Spatial Statistics: Use and misuse of analytical procedures – an obstacle run

Pierre R. L. Dutilleul, Prof.
McGill University
Pierre.Dutilleul@McGill.CA

A preliminary note/question

Spatial Statistics – or – Geostatistics?

Are they synonyms?

Is one of these two branches of Statistical Sciences included in the other?

Can they be seen as two branches with a non-empty intersection?

If the third option, it would be justified by the facts that spatial correlograms like those based on Moran's I and Geary's c statistics (with tests of significance of the ordinates) are not used in geostatistics; variograms are used, analyzed and modeled in both branches; and the concept of regionalized variable (instead of stochastic process) seems to be specific to Geostatistics (in simple terms).

The five major points/questions of the day

In Spatial Statistics,

. A simple random sampling is not recommended!

. Our eyes are OLS! – What does that mean? Is it 'good' to 'be' OLS, or use an OLS procedure?

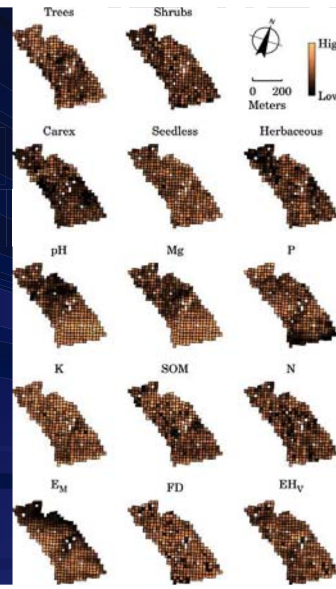
. In general, there is not one mean, but a mean function for the random variable of interest.

. There is autocorrelation in data to be used for correlation analysis. – What is the effect? Should we be concerned?

. There may be correlation at more than one scale; see the concept of structural correlations.

…

Plant-soil example used for illustration (Pelletier et al., 2009a, b). where sampling was more systematic than random.



In Spatial Statistics, a simple random sampling is not recommended!

9.1.1 Basic sampling designs

The four sampling grids (n = 100) of Fig. 9.1 represent as many main types of sampling design in 2-D space:

- simple systematic [panel (a)] – the sampling grid is square (10 × 10) here, but it could be rectangular, circular, or elliptic, depending on the shape of the sampling domain (i.e., a square of side length 10 in Fig. 9.1);
- multiple systematic, with one grid that covers the sampling domain and on which a number of smaller grids are superimposed, each of the grids corresponding to a systematic sampling in the entire sampling domain or a portion of it [panel (b)] – the former grid is 6 × 6 and the latter are four smaller 4 × 4 grids here, but one 8 × 8 grid and four smaller 3 × 3 grids would have been possible too (n = 100);
- simple random [panel (c)], for which the generator of uniform pseudo-random numbers in a computer program (e.g., SAS, Matlab) can be used to define randomly the spatial coordinates of sampling locations, without any constraint within the sampling domain – the expressions "completely random" and "completely randomized" are reserved for a given type of point pattern (Chapters 3 to 5) and of experimental design (Section 9.2), respectively;
- stratified random [panel (d)], in which the sampling domain is divided into a number of plots of same shape and size, or strata, and a given number of sampling locations are randomly defined in each stratum – 100 squares of side length 1, with one sampling location in each of them, were used as strata here, but 25 squares of side length 2, with four sampling locations randomly defined inside each square, could have been used instead (n = 100).

From Dutilleul (2011, Chapter 9)

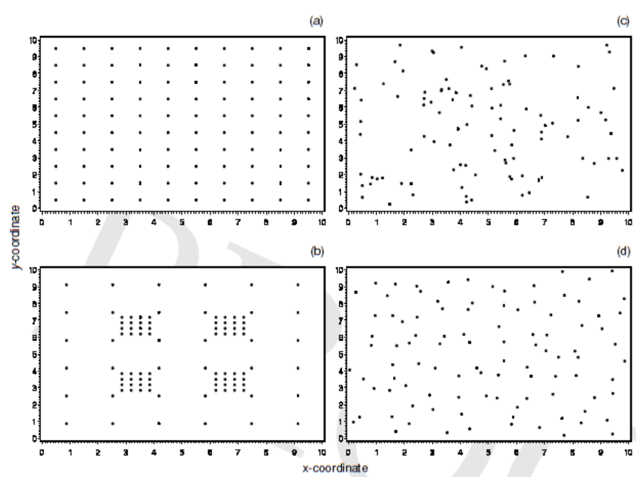


Fig. 9.1. Grids for different sampling designs (n = 100) within the same domain in 2-D space. (a) Simple systematic 10 × 10. (b) Multiple systematic, with one 6 × 6 grid covering the domain and on which four smaller 4 × 4 grids are superimposed. (c) Simple random. (d) Stratified random, with one sampling location inside each of 100 squares of side 1 into which the domain is divided; the spatial coordinates of each sampling location in a square are defined randomly.

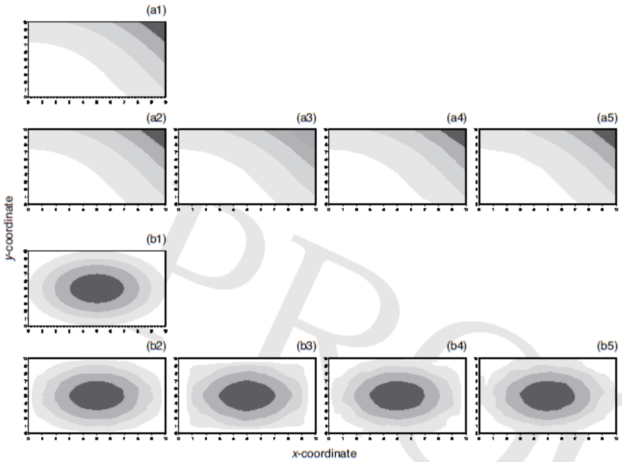


Fig. 9.2. Heterogeneity of the mean in 2-D space. (a1), (b1), and (c1) True deterministic drifts vs. (a2)–(a5), (b2)–(b5), and (c2)–(c5) approximations obtained by interpolation using the four sampling grids of Fig. 9.1. (a1) Purely cubic polynomial trend, $\mu(x, y) = x^3 + y^3 + x^2y + xy^2$. (b1) Bell-shaped trend surface, or single patch, $\mu(x, y) = \exp(-d(x, y)^2)$, with $d(x, y)^2 = \frac{(x-5)^2+(y-5)^2}{6.75}$. (c1) Cosine waves with a "period" of 5 in both directions, or multiple patches, $\mu(x, y) = \cos(\frac{2\pi x}{5}) + \cos(\frac{2\pi y}{5})$. Gray tones go from light (lower values) to dark (higher values).

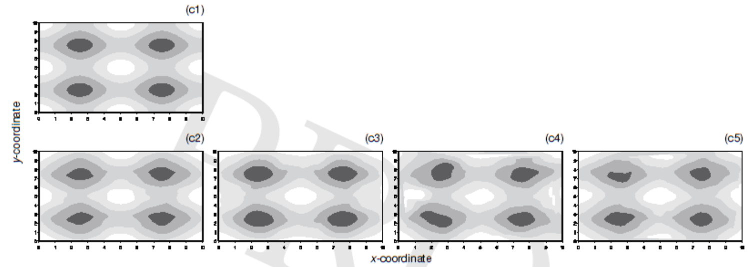


Fig. 9.2. (*cont.*)

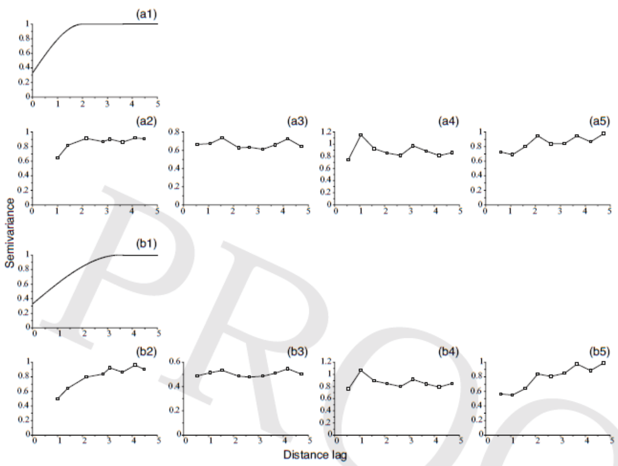


Fig. 9.3. Heterogeneity due to autocorrelation, combined with unequal variance components ($\frac{1}{3}$ against $\frac{2}{3}$) for the nugget effect (non-spatial) and the spherical basic function (small-scale, spatially autocorrelated) in the variogram model. (a1) Theoretical variogram model with a range of 2 for spatial autocorrelation and (a2)–(a5) experimental variograms computed from partial realizations obtained using the four sampling grids of Fig. 9.1 and the same "seed" in simulations. Similarly, the theoretical variogram model with range 3.5 and the four experimental variograms are plotted in (b1) and (b2)–(b5), respectively.

Our eyes are OLS!

What does "OLS" mean?

What is the difference between "OLS" and "WLS", "GLS", "EGLS"?

LS = Least Squares

It is a family of estimation methods in Statistics, based on the minimization of the squared distance between the values of a variable or a function provided by a certain model and the data or statistic values to which that model is fitted.

The OLS, WLS, GLS, and EGLS procedures essentially differ by the metric used to calculate the squared distance between the values predicted by the model and the observed values or preliminary estimates.

Our eyes are OLS! (continued)

OLS: Ordinary Least Squares
Assumes independence and homoscedasticity of 'the data'
Metric: Euclidean, squared distance between **y** and **y**-hat

$$\Sigma(y-\hat{y})^2$$

WLS: Weighted Least Squares
Accounts for heteroscedasticity of 'the data' through weights
Metric: weighted squared distance between **y** and **y**-hat

$$\Sigma\, w(y-\hat{y})^2$$

GLS: Generalized Least Squares
EGLS: Estimated Generalized Least Squares
Aimed to account for heteroscedasticity and autocorrelation in 'the data'

$$(y-\hat{y})'S^{-1}(y-\hat{y})$$

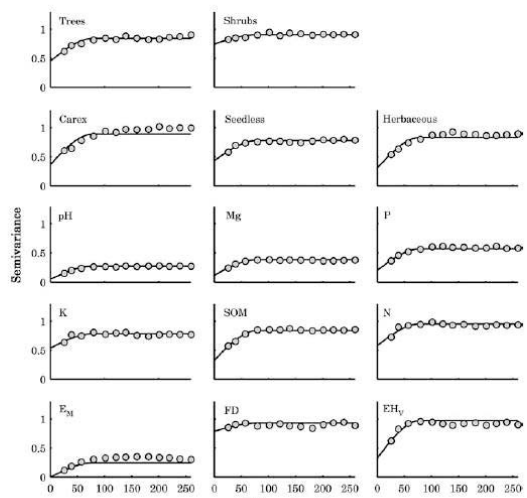


Fig. 5 Direct experimental variogram of residuals for each variable, after removal of the L_1 drift estimates in the example. The solid line represents the variogram model obtained by fitting the LMC

From Pelletier et al. (2009b, EEST)

Our eyes are OLS! (last page)

For an example of EGLS estimation procedure with variograms, see Pelletier et al. (2004)

Back to the question “Is it ‘good’ to ‘be’ OLS, or use an OLS procedure in Spatial Statistics?”, the answer is generally “No” because spatial data and derived coefficients (to which a model needs to be fitted) tend to be autocorrelated and/or show heterogeneity of the variance.

What are the consequences of using an OLS estimation procedure when the conditions for its application are not satisfied?

. In estimation, a bias in the estimated variance of the estimator

. In testing, an inflated Type I error risk if the test statistic and its distribution are not modified accordingly

In Spatial Statistics, there is not one mean, but generally a mean function (alias ‘trend’ or ‘drift’) for the random variable of interest.

Two main options for drift modeling and estimation:

. Global, using a trend surface model (e.g., 2nd or 3rd degree polynomial in spatial coordinates)

. Local, using a moving window with optimized size and a 0, 1st or 2nd degree polynomial in spatial coordinates inside

What is better?

. The local drift estimation approach, with 1st degree polynomial in spatial coordinates inside the window with optimized size

Note: This answer is given on the basis of theoretical and simulation results; see Pelletier et al. (2009a), Phase 1 of the CRAD method.

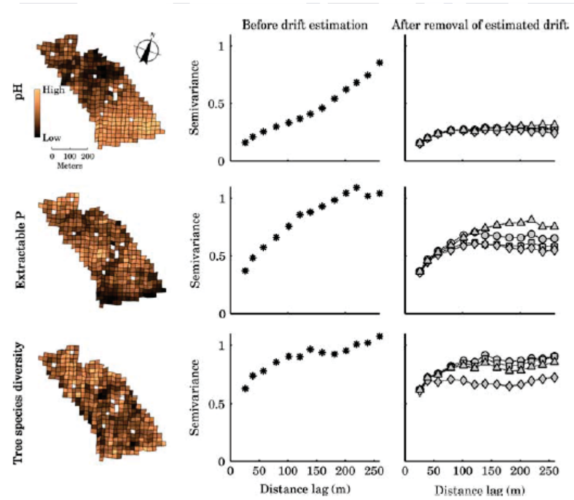


Fig. 7 Maps and the corresponding direct experimental variograms, before and after removal of the estimated drift, for the three variables in the example: pH, extractable P, and tree species diversity. In the direct experimental variograms of residuals (i.e., after drift estimation), symbols for drift estimation procedures are the same as in Fig. 5

From Pelletier et al. (2009a, EEST)

There is autocorrelation in spatial data to be used for correlation analysis. So what?

With the exception of a few very particular cases, the presence of autocorrelation in spatial data is always a problem for correlation analysis and the outcome of the test of significance.

What is the problem?

Spatial autocorrelation introduces a bias in the variance of the correlation estimator (e.g., Pearson’s r statistic), leading to rejection of the null hypothesis of absence of correlation more often than excepted at a given significance level.

A recommended solution: A modified t-test with a number of degrees of freedom (M – 2, instead of N – 2) adjusted for spatial autocorrelation
Reference: Dutilleul (1993)

$$\frac{r}{\sqrt{\dfrac{1-r^2}{M-2}}}$$

Table 1

Theoretically expected numbers of degrees of freedom for the modified t test in assessing the correlation between two first-order simultaneous autoregressive lattice processes X and Y, without (df_1) and with (df_2) complete modification, relative to the number of locations and the values of parameters a_X and a_Y

	$a_X = a_Y = a$							
	$a = .05$		$a = .10$		$a = .15$		$a = .20$	
Number of locations	df_1	df_2	df_1	df_2	df_1	df_2	df_1	df_2
4	2.68	1.92	1.80	1.79	.81	1.65	.19	1.54
	(28.19)		(.84)		(50.97)		(87.74)	
9	7.45	6.67	5.79	5.78	3.33	4.57	.97	3.33
	(10.45)		(.25)		(21.17)		(70.81)	
16	14.12	13.33	11.41	11.35	7.18	8.32	2.54	4.91
	(5.59)		(.50)		(13.71)		(48.32)	
25	22.71	21.92	18.67	18.62	12.11	13.24	4.68	6.86
	(3.48)		(.26)		(8.52)		(31.78)	
36	33.22	32.42	27.59	27.57	18.20	19.39	7.28	9.34
	(2.38)		(.10)		(6.14)		(22.10)	
49	45.64	44.85	38.18	38.17	25.48	26.75	10.32	12.38
	(1.73)		(.02)		(4.76)		(16.68)	
64	59.99	59.20	50.43	50.44	33.95	35.31	13.84	15.98
	(1.31)		(.02)		(3.85)		(13.41)	
100	94.44	93.65	79.90	79.94	54.44	55.96	22.38	24.78
	(.83)		(.05)		(2.71)		(9.70)	
144	136.57	135.77	116.03	115.98	79.74	80.96	32.98	33.95
	(.59)		(.04)		(1.50)		(2.85)	
256	243.87	243.05	208.24	208.21	144.50	145.45	60.57	61.42
	(.34)		(.01)		(.65)		(1.38)	
400	381.94	381.12	326.98	326.97	228.22	229.07	96.19	97.16
	(.21)		(<.01)		(.37)		(1.00)	

	$a_X = -a_Y = a$							
	$a = .05$		$a = .10$		$a = .15$		$a = .20$	
Number of locations	df_1	df_2	df_1	df_2	df_1	df_2	df_1	df_2
4	3.35	2.11	4.78	2.51	9.85	3.44	41.78	5.42
	(36.89)		(47.54)		(65.06)		(87.04)	
9	8.57	7.36	10.69	8.64	16.46	11.58	34.92	19.06
	(14.05)		(10.21)		(29.63)		(45.41)	
16	15.90	14.71	19.24	17.25	28.35	23.56	64.43	42.43
	(7.49)		(10.33)		(16.89)		(34.15)	
25	25.32	24.13	30.16	28.20	43.08	38.45	89.98	70.65
	(4.69)		(6.51)		(10.74)		(21.48)	
36	36.82	35.63	43.46	41.51	61.04	56.46	124.47	104.84
	(3.22)		(4.49)		(7.50)		(15.77)	
49	50.40	49.22	59.15	57.21	82.12	77.58	163.41	144.45
	(2.35)		(3.29)		(5.53)		(11.60)	
64	66.06	64.88	77.22	75.28	106.34	101.82	208.24	189.52
	(1.79)		(2.51)		(4.24)		(8.99)	
100	103.65	102.47	120.51	118.58	164.18	159.71	314.21	296.16
	(1.14)		(1.60)		(2.72)		(5.74)	
144	149.58	148.39	173.37	171.40	234.67	229.69	442.60	419.49
	(.80)		(1.14)		(2.12)		(5.22)	
256	266.44	265.25	307.69	305.63	413.50	407.79	767.52	740.32
	(.45)		(.67)		(1.38)		(3.54)	
400	416.68	415.46	480.47	478.30	643.37	637.23	1182.25	1152.65
	(.29)		(.45)		(.95)		(2.50)	

Numbers in parentheses are percentages $100 \times |df_1 - df_2| / \max(df_1, df_2)$.

There is autocorrelation in spatial data to be used for correlation analysis. So what? (bis)

For partial correlations, see Alpargu and Dutilleul (2006).

For the multiple-correlation case, see Dutilleul et al. (2008).

For the case of structural correlations (see the next and last major point/question of the day), the reference is Dutilleul and Pelletier (2011).

And there is more work in progress and recent results are submitted for publication.

Note: Computer programs (Matlab and non-Matlab versions) are available; feel free to visit http://environmetricslab.mcgill.ca.

There may be correlation at more than one scale in spatial data; see the concept of structural correlations

Let $Z_1(x, y) = Z_{11}(x, y) + Z_{12}(x, y)$, $Z_2(x, y) = Z_{21}(x, y) + Z_{22}(x, y)$ be two 2-D spatial processes with a random non-spatial component and a random spatially autocorrelated component,

such that $\mathrm{Cov}(Z_{11}(x, y), Z_{12}(x, y)) = 0.0$,
$\mathrm{Cov}(Z_{21}(x, y), Z_{22}(x, y)) = 0.0$,
$\mathrm{Cov}(Z_{11}(x, y), Z_{21}(x, y)) = -0.5$,
$\mathrm{Cov}(Z_{12}(x, y), Z_{22}(x, y)) = +0.5$,

so that $\mathrm{Cov}(Z_1(x, y), Z_2(x, y)) = 0.0!$

How to have a chance to find the non-spatial and spatial correlations of −0.5 and +0.5? Answer: Through the analysis of cross-variograms and the EGLS fitting of a linear model of coregionalization to experimental variograms.

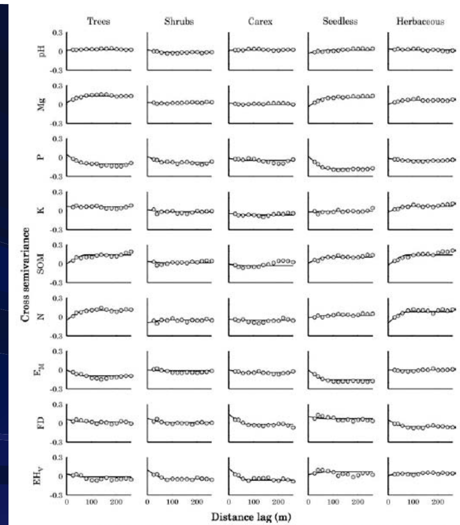


Fig. 6 Cross experimental variogram of residuals for each species group diversity with each of the explanatory variables (soil and topographical indices), after removal of the L_1 drift estimates. The solid line represents the variogram model obtained by fitting the LMC

From Pelletier et al. (2009b, EEST)

Table 6 Estimates of structural and pseudo correlations between diversity and each explanatory variable, obtained with L_1 in the example

	Correlations								
	pH	Mg	P	K	SOM[a]	N	E_M^b	FD	EH_V
Nugget									
Trees	0.11	0.13	0.16	0.14	−0.01	−0.11	–	0.10	0.09
Shrubs	0.12	0.11	0.04	0.03	0.07	−0.14	–	0.10	0.21
Carex	0.11	0.11	−0.05	−0.10	−0.03	−0.08	–	0.26	0.36
Seedless	−0.19	−0.17	0.06	−0.05	0.03	−0.03	–	0.17	0.06
Herbaceous	0.28	0.02	−0.06	−0.06	−0.07	−0.23	–	0.10	0.04
Spherical (79 m)									
Trees	0.04	0.32	−0.40	−0.03	0.32	0.42	−0.30	−0.17	−0.09
Shrubs	−0.21	−0.04	−0.39	−0.19	−0.08	0.17	−0.06	−0.45	−0.44
Carex	0.02	−0.07	−0.07	−0.07	−0.04	−0.04	−0.12	−0.58	−0.32
Seedless	0.18	0.48	−0.55	0.05	0.23	0.11	−0.54	−0.14	0.10
Herbaceous	−0.04	0.17	−0.07	0.27	0.31	0.40	0.00	−0.38	0.06

Note: The estimated value of a structural correlation is calculated following the formula of Pearson's r, by using the nugget effects estimated from the direct and cross variograms for the non-spatial correlation and by using the partial sills of the same direct and cross variograms for the spatial correlation.

Closing Remark

In the Statistical Sciences, which include Spatial Statistics, many (good) things can be discussed simply in terms of means, variances, covariances and correlations. The particle "auto" in "autocorrrelation" and "autocovariance" is specific to Temporal and Spatial Statistics, where Heterogeneity can be the source of 'obstacles' before the data analyst can reach The Truth…

References

Alpargu, G., Dutilleul, P. 2006. Stepwise regression in mixed quantitative linear models with autocorrelated errors. *Communications in Statistics* 35, 79–104.

Dutilleul, P., 1993. Modifying the *t*-test for assessing the correlation between two spatial processes. Biometrics 49, 305–314.

Dutilleul, P., 2011. *Spatio-Temporal Heterogeneity: Concepts and Analyses*. Cambridge: Cambridge University Press.

Dutilleul, P., Pelletier, B., 2011. Tests of significance for structural correlations in the linear model of coregionalization. Mathematical Geosciences 43, 819–846.

Dutilleul, P., Pelletier, B., Alpargu, G., 2008. Modified *F*-tests for assessing the multiple correlation between one spatial process and several others. Journal of Statistical Planning and Inference 138, 1402–1415.

Pelletier, B., Dutilleul, P., Larocque, G., Fyles, J.W., 2004. Fitting the linear model of coregionalization by generalized least squares. Mathematical Geology 36, 323–343.

Pelletier, B., Dutilleul, P., Larocque, G., Fyles, J.W., 2009a. Coregionalization analysis with a drift for multi-scale assessment of spatial relationships between ecological variables 1. Estimation of drift and random components. Environmental and Ecological Statistics 16, 439–466.

Pelletier, B., Dutilleul, P., Larocque, G., Fyles, J.W., 2009b. Coregionalization analysis with a drift for multi-scale assessment of spatial relationships between ecological variables 2. Estimation of correlations and coefficients of determination. Environmental and Ecological Statistics 16, 467–494 .