely to share the same disease status than are plants separated by larger distances.

It is possible that the calculated values can not be fitted to a transitional model. There are numerous examples in the literature and in the articles listed at the end of this chapter, where the calculated semivariances did not conform to any theoretical transitional model. There may be may reason, but one of them is inadequate sample of data. In exploring spatial relationships, experimental variograms can be calculated at various times. It is important to note apparent changes in variograms may be due to exogenous effects, particularly if there are pathways for disease to enter beyond the area under study. The appearance of disease at distant locations at different times, may be the result of disease introduction with delays in identification resulting from discovery lags and latency. Also, disease expansion may be a complex process, with interacting factors helping causing local and long distance dissemination, simultaneous with other factors limiting the dissemination process.

Spatial descriptive analysis is very useful. The Z variable data can be displayed using different colored circles on a map. The variogram with both the data and the best fit form, are another means of graphically visualizing the data and potential relationships.

2. Selected Relevant Excerpts From The Published Article

Selected Excepts from Gottwald, T.R., X. Sun, Riley, T. Graham, J.H., Ferrandino, F. and Taylor, E., 2002, Geo-Referenced Spatiotemporal Analysis of the Urban Citrus Canker Epidemic in Florida, Phytopathology, Vol 92, No. 4.

Every effort has been taken to transcribe the relevant passages exactly as published. There are no tables or figures which accompanied the text. The original article should always be consulted, to see the discussion in the context presented by the author. The article may be downloaded free of charge from a number of websites, including this one.

--- Abstract, page 361

Spatiotemporal analyses between periods over areas larger than previously examined were accomplished via spatiotemporal semivariogram analysis. These methods in combination demonstrated rapid increases in range of spatial dependency and range of spatiotemporal dependency for all study sites.

--- Page 363, second paragraph, left column

Semivariance analysis has gained considerable popularity in recent years for the exploration of geostatistical data. Although variogram estimation has been accomplished using spatial point pattern data for plant pathogens, it is perhaps more appropriately applied to quantitative data.

--- Page 365, first paragraph, right column

Spatiotemporal analysis of the spatial point pattern. Semi- variance analysis was used to further examine the spatiotemporal relationships among ACC-affected trees for each urban study site. Data sets for each site were prepared for analysis by assigning diseased trees a temporal value (*t*) based on the number of 30-day periods infected relative to the final assessment period. That is, if a tree remained healthy over the majority of time during which the epidemic was monitored but its disease status changes during the last four 30-day periods, then $t = 4$. Whereas if a tree remained infected for the duration of the time during which the epidemic was monitored, then $t = 25$. Thus, an individual tree was quantitatively weighted more heavily (by a larger *t* value) if it became infected earlier in the epidemic. Semivariogram analysis was performed using GS+ geostatistical software (version 5.1, Gamma Design Software, Plainwell, MI) for 0° (omnidirectional with an angle of inclusion of 180°) and 0, 45, 90, and 135° relative to North, and each site with an angle of inclusion of 90°. The semivariance $\gamma(h)$ (the variance about the mean difference in disease between all sampling units for a given distance = *h*) versus distance in meters was plotted. By convention, the distance between locations is designated as *h* for semivariance analyses, whereas it is designated as *d* for the Ripley's *K*-function. Linear, linear to sill, exponential, spherical, and Gaussian transitional models were fitted to the semivariance $\gamma(h)$ versus distance data by means of nonlinear regression analysis performed via a model-fitting subroutine. The resulting two-dimensional spatiotemporal structure was considered anisotropic when directional semivariograms diverged from one another over distance (15). The range of spatiotemporal dependency (RSTD = A_o) was estimated using the chosen model as the point at which semivariance reached a plateau for each comparison among temporal periods.

--- Page 370, first paragraph, right column

Analyses of the spatiotemporal relationships of the SPP of ACC were conducted to compare consecutive 30-day periods over the duration of all 25 temporal periods. The spherical model was the best descriptor of the spatiotemporal structure of the point patterns for sites D1, D2, and D3 through time based on residuals of regression and the r^2 of regression. This model provided a definitive estimation of the RSTD = the range of spatiotemporal dependency. That is, for the associations of diseased trees, the model was capable of expressing the significant range of distance over which this association occurred and the dynamics of this range through time. The isotropic RSTD increased rapidly for site D1. The spatiotemporal structure had a RSTD of 0.119 km for the comparison of the first versus the second 30-day period. The RSTD exceeded the maximum active lag distance (80% of the longest diagonal axes of the site) with the comparison of the second and third 30-day periods, which corresponded to spread of citrus canker across the entire site by T3. For site D2, no focal trees existed during T1, but at T2, ACC-affected trees were widely dis-persed across the site. Thus, the first comparison that could be made was between T2 and T3 for which the RSTD exceed the maximum active lag distance. For site D3, disease started with a single isolated focal tree and increased intermittently through time. Only very few diseased trees existed until T6 and thus no spatiotemporal comparisons could be made until that time. Subsequent periods of disease increase were related to T13, T14, T17, and T25. RSTD values associated with T1 to T6, T1 to T13, and T1 to T14 were 0.037, 0.009, and 0.033 km, respectively. Spatio-temporal comparisons of T1 to T17 and T1 to T25 resulted in RSTD values that exceeded the active lag distance extents of the plot.

--- Page 376, first paragraph, left hand column (Discussion section)

To our knowledge, the regional increase and spread of ACC has not been previously examined. Spatiotemporal autocorrelation was used previously to examine relationship among ACC-affected citrus trees in small experimental nursery and orchard plantings (11–13). In the present study, the modified Ripley's *K*-function and semivariance analyses were used to examine the SPP of citrus canker for each study site, temporally for each 30-day period and spatiotemporally between periods over areas larger than previously examined. In sites D1 and D2, disease spread rapidly across the extents of the sites during the second and third 30-day periods (T2 and T3), respectively. Site D3 provided the most insight into the spatiotemporal structure and dynamics of ACC in an urban setting. Disease increase was recorded relative to only a few 30-day temporal periods, and was aggregated and constrained to the area immediately surrounding the initial focal tree until T17. However, during T17, disease spread extended over the extents of the study site. These disease dynamics were well reflected in the rapid increase of the RSD_{eff} associated with the Ripley's analyses of sites D1 and D2 versus the more gradual increase of the RSD_{eff} associated with site D3 through time. A rapid increase of RSTD was also associated with spatiotemporal disease dynamics by semi- variance analyses for D1 and D2 versus a less rapid increase of RSTD for site 3. This represents a rapid increase in the association of disease in both spatial and temporal scales, simultaneously. Examination of the SPP maps for each plot revealed that for D1, D2, and D3, the distribution of infected trees over the extents of each plot occurred at T3, T2, and T17, respectively, and related to $RSD_{eff,}$ spatiotemporal distance relationships of 1.5, 2.2, and 1.6 km, respectively.

Note: On page 376, the authors are citing statistics from both spatial point pattern and semi-variance analyses. Only the statements on RSTD statistics, represent semi-variance analysis. The statements on RSD_{eff} are based on spatial point pattern analysis, as reviewed in Appendix F.

3. Review of the Semivariance analysis

What was done

The semivariance analysis was used to evaluate spatial characteristics of a "temporal value", denoted as *t* and assigned to every infected tree within a site. This temporal value is the number of 30-day time periods in which a citrus tree was designated as either a newly infected or a prior infected tree. Variograms were calculated for all 25 time periods (T1 to T25) for Sites D1, D2 and D3, but not for B1 and B2.

Why the Broward sites were excluded is not stated in the article. In contrast, the abstract seems to indicate that semivariance analyses were conducted for to all sites.

There are 25 time periods with sequence numbers *n = {1, 2, ... 25}* or *{T1, T2 ... T25}* for 30-day scenarios based on the discussion on page 363. The temporal value, *t,* reverses this sequence in descending order, or *t = {25, 24 ... 1} or t = 26 - n.*

There is a T0 period implied in the procedure, which would include all newly infected trees prior to 10/26/97 in sites D1 and D3. The duration of this time period is not stated in the article. Presumably *t = 26* for this time period, or the infected trees in this time period were not included.

For example, for Site D1, time period T3, using data from Table 1, page 365, there would be 10 trees with *t = 25*, 9 trees with *t = 24*, and 17 trees with *t = 23*. In this case, possible delta Z-squared values are {4, 1, 0}. This assumes the infected trees assigned to time period T0 are excluded from the calculations.

The actual number of semivariances calculated would depend on the lag interval and tolerance. No variograms are presented. Instead, select range values (RSTD) are presented on presented on pages 370 to 371. These were based on fitted spherical models to the variogram.

The data (temporal values) were analyzed by the GS+ geostatistical software. The software is capable of correcting for anisotropic effects. The statement on page 365, "The resulting two-dimensional spatiotemporal structure was considered anisotropic when directional semivariograms diverge from one another over distance (15)" gives the impression that anisotropic corrections were done, however reference 15 is a standard reference on geostatistics. This sentence is therefore interpreted as simply a general statement, and the verb "was" should have been "is."

Additional results are discussed under the general heading of "Ripley's K-function" although neither this analysis nor the spatial point pattern analysis as discussed in Appendix F, are Ripley's K-function analysis.

On page 370, it is stated that the spherical model provided the best descriptor of the spatiotemporal structure of the point patterns for the 3 sites in Miami-Dade based on residuals of regression and correlation correlation of regression. There is no statistical values given for this statement.

Results

For site D1, it is stated, “The spatiotemporal structure has a RSTD of 0.119 km for comparison of the first verses the second 30-day period.” Did the range increase by 0.119 km from the first to the second period? It is hard to imagine an interpretable variogram with 14 points, and z-squared values of zero or one in time period 1, assuming the zero time period was included. Thus, it is interpreted this result as the variogram range based on T2.

For site D2, no numerical results are provided. For site D3, the authors note the lack of infected trees until the sixth period (180 days). The results for ranges T1 to T6, T1 to T13 and T1 to T14 are 0.037, 0.009 and 0.033 km. This is interpreted to mean the ranges based on temporal values in periods T6, T13 and T14. The range value changes described on pages 370 and 371, stated as “disease increases.”

Further, it is stated on page 371 in discussion of the difference of RSTD in Site D3, it is stated that the T1 to T17

On page 376, the article concludes, “A rapid increase of RSTD was also associated with spatiotemporal disease dynamics by semivariance analysis for D1 and D2, with less rapid increase of RSTD for site 3.” Validity of Semivariance Analysis

Validity of Analysis

The resequencing of the time periods from ascending order, {1, 2, 3 .. }. to descending order, { 25, 24, 23 ...} is was done so, according to the authors, “an individual tree was quantitatively weighted more heavily (by a larger t value) if it became infected earlier in the epidemic” (page 365). This seems to be an odd statement, as the Z-squared values would be identical if time periods were referenced with increasing or decreasing indexes.

No variograms presented with calculated data, it is not possible to provide a proper evaluation. The ranges are strictly based on fitted transitional models. So, even if one accepts the validity of the application of the semivariance analysis, there is no way to evaluate adequacy of the transitional model fit, or any of the model parameters such as range, given the inadequate presentation of results.

In the eight articles published in Phytopathology and Plant Disease articles and referenced at the end of this appendix, there are numerous examples of experimental variograms which do not exhibit well defined transitions from nugget to sill values (corresponding to lag distances from 0 to the range value). Yet, every experimental variogram, no matter how “noisy” can be fitted to a variogram.

Citrus canker can only be found in citrus trees. Leaf drop or pruning can remove symptoms of citrus canker. Canker symptoms can be misdiagnosed in their early stages. Within the site, non-citrus areas exist (lakes, canals, parking lots, shopping centers) and others where citrus is unlikely (parks, schools, commercial buildings). Thus, application of the semivariance model would be a violation of the basic assumptions inherent in the method.

The RSTD values are not considered meaningful to the dissemination of canker as the analysis is severely flawed. However, if one were to accept these values as valid support of "distances of spread" within the time periods, they are all considerably less than the adopted 1900 ft (579 m) rule. All ranges are in fact, closer to the 125 ft (38.1 m) rule.

4. Conclusions

1. No experimental variogram data with the fitted transitional model curve were presented. The reliability of range estimates is unknown as no information on the curve tit is given. Review of eight other published articles demonstrates this information would normally be included if transitional model parameters (nugget, sill and range) are part of the interpretation of results.

2. Due to the large non-citrus areas, the assumption of a continuum for the transitional model is violated.

3. The results of semivariance analysis are interpreted in the published article to support the rapid spread of canker. However, this is not supported by the numerical results. In fact, the five numerical ranges given in the article are from 9 to 119 m, which is closer to the existing 38.1 m (125-ft) eradication policy, than 579 m (1900-ft) , the post Jan 2000 policy.

Based on the following, it is suggested that the semivariance analysis does not demonstrate meaningful results on the dissemination of citrus canker in the Miami-Dade sites.

No results for the Broward Sites were presented.

References

1. Gottwald, T.R., Hughes, G., Graham, J.H, Sun, X., Riley, T., 2001, The Scientific Basis of Regulatory Eradication Policy for an Invasive Species, Phytopathology, 91:30-34.

2. Madden, L.V., Hughes, G., and van den Bosch, F., 2007, The Study of Plant Disease Epidemics, American Phytopathology Society, Minnesota. ISDN 978-0-89054-354-2.

Articles on Semivariance Analysis

Byamukama, E., Eggenberger, S.K., Coelho-Netto, R.A. Robertson, A. E., Nutter, F.W., Jr. 2014. Geospatial and Temporal Analyses of *Bean pod mottle virus* Epidemics at Three Spatial Scales, Phytopathology 104: 365- 378.

Dandurand, L.M., Knudsen, G.R, and Schotzko, D.J. 1995, Quantification of *Pythium ultimum* var. *sporangiiferum* Zoospore Encystment Patterns Using Geostatistics. Phytopathology 85:186-190.

Gavossoni, W. L., Tylka, G. L, Munkvold, G. P. 2001. Relationships Between Tillage and Spatial Patterns of *Heterodera glycines*. Phytopathology 91:534-545.

Gottwald, T. R., LLácer, G., Hermoso de Mendoza, A., Cambra, M. 1995. Analysis of the Spatial Spread of Sharka (Plum Pox Virus) in Apricot and Peach Orchards in Eastern Spain, Plant Dis., 79:266-278.

Groves, R. L., Chen, J., Civerolo, E. L., Freemond, M.W. and Viveros, M. A. 2005. Spatial Analysis of Almond Leaf Scorch Disease in the San Joaquin Valley of California: Factors Affecting Pathogen Distribution and Spread. Plant Dis. 89:581-589.

Johnson, D. A, Alldredge, J. R., Allen, J. R. and Allwine, R. 1991. Spatial Pattern of Downy Mildew in Hop Yards During Severe and Mild Disease Epidemics. Phythopathology 81:1369-1374.

Leocoustre, R., Fargette, C., Fauquet, C., de Reffye, P. 1989. Analysis and Mapping of Spatial Spread of African Cassava Mosaic Virus Using Geostatistics and the Kriging Technique. Phytopathology 79:8913-920.

Nelson, M.R., Felix-Gastelum, R., Orum, T. V., Stowell, I. J., and Myers, D. E. 1994. Geographic Information Systems and Geostatistics in the Design and Validation of Regional Plant Virus Management Programs. Phytopathology 84:898-905.

TEXTBOOK REFERENCES ON SEMIVARIANCE ANALYSIS

Chilés, J.P. and P. Delfiner, 1999. Geostatistics, Modeling Spatial Uncertainty, John Wiley & Sons, New York.

Cressie, N. A. C. 1991, Statistics for Spatial Data (Revised Edition 1993). John Wiley and Sons.

Isaaks, E.H. and R. Mohan Srivastava, An Introduction to Applied Geostatistics, Oxford University Press.

Kelkar, M. and Godofredo, P., 2002, Applied Geostatistics for Reservoir Characterization, Society of Petroleum Engineers, Inc., Richardson, Texas (ISBN 1-555630095-2).

These represent a few of the many books on the subject.