Math Geosci (2011) 43:735–740
DOI 10.1007/s11004-011-9348-3

SHORT NOTE

Variogram or Semivariogram? Variance or Semivariance? Allan Variance or Introducing a New Term?

Martin Bachmaier · Matthias Backes

Received: 30 September 2010 / Accepted: 24 March 2011 / Published online: 1 July 2011

Abstract There is a confusing situation in geostatistical literature: Some authors write variogram, and some authors write semivariogram. Based on a formula for the empirical variance that relates to pairwise differences, it is shown that the values depicted in a variogram are entire variances of observations at a given spatial separation (lag). Therefore, they should not be called semivariances, and the term semivariogram should also be avoided. To name a variogram value, we suggest the use of the term gammavariance instead of the misleading semivariance.

Keywords Structure function · Variogram · Semivariogram · Variance · Semivariance · Allan variance · Gammavariance · Spatial variability

1 Introduction

The definition of the theoretical variogram, γ, is based on regionalized random variables $Z(\vec{x})$ and $Z(\vec{x}+\vec{h})$ where $\vec{x}$ and $\vec{x}+\vec{h}$ represent the spatial positions separated by a vector $\vec{h}$

$$\gamma(\vec{h}) = \frac{1}{2}\,\mathrm{E}\big[\big[Z(\vec{x}+\vec{h}) - Z(\vec{x})\big]^2\big] = \frac{1}{2}\,\mathrm{Var}\big[Z(\vec{x}+\vec{h}) - Z(\vec{x})\big]. \tag{1}$$

M. Bachmaier (✉)
Agricultural Systems Engineering, Technische Universität München, Freising-Weihenstephan, Germany
e-mail: bachmai@wzw.tum.de

M. Backes
Institut für Kartographie und Geoinformation, Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn, Germany
e-mail: backes@ikg.uni-bonn.de

The $Z(\vec{x})$ and $Z(\vec{x}+\vec{h})$ denote random variables. According to the intrinsic hypothesis, $\gamma(\vec{h})$ is assumed to depend only on the separation vector, the lag $\vec{h}$, but not on the location $\vec{x}$. Further, the increments $Z(\vec{x}+\vec{h})-Z(\vec{x})$ are assumed to have no drift: $\mathrm{E}[Z(\vec{x}+\vec{h})-Z(\vec{x})]=0$ for all $\vec{h}$ and all $\vec{x}$; otherwise, the last identity in (1) would not hold. It follows from the variance addition theorem that $\operatorname{Var}(Z(\vec{x}+\vec{h})-Z(\vec{x}))=\operatorname{Var}(Z(\vec{x}+\vec{h}))+\operatorname{Var}(Z(\vec{x}))-2\operatorname{Cov}(Z(\vec{x}+\vec{h}),Z(\vec{x}))$, and this is equal to $2\operatorname{Var}(Z(\vec{x}))-2\operatorname{Cov}(Z(\vec{x}+\vec{h}),Z(\vec{x}))$ because of the assumed second-order stationarity, which says that mean, variance, and the mentioned covariance do not depend on the position $\vec{x}$. Thus, using $\sigma^2=\operatorname{Var}(Z(\vec{x}))$ and the covariance function, $C(\vec{h})=\operatorname{Cov}(Z(\vec{x}+\vec{h}),Z(\vec{x}))$, yields the writing

$$\gamma(\vec{h})=\sigma^2-C(\vec{h}). \tag{2}$$

σ^2 is identical with the sill of the variogram. The sill is the limit of $\gamma(\vec{h})$ for $\|\vec{h}\|\to\infty$, so that the covariance function tends to zero. The covariance function is also the centered second-order spatial cumulant of the regionalized variable Z. Spatial cumulants of higher order are treated in Mustapha and Dimitrakopoulos (2010).

2 Confusion in the Literature

As already mentioned in Bachmaier and Backes (2008), there is great confusion in the geostatistical literature with regard to the terminology. Some authors call the function γ a variogram (Wackernagel 2003; Worboys 1995; Gneiting et al. 2001), several authors call it a semivariogram (Journel and Huijbregts 1978; Cressie 1991; Goovaerts 1997; Burrough and McDonnell 1998; Olea 1999; Stein 1999; Gringarten and Deutsch 2001), stating that a semivariogram is half a variogram, and yet others use the terms variogram and semivariogram synonymously (Isaaks and Srivastava 1989; Webster and Oliver 2007). To explain what is depicted in a variogram, authors of geostatistical books and articles often take refuge in different phrases. Gringarten and Deutsch (2001), for example, denote $\gamma(\vec{h})$ by semivariogram value and seem to want to avoid the terms variance or semivariance, writing: "The variogram is a measure of variability." Most authors who denote a variogram by 2γ usually write semivariance when referring to $\gamma(\vec{h})$ (Burrough and McDonnell 1998), but they do not mention the semivariance of what they mean. Phrases such as "$\gamma(\vec{h})$ is known as the semivariance" are usual. Evidently there is great uncertainty with regard to terminology and the interpretation of variograms.

The confusion concerning the prefix semi might have arisen because Matheron (1965) had the variance of differences in mind in his seminal thesis, $\operatorname{Var}[Z(\vec{x}+\vec{h})-Z(\vec{x})]$, but the quantity we want in practice is the half (semi) of this, because this gives the "variance per point when the points are considered in pairs" (Webster and Oliver 2007). This formulation should emphasize that not the variance of the difference of those points is meant, but of the points themselves; applied to independent identically distributed random variables X and Y, this would mean that we are not interested in $\operatorname{Var}(X-Y)$, but in $\operatorname{Var}(X)$, which gives the half of the former. Further, if the variances of the difference were used to find the weights of the kriging estimator by minimizing its variance, one would constantly face the factor 2 in the system of equations to be solved. This could be the most important reason to avoid referring to the variances of the difference. $\gamma(\vec{h})$ can be interpreted as the variance of the regionalized variable (e.g., of the yield data) at the given separation vector $\vec{h}$, which means that we consider only pairs that are spatially separated by the lag $\vec{h}$. The spatial closeness of such data to each other makes them correlated so that their variance is, by their covariance, smaller than if they were independent. This is, what has been expressed in (2). $\gamma(\vec{h})$ should not be called a semivariance since this term originates from the variance of the differences, which is not the actual quantity of interest. And if it were, one should not compute the half of it, but the whole variance. No one says that the semiheight of his or her body is 86 cm.

3 Understanding the Empirical Variances in a Variogram

The empirical variance of measured values z_i can be computed in two different ways

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z})^2 = \frac{1}{2}\cdot\frac{1}{n(n-1)}\sum_{\text{all } i \neq j}(z_j - z_i)^2, \tag{3}$$

where $n(n-1)$ is the number of pairs in the latter sum. It would suffice to consider only all pairs with $i < j$ since $(z_i - z_j)^2 = (z_j - z_i)^2$. This would halve its number. The variances in an experimental variogram, $\hat{\gamma}(\vec{h})$, arise by restricting the latter expression in (3) to pairs of measured values, $z_i = z(\vec{x}_i)$ and $z_j = z(\vec{x}_j)$, whose positions, $\vec{x}_i$ and $\vec{x}_j$, are separated by a spatial vector $\vec{h}$

$$\hat{\gamma}(\vec{h}) = \frac{1}{2}\cdot\frac{1}{N(\vec{h})}\sum_{\substack{\text{all } i \neq j \\ \text{with } \vec{x}_j - \vec{x}_i = \vec{h}}}(z_j - z_i)^2, \tag{4}$$

where $N(\vec{h})$ is the number of pairs whose positions, $\vec{x}_j$ and $\vec{x}_i$, are separated by the vector $\vec{h}$. Thus, $\hat{\gamma}(\vec{h})$ should simply be referred to as the (empirical) variance of the measured values at the given separation vector $\vec{h}$. The restriction at the given separation vector $\vec{h}$ corresponds in a certain way to the substraction of the covariance in the theoretical $\gamma(\vec{h})$ in (2), because a small separation vector, for example, causes high correlations in the data, which diminish their variance. When referring to isotropic variation, $\hat{\gamma}(h)$ denotes this variance at a given separating distance $h = \|\vec{h}\|$. We do not need the term semivariance unless we want to cite references where it is used. It should be added that it is the semivariance of the difference of random variables or measured values.

Obtaining $\hat{\gamma}(\vec{h})$ by an analogous restriction of the former, more commonly known variance expression in (3) is not completely possible, because the restriction refers to pairs and not to deviations from the mean. If such a direct analogon existed, it would have been used, and the confusion concerning the prefix semi would never have arisen. However, it is possible to write $\hat{\gamma}(\vec{h})$ as an average of such mean-based variances of only two points with each fulfilling the separation condition $\vec{x}_j - \vec{x}_i = \vec{h}$

$$\hat{\gamma}(\vec{h}) = \overline{s^2(\vec{h})} = \frac{1}{N(\vec{h})} \sum_{\substack{\text{all } i \neq j \\ \text{with } \vec{x}_j - \vec{x}_i = \vec{h}}} s_{ij}^2(\vec{h}), \tag{5}$$

where

$$s_{ij}^2(\vec{h}) = \frac{1}{2-1}\left[\left(z_i - \frac{z_i + z_j}{2}\right)^2 + \left(z_j - \frac{z_i + z_j}{2}\right)^2\right] = \frac{(z_j - z_i)^2}{2}. \tag{6}$$

The first expression in (6) describes this variance as a special case of the usual variance formula for $n = 2$ (i.e., the first expression in (3)), whereas the last expression in (6) already points to the difference-based writing of $\hat{\gamma}(\vec{h})$ in (4).

Many readers will find the writing of $\hat{\gamma}(\vec{h})$ in (4) somewhat unusual, as they are accustomed to the notation

$$\hat{\gamma}(\vec{h}) = \frac{1}{2}\frac{1}{N(\vec{h})} \sum_{i=1}^{N(\vec{h})} \left[z(\vec{x}_i + \vec{h}) - z(\vec{x}_i)\right]^2. \tag{7}$$

Here, the positions $\vec{x}_i + \vec{h}$ and $\vec{x}_i$ immediately indicate that they are separated by the vector $\vec{h}$. However, one must be aware that the index i used in this sum cannot serve as a counter of the different positions $\vec{x}_i$, unless the separation of $\vec{h}$ is required to be exact—in practice, it suffices to meet a class around $\vec{h}$—and the variogram is anisotropic. To compute $\hat{\gamma}$(50 m) in an isotropic variogram, for example, the positions (10 m, 20 m) and (10 m, 70 m) match, but the same position (10 m, 20 m) is used once again if it is paired with the position (60 m, 20 m). Therefore, the position $\vec{x}_i =$ (10 m, 20 m) must receive at least two different indices i in order to apply the writing of $\hat{\gamma}(\vec{h})$ in (7), where i counts pairs, not single positions. Equations (3) and (7) also show that the connection between the empirical variance of all measured values, s^2, and the variances $\hat{\gamma}(\vec{h})$ in an experimental variogram can well be demonstrated by writing the total variance s^2 as a weighted mean of the latter

$$s^2 = \sum_{\text{all classes}} \frac{N(\vec{h})}{N} \hat{\gamma}(\vec{h}) \tag{8}$$

(Bachmaier 2007), where the separation vector $\vec{h}$ is usually classified.

4 Discussion and Conclusions: The Introduction of a New Term to Denote Variogram Values

Auerswald et al. (2009) and Wittmer et al. (2010) adopted Bachmaier's and Backes' (2008) variogram interpretation as follows: "The semivariance (γ) equals the variance for values at points which are separated by a certain distance called lag". In this way, the terms variance and semivariance clash, and Bachmaier's and Backes' interpretation is misrepresented. In this unfortunate citation, the prefix "semi" was intended to denote a restriction to certain distances, and not to designate the factor $\frac{1}{2}$, as introduced by Matheron (1965). Although this misinterpretation has appeared in three different journals, only one editor accepted a rectification by Bachmaier (2010).

The preceding citation also reveals that scientists would like to name variogram values. They may find it a bit cumbersome to call them variances at a given lag, because the variance is known as a single-valued statistic and not as a function of a lag. We have the name variogram to summarize or visualize these variances as a function of the lag, but we do not have a proper name for these variances themselves, the single values of the variogram. Intended and factual meaning of the denotation semivariance diverge, and the term variogram value, which we used here several times, is only a deduction of variogram.

In contrast to this, an unambiguous name for single variances of this kind has been used in the analysis of time series. It is the term Allan variance named after Allan (1986, 1988). Under conditions that are analogous to the intrinsic hypothesis, the Allan variance denotes the variance of measured values at a given time lag τ

$$\sigma_z^2(\tau) = \frac{1}{2}\,\mathrm{E}\bigl([Z_{t+\tau} - Z_t]^2\bigr). \tag{9}$$

This formula also contains the factor $\frac{1}{2}$, but there is nothing in the name Allan variance that reminds one of half a variance. The Allan variance is only applied to time lags, and further, the values measured are always frequencies (Arpaia et al. 2003; Greenhall 1997). Therefore, an expansion of the meaning of Allan variance to geostatistical areas is not adequate, as it would destroy its strong association with this specialized research.

It seems more reasonable to adopt the principle of naming something after its inventor. Likewise to the Allan variance, geostatistical variogram values could be named after authors who introduced the corresponding definitions and worked with them. However, there are many researchers who did pioneering work in this area. Independently of Matheron (1965), Gandin (1963) also developed the method of kriging. He called the variogram a structure function, as already did Yaglom (1957), and even Kolmogorov (1941) and Jowett (1952) did some research in this area. As such, it is hard to make a decision in whose honor variogram values should be named, and we do not want to do this. Despite the confusing situation in geostatistical literature with regards to the prefix semi, however, all geostatistical authors use the same designation for a variogram or its single values, namely the Greek lower-case letter γ. Therefore, it appears reasonable to simply denote a variogram value γ-variance or, to express it in Latin letters, gammavariance. Everyone would know what is meant, and no one is mislead to believe that it is only half a variance of measured values at the given lag.

Acknowledgement We thank Karl Auerswald, Lehrstuhl für Grünlandlehre, Technische Universität München, for giving us references to early articles on this research.

References

Allan DW (1986) Should the classical variance be used as a basic measure in standards metrology? IEEE Trans Instrum Meas 36(2):646–654

Allan DW (1988) Time and frequency (time-domain) characterization, estimation, and prediction of precision clocks and oscillators. IEEE Trans Ultrason Ferroelectr Freq Control 34(6):647–654

Arpaia P, Daponte P, Rapuano S (2003) Characterization of digitizer timebase jitter by means of the Allan variance. Comput Stand Interfaces 25(1):15–22

Auerswald K, Wittmer MHOM, Männel TT, Bai YF, Schäufele R, Schnyder H (2009) Large regional scale-variation in C3/C4 distribution pattern of Inner Mongolia steppe is revealed by grazer wool carbon isotope composition. Biogeosciences 6(1):795–805

Bachmaier M (2007) Using a robust variogram to find an adequate butterfly neighborhood size for one-step yield mapping using robust fitting paraboloid cones. Precis Agric 8(1/2):75–93

Bachmaier M (2010) On the misleading use of the term 'semivariance' in recent articles. Rapid Commun Mass Spectrom 24(7):1111

Bachmaier M, Backes M (2008) Variogram or semivariogram?—understanding the variances in a variogram. Precis Agric 9(3):173 –175

Burrough PA, McDonnell RA (1998) Principles of geographical information systems. Spatial information systems and geostatistics. Oxford University Press, Oxford

Cressie NAC (1991) Statistics for spatial data. Wiley, New York

Gandin LS (1963) Objective analysis of meteorological fields. Gidrometeorologicheskoe Izdatel'stvo (GIMIZ), Leningrad (1965 ed. available from the Israel Program for Scientific Translations, Jerusalem)

Gneiting T, Sasvári Z, Schlather M (2001) Analogies and correspondences between variograms and covariance functions. Adv Appl Probab 33(3):617–630

Goovaerts P (1997) Geostatistics for natural resource evaluation. Oxford University Press, Oxford

Greenhall CA (1997) The third-difference approach to modified Allan variance. IEEE Trans Instrum Meas 46(3):696–703

Gringarten E, Deutsch CV (2001) Teacher's aide—variogram interpretation and modeling. Math Geol 33(4):507–534

Isaaks EH, Srivastava RM (1989) An introduction to applied geostatistics. Oxford University Press, Oxford

Journel AG, Huijbregts CJ (1978) Mining geostat. Academic Press, London

Jowett GH (1952) The accuracy of systematic sampling from conveyor belts. Appl Stat 1(1):50–59

Kolmogorov AN (1941) The local structure of turbulence in an incompressible fluid at very large Reynolds numbers. Dokl Akad Nauk SSSR 30(4):301–305

Matheron G (1965) Les variables régionalisées et leur estimation. Masson, Paris

Mustapha H, Dimitrakopoulos R (2010) A new approach for geological recognition using high-order spatial cumulants. Comput Geosci 36(3):313–334

Olea RA (1999) Geostatistics for engineers and earth scientists. Kluwer Academic, Boston

Stein ML (1999) Interpolation of spatial data: some theory for kriging. Springer, Berlin

Wackernagel H (2003) Multivariate geostatistics. Springer, Berlin

Webster R, Oliver MA (2007) Geostatistics for environmental scientists. Wiley, New York

Wittmer MHOM, Auerswald K, Bai YF, Schäufele R, Schnyder H (2010) Changes in the abundance of C3/C4 species of Inner Mongolia grassland: evidence from isotopic composition of soil and vegetation. Glob Change Biol 16(2):605–616

Worboys MF (1995) GIS—a computing perspective. Taylor & Francis, London

Yaglom AM (1957) Some classes of random fields in n-dimensional space, related to stationary random processes. Theory Probab Appl 2(3):273–320