10 Extension and Dispersion Variance

Measurements can represent averages over volumes, surfaces or intervals, called their *support*. The computation of variances depends intimately on the supports that are involved as well as on a theoretical variogram associated to a pointwise support. This is illustrated with an application from industrial hygienics. Furthermore, three simple sampling designs are examined from a geostatistical perspective.

Support

In the investigation of regionalized variables the variances are a function of the size of the domain. On Table 10.1 the results of computations of means and variances in nested 2D domains $\mathcal{D}_n$ are shown.

	Size	Mean $m(\mathcal{D}_n)$	Variance $\sigma^2(\cdot\|\mathcal{D}_n)$
$\mathcal{D}_1$	32×32	20.5	7.4
$\mathcal{D}_2$	64×64	20.1	13.8
$\mathcal{D}_3$	128×128	20.1	23.6
$\mathcal{D}_4$	256×256	20.8	34.6
$\mathcal{D}_5$	512×512	18.8	45.0

Table 10.1: Nested 2D domains $\mathcal{D}_n$ for which the variance increases with the size of the domain (from a simulation of an intrinsic random function by C LAJAUNIE)

In this example the variance $\sigma^2(\cdot|\mathcal{D}_n)$ of point samples in a domain $\mathcal{D}_n$, increases steadily with the size of the domain whereas the mean does not vary following a distinctive pattern. This illustrates the influence that a change in the size of a support (here the domain $\mathcal{D}_n$) can have on a statistic like the variance.

In applications generally two or more supports are involved as illustrated by the Figure 10.1. In mining the samples are collected on a support that can be considered pointwise (only a few cm^3); subsequently small blocs v (m^3) or larger panels V (100m^3) have to be estimated within deposits $\mathcal{D}$. In soil pollution small surface units s are distinguished from larger portions S. In industrial hygiene the problem may be set in terms of time supports: with average measurements on short time intervals Δt the excess over a limit value defined for a work day T should be estimated.

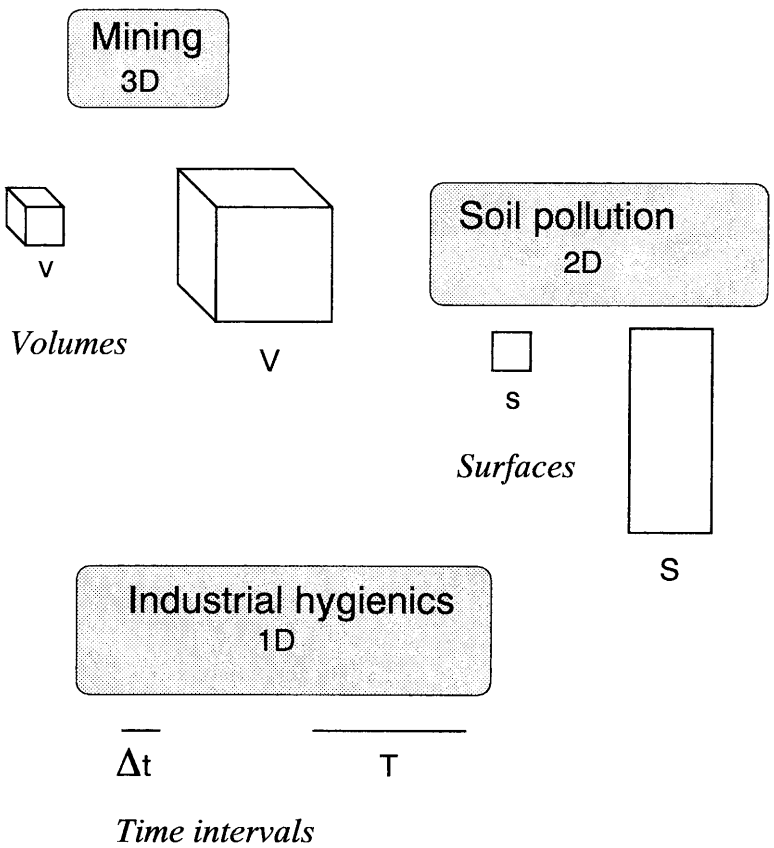

Figure 10.1: Supports in 1, 2, 3D in different applications.

Extension variance

With regionalized variables it is necessary to take account of the spatial disposal of points, surfaces or volumes for which the variance of a quantity should be computed.

The *extension variance* of a point $\mathbf{x}$ with respect to another point $\mathbf{x}'$ is defined as twice the variogram

$$\sigma_{\mathrm{E}}^2(\mathbf{x},\mathbf{x}') \;=\; \mathrm{var}(Z(\mathbf{x}) - Z(\mathbf{x}')) \;=\; 2\,\gamma(\mathbf{x}{-}\mathbf{x}'). \tag{10.1}$$

It represents the square of theoretical error committed when a value at a point $\mathbf{x}$ is "extended" to a point $\mathbf{x}'$.

The extension variance of a small volume v to a larger volume V at a different location (see Figure 10.2) is obtained by averaging the differences between all positions of a point $\mathbf{x}$ in the volume v and a point $\mathbf{x}'$ in V

$$\begin{aligned}\sigma_{\mathrm{E}}^2(v,V) \;&=\; \mathrm{var}(\,Z(v) - Z(V)\,) \\ &=\; 2\,\frac{1}{|v|\,|V|}\int\limits_{\mathbf{x}\,\in v}\;\int\limits_{\mathbf{x}'\in V}\gamma(\mathbf{x}{-}\mathbf{x}')\,d\mathbf{x}\,d\mathbf{x}'\end{aligned} \tag{10.2}$$

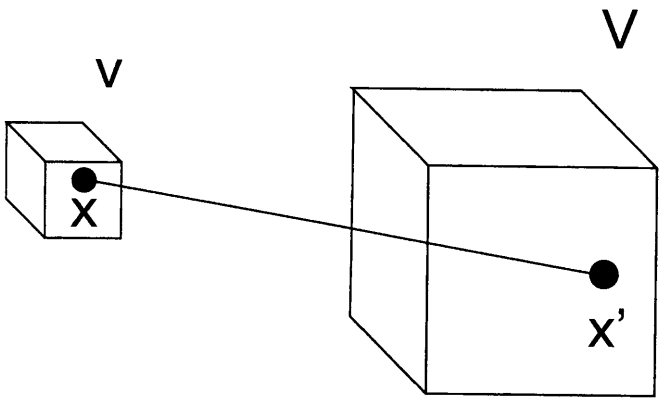

Figure 10.2: Points $\mathbf{x} \in v$ and $\mathbf{x}' \in V$.

$$-\frac{1}{|v|^2} \int_{\mathbf{x} \in v} \int_{\mathbf{x}' \in v} \gamma(\mathbf{x}-\mathbf{x}')\, d\mathbf{x}\, d\mathbf{x}'$$

$$-\frac{1}{|V|^2} \int_{\mathbf{x} \in V} \int_{\mathbf{x}' \in V} \gamma(\mathbf{x}-\mathbf{x}')\, d\mathbf{x}\, d\mathbf{x}'. \tag{10.3}$$

Denoting

$$\overline{\gamma}(v,V) \;=\; \frac{1}{|v|\,|V|} \int_{\mathbf{x} \in v} \int_{\mathbf{x}' \in V} \gamma(\mathbf{x}-\mathbf{x}')\, d\mathbf{x}\, d\mathbf{x}', \tag{10.4}$$

we have

$$\sigma^2_{\mathrm{E}}(v,V) \;=\; 2\,\overline{\gamma}(v,V) - \overline{\gamma}(v,v) - \overline{\gamma}(V,V). \tag{10.5}$$

The extension variance depends on variogram integrals $\overline{\gamma}(v,V)$, whose values can either be read in charts (see JOURNEL & HUIJBREGTS [156], chap. II) or integrated numerically on a computer.

Dispersion variance

Suppose a large volume V is partitioned into n smaller units v of equal size. The experimental *dispersion variance* of the values z_v^α of the small volumes v_α building up V is given by the formula

$$s^2(v|V) \;=\; \frac{1}{n} \sum_{\alpha=1}^{n} (z_v^\alpha - z_V)^2, \tag{10.6}$$

where

$$z_V = \frac{1}{n} \sum_{\alpha=1}^{n} z_v^\alpha. \tag{10.7}$$

Considering all possible realizations of a random function we write

$$S^2(v|V) = \frac{1}{n}\sum_{\alpha=1}^{n}(Z_v^\alpha - Z_V)^2. \tag{10.8}$$

The theoretical formula for the dispersion variance is obtained by taking the expectation

$$\begin{aligned}\sigma^2(v|V) &= \mathrm{E}[\,S^2(v|V)\,] \\ &= \frac{1}{n}\sum_{\alpha=1}^{n}\mathrm{E}\Big[\,(Z_v^\alpha - Z_V)^2\,\Big],\end{aligned} \tag{10.9}$$

in which we recognize the extension variances

$$\sigma^2(v|V) = \frac{1}{n}\sum_{\alpha=1}^{n}\sigma_{\mathrm{E}}^2(v_\alpha, V). \tag{10.10}$$

Expressing the extension variances in terms of variogram integrals

$$\begin{aligned}\sigma^2(v|V) &= \frac{1}{n}\sum_{\alpha=1}^{n}(2\,\overline{\gamma}(v,V) - \overline{\gamma}(v,v) - \overline{\gamma}(V,V)) \\ &= -\overline{\gamma}(v,v) - \overline{\gamma}(V,V) + \frac{2}{n}\sum_{\alpha=1}^{n}\frac{1}{|v_a|\,|V|}\int_{\mathbf{x}\in v_\alpha}\int_{\mathbf{x}'\in V}\gamma(\mathbf{x}-\mathbf{x}')\,d\mathbf{x}\,d\mathbf{x}' \\ &= -\overline{\gamma}(v,v) - \overline{\gamma}(V,V) + \frac{2}{n|v|\,|V|}\underbrace{\sum_{\alpha=1}^{n}\int_{\mathbf{x}\in v_\alpha}}_{\mathbf{x}\in V}\int_{\mathbf{x}'\in V}\gamma(\mathbf{x}-\mathbf{x}')\,d\mathbf{x}\,d\mathbf{x}' \\ &= -\overline{\gamma}(v,v) - \overline{\gamma}(V,V) + 2\,\overline{\gamma}(V,V),\end{aligned} \tag{10.11}$$

so that we end up with the simple formula

$$\sigma^2(v|V) = \overline{\gamma}(V,V) - \overline{\gamma}(v,v). \tag{10.12}$$

The theoretical determination of the dispersion variance reduces to the computation of the variogram integrals $\overline{\gamma}(v,v)$ and $\overline{\gamma}(V,V)$ associated to the two supports v and V.

Krige's relation

Starting from the formula of the dispersion variance, first we see that for the case of the point values (denoted by a dot) the dispersion formula reduces to one term

$$\sigma^2(\cdot|V) = \overline{\gamma}(V,V). \tag{10.13}$$

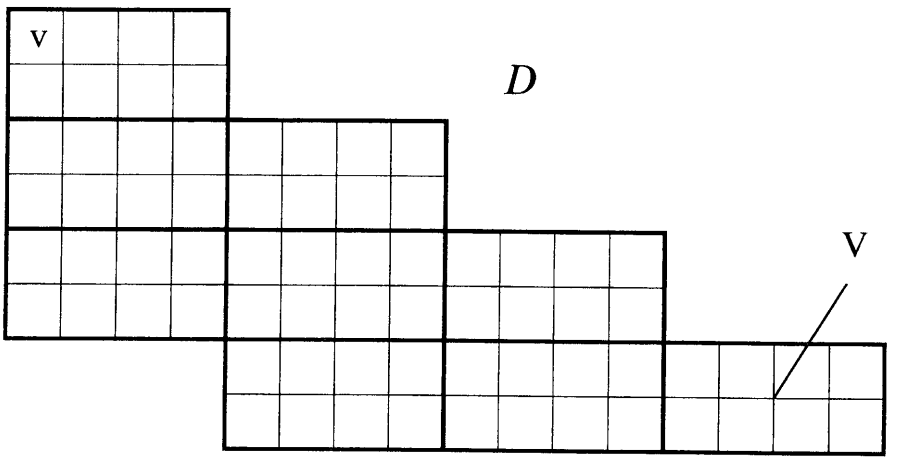

Figure 10.3: A domain $\mathcal{D}$ partitioned into volumes V which are themselves partitioned into smaller volumes v.

Second, we notice that $\sigma^2(v|V)$ is the difference between the dispersion variances of point values in V and in v

$$\sigma^2(v|V) = \sigma^2(\cdot|V) - \sigma^2(\cdot|v). \tag{10.14}$$

Third, it becomes apparent that the dispersion variance of point values in V can be decomposed into

$$\sigma^2(\cdot|V) = \sigma^2(\cdot|v) + \sigma^2(v|V). \tag{10.15}$$

This decomposition can be generalized to non point supports. Let $\mathcal{D}$ be a domain partitioned into large volumes V which are themselves partitioned into small units v as represented on Figure 10.3. Then the relation between the three supports v, V and $\mathcal{D}$ can be expressed theoretically by what is called *Krige's relation*

$$\sigma^2(v|\mathcal{D}) = \sigma^2(v|V) + \sigma^2(V|\mathcal{D}). \tag{10.16}$$

As the dispersion variances are basically differences of variogram averages over given supports, the sole knowledge of the pointwise theoretical variogram model makes dispersion variance computations possible for any supports of interest.

Change of support effect

In the early days of ore reserve estimation, mining engineers used a method called the *polygon method.* It consists in defining a polygon around each sample, representing the area of influence of the sample value, in such a way that the ore deposit is partitioned by the polygons. The reserves are estimated as a linear combination of the grades with the corresponding areas of influence. In the polygon method each sample value is extended to its area of influence, neglecting the fact that the samples are obtained from pointwise measurements while the polygons represent a much larger support.

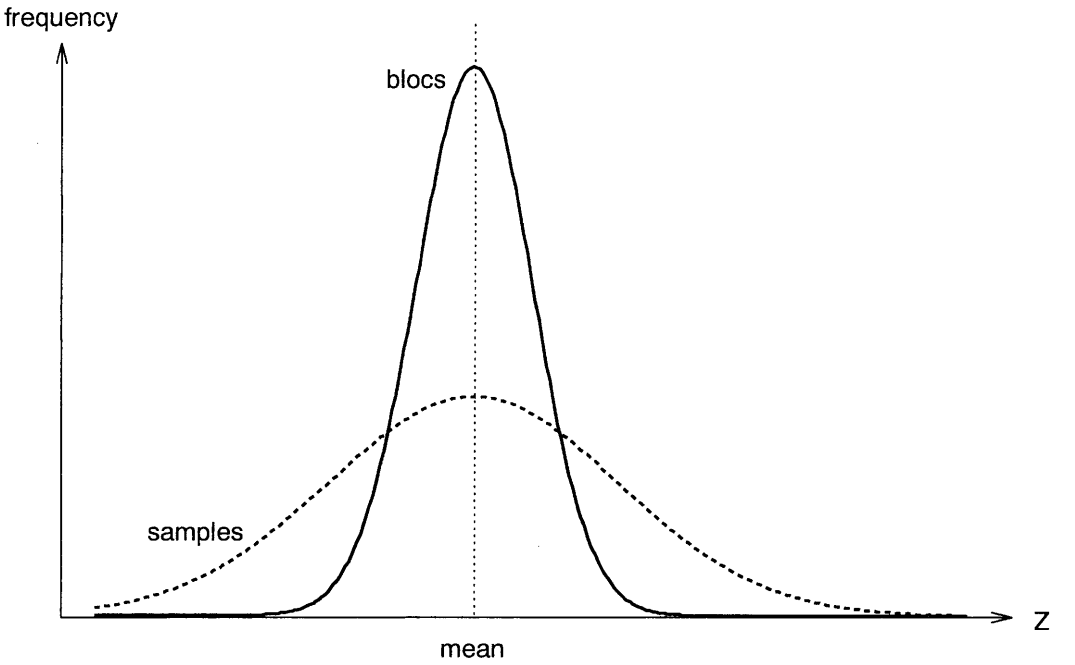

Figure 10.4: The distribution of block values is narrower than the distribution of values at the sample points.

In the case of a square grid the polygons are square blocks v which partition the exploration area. The value at each grid node is extended to each area of influence v. The method implies that the distribution of average values of the blocks is the same as the distribution of the values at the sample points. From Krige's relation we know that this cannot be true: the distribution of the values for a support v is narrower than the distribution of point values (as represented on Figure 10.4) because the variance $\sigma^2(\cdot|v)$ of the points in v generally is not negligible.

In mining, the *cut-off value* defines a grade above which a mining block should be sent to production. Mining engineers are interested in the proportion of the values above the cut-off value which represent the part of a geological body which is of economical interest. If the cut-off grade is a value substantially above the mean, the polygon method will lead to a systematic overestimation of the ore reserves as shown on Figure 10.5. To avoid systematic over- or underestimation the *support effect* needs to be taken into account.

Change of support: affine model

In this section we consider a stationary random function $Z(\mathbf{x})$ with a mean m and a variance σ^2. The mean m is not changed by a change of support and, whatever the distribution, we have the physical fact,

$$\mathrm{E}[\,Z(\mathbf{x})\,] \;=\; \mathrm{E}[\,Z_v(\mathbf{x})\,] = m, \tag{10.17}$$

i.e. the mean of the point variable $Z(\mathbf{x})$ is the same as that of the block variable $Z_v(\mathbf{x})$.

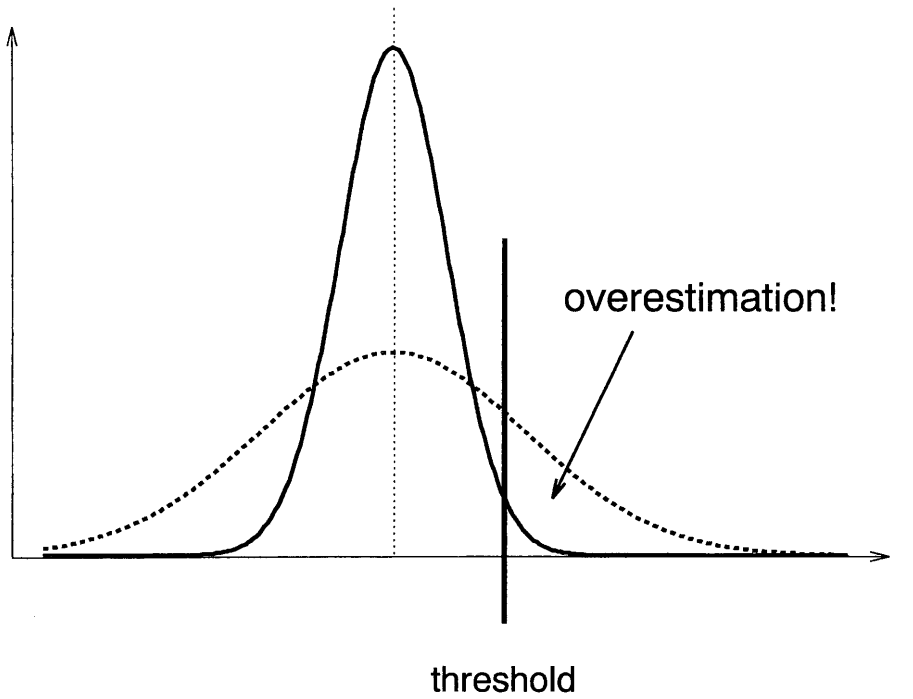

Figure 10.5: The proportion of sample values above the cut-off value is greater than the proportion of block values: the polygon method leads to a systematic overestimation in this case.

The *affine model* is based on the assumption that the standardized point variable follows the same distribution as the standardized block variable. This model is suitable for a Gaussian random function as the distribution of block grades is also Gaussian, i.e. if $Z(\mathbf{x}) \sim \mathcal{N}(m, \sigma^2)$,then $Z_v(\mathbf{x}) \sim \mathcal{N}(m, \sigma_v^2)$ and

$$\frac{Z(\mathbf{x}) - m}{\sigma} \stackrel{\mathcal{L}}{=} \frac{Z_v(\mathbf{x}) - m}{\sigma_v} \sim \mathcal{N}(0, 1), \tag{10.18}$$

where $\stackrel{\mathcal{L}}{=}$ means that the two quantities are identically distributed.

The distribution of the block values is therefore simply obtained from the distribution of the point values by an affine transformation,

$$Z_v(\mathbf{x}) \stackrel{\mathcal{L}}{=} m + r\,(Z(\mathbf{x}) - m) \sim \mathcal{N}(m, \sigma_v^2), \tag{10.19}$$

where $r = \sigma_v/\sigma$ is the change of support coefficient.

In practice if the point variance σ^2 and the variogram $\gamma(\mathbf{h})$ are known, the bloc variance is computed by the formula (10.12) of the dispersion variance,

$$\sigma_v^2 = \overline{C}(v, v) = \sigma^2 - \overline{\gamma}(v, v) \tag{10.20}$$

as, assuming a large domain in comparison to the range of the variogram, $\sigma^2 = \overline{\gamma}(\infty, \infty)$. The change of support coefficient can then readily be computed. The affine change of support model is appropriate only if the data comply with the Gaussian distributional assumption. Otherwise the affine model should only be used for v relatively small as compared to the range.

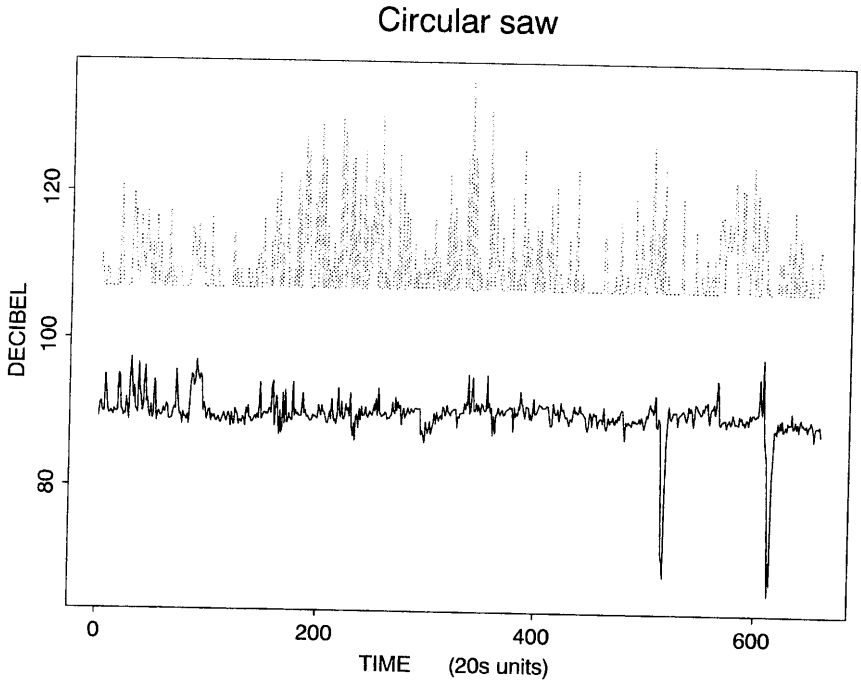

Figure 10.6: Measurements of maximal noise level L_{max} (dots) and average noise L_{eq} (plain) on time intervals of 20 seconds during 3 hours and 40 minutes. They represent the exposure of a worker to the noise of a circular saw.

Application: acoustic data

A series of 659 measurements of equivalent noise levels L_{eq} (expressed in dB_A) averaged over 20 seconds were performed on a worker operating with a circular saw. The problem is to evaluate whether a shorter or larger time integration interval would be of interest.

The $L_{eq}(t)$ are not an additive variable and need to be transformed back to the sound exposure $V_{eq}(t)$. The average sound exposure $V_{eq}(t)$ is defined as the integral (over time interval Δt) of the squared ratio of the instant acoustic pressures $p(x)$ against the reference acoustic pressure p_0

$$V_{eq}(t) = \frac{10^{-9}}{\Delta t} \int_{t-\Delta t/2}^{t+\Delta t/2} \left(\frac{p(x)}{p_0} \right)^2 dx \tag{10.21}$$

$$= \exp(\alpha\, L_{eq}(t) - \beta), \tag{10.22}$$

where $\alpha = (\ln 10)/10$ and $\beta = \ln 10^9$.

The measurements were taken continuously during a period of 3 hours and 40 minutes. The Figure 10.6 shows with a continuous line the time series (in dB_A) of the equivalent noise levels L_{eq} integrated over intervals of 20 seconds. The maximal noise levels L_{max} within these time intervals are plotted with a dotted line (when they are above 107 dB). We observe in passing that the averaging over 20 seconds has

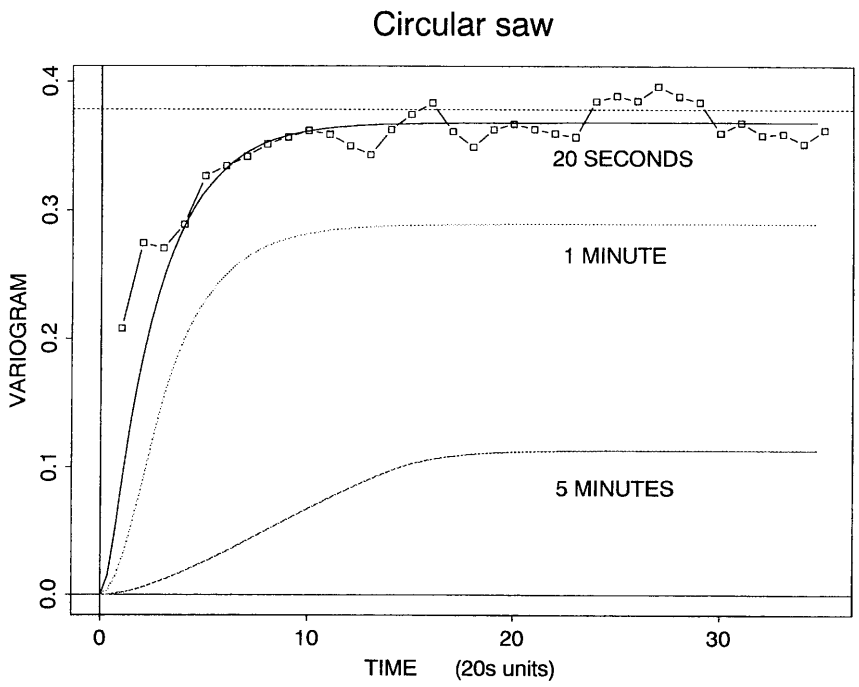

Figure 10.7: Experimental variogram of the sound exposure V_{eq} and a regularized exponential variogram model for time intervals of $\Delta t = 20$s, 1mn and 5mn.

enormously reduced the variation.

The theoretical variogram of the sound exposure was modeled with a pointwise exponential model

$$\gamma(h) = b\left(1 - \mathrm{e}^{-|h|/a}\right) \qquad \text{with } a, b > 0. \tag{10.23}$$

The sill is $b = .42$ and the range parameter is $a = 2.4$. It corresponds to a practical range of $3a = 7.2$ time units, i.e. 2.4 minutes, which is the time of a typical repetitive working operation.

The support of the sound exposure, i.e. the integration time, has an impact on the shape of the theoretical variogram: it alters the behavior at the origin, reduces the value of the sill and increases the range. The exponential variogram regularized over time intervals Δt is defined by the formula ([156], p84)

$$\gamma_{\Delta t}(h) = \begin{cases} \dfrac{b\,a^2}{(\Delta t)^2}\left(2\mathrm{e}^{-\Delta t/a} - 2 + \dfrac{2h}{a} + \mathrm{e}^{-h/a}\left(2 - \mathrm{e}^{-\Delta t/a}\right) - \mathrm{e}^{(h-\Delta t)/a}\right) \\ \qquad\qquad \text{for } 0 \le h \le \Delta t, \\ \dfrac{b\,a^2}{(\Delta t)^2}\left(\mathrm{e}^{-\Delta t/a} - \mathrm{e}^{\Delta t/a} + \left(\mathrm{e}^{-\Delta t/a} + \mathrm{e}^{\Delta t/a} - 2\right)\cdot\left(1 - \mathrm{e}^{-h/a}\right)\right) \\ \qquad\qquad \text{for } h > \Delta t. \end{cases} \tag{10.24}$$

The Figure 10.7 shows the experimental variogram together with the exponential

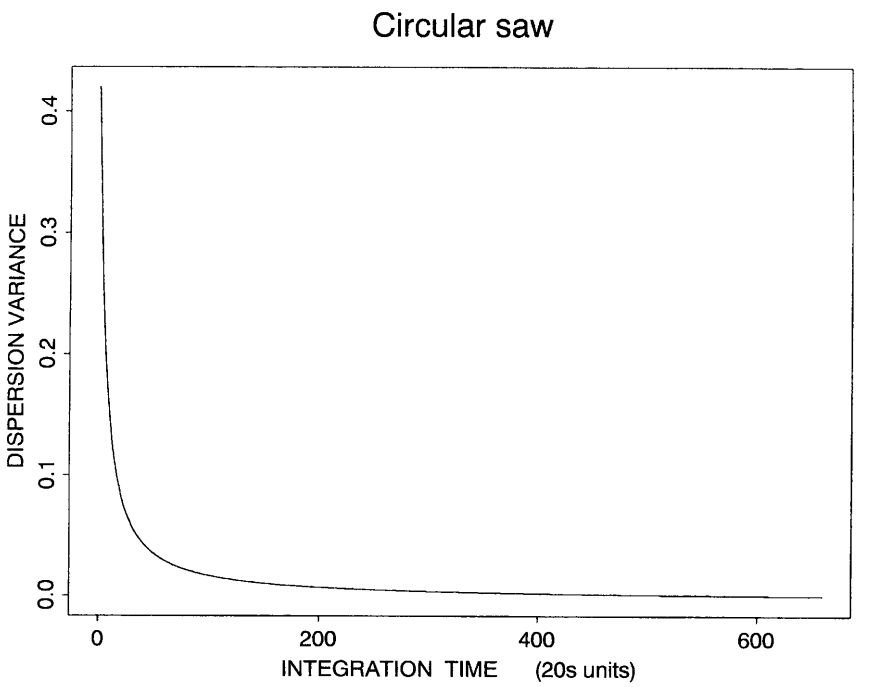

Figure 10.8: Curve of dispersion variances $\sigma^2(\Delta t|\mathcal{D})$ as a function of the integration time Δt with fixed $\mathcal{D}$.

model regularized over time lags of 20 seconds, 1 and 5 minutes illustrating the effect of a modification of the support on the shape of the theoretical variogram.

Finally a curve of the dispersion variance of the sound exposure as a function of the integration time Δt is represented on Figure 10.8. The dispersion variance for an exponential model is calculated with the formula

$$\begin{aligned} \sigma^2(\Delta t|\mathcal{D}) &= \overline{\gamma}(\mathcal{D},\mathcal{D}) - \overline{\gamma}(\Delta t, \Delta t) \\ &= F(\mathcal{D}) - F(\Delta t), \end{aligned} \tag{10.25}$$

where, for $L = \Delta t, \mathcal{D}$,

$$F(L) = b\left(1 + \frac{2a}{L}\left(\frac{a}{L} - 1\right) - \frac{2a^2}{L^2}\exp\left(-\frac{L}{a}\right)\right). \tag{10.26}$$

As the practical range of the variogram is relatively short (2.4 minutes), it can be learned from Figure 10.8 that for a time integration support of less than 1/2 hour (90 time units) a small increase of the support leads to large dropping of the dispersion variance. Conversely it does not seem to make much difference if the integration is changed from 1 hour to 2 hours. With a short practical range the essential part of the variability can only be recovered using an integration time much shorter than 1/2 hour.

Comparison of sampling designs

The concepts of estimation and dispersion variance can be used to compare three sampling designs with n samples

A - regular grid: the domain $\mathcal{D}$ is partitioned into n cubic cells v at the center of which a sample $z(\mathbf{x}_\alpha)$ has been taken;

B - random grid: the n samples are taken at random within the domain $\mathcal{D}$;

C - random stratified grid: the domain $\mathcal{D}$ is partitioned into n cubic cells v inside each of which one sample is taken at random.

For ***design A***, with a regular grid the global estimation variance σ^2_{EG} is computed as

$$\begin{aligned}\sigma^2_{\mathrm{EG}} &= \mathrm{var}\Big(Z^\star_{\mathcal{D}} - Z_{\mathcal{D}}\Big) \\ &= \mathrm{E}\Big[\Big(\frac{1}{n}\sum_{\alpha=1}^{n} Z(\mathbf{x}_\alpha) - \frac{1}{n}\sum_{\alpha=1}^{n} Z(v_\alpha)\Big)^2\Big] \\ &= \mathrm{E}\Big[\Big(\frac{1}{n}\sum_{\alpha=1}^{n} (Z(\mathbf{x}_\alpha) - Z(v_\alpha))\Big)^2\Big]. \end{aligned} \tag{10.27}$$

If we consider that the elementary errors $Z(\mathbf{x}_\alpha) - Z(v_\alpha)$ are independent from one cell to the other

$$\begin{aligned}\sigma^2_{\mathrm{EG}} &= \frac{1}{n^2}\sum_{\alpha=1}^{n} \mathrm{E}\Big[\Big(Z(\mathbf{x}_\alpha) - Z(v_\alpha)\Big)^2\Big] \\ &= \frac{1}{n^2}\sum_{\alpha=1}^{n} \sigma^2_{\mathrm{E}}(\mathbf{x}_\alpha, v_a). \end{aligned} \tag{10.28}$$

As the points $\mathbf{x}_\alpha$ are at the centers $\mathbf{x}_c$ of cubes of the same size we have for design A

$$\sigma^2_{\mathrm{EG}} = \frac{1}{n}\sigma^2_{\mathrm{E}}(\mathbf{x}_c, v). \tag{10.29}$$

For ***design B***, the samples are supposed to be located at random in the domain (Poisson points). We shall consider one realization z with random coordinates X_1, X_2, X_3. The expectation will be taken on the coordinates. The global estimation variance is

$$\begin{aligned}s^2_{\mathrm{EG}} &= \mathrm{E}_X[(z^\star_{\mathcal{D}} - z_{\mathcal{D}})^2] \\ &= \mathrm{E}_X\Big[\Big(\frac{1}{n}\sum_{\alpha=1}^{n} z(X_1^\alpha, X_2^\alpha, X_3^\alpha) - z(\mathcal{D})\Big)^2\Big]. \end{aligned} \tag{10.30}$$

Assuming elementary errors to be independent, we are left with

$$s^2_{\mathrm{EG}} = \frac{1}{n^2}\sum_{\alpha=1}^{n} \mathrm{E}_X[\,(z(X_1^\alpha, X_2^\alpha, X_3^\alpha) - z(\mathcal{D}))^2\,]. \tag{10.31}$$

We now write explicitly the expectation over the random locations distributed with probabilities $1/|\mathcal{D}|$ over the domain

$$\begin{aligned} s^2_{\mathrm{EG}} &= \frac{1}{n^2}\sum_{\alpha=1}^{n} \int_{x_1}\int_{x_2}\int_{x_3} p(x_1^\alpha, x_2^\alpha, x_3^\alpha)\cdot\Big(z(x_1^\alpha, x_2^\alpha, x_3^\alpha) - z(\mathcal{D})\Big)^2 dx_1\, dx_2\, dx_3 \\ &= \frac{1}{n^2}\sum_{\alpha=1}^{n} \frac{1}{|\mathcal{D}|}\int_{x_1}\int_{x_2}\int_{x_3} \Big(z(x_1^\alpha, x_2^\alpha, x_3^\alpha) - z(\mathcal{D})\Big)^2 dx_1\, dx_2\, dx_3 \\ &= \frac{1}{n^2}\sum_{\alpha=1}^{n} s^2(\cdot|V) \\ &= \frac{1}{n}\, s^2(\cdot|V). \end{aligned} \tag{10.32}$$

Generalizing the formula from one realization z to the random function Z and taking the expectation (over Z), we have for design B

$$\sigma^2_{\mathrm{EG}} = \frac{1}{n}\,\sigma^2(\cdot|V). \tag{10.33}$$

For ***design C***, each sample point is located at random within a cube v_α and the global estimation variance for one realization z is

$$\begin{aligned} s^2_{\mathrm{EG}} &= \mathrm{E}_X[\,(z^\star_{\mathcal{D}} - z_{\mathcal{D}})^2\,] \\ &= \mathrm{E}_X\Big[\,\Big(\frac{1}{n}\sum_{\alpha=1}^{n}\big(\,z(X_1^\alpha, X_2^\alpha, X_3^\alpha) - z(v_\alpha)\,\big)\Big)^2\,\Big] \\ &= \frac{1}{n}\, s^2(\cdot|v). \end{aligned} \tag{10.34}$$

Randomizing z to Z and taking the expectation, we have for design C

$$\sigma^2_{\mathrm{EG}} = \frac{1}{n}\,\sigma^2(\cdot|v). \tag{10.35}$$

Comparing the random grid B with the random stratified grid C we know from Krige's relation that

$$\sigma^2(\cdot|v) \leq \sigma^2(\cdot|V), \tag{10.36}$$

and thus design C is a better strategy than design B.

To compare the random stratified grid C with the regular grid A we have to compare the extension variance of the central point in the cube

$$\sigma^2_{\mathrm{E}}(\mathbf{x}_c, v) = 2\,\overline{\gamma}(\mathbf{x}_c, v) - \overline{\gamma}(v, v), \tag{10.37}$$

with the dispersion variance of a point in the cube

$$\sigma^2(\cdot|v) = \overline{\gamma}(v, v). \tag{10.38}$$

It turns out that the former is usually lower than the latter (see [51], p136, for a numerical example). The regular grid is superior to the random stratified grid from the point of view of global dispersion variance as the samples cover evenly the region.

However for computing the experimental variogram, an advantage can be seen in using unequally spaced data: they will provide more information about small-scale variability than evenly placed samples. This helps in modeling the variogram near the origin.