OPTIMAL ESTIMATION OF THE AVERAGE AREAL RAINFALL OVER A BASIN AND OPTIMAL SELECTION OF RAIN-GAUGE LOCATIONS

B. Lorent, G. Bastin, C. Duqué and M. Gevers

Louvain University, Bâtiment Maxwell, Place du Levant, B-1348 Louvain-la-Neuve, Belgium

ABSTRACT. Using a linear minimum variance unbiased estimation procedure for the estimation of spatially distributed random variables (called "kriging" by geostatisticians) we solve the following problems : 1) estimate the average areal rainfall over a catchment area from measurements in a few rain-gauges, and 2) find the measurement locations that will give the most accurate estimate of this areal rainfall. Furthermore we show that the estimated areal rainfall depends only on the location of the rain-gauges, and that one can find the measurement locations leading to the smallest estimation error variance.

KEYWORDS. Estimation ; 2-dimensional interpolation ; optimal location ; rainfall ; random fields.

1. INTRODUCTION.

The estimation of the average areal rainfall over a catchment area is an important step in many hydrological applications. For example, the average rainfall over a river basin is the main input to any rainfall-riverflow model [1].

Following previous contributions [2,3], we propose a probabilistic method for the estimation of the average areal rainfall. The rainfall over a basin is modelled as a 2-dimensional random field. This approach allows us to take into account, in a rigorous and systematic way, the seasonal and spatial variability of the rainfall process.

The paper is organized as follows. In section 2 we introduce the notations, we define the average areal rainfall, and we show how to compute an optimal (unbiased, minimum variance) estimation of this average rainfall. The optimal estimator requires the knowledge of the covariance function or, alternatively, of the variogram of the rainfall process as a function of space. We argue that the variogram is preferable. The estimation of a model for this variogram turns out to be the most difficult and critical part of the spatial extrapolation procedure. Sections 3 and 4 are concerned with this problem and are the main contribution of our paper : we propose estimators for the variogram under different sets of assumptions on the rainfall process, and we study the influence of seasonal variations and the rainfall intensity on the estimators of the variogram. This leads to a systematic procedure for a practical implementation of the areal rainfall estimation. This is the subject of section 4, where we also present an application of our procedure to real data. Finally, in section 5 we show how the optimal extrapolation method developed earlier can also be used to optimally select the location of rainfall gauges in the catchment area.

2. OPTIMAL ESTIMATION OF THE AREAL AVERAGE RAINFALL.

The rainfall random field.

The "rainfall function" is denoted $p(k,z)$. It is the volume per unit area of precipitated water at the point z during the time period of index k. $p(k,z)$ is thus a real-valued function on $(N \times R^2)$ with $k \in N$: the discrete time coordinate, and $z = (x,y) \in R^2$: the continuous cartesian space coordinate. For a fixed k, the function $p(k,z)$ is viewed as a realisation of a 2-dimensional random field (RF) denoted $P(k,z)$. The mean and the variance of this random field are assumed to be space-stationary (i.e. independent of z) and are written :

$$m_p(k) = E\,[P(k,z)] \tag{1}$$

$$\sigma_p^2(k) = E[\{P(k,z) - m_p(k,z)\}^2] \tag{2}$$

For a fixed k, one defines the spatial covariance of the RF :

$$C(k,i,j) = E[\{P(k,z_i)-m_p(k)\}\{P(k,z_j)-m_p(k)\}] \tag{3}$$

with (z_i,z_j) a pair of points in R^2. One also defines the spatial variogram (which is the name given by geostatisticians [4] to the semi-variance of increments of the RF) :

$$\gamma(k,i,j) = \frac{1}{2}\,E[\{P(k,z_i)-P(k,z_j)\}^2] \tag{5}$$

It is easy to show that :

$$\gamma(k,i,j) = \sigma_p^2(k) - C(k,i,j) \tag{6}$$

The measurement stations

Consider a catchment area (it is most often a river basin) $\Omega \in R^2$ with N rainfall measurement stations numbered 1 to N. For the time period k, the measurements are thus specific numerical values of the function $p(k,z)$: $p(k,z_1)$, $p(k,z_2)$, ..., $p(k,z_N)$.

As a matter of illustration, we present later in the paper two applications : the Semois river basin (fig. 1, 1230 km^2) with 17 stations and a time period of 1 day, and the Dyle river basin with 16 stations and a time period of six hours.

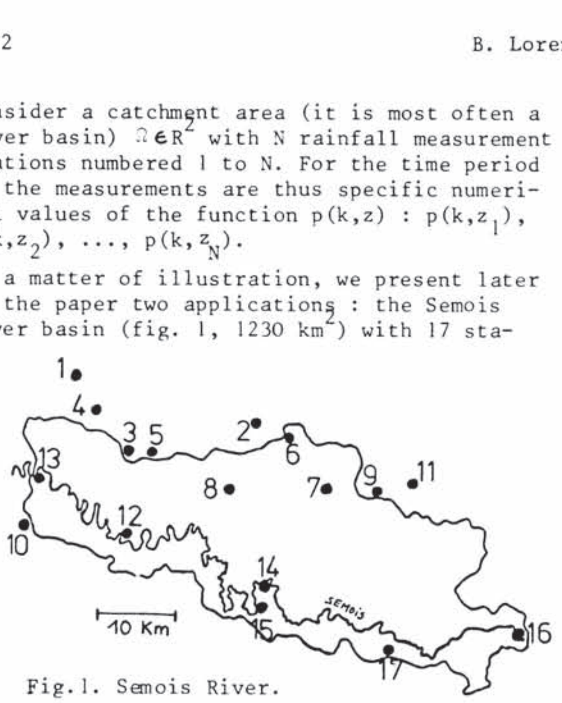

Fig.1. Semois River.

The average areal rainfall

A discretisation square grid of M nodes is superimposed to the catchment area Ω . The M nodes are numbered N+1 to N+M. The average rainfall for the period k is defined as the average rainfall, taken over all the grid nodes, for the time period k :

$$A(k) = \frac{1}{M} \sum_{j=1}^{M} p(k,z_{N+j}) \qquad (7)$$

Note that A(k) is a space-average, not a time-average; it is a discrete 1-D stochastic process with variance :

$$\sigma_A^2(k) = \frac{1}{M^2} \sum_{i=1}^{M} \sum_{j=1}^{M} c(k,N+i,N+j)$$

$$= \sigma_p^2(k) - \delta_A(k) \qquad (8)$$

where

$$\delta_A(k) = \frac{1}{M^2} \sum_{i=1}^{M} \sum_{j=1}^{M} \gamma(k,N+i,N+j) \qquad (9)$$

Clearly $\sigma_A^2(k)$ is smaller than $\sigma_p^2(k)$, since $\delta_A(k)$ is positive. Expression (7) is the definition of the average areal rainfall; it is not computable, however, since the rainfall is in general unknown at the M grid nodes. Therefore we now seek a minimum variance unbiased linear estimate of A(k), obtained from the set of N rainfall observations $\{p(k,z_1),\dots,p(k,z_N)\}$ during the time period k. The estimator will thus have the following form :

$$\hat{A}(k) = \sum_{i=1}^{N} \lambda_i \, p(k,z_i) \qquad (10)$$

Following the classical linear minimum-variance estimation theory, it is easy to show that the λ_i's are the solution of the following linear system (the "kriging" system in the esoteric geostatistics language [5]) :

$$\sum_{j=1}^{N} \lambda_j \gamma(k,i,j) + \mu = \sum_{j=1}^{M} \gamma(k,i,N+j)$$

$$i = 1, \dots, N \qquad (11)$$

$$\sum_{i=1}^{N} \lambda_i = 1$$

where μ is a Lagrange coefficient. The estimation variance $\sigma_E^2(k)$ is written :

$$\sigma_E^2(k) = - \delta_A(k) + \mu + 2 \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \gamma(k,i,N+j) \qquad (12)$$

where μ and the λ_i are the solutions of the linear system (11).

3. IDENTIFICATION OF A VARIOGRAM MODEL.

The optimal λ_i are computed by the linear system (11) from the knowledge of the variogram $\gamma(k,i,j)$. The optimal λ_i can also be expressed as functions of the covariance functions C(k,i,j) instead of the variogram. We prefer the variogram formulation for two main reasons :

a) the variogram $\gamma(i,j,k)$ of the random field can be identified from the available data without any preliminary knowledge or estimation of the mean m_p of the rainfall process.

b) the class of admissible RF models is wider with the variogram formulation. For example, 2-D Wiener models can be used : for such models the variance σ_p^2 is infinite but the variogram exists.

In practice the variogram is not given and must be inferred from the data. We now study the estimation of the variogram under different sets of assumptions.

First set of assumptions

Assume that the RF P(k,z) is ergodic and stationary in time and space. Then :

a) $\gamma(k,i,j)$ is independent of k : $\gamma(k,i,j) = \gamma(i,j)$

b) $\gamma(i,j)$ depends only on the Euclidean distance d_{ij} between z_i and z_j :

$\gamma(i,j) = \gamma(d_{ij})$

Now assume that, in a catchment area, there are K time periods with a rainfall event. From the observations in the rain-gauges located at points z_i and z_j, the following unbiased estimate of $\gamma(d_{ij})$ is obtained :

$$\hat{\gamma}(d_{ij}) = \frac{1}{2K} \sum_{k=1}^{K} \{p(k,z_i) - p(k,z_j)\}^2 \qquad (13)$$

Such an estimate has been computed for every pair of rain-gauges in the Semois river basin (7 years of daily observations) and in the Dyle river basin (3 years of six hourly observations). The results are graphically presented on figs 2 and 3.

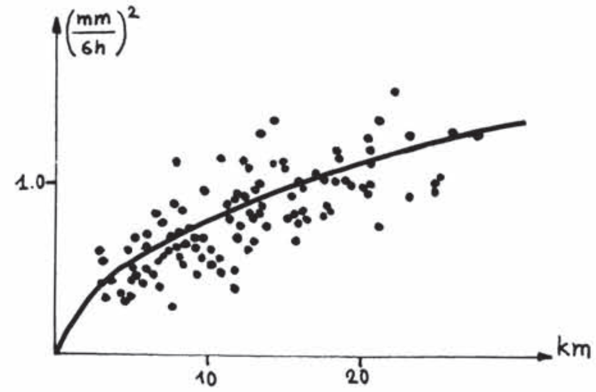

Fig.2. Dyle river: Experimental and theoretical variograms.

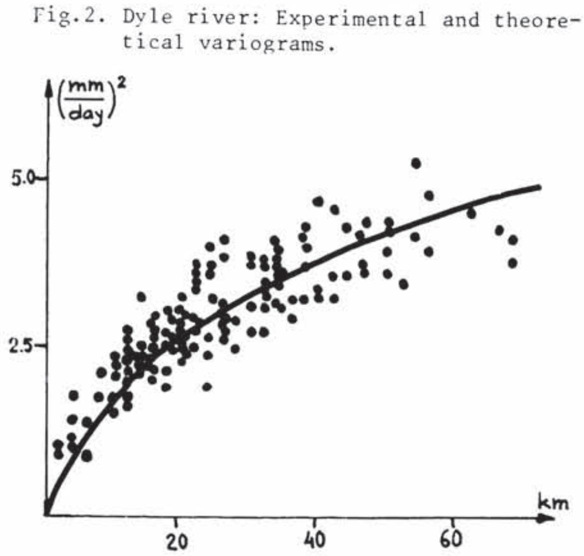

Fig.3. Semois river: Experimental and theoretical variograms.

It appears clearly that, even with several thousands of observations at each maesurement point, the experimental variogram $\hat{\gamma}(d_{ij})$ takes the form of a somewhat extended cluster op points. Therefore the identification of a theoretical variogram model, which conforms with the experimental cluster, is required in order to compute the $\gamma(k,i,j)$ needed in (11)-(12). Based upon many experimental results such as those of figs 2 and 3, and in line with common practice in the geostatistical litterature, we adopt the following very simple model :

$$\gamma(d_{ij}) = \alpha\, d_{ij}^{\beta} \qquad (14)$$

By a least squares fit, the following values were obtained (for d_{ij} expressed in kilometers) :

Semois river : $\alpha = 1.12$ $\beta = 0.51$
Dyle river : $\alpha = 0.204$ $\beta = 0.56$

Second set of assumptions

The time-stationary assumption of the RF is most probably unrealistic because :

- it does not take into account the potential seasonal trends of the phenomenon;
- it yields a unique estimation standard deviation σ_E^2 of the average areal rainfall (see (12)) for all rainfall events, whatever the meteorological conditions and the rainfall intensity. This is not very plausible.

In order to refine the analysis we therefore assume a piecewise stationary seasonal trend (on a monthly basis) for the RF. More precisely we assume that the variance σ_p^2 (if it exists) and the variogram $\gamma(k,i,j)$ are stationary and ergodic in space and time during a month but not necessarily from one month to another.

Typical examples of monthly experimental variograms are presented on figs 4 and 5. For graphical clarity, the clusters of points have been approximated by a broken line which is obtained by dividing the d_{ij} axis into a number of classes and by computing the mean of $\hat{\gamma}(d_{ij})$ for all points d_{ij} which belong to the same class.

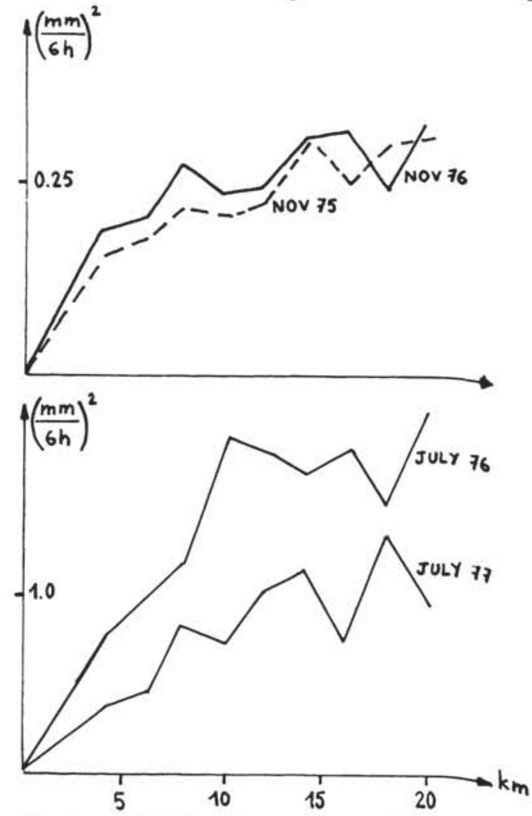

Fig.4 and 5,Experimental monthly variograms.

These figures clearly show the seasonal behaviour of the spatial variability of the rainfall process : the variogram $\gamma(d_{ij})$ nears much larger in the Summer than in Autumn.

Note that the scales of figs. 4 and 5 are different.

Similar trends have been observed all through every year for which we had data. Therefore we have divided all the available rainfall data into 12 classes, one for each month ; for example, the data of November 1975 and

and November 76 are taken in the same class and processed together. We have then computed a theoretical variogram of the form αd_{ij}^{β} for each month by least-squares fitting, as described above.

Typical results are shown on fig. 6 and 7, and in Table 1. The results clearly show the seasonal patterns of the variograms.

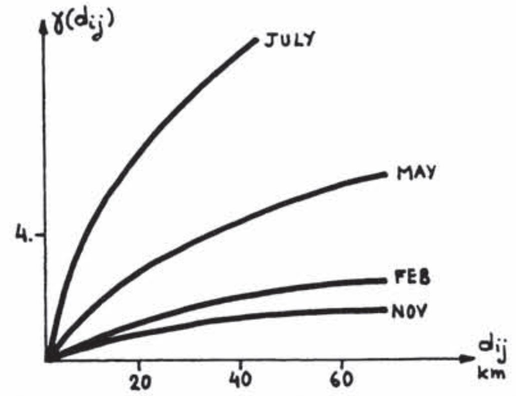

Fig.6. Semois river:Estimated theoretical monthly variograms.

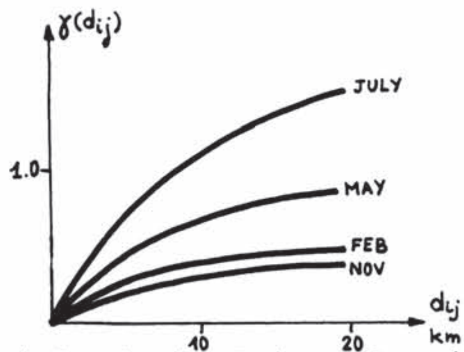

Fig.7. Dyle river:Estimated theoretical monthly variograms.

	Semois		Dyle	
	α	β	α	β
January	0.29	0.63	0.067	0.44
February	0.55	0.54	0.063	0.59
March	0.45	0.62	0.072	0.60
April	0.69	0.47	0.221	0.29
May	1.02	0.59	0.362	0.52
June	3.14	0.40	0.673	0.30
July	3.17	0.51	0.368	0.54
August	2.22	0.51	0.505	0.51
September	1.06	0.53	0.144	0.54
October	0.41	0.74	0.042	0.56
November	0.25	0.62	0.105	0.49
December	0.64	0.53	0.090	0.61
Global	1.12	0.51	0.206	0.56

Table 1.

Influence of the rainfall intensity

One might wonder whether the seasonal variations in the variogram are not greatly amplified by the differences between the mean rainfall intensity in Summer and Winter. More specifically, are the larger values of the variogram in the Summer not caused by the higher intensity of the rainfalls during that season rather than by a truly larger spatial variability ? In order to answer that question we compute the seasonal variation of the experimental rainfall variance for different intensity ranges. We define the rainfall intensity during period k as the average rainfall taken over the N measurement stations during that period :

$$\hat{\mu}_p(k) = \frac{1}{N}\sum_{i=1}^{N} p(k,z_i) \qquad (15)$$

We also define the experimental variance :

$$\hat{\sigma}_p^2(k) = \frac{1}{N}\sum_{i=1}^{N} [p(k,z_i) - \hat{\mu}_p(k)]^2 \qquad (16)$$

Fig. 8 represents the histogram of the intensities for the Dyle river basin over the 3 year period mentioned earlier. The wide range

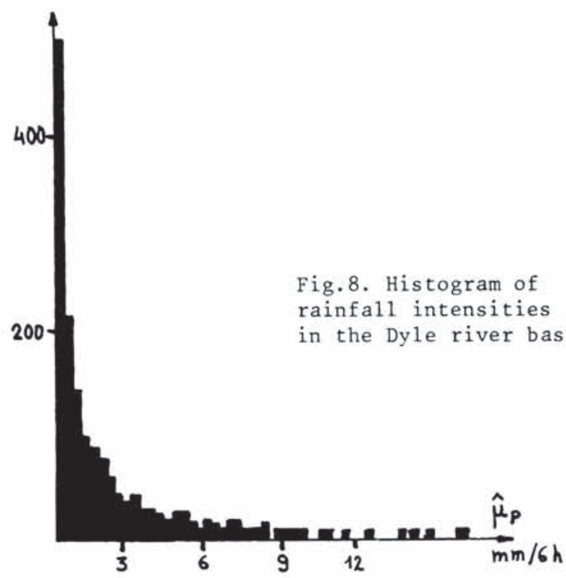

Fig.8. Histogram of rainfall intensities in the Dyle river basin.

of intensities observed in fig. 8 certainly increases the relevance of our question. In order to study the relation between the rainfall intensities and the standard deviation of these rainfalls in different seasons, we have partitioned all the available data into 4 seasons and a reasonable number of intensity ranges. Fig. 9 shows the normalized standard deviation of the rainfall $\hat{\sigma}_p/\hat{\mu}_p$ versus the rainfall intensity $\hat{\mu}_p$ for each of the four seasons, with a logarithmic scale for $\hat{\mu}_p$. The Fall and Winter curves are so close that they have been drawn together. The figure shows that, even when rainfalls of same intensity levels are considered, a seasonal trend is still clearly present in $\hat{\sigma}_p$. It also shows that the normalized standard deviation (and therefore also the variogram) is a function of the rainfall

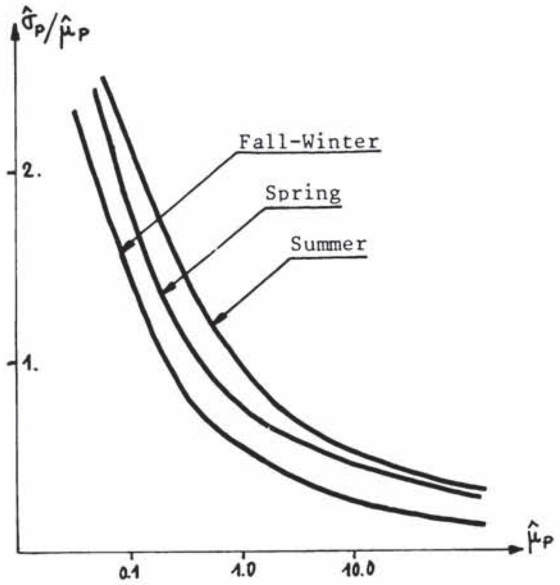

Fig.9. Seasonal behavior of the RF.

intensity. This is not surprising, but it suggests that choosing a unique variogram model within a given season would lead to a systematic underevaluation of σ_E^2 for high-intensity rainfalls, and an overevaluation for low-intensity rainfalls.

4. PRACTICAL IMPLEMENTATION OF THE AVERAGE AREAL RAINFALL ESTIMATOR.

The optimal estimation of the average areal rainfall estimator during period k involves the following steps :

1) estimate the parameters α and β of a variogram model of the form

$$\gamma(k,d) = \alpha(k)\ d^{\beta(k)} \qquad (17)$$

2) compute the coefficients (k,i,j) for all grid points and all measure points
3) compute λ_i, i=1,...,N and μ by solving the kriging system (11)
4) compute $\hat{A}(k)$ by (10) and $\sigma_E^2(k)$ by (12).

Steps 2), 3) and 4) of this procedure pose no particular problem (see section 2). As for the variogram model, we have demonstrated in section 3 the interest of relating the time nonstationarity of the raondom field not only to the season but also to the rainfall intensity.

Variogram estimation : procedure 1.

It follows from Table 1 that the monthly variograms differ much more in the coefficient α than in β. Following this observation, the first procedure for the estimation of the variogram is as follows :

a) estimate a unique coefficient β by a least-squares fit over all available data points (see section 3)
b) for each period k, estimate $\alpha(k)$ by a least-squares fit to the cluster of points corresponding to that particular period. We have shown in [6] that it is actually better to use a weighted least-squares procedure for the estimation of α, where the weighting matrix takes into account the geometrical distribution of the measurement stations. See [6] for more details.

Using this first procedure a nonstationary variogram of the following form is obtained.

$$\gamma(k,d) = \alpha(k)\ \ \gamma^*(d), \text{ with } \ \gamma^*(d)=d^{\beta} \qquad (18)$$

The nonstationarity is concentrated in the scaling factor $\alpha(k)$, which can be interpreted as a measure of the spatial variability of the random field during period k.

For the Semois river basin, application of procedure 1 leads to β = 0.51 (see Table 1), and the nonstationary variogram has the form $\gamma(k,d) = \alpha(k)d^{0.51}$.

Datum	26/1	26/4	18/6	16/8
$\hat{\alpha}$	209.1	208.1	639.1	830.2
$\hat{\mu}_P$	320.4	182.0	306.2	139.1
$\hat{A}(k)$	318.8	194.6	308.6	127.4
$\hat{\sigma}_E(k)$	9.9	9.9	17.3	19.7
$\hat{\sigma}_E(k)/\hat{A}(k)$	0.03	0.05	0.06	0.15

Table 2.

Table 2 shows the results of the estimation of $\hat{A}(k)$ for a few days chosen in the year 1971. (Rainfall data : 0.1 mm/day)

Variogram estimation : procedure 2.

The theorical variogram is now written as

$$\gamma(k,d) = \alpha(k)\ \ \gamma^*(d),$$
$$\text{with } \gamma^*(d) = \gamma_o d^{\beta}$$

The coefficients γ_o, β and $\alpha(k)$ are determined as follows :

a) compute the parameters α and β of a global variogram of the form $\gamma(d) = \alpha d^{\beta}$ by least-squares fit over all available data. This determines β.
b) compute the global experimental variance $\hat{\sigma}_p^2$ using all available data. This determines $\gamma_o = \alpha/\hat{\sigma}_p^2$.
c) now for each period k compute the rainfall intensity using (15).
d) from the graphs relating $\hat{\sigma}_p^2/\hat{\mu}_p$ to $\hat{\mu}_p$, compute the value of $\hat{\sigma}_p^2(k)$ corresponding to the value $\hat{\mu}_p(k)$ computed in c) using the appropriate seasonal curve (see fig. 9). This determines $\alpha(k) = \hat{\sigma}_p^2(k)$ and finally

$$\gamma(k,d) = \alpha(k)\gamma_o d^{\beta} = \frac{\hat{\sigma}_p^2(k)}{\hat{\sigma}_p^2}\ \alpha d^{\beta} \qquad (20)$$

Notice that this variogram is nonstationary, and that the nonstationary scaling factor $\alpha(k)$ takes into account both the seasonal variations and the effects of the intensity through the use of the seasonal graph

($\hat{\sigma}_p/\hat{\mu}_p-\hat{\mu}_p$) with $\hat{\sigma}_p$ computed by (15).

The motivation for this procedure will be best illustrated by applying it to the Dyle river data. Recall (see Table 1) that we have identified a mean global variogram $\gamma(d) = 0.206\ d^{0.56}$. The experimental variance $\hat{\sigma}_p^2$ over the 3 years of available measurements is $\hat{\sigma}_p^2 = 1.264$. Therefore, in accordance with (6), one can write

$$\gamma(d) = \hat{\sigma}_p^2\ \gamma^*(d) \quad \text{for } d \leq 26 \text{ km} \qquad (21.a)$$

$$\gamma(d) = \hat{\sigma}_p^2 \quad \text{for } d > 26 \text{ km} \qquad (21.b)$$

with $\gamma^*(d) = 0.1614\ d^{0.56}$. In practice only (21.a) is of interest since no two measurement stations are separated by more than 26 km. It makes sense, therefore, to take into account the nonstationarities due to seasonal and intensity variations through the scaling factor $\alpha(k)$ by writing $\gamma(k,d) = \alpha(k)\gamma^*(d) = \alpha(k)\ 0.1614\ d^{0.56}$. Hence $\gamma_o = 0.1614$, $\beta = 0.56$, and $\alpha(k) = \hat{\sigma}_p^2(k)$, determined by steps c) and d) above.

Comment

In the 2 procedures the variogram takes the form $\gamma(k,d) = \alpha(k)\ \ \gamma^*(d)$, where $\alpha(k)$ is time varying and determined in "real time" for each time period k, while $\gamma^*(d)$ is time-invariant. This has the following important implications :

a) the coefficients λ_i in (10) are independent of k; they depend only upon the geometrical location of the rain-gauges and can be computed once and for all.

b) the variance of the estimate can be written

$$\sigma_E^2(k) = \alpha(k)\ (\sigma_E^*)^2 \qquad (22)$$

where σ_E^* is independent of k and can also be computed once and for all.

n°	$(\sigma_E^*)^2$	n°	$(\sigma_E^*)^2$	n°	$(\sigma_E^*)^2$	n°	$(\sigma_E^*)^2$
18	1.86	13	0.64	7	0.44	11	0.40
12	1.31	17	0.58	14	0.43	4	0.39
19	0.93	15	0.53	6	0.42	2	0.39
8	0.79	3	0.50	10	0.41	1	0.39
9	0.71	16	0.47	5	0.41		

Table 3.

5. OPTIMAL SELECTION OF RAINFALL GAUGES LOCATIONS.

One practically important problem is that of determining, among a set of possible rainfall measurement locations, those that are most representative, in the sense that they will lead to the smallest estimation error variance for the estimation of the average areal rainfall.

Now recall that the normalized standard deviation error, σ_E^*, depends only upon the location of the rainfall gauges and upon the structure of the normalized variogram $\gamma^*(d_{ij})$, and is independent of the rainfalls or their intensities. It is therefore easy, using the methods described in this paper, to select among the available measurement stations the one that leads to the smallest σ_E^*. Next one can add to this first station a second station which, combined with the first one, leads to a minimum σ_E^* again. This procedure can be continued, adding more stations and monitoring the decrease of the normalized standard deviation σ_E^* of the estimation error, until the obtained precision is judged satisfactory.

This method has been applied to the Semois river basin in order to chooze the "most representative" locations among 17 existing measurement stations and two potential supplementary locations (numbered 18 and 19 on Fig. 1). The result of this successive selection procedure is illustrated in table 3. Notice that the last 7 stations chozen (Nrs 6, 10, 5, 11, 4, 2, 1) by our successive selection procedure are obviously superfluous, since including them in the optimal estimator does not result in any significant decrease of σ_E^*. With only the"best" 3 stations (18, 12, 19) the coefficients of the optimal estimator are $\lambda_{18}=0.356$, $\lambda_{12}=0.327$, $\lambda_{19}=0.317$, and $\sigma_E^*=0.97$ as compared with $\sigma_E^*=0.65$ when 17 stations are used.

6. REFERENCES.

[1] Lorent B. and Gevers M., Identification of a Rainfall-Runoff process., 4th IFAC Symp. on Identification & System Parameter Estimation, Tbilisi, USSR, Sept. 21-27, 1976.

[2] Creutin J.D. and Obled C., Modelling Spatial and Temporal Characteristics of Rainfall as Input to Flood Forecasting Model, Proceedings of the Oxford Symp., April 1980.

[3] Rodriguez-Iturbe I. and Mejia J.M., The Design of Rainfall Network in Time and Space, Water Resources Research, vol.10 (4), August 1974.

[4] Matheron G., La théorie des variables régionalisées et ses applications, Cahiers du Centre de Morphologie Mathématique, Fasc. 5, Fontainebleau, 1970.

[5] Delfiner P. and Delhomme J.P., Optimum Interpolation by Kriging, in "Display and Analysis of Spatial Data", Davis and Mc Cullagh Ed. J. Wiley, 1975.

[6] Gevers M. and BASTIN G., On the estimation of the Variogram in Spatial Interpolation Methods, in "Applications of Information and Control Systems", Lainiotis and Tzannes Ed., D. Reidel Publ. Cy, 1980.