Geostatistical methods for double sampling schemes

application to combined forest inventories

Habilitation Thesis

Author(s):
Mandallaz, Daniel

Publication date:
1993

Permanent link:
https://doi.org/10.3929/ethz-a-000943897

ETH Habilitationsschrift

Daniel Mandallaz

Geostatistical Methods for Double Sampling Schemes: Application to Combined Forest Inventories

Chair of Forest Inventory and Planning
Swiss Federal Institute of Technology (ETH), Zürich
1993

Habilitation thesis

Swiss Federal Institute
of Technology, Zürich

Habilitationsschrift

Eidgenössische
Technische Hochschule
Zürich

Published by:

Chair of Forest Inventory and Planning
Department of Forest and Wood
Sciences
ETH-Zentrum
CH-8092 Zürich

Herausgeber:

Professur für Forsteinrichtung und
Waldwachstum
Departement Wald-und Holzforschung
ETH-Zentrum
CH-8092 Zürich

Aknowledgements

I would like to express my gratitude to Professor P. Bachmann (Chair of Forest Inventory and Planning, ETH Zürich) for his generous support and encouragement throughout this work, as well as for the working environment and spirit he succeeded to create.

I also express my thanks to Professor R. Schlaepfer (Director of the Swiss Federal Institute for Forest, Snow and Landscape Research, Birmensdorf, and ETH Zürich) who systematically promotes forest biometry in Switzerland.

Very special thanks are due to Professor H.R. Künsch (Seminar für Statistik, ETH Zürich) for clarifying and improving many key mathematical arguments, and to Professor M. Maignan (Institute of Mineralogy and Petrography, University of Lausanne) for his expert advice in the intricacies of geostatistics.

Thanks are also due to Professor R. Webster and Dr. O. Smith who carefully scrutinized the manuscript.

Many thanks go to E. Ildefonso and J.P. Béhaxétéguy for their invaluable expertise of the software Bluepack, to A. Lanz and M. Irmay for the data management tasks, as well as to the Ecole Nationale Supérieure des Mines de Paris, in particular to Mrs C. de Fouquet, for many helpful suggestions.

Further thanks go to the forest administration of the Canton of Zürich for its cooperation in the case study, and to the Swiss Federal Office of Forestry who financed the research project no. 3 "Optimization of inventory techniques" within the SANASILVA program and subsidized the printing costs of this work.

Seite Leer /
Blank leaf

Contents

List of Tables

List of Figures

Seite Leer /
Blank leaf

Summary

This work presents different geostatistical estimation methods for double sampling schemes as they are used in combined forest inventories. The objective is to provide efficient estimates for spatial means of given quantities on the basis of a small sample of exact but expensive observations, e.g. the terrestrial plots, and on a large and inexpensive sample of related auxiliary information, usually of qualitative nature, e.g. the aerial photographs. The motivation for using geostatistical procedures is that classical sampling theory techniques are often of little value for small area estimations within global surveys; this topic is of growing importance in forest inventory.

The proposed solutions (kriging with errors, mixed, double and universal kriging) are straightforward, if a prediction model given a priori can be used to predict the exact values with the auxiliary information. If the prediction model is based on the actual data set, the estimation of the residual spatial covariance is a non-trivial task: it can be performed by least square or restricted maximum likelihood procedures; the latter being more efficient under strong spatial correlation.

The best procedure, in terms of simplicity, efficiency and reliability, is double kriging, which adds up the kriging point and variance estimates of predictions and residuals.

A combined forest inventory, completed by a full census, illustrates the techniques and gives a first empirical validation. It was found that the geostatistical techniques are all essentially equivalent with respect to point estimation, whereas mixed kriging and kriging with errors tend to underestimate the expected mean square error, in contrast to double and universal kriging. For local estimation, double and universal kriging performed much better than the classical design-based techniques (with respect to empirical bias and standard error), whereas for global estimation all the point estimates are equivalent and the geostatistical standard errors remain slightly smaller.

Résumé

Ce travail présente plusieurs techniques d'estimation géostatistiques dans le cadre des échantillonnage à deux phases, tels qu'on les utilise pour les inventaires forestiers. Le but est d'obtenir des estimateurs efficaces des moyennes spatiales de diverses grandeurs sur la base d'un petit échantillon d'observations exactes mais coûteuses, e.g. les placettes terrestres, et d'un grand échantillon peu coûteux d'observations auxiliaires afférentes, e.g. les photos aériennes. Le recours à la géostatistique est motivé par le fait que les techniques classiques sont de peu d'utilité pour les estimations locales sur la base de sondages à vocation globale, une question d'actualité croissante pour l'inventaire forestier.

Les solutions proposées (krigeages avec erreur, krigeages double et universel) sont simples si l'on dispose d'un modèle donné a priori pour la prédiction des données exactes sur la base de l'information auxiliaire. En revanche, si le modèle de prédiction est estimé avec les données mêmes du sondage, l'estimation de la covariance spatiale résiduelle est une tâche difficile: on peut l'effectuer par des techniques de moindres carrés ou de maximum de vraisemblance restreint, ce qui est plus efficace dans le cas de forte corrélation spatiale.

En terme de simplicité, efficacité et sûreté, le krigeage double s'avère comme la meilleure technique; elle consiste à additionner les estimateurs de krigeage des prédictions et des résidus, de même pour les variances.

Un inventaire forestier complété par un prélèvement exhaustif illustre les méthodes et donne une première validation. Il s'avère que les méthodes géostatistiques proposées sont supérieures, en terme de biais et d'erreur standard, aux méthodes classiques, et ce tout particulièrement pour l'estimation locale.

Zusammenfassung

Die Arbeit stellt verschiedene geostatistische Schätzverfahren für zweiphasige Stichprobenerhebungen vor, wie sie im Rahmen kombinierter Forstinventuren verwendet werden. Das Ziel ist die Herleitung effizienter Schätzungen der räumlichen Mittelwerte bestimmter Grössen. Verknüpft wird eine kleine Stichprobe exakter, teurer Beobachtungen (z.B. terrestrische Proben) mit einer grossen, billigen Stichprobe aus Hilfsbeobachtungen, meistens qualitativer Natur (z.B. Luftbildproben). Der Grund für die Verwendung der Geostatistik liegt darin, dass im Rahmen von Erhebungen auf globaler Ebene die klassischen Verfahren für lokale Schätzprobleme nur bedingt brauchbar sind. Diese Problematik gewinnt immer mehr an Bedeutung in der Forstinventur.

Die vorgeschlagenen Schätzverfahren (Kriging mit Fehlern, Doppel- und Universal-Kriging) sind einfach einzusetzen, sofern man a priori über ein Modell verfügt, welches mittels der Hifsbeobachtungen Prognosen für die exakten Beobachtungen liefert. Hingegen entstehen mathematische Schwierigkeiten für die Schätzung der residuellen räumlichen Kovarianz, falls das Modell mit den erhobenen Daten angepasst werden muss; die residuelle Kovarianz kann dann mit Verfahren der kleinsten Quadrate oder des eingeschränkten Maximum Likelihood geschätzt werden, wobei letzteres bei stärkerer räumlicher Korrelation effizienter ist.

In bezug auf Einfachheit, Effizienz und Zuverlässigkeit schneidet das Doppel-Kriging Verfahren am besten ab; es besteht aus getrenntem Kriging der Modellprognosen und der Residuen mit anschliessender Addition beider Schätzwerte und Fehlervarianzen.

Eine kombinierte Forstinventur, ergänzt durch eine Vollerhebung, illustriert die Verfahren und erlaubt eine erste Validierung. Es stellt sich heraus, dass die geostatistischen Verfahren den klassischen überlegen sind, sowohl hinsichtlich des Bias als auch der Fehlervarianz, vor allem für lokale Schätzprobleme.

Seite Leer /
Blank leaf

1 Introduction

This work has been accepted as a habilitation thesis (*Habilitationsschrift*) by the Swiss Federal Institute of Technology, Zürich, on the recommendation of professors P. Bachmann, R. Schlaepfer, H.R. Künsch and M. Maignan. It is part of the project no 3 "Optimization of Inventory Techniques" of the Swiss SANASILVA research program, Phase II (1988-1993), directed by the author.

The primary objective was to adapt and test geostatistical techniques in the context of combined forest inventories, i.e. inventories using information coming from different sources, at the enterprise level before using them at regional or national levels. The motivation was threefold:

(1)
Geostatistical techniques have been widely and successfully used in other fiels of natural ressource assessment, primarily mining and petrology, but rarely in forestry.

(2)
The growing importance of remote sensing and geographical information systems, which provide a large amount of cheap auxiliary information, calls for a development of efficient geostatistical techniques combining this information with terrestrial data, to provide alternatives to the classical designed-based regression estimates.

(3)
There is an increasing demand from local authorities to use regional or national inventory data for local purposes, a task generally problematic in the classical framework.

This work gives a self-contained presentation of the required geostatistical concepts and techniques at an intermediate mathematical level. The mathematical appendix investigates the asymptotic properties of the estimates of the spatial correlation. The aspects specific to double sampling entail many new results.

It is written primarily for statistically oriented forest inventorists, but should also be of interest to other scientists, particularly in soil science.

A detailed case study illustrates the theoretical developments on the basis of a data set of unmatched quality to date; it is hoped that the astounding accuracy of the geostatistical procedures will motivate forest inventorists to include them in their tool-kit. Those still unimpressed by mean square error and bias might be convinced by Mark Twain's statement:

"If all you have got is a hammer, then everything looks like a nail".

2 Historical background

By and large, forest inventory methodology rests upon classical sampling theory, that is upon the randomization principle and the design-based approach (D. Mandallaz, 1991). The major objection to this has been known nearly from the origin: inventories are performed in their overwhelming majo-rity with systematic grids, whereas the errors are estimated under assumptions of a different type (i.e. essentially that more than one randomization is available); as a result, point estimation is legitimate but not error estimation. Of course, ad hoc procedures have been proposed, and occasional works have drawn attention to other approaches, such as the model-dependent inference, but without much impact. Fortunately, empirical evidence suggests that the situation is generally not dangerous from a practical point of view, even if efficiency is lost. The issue is simple: either the forest is a fixed entity and the inventory needs more than one randomization to allow for inference, or the inventory sample is fixed and the actual forest must be viewed as the unique realization of a stochastic process with some properties (implying a spatial stationarity of some kind) to allow for statistical inference. Matérn's work on spatial variation (B. Matérn, 1960) was the first significant contribution to this stochastic approach in forest inventory, followed by Giudicelli et al (1972). The general theoretical breakthrough is due to G. Matheron (1962, 1963, 1965, 1970), who introduced the concept of regionalized variables and gave a sound mathematical foundation to a fairly empirical estimation technique (i.e. kriging, first proposed by D.G Krige in his 1951 MSc thesis; see N. Cressie, 1990, for an historical review).

From 1960 onwards, the development of geostatistical techniques (a rather misleading but popular word: statistical estimation for spatial stochastic processes is really what it is about, but obviously unmarketable!) has been soaring, particularly under the impulse of G. Matheron and his school, (primarily in mining, petrology, geology but also oceanography, meteorology, hydrology and soil science). The first and so far most significant contributions to forest inventory are due to

D. Guibal (1973) and P. Marbeau (1976) who applied kriging to terrestrial forest inventory. P. Duplat and G. Peyrotte (1981), F. Houiller (1986) are further important, though more general references, whereas H. Ramirez-Maldonado (1988) is essentially descriptive. J. Bouchon (1979) used geostatistics for surface area estimation and structural analysis of forest stand; A. Jost (1993) compared, under systematic sampling, the classical error estimates with their geostatistical counterparts.

The reasons for the paucity of concrete geostatistical results in forest inventory are various:

(1)- The mathematical difficulty and the esoterism of the matter, with a scattered and not always easily accessible literature.
(2)- The lack of affordable software.
(3)- Doubts about the final cost-benefits ratios.
(4)- Inadequacies in the theory due to particularities of forest inventory.
(5)- The complexity of the sampling schemes, particularly in combined forest inventories.
(6)- Last not least, a touch of "not invented here syndrome".

The latter seems also to affect geostatisticians disdaining sampling theory (de Gruijter, ter Braak, 1990)

.

3 Formulation of the problem

3.1 Terrestrial Forest Inventory

We consider a forest area $V \subset R^2$ and a finite population P of N trees whose centres are at the points $u_i \in V \subset R^2$. The population P of interest does not, in practice, include all the trees, but only part of them (nearly always a subset of trees with diameter at breast height, dbh, larger than a threshold value determined by the inventorist). Scalar numerical variables $Y_i^{(k)}, k = 1,2...p$ are assigned to each tree in P. They can be either a set of binary 0-1 codes of nominal characteristics (i.e. species, state of health, or just any sub-population), or quantitative measurements (diameter, basal area, timber volume etc), assumed to be error free. The trivial variable $Y_i \equiv 1$ simply counts the number of trees. The objectives of forest inventory, in the restricted sense, are the estimation of quantities of the form:

$$\overline{Y}_{V_o}^{(k)} = \frac{1}{\lambda(V_o)} \sum_{u_i \in V_o} Y_i^{(k)}, \quad R_{k,l} = \frac{\overline{Y}_{V_o}^{(k)}}{\overline{Y}_{V_o}^{(l)}} \tag{3.1}$$

for arbitrary domain $V_o \subset V$, $\lambda(V_o)$ denoting the surface area (in ha). From now, on the index (k) will be omitted whenever no confusion occurs.

In this work, we shall use the infinite population approach (D. Mandallaz, 1991), which offers many mathematical advantages over the classical finite framework. Most of the inventory schemes used in practice can be described in terms of a function defined on R^2 by

$$z(x) = \sum_{u_i \in V} Y_i \phi(x - u_i, d_i) \tag{3.2}$$

where $d_i \in D$ is a variable which, in the design-based approach, determines, for a given value of $x - u_i$, the inclusion probability of the i-th tree; D can be fairly general, for instance a cartesian product of sets, but usually d_i is simply the diameter and D a bounded interval (for details see Mandallaz, 1991). The function $\phi(u,d)$ is positive and must satisfy the condition:

$$\int_{R^2} \phi(u,d)du = 1 \quad \forall d \in D \tag{3.3}$$

In most applications $\phi(u,d)$ depends, for a given d, only on $|u|$, but the formalism can cope with far more complex situations, e.g. the structure of ϕ could depend also on the species.
In practice $\phi(u,d) \neq 0$ when u lies in a bounded set depending on d so that the sum defining $z(x)$ extends only over trees in a neighbourhood of x. For instance

$$\phi(u,d) = (\pi r^2)^{-1} \text{ if } |u| \leq r, 0 \text{ otherwise.}$$

yields the simple circular plot technique, whereas

$$\phi(u,d) = \left(\frac{\pi d^2}{4}\right)^{-1} \sin^2\frac{\alpha}{2} \text{ if } |u| \leq \frac{d}{2\sin \alpha/2}, 0 \text{ otherwise.}$$

gives the famous angle count technique with angle α (de Vries, 1986).
One has the fundamental property

$$\int_{R^2} z(x)dx = \sum_{u_i \in V_o} Y_i = \int_A z(x)dx \quad \text{for some} \quad A \supset V_o \tag{3.4}$$

Neglecting or adjusting for boundary effects at the forest edge, one can write (D. Mandallaz, 1991):

$$\int_{V_o} z(x)dx = \sum_{u_i \in V_o} Y_i \tag{3.5}$$

We emphasize the fact that 3.2 is far more general than a standard regularization with kernel methods.
Thus, the summation over a finite population of trees is equivalent to the integration of a regionalized variable (in the sense of G. Matheron, 1970) over a domain of the plane. Note that the function is defined pointwise and that the support of the function ϕ will never occur explicitly in the calculations thereafter. This allows for absolute generality and simplicity, at a negligible price (no sampling fraction correction in the case where $\phi(u,d)$ only depends on $|u|$, D. Mandallaz, 1991).
The formulation given here differs therefore from previous work (P. Marbeau, 1976; P. Duplat and G. Peyrotte, 1981).
In the classical design-based approach, $z(x)$ is the realization of a random variable because one draws the point x according to some random mechanism, whereas in the geostatistical

approach $z(x)$ is considered as the realization of a stochastic process $Z(x)$ at the point x. The case where the function $z(x)$ is estimated by some random sampling mechanism at the point x, as for instance in Poisson or Probability Proportional to Size (PPS) Sampling (D. Mandallaz, 1991), will be briefly considered for one-phase sampling in section 5.2 (Kriging with sampling errors).

3.2 Auxiliary Information

The terrestrial inventory yields for a finite number of points $x, x \in S_2$, the value $z(x)$. The costs $c(x)$ of observing $z(x)$ are generally a complicated function of x and $z(x)$, and are nearly always a substantial part of the overall inventory cost. For this reason, forest inventorists have been using other cheaper sources of information for a long time. This auxiliary information must be correlated with $z(x)$; it can be based on remote sensing, previous inventories, thematic maps (describing geological or stand structure for instance), and is available at points $x \in S_1$; in most instances $S_2 \subset S_1$. We shall assume that the auxiliary information can adequately be described by a p-dimensional vector $A(x) \in R^p$. In practice, most components of $A(x)$ are 0-1 variables, coding purely qualitative information.

This primarily qualitative information is transformed, by means of a prediction model, into a quantitative regionalized variable $\hat{z}(x)$ directly related to $z(x)$, so that the following decomposition holds:

$$z(x) = \hat{z}(x) + \varepsilon(x) \qquad (3.6)$$

where $\hat{z}(x)$ is the prediction and $\varepsilon(x)$ is the residual at x. The prediction can be written, fairly generally, as:

$$\hat{z}(x) = f(A(x), \beta) \qquad (3.7)$$

where f is a function defining the model and β is a vector of parameters. In the linear case, one has:

$$\hat{z}(x) = F(x)'\beta \quad \beta, F(x) \in R^q \qquad (3.8)$$

We shall assume, for the time being that β is given. Of course, in practice, the model $f(.,.)$ and the parameters will often have to be built and estimated from the data

$$\{A(x), z(x); x \in s_2\}$$

This has been the source of great theoretical and practical difficulties: circular procedures, bias, over-optimistic fit of the model (M. Armstrong, 1984). The theory of intrinsic random function of order k, IRF-k, (G. Matheron, 1973), provides a correct solution when the $A(x)$ are polynomial in the coordinates. A major result of this work is the adaptation of the IRF-k philosophy to arbitrary $A(x)$, a conditio sine qua non in forest inventory; sections 9.2 and 9.3 give a mathematically rigorous answer to this problem, whereas section 5.6 outlines the main points from a practical point of view.
However, the assumption that $f(.,.)$ and β are given is perfectly justified when they are external to the inventory, for instance, if they rest upon large and independent inventories, expert judgement or more frequently on yield tables. In such a case, double kriging (section 5.5) is a perfectly legitimate and elegant procedure; kriging with errors and mixed kriging (sections 5.2 and 5.3) are further and simpler alternatives, which are satisfactory for point estimation but which tend to underestimate the error because of mathematical inconsistencies or approximations.

3.3 Combined Forest Inventory

Combined forest inventory can be mathematically described in the following geostatistical framework:

The realization of a stochastic process $\hat{Z}(x)$ is observed at n_1 points $x \in s_1 \subset V$. These observations $\hat{z}(x)$ can be interpreted as the crude predictions of another stochastic process $Z(x)$ resulting from the ground inventory, whose realization $z(x)$ is observed at n_2 points $x \in s_2$; the points in s_2 are a subset of the points in s_1, i.e. $s_2 \subset s_1$.
The problem is to estimate quantities of the form:

$$z(V_o) = \frac{1}{\lambda(V_o)} \int_{V_o} z(x)dx \quad \text{for some} \quad V_o \subset V \tag{3.9}$$

or ratios thereof, on the basis of the data

$$\{z(x), x \in s_2; \hat{z}(x), x \in s_1\}$$

s_2 is always finite, whereas s_1 may in some instances be a domain of the plane.

The integral in (3.9) is the realization of a stochastic integral. As a random variable, this integral must be interpreted as the limit, in the mean square sense, of a Riemanian sum. This limit exists if the covariance function $B(x,y) = E\{Z(x)Z(y)\}$ is continuous on the diagonal $x = y$ (see I. Guikman, A. Skorohod, 1980, chapter 5). It must be emphasized that the auxiliary information $A(x)$, and consequently $\hat{z}(x)$, is viewed as the realization of a stochastic process when $n_1 < \infty$, and as a deterministic function when $n_1 = \infty$ (like in the model-dependent approach, D. Mandallaz, 1991 or in Universal Kriging, see section 5.6 below). Likewise, the relation (3.6) actually defines the residual process, so that the prediction model is not, at this stage, assumed to be "true".

The above formulation differs slightly (more in its interpretation than in the mathematical tools required) from the standard geostatistical context. The reason for this is due to a shift of emphasis towards efficient use of auxiliary information; the question of stationarity of the underlying processes being only a side aspect of the problem, even if technically important.

3.4 Other applications

In other fields of application, the function $z(x)$ is generally given by the problem at hand (e.g. density of metals, oil, water or solar energy etc.) and does not have to be constructed first as in forest inventory; the only requirement being that $\int_V z(x)dx$ must be a meaningful physical quantity. In forestry, the "field" V (i.e. the forest area) is generally well defined and must be explicitly taken into account by means of sets of polygons. In other applications, the field V is often not

known beforehand. The concept of auxiliary information should be straightforward in any particular situation. In most instances, the function $z(x)$ is not directly observable, but only a regularization thereof, i.e. a function $z_p(x)$ defined by

$$z_p(x) = \int z(x+y)p(y)dy \quad \text{with} \quad p(y) \geq 0 \quad \text{and} \quad \int p(x)dx = 1$$

Neglecting boundary effects one has

$$\int_V z_p(x)dx = \int_V z(x)dx$$

(this equality is exact for $V = R^2$)

Therefore, all the concepts and results given below for $z(x)$ can be transposed mutatis mutandis to $z_p(x)$. In some instances, the explicit knowledge of the relation between $z(x)$ and $z_p(x)$ can offer a slight improvement (convolution or deconvolution of variograms), a topic we shall not deal with.

4 Variograms and cross-variograms

Statistical inference for the spatial processes $Z(x), \hat{Z}(x), \varepsilon(x)$ is not possible without further assumptions, as only one realization is available. These assumptions require essentially some form of stationarity, either of the processes themselves or of derived quantities. Following G. Matheron (1970) we shall assume that the processes are **intrinsic,** i.e.:

(4.1)

$$E\{Z(x+h)-Z(x)\}=0 \quad \forall x,h, \text{ and similarly for } \hat{Z}(x),\varepsilon(x)$$

(4.2)

$$E\{Z(x+h)-Z(x)\}^2 = var\{Z(x+h)-Z(x)\} = 2\gamma_z(h) \quad \forall x,h$$

and similarly for $\gamma_{\hat{z}}, \gamma_{\varepsilon}$.

The functions $\gamma(h)$ are called **semi-variograms**, for short **variograms** thereafter. Note that $\gamma(h)=\gamma(-h)$.

Sometimes a more general definition is used, by allowing a $E\{Z(x+h)-Z(x)\}=m(h)$, independent of x. The drift $m(h)$ is linear in h; if $m(h)$ is known, it can be assumed to be zero by subtracting it.

Though the intrinsic hypothesis does not require finite expectation and variance for the process itself but only for its increments, we shall always assume this to be the case (it is per definition in practice); the main advantage of the intrinsic hypothesis is that it requires the stationarity in mean and variance of the increments only: for instance, the brownian motion is intrinsic but not stationary (non-constant variance)

Variograms satisfy the following conditions (Matheron, 1970)

(4.3)

(1) $\gamma(0) = 0$

(2) $\lim_{h\to\infty} \frac{\gamma(h)}{h^2} = 0$

(3) $\sum_{i=1}^{n}\lambda_i\lambda_j\gamma(x_i-x_j)\le 0 \quad \forall n, \forall x_i, x_j$, provided that $\sum_{i=1}^{n}\lambda_i = 0$.

and allow for a straightforward calculation of the variance according to:

(4.4)

$$\text{var}\left(\sum_{i=1}^{n}\lambda_i Z(x_i)\right) = \sum_{i,j=1}^{n}\lambda_i\lambda_j\gamma_Z(x_i-x_j)$$

(4.4) is valid for all **authorized linear combinations**, i.e. satisfying the constraint $\sum\lambda_i = 0$.

This can be seen by noting first that for any x_o

$$\mathrm{var}\left(\sum_{i=1}^{n}\lambda_i(Z(x_i)-Z(x_o))\right)=\mathrm{var}\left(\sum_{i=1}^{n}\lambda_i Z(x_i)\right)$$

and second that

$$\mathrm{cov}\{Z(x_1)-Z(x_0),Z(x_2)-Z(x_0)\}=\gamma(x_1)+\gamma(x_2)-\gamma(x_1-x_2).$$

If $\gamma(h)$ is continuous at the origin, the process is continuous in the mean square sense, otherwise the discontinuity at the origin, i.e. $c_o=\gamma(0^+)-\gamma(0)>0$ is called the **nugget effect** and implies an irregular behaviour.

The variogram is called **isotropic** if $\gamma(h)=\gamma(|h|)$, otherwise **anisotropic.**

If the underlying process itself is **second order stationary**, i.e.

(4.5)

$$\mathrm{E}\{Z(x)\}=m$$
$$\mathrm{cov}\{Z(x+h),Z(x)\}=C(h)$$

then

(4.6)

$$\gamma(h)=C(0)-C(h)$$

If $\lim_{h\to\infty}C(h)=0$ then $C(0)=\gamma(\infty)$ is called the **sill**. The smallest vector r_o for which $\gamma(r_o(1+\varepsilon))=C(0)$ for any $\varepsilon>0$ is called the **range** in the direction r_o .

If $Z_1(x)$, $Z_2(x)$ are two intrinsic processes, the cross-variogram $\gamma_{Z_1,Z_2}(h)$ is defined by

(4.7)

$$2\gamma_{Z_1,Z_2}(h)=\mathrm{cov}\{Z_1(x+h)-Z_1(x),Z_2(x+h)-Z_2(x)\}$$

(A.G. Journel, C. Huijbregts, 1978)

Sometimes an alternative definition is used, namely:

$$2\Gamma_{Z_1,Z_2}(h)=\mathrm{var}\{Z_1(x+h)-Z_2(x)\}$$

(I. Clark et al, 1989; N. Cressie 1991), so that caution is required.

In general $\gamma_{Z_1,Z_2}(h)\neq\gamma_{Z_1,Z_2}(-h)$. However, because of the decomposition $Z(x)=\hat{Z}(x)+\varepsilon(x)$, we have:

$$\gamma_Z(h) = \gamma_{\hat{Z}}(h) + \gamma_\varepsilon(h) + 2\gamma_{\hat{Z},\varepsilon}(h) \tag{4.8}$$

and therefore:

$$\gamma_{\hat{Z},\varepsilon}(h) = \gamma_{\hat{Z},\varepsilon}(-h) \tag{4.9}$$

If $\hat{Z}(x_1), \varepsilon(x_2)$ are uncorrelated $\forall x_1, x_2$, then:

$$\gamma_{Z,\varepsilon}(h) = \gamma_\varepsilon(h) = \gamma_{Z,\varepsilon}(-h) \tag{4.10}$$

Let us briefly discuss the important but sometimes confusing concept of nugget effect $c_o = \gamma(0^+) - \gamma(0) > 0$. Mathematically, this cannot happen for L_2-continous processes. Note that by construction, each realization $z(x)$ of $Z(x)$ is piecewise constant (over cells of a tremendously complex tesselation of the plane determined by the relative position of all trees), the set of discontinuities having zero Lebesgue measure. On the other hand, one should recall that averaging a bi-dimensional white noise over a circle, no matter how small, yields a continuous variogram (I. Gelfand, N. Vilenkine, 1964), likewise for a Poisson marked point process (P. Marbeau, 1976; D. Stoyan et al, 1987; E. Tomppo, 1986). Therefore, the only way for a nugget effect to occur is by the presence of measurement or sampling errors; for instance, if $z(x)$ is itself estimated at x by some sampling technique (like PPS sampling, D. Mandallaz, 1991), a situation we shall deal with in section 5.2. Thus, in principle, the random process $Z(x)$ has a continous variogram. However, in applications, ad hoc variograms with a nugget effect will be occasionally used, simply to reflect our numerical ignorance with respect to the behaviour of the process at the microscale, because observations are available only for points lying at a given minimal distance from each other. This is determined by the inventory scheme. The value of this pseudo nugget effect is often inversely proportional to the surface area of the support (in R^2) of the sampling function ϕ (for a Poisson-type forest this follows directly from P. Marbeau, 1976 and G. Matheron, 1970). For a better intuitive understanding, it is useful to note that in one dimension, the variogram of the Brownian motion is a linear function, i.e. $\gamma(h) = \alpha|h|$.

For future use, we note the important formula (extension variance of V_2 to V_1)

(4.11)

Let $z(V_i)=\frac{1}{\lambda(V_i)}\int_{V_i} z(x)dx$, then

$$\mathrm{var}\{z(V_1)-z(V_2)\}=$$

$$-\frac{1}{\lambda^2(V_1)}\int_{V_1}\int_{V_1}\gamma_Z(x-y)dxdy-\frac{1}{\lambda^2(V_2)}\int_{V_2}\int_{V_2}\gamma_Z(x-y)dxdy+2\frac{1}{\lambda(V_1)\lambda(V_2)}\int_{V_1}\int_{V_2}\gamma_Z(x-y)dxdy=$$

$-\bar{\gamma}_Z(V_1,V_1)-\bar{\gamma}_Z(V_2,V_2)+2\bar{\gamma}_Z(V_1,V_2)$, in obvious notation.

Indeed, the difference is an authorized linear combination since both integrals can be approximated by discrete sums (with constant weights summing up to 1), for which 4.4 can be applied and then the limit taken. Similarly, one has:

(4.12)

If $\sum\lambda_i=1$ then

$$\mathrm{var}\left\{\sum_i\lambda_i Z(x_i)-\frac{1}{\lambda(V)}\int_V Z(x)dx\right\}=$$

$$-\sum_{i,j}\lambda_i\lambda_j\gamma_Z(x_i-x_j)-\frac{1}{\lambda^2(V)}\int_V\int_V\gamma_Z(x-y)dxdy+2\sum_i\lambda_i\frac{1}{\lambda(V)}\int_V\gamma_Z(x_i-y)dy$$

$$:=-\sum_{i,j}\lambda_i\lambda_j\gamma_Z(x_i-x_j)-\bar{\gamma}_Z(V;V)+2\sum_i\lambda_i\bar{\gamma}_Z(x_i,V)$$

Various parametric variogram models are available (see A. G. Journel, C. Huijbregts, 1978, for a review). The following are frequently used, in particular in the case study.

i) Linear model, valid in R^d , $d\geq 1$

(4.13)

$$\gamma(h)=\sigma^2|h|$$

unbounded variogram, in R induced by the Brownian motion.

ii) Spherical model, valid in $R^d, d\leq 3$

(4.14)

$$\gamma(h)=\sigma^2\left\{\frac{3|h|}{2r}-0.5\left(\frac{|h|}{r}\right)^3\right\} \quad if \quad |h|\leq r \quad ,=\sigma^2 \text{ otherwise.}$$

The sill is σ^2 and the range r. This variogram results from averaging a white noise over spheres of radius $r/2$.

iii) Exponential variogram, valid in $R^d, d \leq 3$

(4.15)

$$\gamma(h) = \sigma^2 \left\{ 1 - \exp\left(-\frac{|h|}{a} \right) \right\}$$

The sill is σ^2, the range is infinite, though in practice it is defined to be $3a$.

iv) Circular variogram, valid in $R^d, d \leq 2$

(4.16)

$$\gamma(h) = \sigma^2 \left\{ 1 - \frac{2}{\pi} \left(ar\cos\left(\frac{|h|}{r}\right) - \left(\frac{|h|}{r}\right) \sqrt{1 - \left(\frac{|h|}{r}\right)^2} \right) \right\}, \text{ for } |h| \leq r$$

$$\gamma(h) = \sigma^2, \text{ for } |h| > r$$

This variogram results from averaging a white noise over a circle of radius $r/2$.

v) Pure nugget effect

(4.17)

$$\gamma(0) = 0, \quad \gamma(h) = \sigma^2 \text{ for } |h| > 0$$

Given p variograms $\gamma_i(h)$, one can define a further variogram by setting:

(4.18)

$$\gamma(h) = \sum_{i=1}^{p} \gamma_i(h)$$

Finally, an anisotropic variogram is geometrically anisotropic if

(4.19)

$$\gamma(h) = \gamma_o(|Ah|)$$

where $\gamma_o(h)$ is a valid one-dimensional variogram and A is a regular (d,d) matrix.

In practice, the underlying variograms $\gamma(h)$ have to be estimated from the data, a problem we now briefly discuss. Cross-variograms can be dealt with in a similar way. Under the intrinsic assumption, a natural and non-parametric estimator based on the methods of moments, due to Matheron (1962), is to set:

(4.20)

$$\hat{\gamma}(h) = \frac{\sum_{i,j \in N(h)} \left(z(x_i) - z(x_j) \right)^2}{|N(h)|}$$

where $N(h)=\{(x_i,x_j):|x_i-x_j|=h\}$ and $|N(h)|$ is the number of pairs in $N(h)$.

In most instances, because of the positions of the sample points, only smoothed versions of (4.20) are available, namely:

$$\hat{\gamma}(h_l)=average\{(z(x_i)-z(x_j))^2:(x_i,x_j)\in N(h),h\in T(h_l)\} \quad \textbf{(4.21)}$$

where $T(h_l)$ is some specified tolerance region around h_l.

Tolerance regions should be as small as possible to retain spatial resolution, yet large enough to ensure stability of the estimates (say at least 60 pairs).

It is better to directly estimate the variogram rather than the corresponding autocorrelation function and the relation (4.6), (N. Cressie, 1991). There exist more sophisticated estimation procedures, e.g. based on moving-windows or relying on robust techniques. If the process is gaussian, results for the exact distribution of $\hat{\gamma}(h)$ are available (see N. Cressie, 1991, for a review and further references).

After the empirical variogram $\hat{\gamma}(h)$ has been obtained via (4.21) or otherwise, one has still to fit a model to it (i.e. to choose a valid variogram and to estimate its parameters like sill and range). Again, several methods are possible, among others non-linear weighted least squares (for a review see N. Cressie, 1991). In this work, we primarily used the techniques implemented in the softare BLUEPACK (essentially interactive "fitting by eye" procedures) and the least square technique described in section 5.6.

If the intrinsic assumption does not hold, then $\hat{\gamma}(h)$ does not, in general, estimate $\gamma(h)$. For instance, $\hat{\gamma}(h)$ will show a quadratic behaviour if there is a linear drift in $Z(x)$, a useful fact for graphical inspection. The estimation of the variogram of the residual process $\varepsilon(x)$ can be a source of concern when the drift-model is fitted simultaneously: vario-grams based directly on the empirical residuals can be seriou-sly biased, especially for large lags h (G. Matheron, 1970; J.P. Chilès, 1977; N. Cressie, 1991). This problem actually led G. Matheron to develop, for polynomial drifts, the theory of intrinsic

random functions of order k , **IRF-k** for short (G. Matheron, 1973).

We shall present in sections 5.6, 9.1, 9.2 modified least squares and maximum likelihood techniques to solve this problem for arbitrary drifts, as they occur in forest inventory. Like the IRF-k, these methods filter out the drifts before estimating the residual covariance. They give consistent estimates under fairly general conditions, do not require prior smoothing as (4.21), but are more difficult to implement numerically. Obviously, they can also be used in the stationary case.

5 Estimation Techniques

Assuming that the underlying variograms and cross-variograms are known or have been adequately modelled, it remains to estimate the quantity

$$z(V)=\frac{1}{\lambda(V)}\int_V z(x)dx$$

We shall now discuss this problem under the assumption that the uncertainties with respect to the variograms can be neglected. The impact of variogram estimation on point and variance estimation is extremely difficult to assess (P. Diamond, M. Armstrong,1984; D.L Zimmermann, N. Cressie, 1992).

5.1 Ordinary Kriging

Ordinary Kriging yields the **B**est **L**inear **U**nbiased **E**stimate (**BLUE**) of

(5.1)

$$z(V_o)=\frac{1}{\lambda(V_o)}\int_{V_o} z(x)dx \quad ,V_o \subset V$$

on the basis of the terrestrial inventory data only, i.e. $z(x_i)$ $,x_i \in s_2$. For convenience we shall write indifferently $i \in s_2$ or $x_i \in s_2$ thereafter.

One looks therefore for an estimate of the form

(5.2)

$$z^*(V_o)=\sum_{i\in s_2}\lambda_i z(x_i)$$

satisfying the following conditions:

$\mathrm{E}\left(Z^*(V_o)-Z(V_o)\right)=0$, i.e. unbiasedness.

$\mathrm{E}\left(Z^*(V_o)-Z(V_o)\right)^2$ =minimum.

The first condition also requires that the error $z^*(V_o)-z(V_o)$ is an authorized linear combination, which implies $\sum\lambda_i=1$.

Using the expression (4.12) for the variance and the Lagrange's technique of optimization under constraint, one must minimize the function

$$L(\lambda_i;\mu;i\in s_2)=-\sum_{i,j}\lambda_i\lambda_j\gamma_Z(x_i-x_j)-\bar{\gamma}_Z(V_o,V_o)+2\sum_i\lambda_i\bar{\gamma}_Z(x_i,V_o)-2\mu\left(\sum_i\lambda_i-1\right)$$

This leads immediately to the well known kriging equations:

(5.3)

$$\sum_i \lambda_i \gamma_Z(x_i - x_j) + \mu = \bar{\gamma}_Z(x_j, V_o)$$
$$\sum_i \lambda_i = 1$$

i.e. a linear system of $n_2 + 1$ equations with $n_2 + 1$ unknowns we shall always assume to be regular (this is not the case if for example the same point occurs twice). Using (4.12) and (5.3), the expected mean square error (**MSE**) is found to be

(5.4)

$$E\left\{Z^*(V_o) - Z(V_o)\right\}^2 = \sum_i \lambda_i \bar{\gamma}_Z(x_i, V_o) - \bar{\gamma}_Z(V_o, V_o) + \mu$$

If $V_o = \{x_o\}$ (i.e. one considers punctual estimation) and if $x_o \in s_2$, then it is readily verified that $z^*(x_o) = z(x_o)$. In this sense, kriging is an exact interpolator. Punctual kriging is hardly ever relevant in forest inventory. Variograms with a nugget effect lead to a mathematical inconsistency (frequently overlooked) because punctual kriging cannot be considered, in this case, as the limit of domain kriging when the domain shrinks to a point. Indeed, with a pure nugget effect, it is readily seen from (5.3-5.4) that the **MSE** is $\sigma^2 + \sigma^2/n_2$ for a point, and σ^2/n_2 for any true domain ($\lambda_i \equiv n_2^{-1}$ in both cases). The correct interpretation of this paradox is, again, that a nugget effect is a coarse numerical approximation of the true underlying variogram short range behaviour and is valid as long as this range is negligible with respect to the size of the domain. In applications, s_2 is often reduced to a so called kriging neighbourhood $U \subset V$, $V_o \subset U$, by simply restricting s_2 to $s_2 \cap U$; (5.3) and (5.4) remaining valid with $s_2 \cap U$ in place of s_2.

To illustrate ordinary kriging and compare it with classical sampling, let us consider a process whose variogram is the sum of a spherical variogram of range r and sill σ^2 and of a pure nugget effect c_o. Suppose further that all pairs of data points satisfy $|x_i - x_j| > r$ and that all points x_i are at least at a distance r from the boundary of the forest area V. Then, using polar co-ordinates, it is easily verified that

$$\bar{\gamma}_Z(x_i, V_o) = c_o + \sigma^2 \left\{1 - \frac{\pi r^2}{5\lambda(V)}\right\}$$

Furthermore, as a first approximation,

$$\bar{\gamma}_Z(V,V) \cong \bar{\gamma}_Z(x_t,V) \quad \forall x_t \quad \left(\text{up to } 0\left(\frac{\pi r^2}{5\lambda(V)}\right)\right)$$

It turns out that the λ_t must be constant and therefore equal to n^{-1}; the Lagrange's multiplier is given by

$$\mu = \frac{c_o}{n} + \frac{\sigma^2}{n}\left\{1 - \left(\frac{n\pi r^2}{5\lambda(V)}\right)\right\} = \mathrm{E}\left\{Z^*(V) - Z(V)\right\}^2$$

In other words, the ordinary kriging estimate is the sample mean, and its variance decreases as n^{-1} up to a correction term for finite sampling. To be more explicit, suppose that $c_o = 0$ and that sampling is performed with circular plots of radius $r/2$ and surface area a (i.e. disjoint plots have independent $Z(x)$), then

$$z^*(V) = \frac{1}{n}\sum_i z(x_i) \quad , \quad \mathrm{var}\, Z^*(V) = \frac{\sigma^2}{n}\left\{1 - \left(\frac{4an}{5\lambda(V)}\right)\right\}$$

as compared with $\frac{\sigma_s^2}{n}\left\{1 - \left(\frac{an}{\lambda(V)}\right)\right\}$

given by the sampling theory for a finite population of disks; note that σ^2, σ_s^2 do not have exactly the same meaning.

It is commonly said that geostatistics is useless as soon as the distance between the points exceeds the range of the spatial correlation. This is true for the point estimate, but not for the variance. More generally, it is wrong that the size of the kriging neighbourhood is essentially determined by the variogram range, since the λ_t depend primarily on the inverse of the kriging matrix $\gamma_Z(x_i - x_j)$ (e.g. with a pure nugget effect of zero range the kriging neighborhood is in principle infinite).

In some rare instances, the true mean of the stationary process is known and can be assumed to be zero by subtraction. In such a case, the constraint $\sum_t \lambda_t = 1$ can be ignored and 5.3-5.4 remain valid with $\mu = 0$, leading to the so called **simple kriging** procedure (G. Matheron, 1970).

Finally, it is often instructive to have the linear variogram in mind. In one dimension and without nugget effect, this corresponds to the Brownian motion. In this case, kriging at a point x_o amounts to perform a linear interpolation between the left and right neighbours of x_o, (the other points playing no role): this property is called the screen effect and is no longer valid in the presence of a nugget effect. This is often

an aid for the interpretation of the kriging weights λ_i. However, it is in general difficult to have an intuitive insight into the behaviour of the kriging weights in arbitrary situations (J. Rivoirard, 1984).

5.2 Kriging with sampling or measurement errors

In some instances one does not, or cannot, directly observe the process $Z(x)$, but only a randomly disturbed version thereof; i.e. one observes $S(x_i)$ with

$$S(x_i) = Z(x_i) + \delta(x_i) \quad \text{,for} \quad x_i \in s \tag{5.5}$$

$\delta(x_i)$ denoting the random disturbance.

The decomposition (5.5) is useful when $z(x_i)$ is estimated at x_i by some second stage sampling procedure at the tree level, $s(x_i)$ being usually a generalized Horwitz-Thompson estimate of $z(x_i)$ (see D. Mandallaz, 1991, p.32, for the theory, and E. Kaufmann, 1992, for an application thereof to the Swiss National Forest Inventory). In this context the following assumptions are meaningful:

(5.6)

i) $Z(x)$ is second order stationary

ii) $E_\delta\{\delta(x_i)|Z(x_j), x_j \in s\} = 0$

iiia) $Var_\delta\left(\delta(x_i)\middle|Z(x_j), x_j \in s\right) = \sigma^2(Z(x_i), x_i), \quad E_Z(\sigma^2(Z(x_i), x_i) = \sigma^2(x_i)$

iiib) $\sigma^2(Z(x_i), x_i) = \sigma^2(Z(x_i)), \quad E_Z\left(\sigma^2(Z(x_i))\right) \equiv \sigma^2$

iv) $\mathrm{cov}_\delta\left(\delta(x_i), \delta(x_j)\middle|Z(x_k), x_k \in s\right) = 0 \quad \forall x_i \neq x_j$

Condition (iiia) allows for the possibility to observe exactly at the point x_i, i.e. $\sigma^2(x_i) = 0$. Usually the 2nd stage sampling is performed at all points, and the sampling variance at x_i is only a function of the true value $z(x_i)$; in such a case, because of (i), it is natural to then require (iiib). Conditions (ii) and (iv) can always be inforced by the design of the 2nd stage procedure.
The kriging estimate is defined as:

$$z^*(V_o) = \sum_{j \in s} \lambda_j s(x_j) \tag{5.7}$$

and the error is

$$z^*(V_o) - z(V_o) = \sum_{j \in s} \lambda_j z(x_j) + \sum_{j \in s} \lambda_j \delta(x_j) - \frac{1}{\lambda(V_o)} \int_{V_o} z(x)dx$$

If $\sum_{j\in s}\lambda_j = 1$, then $E_{Z,\delta}\left(Z^*(V_o) - Z(V_o)\right) = 0$; the anticipated mean square error (i.e. taking the expectation with respect to the process and the 2nd stage sampling procedure) is, under (5.6), easily found to be:

$$\mathrm{E}\left\{Z^*(V_o) - Z(V_o)\right\}^2 = -\sum_{i,j\in s}\lambda_i\lambda_j\gamma_Z(x_i - x_j) - \bar{\gamma}_Z(V_o,V_o) + 2\sum_{i\in s}\lambda_i\bar{\gamma}_Z(x_i,V_o) + \sum_{i\in s}\lambda_i^2\sigma^2(x_i)$$

Minimizing this expression under the constraint yields the kriging equations with sampling errors:

(5.8)

$$\sum_{j\in s}\lambda_j\gamma_Z(x_i - x_j) - \lambda_i\sigma^2(x_i) + \mu = \bar{\gamma}_Z(x_i,V_o) \quad \text{for} \quad x_i \in s$$

$$\sum_{j\in s_1}\lambda_j = 1$$

$$\mathrm{E}_{Z,\delta}\left(Z^*(V_o) - Z(V_o)\right)^2 = \sum_{i\in s}\lambda_i\bar{\gamma}_Z(x_i,V_o) - \bar{\gamma}_Z(V_o,V_o) + \mu$$

To apply (5.8), one generally will have to estimate somehow the underlying variogram γ_Z and the variances $\sigma^2(x_i)$ (note that 5.8 is also valid for intrinsic $Z(x)$). Under assumptions (i, iiib) this is not a major difficulty; indeed one first notes that

$$\gamma_S(h) := \frac{1}{2}E_{Z,\delta}\left(S(x+h) - S(x)\right)^2 = \gamma_Z(h) + \sigma^2 \quad \text{for} \quad h \neq 0.$$

Furthermore, let $\hat{\sigma}^2(x_i)$ be a design-unbiased variance estimate of $\sigma^2(z(x_i))$ (usually an Horwitz-Thompson estimate of a quadratic form) and set $\hat{\sigma}^2 = \frac{1}{n}\sum_{i\in s}\hat{\sigma}^2(x_i)$; under (iiib) $E_Z E_{\delta|Z}(\hat{\sigma}^2) = \sigma^2$.

Hence, it suffices to shift the empirical variogram γ_S of the observed process downwards by $\hat{\sigma}^2$ to obtain an estimate of the variogram γ_Z of the unobserved process, while retaining the definition $\gamma_Z(0) = 0$. For punctual kriging at $x_o \notin s$, it can be checked that setting $\gamma_Z(h) = \gamma_S(h) - \sigma^2$ for $h \neq 0$, $\gamma_Z(0) = 0$ into (5.8) is equivalent, for point estimation, to ordinary kriging with $\gamma_S(h), \gamma_S(0) = 0$, whereas the resulting mean square error must be reduced by σ^2 to give the correct answer. If $x_o \in s$, this is no longer true and it is easily verified that (5.8) does not give an exact interpolator anymore, but instead smoothes even the observation point (see N. Cressie, 1991, p.128, for similar findings in a slightly different context). For domain kriging, it is necessary to use (5.8) directly. In short, it is wrong, for kriging purposes, to treat sampling error as an ordinary nugget effect.

The kriging equations (5.8) can be applied in the context of measurement errors (A. Galli et al, 1987), where the following unconditional assumptions are made:

(5.9)

i) $Z(x)$ is intrinsic

ii) $E(\delta(x_i)) = 0, Var(\delta(x_i)) = \sigma^2(x_i) \quad x_i \in s$

iii) $\forall x_i \neq x_j \quad \mathrm{cov}(\delta(x_i), \delta(x_j)) = 0 \quad , \quad \forall x_i, x_j \quad \mathrm{cov}(\delta(x_i), Z(x_j)) = 0$

Condition (iii) requires that the true values and the measurement errors are uncorrelated. For this reason, the kriging equations (5.8) will also be referred to as kriging with (uncorrelated) measurement errors.

To incorporate auxiliary information into kriging it is tempting, by analogy, to set formally:

(5.10)

$$S(x_i) = Z(x_i), \delta(x_i) = \sigma^2(x_i) = 0 \quad \text{for} \quad x_i \in s_2$$
$$S(x_i) = \hat{Z}(x_i), \delta(x_i) = \hat{Z}(x_i) - Z(x_i) \quad \text{for} \quad x_i \in s_1 - s_2$$
$$z^*(V_o) = \sum_{i \in s_2} \lambda_i z(x_i) + \sum_{i \in s_1 - s_2} \lambda_i \hat{z}(x_i)$$

with the kriging equations

(5.11)

$$\sum_{j \in s_1} \lambda_j \gamma_Z(x_i - x_j) + \mu = \bar{\gamma}_Z(x_i, V_o) \quad \text{for} \quad i \in s_2$$
$$\sum_{j \in s_1} \lambda_j \gamma_Z(x_i - x_j) - \lambda_i \sigma^2(x_i) + \mu = \bar{\gamma}_Z(x_i, V_o) \quad \text{for} \quad i \in s_1 - s_2$$
$$\sum_{j \in s_1} \lambda_j = 1$$
$$\mathrm{E}\left\{Z^*(V_o) - Z(V_o)\right\}^2 = \sum_{i \in s_1} \lambda_i \bar{\gamma}_Z(x_i, V_o) - \bar{\gamma}_Z(V_o, V_o) + \mu$$

i.e. to view the predictions as observations with errors. The variogram γ_Z can be estimated on the basis of the small sample only, and $\sigma^2(x_i)$ by some ad-hoc procedure, e.g. residual sum of squares if $\sigma^2(x_i) \equiv \sigma^2$, which we shall assume in the following.

Unfortunately, this procedure is mathematically incorrect: indeed, under the decomposition (3.6) $\delta(x_i) = -\varepsilon(x_i)$, which implies $\mathrm{cov}(Z(x_i), \delta(x_i)) = -\mathrm{var}\,\varepsilon(x_i) \equiv -\sigma^2 \neq 0$ and (5.9) is violated.

However (5.10) still yields an unbiased estimate, and its empirical performance, as shown in the case study of chapter 7 is

satisfactory, with a tendency to underestimate the mean square error.

To proceed further we assume that the correlation range of the residual process $\varepsilon(x)=-\delta(x)$ is small in comparison with the domain and the distance between the sample points, which implies $\mathrm{cov}(\delta(x_i),\delta(x_j))\cong 0$ for $x_i \neq x_j$, $\mathrm{cov}(Z(V_o),\delta(x_j))\cong 0$. The error can be rewritten as $z^*(V_o)-z(V_o)=\sum_{j\in s_1}\lambda_j z(x_j)-\frac{1}{\lambda(V_o)}\int_{V_o} z(x)dx+\sum_{j\in s_1-s_2}\lambda_j\delta(x_j)$; using the equation (4.12) and taking the extra covariance term into account the expected mean square error is found to be:

$$\mathrm{E}\left\{Z^*(V_o)-Z(V_o)\right\}^2 =$$
$$-\sum_{i,j\in s_1}\lambda_i\lambda_j\gamma_Z(x_i-x_j)-\bar{\gamma}_Z(V_o,V_o)+2\sum_{i\in s_1}\lambda_i\bar{\gamma}_Z(x_i,V_o)-\sum_{i\in s_1-s_2}\lambda_i^2\sigma^2(x_i)$$

Minimization under the unbiasedness constraint yields the following kriging equations:

(5.12)

$$\sum_{j\in s_1}\lambda_j\gamma_Z(x_i-x_j)+\mu=\bar{\gamma}_Z(x_i,V_o) \quad \text{for} \quad x_i\in s_2$$
$$\sum_{j\in s_1}\lambda_j\gamma_Z(x_i-x_j)+\lambda_i\sigma^2+\mu=\bar{\gamma}_Z(x_i,V_o) \quad \text{for} \quad x_i\in s_1-s_2$$
$$\sum_{j\in s_1}\lambda_j=1$$
$$\mathrm{E}\left\{Z^*(V_o)-Z(V_o)\right\}^2=\sum_{i\in s_1}\lambda_i\bar{\gamma}_Z(x_i,V_o)-\bar{\gamma}_Z(V_o,V_o)+\mu$$

The only difference between the kriging systems (5.11) and (5.12) is the sign of the term σ^2 for $x_i\in s_1-s_2$. However, to ensure a positive **MSE** in (5.12) one must require $\gamma_Z(0^+)>\sigma^2$, which will be the case if for instance $\gamma_Z(h)=\gamma_{\hat{Z}}(h)+\sigma^2$ for $h\neq 0$, i.e. with a pure nugget effect for the residual process. In such a case, the kriging equations (5.12) can be rewritten in terms of the variogram of the prediction process, to give:

(5.13)

$$\sum_{j\in s_1}\lambda_j\gamma_{\hat{Z}}(x_i-x_j)-\lambda_i\sigma^2+\mu=\bar{\gamma}_{\hat{Z}}(x_i,V_o) \quad \text{for} \quad x_i\in s_2.$$
$$\sum_{j\in s_1}\lambda_j\gamma_{\hat{Z}}(x_i-x_j)+\mu=\bar{\gamma}_{\hat{Z}}(x_i,V_o) \quad \text{for} \quad x_i\in s_1-s_2$$
$$\sum_{j\in s_1}\lambda_j=1$$
$$\mathrm{E}\left\{Z^*(V_o)-Z(V_o)\right\}^2=\sum_{i\in s_1}\lambda_i\bar{\gamma}_{\hat{Z}}(x_i,V_o)-\bar{\gamma}_{\hat{Z}}(V_o,V_o)+\mu$$

Formally, the kriging system (5.13) can be viewed as the standard kriging system with measurement errors (5.11), while interchanging the roles of $\boldsymbol{z(x)}$ and $\boldsymbol{\hat{z}(x)}$. This intuitively nice result leads to the so called mixed kriging procedure, which is derived in a slightly different way in the next section.

5.3 Mixed Kriging

We assume that $\hat{Z}(x)$ is intrinsic and that $\varepsilon(x)$ is stationary (with zero expectation and variance σ^2) and uncorrelated with $\hat{Z}(x)$.

The starting point is the same as in section 5.2, i.e. we consider an estimate of the form

(5.14)

$$z^*(V_o) = \sum_{j \in s_2} \lambda_j z(x_j) + \sum_{j \in s_1 - s_2} \lambda_j \hat{z}(x_j)$$

Define

(5.15)

$$\hat{z}(v_o) = \frac{1}{\lambda(V_o)} \int_{V_o} \hat{z}(x) dx$$

then

(5.16)

$$z^*(V_o) - z(V_o) = z^*(V_o) - \hat{z}(V_o) + \hat{z}(V_o) - z(V_o)$$
$$= \sum_{j \in s_1} \lambda_j \hat{z}(x_j) - \hat{z}(V_o) + \sum_{j \in s_2} \lambda_j e(x_j) - \frac{1}{\lambda(V_o)} \int_{V_o} e(x) dx$$

The first term in (5.16) is an authorized linear combination ($e(x)$ being the realization of $\varepsilon(x)$). **We now assume that the range r of $\gamma_\varepsilon(h)$ is negligible in comparison with $\lambda(V_o)$**, so that the integral over the residual process yields negligible variance and covariance terms (which excludes punctual kriging). The unbiasedness condition implies as usual the constraint $\sum_{j \in s_1} \lambda_j = 1$ and the mean square error is found to be :

(5.17)

$$\mathrm{E}\left\{Z^*(V_o) - Z(V_o)\right\}^2 = \mathrm{E}\left\{\sum_{j \in s_1} \lambda_j \hat{Z}(x_j) - \hat{Z}(V_o)\right\}^2 + \sum_{j \in s_2} \lambda_j^2 \sigma^2 + O\left(\frac{\pi r^2}{\lambda(V_o)}\right)$$

Using (4.12) for the first term, the Lagrange's technique yields the mixed kriging equations

(5.18)

$$\sum_{j \in s_1} \lambda_j \gamma_{\hat{Z}}(x_i - x_j) - \lambda_i \sigma^2 + \mu = \bar{\gamma}_{\hat{Z}}(x_i, V_o) \quad \text{for} \quad x_i \in s_2$$
$$\sum_{j \in s_1} \lambda_j \gamma_{\hat{Z}}(x_i - x_j) + \mu = \bar{\gamma}_{\hat{Z}}(x_i, V_o) \quad \text{for} \quad x_i \in s_1 - s_2$$
$$\sum_{j \in s_1} \lambda_j = 1$$

and the mean square error is again of the form

$$\mathrm{E}\left\{Z^*(V_o)-Z(V_o)\right\}^2=\sum_{t\in s_1}\lambda_t\bar{\gamma}_{\hat{Z}}(x_t,V_o)-\bar{\gamma}_{\hat{Z}}(V_o,V_o)+\mu \tag{5.19}$$

The interesting aspects of mixed kriging is that it only rests upon the variogram of the predictions, which can usually be fitted on large data sets. Formally, it is equivalent to standard kriging with measurement errors, but with the role of observations and predictions interchanged.

The assumptions made above are, of course, violated in punctual kriging, i.e. $V_o=\{x_o\}$. In this case, (5.17) becomes

$$\mathrm{E}\left\{Z^*(V_o)-Z(V_o)\right\}^2=-\sum_{i,j\in s_1}\lambda_i\lambda_j\gamma_{\hat{Z}}(x_i-x_j)+2\sum_{j\in s_1}\lambda_j(x_j-x_o)$$
$$+\sum_{j\in s_2}\lambda_j^2\sigma^2(x_j)+\sigma^2(x_o)-2\sum_{j\in s_2}\lambda_j\,\mathrm{cov}(\varepsilon(x_j),\varepsilon(x_o))$$

which leads to the modified kriging equations

$$\sum_{j\in s_1}\lambda_j\gamma_{\hat{Z}}(x_t-x_j)-\lambda_t\sigma^2(x_t)I_2(x_t)+I_2(x_t)\mathrm{cov}(\varepsilon(x_t),\varepsilon(x_o))+\mu=\gamma_{\hat{Z}}(x_t-x_o) \tag{5.20}$$
$$\sum_{j\in s_1}\lambda_j=1$$

where $I_2(x)=1$ if $x\in s_2$, $I_2(x)=0$ if $x\notin s_2$.

The expected mean square error is given by:

$$\mathrm{E}\left\{Z^*(V_o)-Z(V_o)\right\}^2=\sum_{j\in s_1}\lambda_j\gamma_{\hat{Z}}(x_j-x_o)+\sigma^2(x_o)-\sum_{j\in s_2}\lambda_j\,\mathrm{cov}(\varepsilon(x_j),\varepsilon(x_o))+\mu \tag{5.21}$$

If we now assume that $|x_t-x_j|>r=\mathit{range}(\gamma_\varepsilon(h))$, it is straightforward to describe the following special cases:

(i) if $x_o\in s_2$ then $z^*(x_o)=z(x_o)$, EMS$=0$. .
(ii) if $x_o\in s_1-s_2$ then $z^*(x_o)=\hat{z}(x_o)$, EMS$=\sigma^2(x_o)$.

In this sense, punctual mixed kriging generalizes the exact interpolation property of ordinary kriging; it gives the true observation at an observation point (with zero kriging variance of course) and the prediction at a pure prediction point (with kriging variance equal then to the prediction variance at this point), which is intuitively appealing. However, punctual kriging is irrelevant for forest inventory; it must also be

emphasized that mixed kriging for domains, as given by equations (5.18-5.19) does not enjoy this property, and that it is a valid technique as long as the range of the correlation of the residual process is small in comparison with the domain to be estimated.

Therefore, kriging with errors and mixed kriging are straightforward procedures to incorporate auxiliary information; they yield unbiased point estimates but do not give reliable mean square errors; kriging with errors neglects the correlation between residual and true value, whereas mixed kriging assumes a very short range of the residual process. Nevertheless, the case study of chapter 7 reveals that they give better point estimates than ordinary kriging, which does not incorporate any auxiliary information. Finally, kriging with errors can be used to treat sampling errors in ordinary kriging.

5.4 Co-Kriging

The estimators (5.7) and (5.14) use the predictions $\hat{z}(x)$ where the observations $z(x)$ are not available, i.e. in $s_1 - s_2$. Obviously, there are no reasons for not using $\hat{z}(x)$ in s_2, which is precisely what co-kriging does. The idea is to find the best linear unbiased estimate in the class:

$$z^*(V_o) = \sum_{i \in s_2} \lambda_i z(x_i) + \sum_{l \in s_1} \mu_l \hat{z}(x_l)$$

Before proceeding further we state explicitly several simple technical facts, which are often either overlooked or a source of confusion in the literature, primarily because co-kriging is usually presented in terms of covariances and not of variograms, an unnecessary restriction (see also D. E. Myers, 1982).

Recall first the two possible definitions of the cross-variograms

$$2\gamma_{Z,\hat{Z}}(h) = \mathrm{cov}(Z(x+h) - Z(x), \hat{Z}(x+h) - \hat{Z}(x)) \quad \text{and} \tag{5.22}$$

$$2\Gamma_{Z,\hat{Z}}(h) = \mathrm{var}(Z(x+h) - \hat{Z}(x)) \tag{5.23}$$

as well as their properties

$$\gamma_{Z,\hat{Z}}(h) = \gamma_{\hat{Z},Z}(h), \text{ but in general } \gamma_{Z,\hat{Z}}(h) \neq \gamma_{Z,\hat{Z}}(-h) \tag{5.24}$$

$$\Gamma_{Z,\hat{Z}}(h) = \Gamma_{\hat{Z},Z}(-h), \text{ but in general } \Gamma_{Z,\hat{Z}}(h) \neq \Gamma_{\hat{Z},Z}(h) \tag{5.25}$$

If the processes $Z(x)$ and $\hat{Z}(x)$ are furthermore second order stationary and if the following symmetry condition holds

$$C_{Z,\hat{Z}}(h) = \mathrm{cov}(Z(x+h), \hat{Z}(x)) = \mathrm{cov}(\hat{Z}(x+h), Z(x)) = C_{\hat{Z},Z}(h) \tag{5.26}$$

($C_{Z,\hat{Z}}(h) = C_{\hat{Z},Z}(-h)$ being always true)

then it is straightforward to check that the following relations hold

$$\gamma_{Z,\hat{Z}}(h) = C_{Z,\hat{Z}}(0) - C_{Z,\hat{Z}}(h) \quad , \quad \gamma_{Z,\hat{Z}}(h) = \Gamma_{Z,\hat{Z}}(h) - \Gamma_{Z,\hat{Z}}(0) \tag{5.27}$$

Let $u_o \in R^2$ arbitrary, then

$$\gamma_{Z,\hat{Z}}(y-x) = 0.5\,\mathrm{cov}\{Z(y) - Z(u_o) + Z(u_o) - Z(x), \hat{Z}(y) - \hat{Z}(u_o) + \hat{Z}(u_o) - \hat{Z}(x)\} =$$
$$\gamma_{Z,\hat{Z}}(y-u_o) + \gamma_{Z,\hat{Z}}(x-u_o) - \mathrm{cov}(Z(x) - Z(u_o), \hat{Z}(y) - \hat{Z}(u_o))$$

provided that the well defined covariance of increments is symmetrical, i.e. if

(5.28)

$$\operatorname{cov}(Z(y)-Z(u_o),\hat{Z}(x)-\hat{Z}(u_o))=\operatorname{cov}(Z(x)-Z(u_o),\hat{Z}(y)-\hat{Z}(u_o))$$

then, by the relation above, this covariance also satisfies

$$\operatorname{cov}(Z(y)-Z(u_o),\hat{Z}(x)-\hat{Z}(u_o))=\gamma_{Z,\hat{Z}}(y-u_o)+\gamma_{Z,\hat{Z}}(x-u_o)-\gamma_{Z,\hat{Z}}(y-x)$$

and consequently this also implies

$$\gamma_{Z,\hat{Z}}(y-x)=\gamma_{Z,\hat{Z}}(x-y)$$

From this we can easily calculate the covariance of two authorized linear combinations $\sum\lambda_i=0\ ,\sum\mu_l=0$, namely

$$\operatorname{cov}(\sum\lambda_i Z(x_i),\sum\mu_l\hat{Z}(y_l))=\operatorname{cov}(\sum\lambda_i(Z(x_i)-Z(u_o)),\sum\mu_l(\hat{Z}(y_l)-\hat{Z}(u_o)))$$

$$\sum_{i,l}\lambda_i\mu_l\left(\gamma_{Z,\hat{Z}}(x_i-u_o)+\gamma_{Z,\hat{Z}}(y_l-u_o)-\gamma_{Z,\hat{Z}}(x_i-y_l)\right)=-\sum_{i,l}\lambda_i\mu_l\gamma_{Z,\hat{Z}}(x_i-y_l)$$

Therefore we can state the important result

(5.29)

If $\forall x,y,u_o$

$$\operatorname{cov}(Z(y)-Z(u_o),\hat{Z}(x)-\hat{Z}(u_o))=\operatorname{cov}(Z(x)-Z(u_o),\hat{Z}(y)-\hat{Z}(u_o))$$

then

$$\gamma_{Z,\hat{Z}}(y-x)=\gamma_{Z,\hat{Z}}(x-y)$$

and for all linear combinations $\sum\lambda_i=0\ ,\sum\mu_l=0$, we have

$$\operatorname{cov}(\sum\lambda_i Z(x_i),\sum\mu_l\hat{Z}(y_l))=-\sum_{i,l}\lambda_i\mu_l\gamma_{Z,\hat{Z}}(x_i-y_l)$$

This generalizes the fundamental property of the variogram to the cross-variogram. The symmetry condition in (5.29) seems natural, but is of course very difficult to check in practice. This result is easier to establish with the second cross-variogram, indeed

$$\Gamma_{Z,\hat{Z}}(x-y)=0.5\operatorname{var}(Z(x)-\hat{Z}(y))=0.5\operatorname{var}Z(x)+0.5\operatorname{var}\hat{Z}(y)-\operatorname{cov}(Z(x),\hat{Z}(y))$$

(this is true also in the non-stationary case) so that

$$\operatorname{cov}(\sum\lambda_i Z(x_i),\sum\mu_l\hat{Z}(y_l))=\sum_{i,l}\lambda_i\mu_l\left(0.5\operatorname{var}Z(x_i)+0.5\operatorname{var}Z(y_l)-\Gamma_{Z,\hat{Z}}(x_i-y_l)\right)$$

$$=-\sum_{i,l}\lambda_i\mu_l\Gamma_{Z,\hat{Z}}(x_i-y_l)\ ,\ \text{by using the constraints.}$$

Hence, we have

(5.30)

if $\sum\lambda_i=0\ ,\sum\mu_l=0$ then

$$\operatorname{cov}(\sum\lambda_i Z(x_i),\sum\mu_l\hat{Z}(y_l))=-\sum_{i,l}\lambda_i\mu_l\Gamma_{Z,\hat{Z}}(x_i-y_l)$$

For future use we need the following results:

(5.31)

if $\sum\lambda_i = 1$, $\sum\mu_l = 0$ and the symmetry condition (5.29) holds then

$$\operatorname{cov}\left(\sum_i \lambda_i Z(x_i) - \frac{1}{\lambda(V_o)}\int_{V_o} Z(x)dx \,,\, \sum_l \mu_l \hat{Z}(y_l)\right) =$$
$$-\sum_{i,l}\lambda_i\mu_l\gamma_{Z,\hat{Z}}(x_i - y_l) + \sum_l \mu_l \bar{\gamma}_{Z,\hat{Z}}(y_l, V_o) = -\sum_{i,l}\lambda_i\mu_l\Gamma_{Z,\hat{Z}}(x_i - y_l) + \sum_l \mu_l \bar{\Gamma}_{Z,\hat{Z}}(y_l, V_o)\,.$$

where as usual

$$\bar{\gamma}_{Z,\hat{Z}}(u, V_o) = \frac{1}{\lambda(V_o)}\int_{V_o}\gamma_{Z,\hat{Z}}(u-v)dv \quad , \quad \bar{\Gamma}_{Z,\hat{Z}}(u, V_o) = \frac{1}{\lambda(V_o)}\int_{V_o}\Gamma_{Z,\hat{Z}}(u-v)dv$$

This can be easily proved by approximating the integral by an arithmetic mean and by using (5.29-5.30). Likewise, one obtains with $(\hat{Z}, \varepsilon)$ in place of $(Z, \hat{Z})$ the following result:

(5.32)

if $\sum\lambda_i = 1$, $\sum\mu_l = 1$ and the symmetry condition (5.29) for $(\hat{Z}, \varepsilon)$ holds then

$$\operatorname{cov}\left(\sum_l \mu_l \hat{Z}(y_l) - \frac{1}{\lambda(V_o)}\int_{V_o}\hat{Z}(x)dx \,,\, \sum_i \lambda_i \varepsilon(x_i) - \frac{1}{\lambda(V_o)}\int_{V_o}\varepsilon(x)dx\right) =$$
$$\sum_{i,l}\lambda_i\mu_l\gamma_{\hat{Z},\varepsilon}(x_i - y_l) - \bar{\gamma}_{\hat{Z},\varepsilon}(V_o, V_o) + \sum_l \mu_l \bar{\gamma}_{\hat{Z},\varepsilon}(y_l, V_o) + \sum_i \lambda_i \bar{\gamma}_{\hat{Z},\varepsilon}(x_i, V_o) =$$
$$\sum_{i,l}\lambda_i\mu_l\Gamma_{\hat{Z},\varepsilon}(x_i - y_l) - \bar{\Gamma}_{\hat{Z},\varepsilon}(V_o, V_o) + \sum_l \mu_l \bar{\Gamma}_{\hat{Z},\varepsilon}(y_l, V_o) + \sum_i \lambda_i \bar{\Gamma}_{\hat{Z},\varepsilon}(x_i, V_o)\,.$$

where

$$\bar{\gamma}_{\hat{Z},\varepsilon}(V_o, V_o) = \frac{1}{\lambda^2(V_o)}\int_{V_o}\int_{V_o}\gamma_{\hat{Z},\varepsilon}(u-v)dudv \quad , \quad \bar{\Gamma}_{\hat{Z},\varepsilon}(V_o, V_o) = \frac{1}{\lambda^2(V_o)}\int_{V_o}\int_{V_o}\Gamma_{\hat{Z},\varepsilon}(u-v)dudv$$

We are now ready to find the BLUE estimate in the class

(5.33)

$$z^*(V_o) = \sum_{i\in s_2}\lambda_i z(x_i) + \sum_{l\in s_1}\mu_l \hat{z}(x_l)$$

The error is

$$z^*(V_o) - z(V_o) = \sum_{i\in s_2}\lambda_i z(x_i) - \frac{1}{\lambda(V_o)}\int_{V_o} z(x)dx + \sum_{l\in s_1}\mu_l \hat{z}(x_l)$$

hence the conditions

$$\sum\lambda_i = 1 \quad , \quad \sum\mu_l = 0$$

ensure first that the error is an authorized linear combination of the underlying processes, and second that it has zero expectation (unbiasedness). The expected mean square error can now easily be calculated by using (4.4), (4.12), (5.31) and is found to be:

(5.34)

$$E\left\{Z^*(V_o)-Z(V_o)\right\}^2 = -\sum_{i,j\in s_2}\lambda_i\lambda_j\gamma_Z(x_i-x_j)-\bar{\gamma}_Z(V_o,V_o)+2\sum_{i\in s_2}\lambda_i\bar{\gamma}_Z(x_i,V_o)$$
$$-\sum_{k,l\in s_1}\mu_k\mu_l\gamma_{\hat{Z}}(x_k-x_l)-2\sum_{i\in s_2,l\in s_1}\lambda_i\mu_l\gamma_{Z,\hat{Z}}(x_i-x_l)+2\sum_{l\in s_1}\mu_l\bar{\gamma}_{Z,\hat{Z}}(x_l,V_o)$$

Minimizing (5.34) under the constraints yields the so called co-kriging equations, which read:

(5.35)

$$\sum_{i\in s_2}\lambda_i\gamma_Z(x_i-x_k)+\sum_{l\in s_1}\mu_l\gamma_{Z,\hat{Z}}(x_l-x_k)+\nu_1=\bar{\gamma}_Z(x_k,V_o) \quad \text{for} \quad k\in s_2$$
$$\sum_{i\in s_2}\lambda_i\gamma_{Z,\hat{Z}}(x_i-x_k)+\sum_{l\in s_1}\mu_l\gamma_{\hat{Z}}(x_l-x_k)+\nu_2=\bar{\gamma}_{Z,\hat{Z}}(x_k,V_o) \quad \text{for} \quad k\in s_1$$
$$\sum_{i\in s_2}\lambda_i=1$$
$$\sum_{l\in s_1}\mu_l=0$$

The mean square error is then found to be:

(5.36)

$$E\left\{Z^*(V_o)-Z(V_o)\right\}^2=\sum_{i\in s_2}\lambda_i\bar{\gamma}_Z(x_i,V_o)+\sum_{l\in s_1}\mu_l\bar{\gamma}_{Z,\hat{Z}}(x_l,V_o)-\bar{\gamma}_Z(V_o,V_o)+\nu_1$$

The same equations apply with $\Gamma_{Z,\hat{Z}}$ in place of $\gamma_{Z,\hat{Z}}$. If furthermore (5.27) holds (i.e. stationary processes with symmetrical covariance), then it is easily verified that the solutions $\lambda_i,\mu_l,\nu_1,\nu_2$ coincide, as well as the mean square error, so that both cross-variogram definitions are then equivalent with respect to co-kriging.

So far $Z,\hat{Z},\varepsilon$ were arbitrary intrinsic or stationary processes and the relation $Z(x)=\hat{Z}(x)+\varepsilon(x)$ has not been used; it implies at once from the definitions the following identities:

(5.37)

$$\gamma_Z(h)=\gamma_{\hat{Z}}(h)+\gamma_\varepsilon(h)+2\gamma_{\hat{Z},\varepsilon}(h)$$
$$\gamma_{Z,\hat{Z}}(h)=\gamma_{\hat{Z}}(h)+\gamma_{\hat{Z},\varepsilon}(h) \quad , \quad \gamma_{Z,\varepsilon}(h)=\gamma_{\hat{Z},\varepsilon}(h)+\gamma_\varepsilon(h)$$

The other cross-variograms $\Gamma_{Z,\hat{Z}}(h)$ and $\Gamma_{\hat{Z},\varepsilon}(h)$ do not have these properties, which are very useful for graphical model checking. Furthermore, if predictions and residuals (or their increments) are uncorrelated, the standard and natural assumption of any model, then $\gamma_{\hat{Z},\varepsilon}(h)\equiv 0$, but $\Gamma_{\hat{Z},\varepsilon}(h)\neq 0$. Furthermore, this also implies $\Gamma_{Z,\hat{Z}}(x-y)=\gamma_{\hat{Z}}(x-y)+0.5\,\mathrm{var}\,\varepsilon(x)$ and consequently a constant residual variance since this should depend only on $x-y$. For these reasons, we prefer to use the cross-variograms

$\gamma_{Z,\hat{Z}}$ and $\gamma_{\hat{Z},\varepsilon}$, which are anyway more germane to the intrinsic hypothesis, as they are based on the covariance of the increments.

We shall now give an alternative formulation of the co-kriging procedure which is more in line with the design-based approach to double sampling as defined in (D. Mandallaz, 1991).

Since $Z(x)=\hat{Z}(x)+\varepsilon(x)$, the co-kriging estimate (5.33) can be rewritten in the form

(5.38)

$$z^*(V_o)=\sum_{i\in s_2}\tilde{\lambda}_i e(x_i)+\sum_{l\in s_1}\tilde{\mu}_l\hat{z}(x_l)$$

where

$\tilde{\lambda}_i=\lambda_i$ for $i\in s_2$, $\tilde{\mu}_l=\lambda_l+\mu_l$ for $l\in s_2$, $\tilde{\mu}_l=\mu_l$ for $l\in s_1-s_2$

which implies $\sum_{l\in s_2}\tilde{\lambda}_l=1$ and $\sum_{l\in s_1}\tilde{\mu}_l=1$

Therefore, finding the best estimate in the class (5.38) under the new constraints is equivalent to the co-kriging set-up as defined in (5.33). However, it is very instructive to work directly with (5.38).

$$z^*(V_o)-z(V_o)=\sum_{i\in s_2}\tilde{\lambda}_i e(x_i)-\frac{1}{\lambda(V_o)}\int_{V_o}e(x)dx+\sum_{l\in s_1}\tilde{\mu}_l\hat{z}(x_l)dx-\frac{1}{\lambda(v_o)}\int_{V_o}\hat{z}(x)dx$$

which is a sum of authorized linear combinations under the new constraints. Using (4.12) and (5.32) the expected mean square error is found to be:

$$\begin{aligned}E\left\{Z^*(V_o)-Z(V_o)\right\}^2 =& -\sum_{i,j\in s_2}\tilde{\lambda}_i\tilde{\lambda}_j\gamma_\varepsilon(x_i-x_j)-\bar{\gamma}_\varepsilon(V_o,V_o)+2\sum_{i\in s_2}\tilde{\lambda}_i\bar{\gamma}_\varepsilon(x_i,V_o)\\ &-\sum_{k,l\in s_1}\tilde{\mu}_k\tilde{\mu}_l\gamma_{\hat{Z}}(x_k-x_l)-\bar{\gamma}_{\hat{Z}}(V_o,V_o)+2\sum_{l\in s_1}\tilde{\mu}_l\bar{\gamma}_{\hat{Z}}(x_l,V_o)\\ &+2\sum_{l\in s_2,l\in s_1}\tilde{\lambda}_l\tilde{\mu}_l\gamma_{\hat{Z},\varepsilon}(x_l-x_l)-2\bar{\gamma}_{\hat{Z},\varepsilon}(V_o,V_o)\\ &+2\sum_{l\in s_1}\tilde{\mu}_l\bar{\gamma}_{\hat{Z},\varepsilon}(x_l,V_o)+2\sum_{l\in s_2}\tilde{\lambda}_l\bar{\gamma}_{\hat{Z},\varepsilon}(x_l,V_o)\end{aligned}$$

Minimizing this expression under the new constraints yields the modified co-kriging equations:

(5.39)

$$\sum_{l\in s_2}\tilde{\lambda}_l\gamma_\varepsilon(x_l-x_k)+\sum_{l\in s_1}\tilde{\mu}_l\gamma_{\hat{Z},\varepsilon}(x_l-x_k)+\tau_1=\bar{\gamma}_\varepsilon(x_k,V_o)+\bar{\gamma}_{\hat{Z},\varepsilon}(x_k,V_o)\quad\text{for}\quad k\in s_2$$

$$\sum_{l\in s_2}\tilde{\lambda}_l\gamma_{\hat{Z},\varepsilon}(x_l-x_k)+\sum_{l\in s_1}\tilde{\mu}_l\gamma_{\hat{Z}}(x_l-x_k)+\tau_2=\bar{\gamma}_{\hat{Z}}(x_k,V_o)+\bar{\gamma}_{\hat{Z},\varepsilon}(x_k,V_o)\quad\text{for}\quad k\in s_1$$

$$\sum_{l\in s_2}\tilde{\lambda}_l=1\ ,\ \sum_{l\in s_1}\tilde{\mu}_l=1$$

The mean square error is then found to be:

(5.40)

$$E\left\{Z^*(V_o)-Z(V_o)\right\}^2 = \sum_{i\in s_2}\tilde{\lambda}_i\bar{\gamma}_\varepsilon(x_i,V_o)-\gamma_\varepsilon(V_o,V_O)+\tau_1$$
$$+\sum_{l\in s_1}\tilde{\mu}_l\bar{\gamma}_{\hat{Z}}(x_l,V_o)-\gamma_{\hat{Z}}(V_o,V_O)+\tau_2$$
$$+\sum_{l\in s_2}\tilde{\lambda}_l\bar{\gamma}_{\hat{Z},\varepsilon}(x_l,V_o)+\sum_{l\in s_1}\tilde{\mu}_l\bar{\gamma}_{\hat{Z},\varepsilon}(x_l,V_o)-2\bar{\gamma}_{\hat{Z},\varepsilon}(V_o,V_O)$$

The equations (5.39) and (5.40) are also valid with $\Gamma_{\hat{Z},\varepsilon}$.

Substituting (5.37) into (5.35), it is straightforward but tedious to check that indeed
$\tilde{\lambda}_l=\lambda_l$ for $i\in s_2$, $\tilde{\mu}_l=\lambda_l+\mu_l$ for $l\in s_2$, $\tilde{\mu}_l=\mu_l$ for $l\in s_1-s_2$
and that the mean square errors are equal, as expected on general grounds. Why then consider the more difficult and non-standard equations (5.39) and (5.40)? The reason is that, if the increments of residuals and predictions are uncorrelated, then $\gamma_{\hat{Z},\varepsilon}(h)\equiv 0$ and (5.39) and (5.40) simply reduce to separate kriging of residuals and predictions, and subsequent addition of the resulting mean square errors (note that this simplification does not occur explicitly with the cross-variogram $\Gamma_{\hat{Z},\varepsilon}$, though under assumption 5.26 the solutions coincide). This is a very appealing procedure, which we shall call **double kriging** and investigate in the next section.

5.5 Double Kriging

Double Kriging yields the best linear unbiased estimate in the class

$$z^*(V_o)=\sum_{i\in s_2}\lambda_i\varepsilon(x_i)+\sum_{l\in s_1}\mu_l\hat{z}(x_l) \qquad (5.41)$$

under the following assumptions:

i) $\varepsilon(x),\hat{Z}(x)$ are intrinsic processes

ii) $\varepsilon(x)$ and $\hat{Z}(x)$ have uncorrelated increments, i.e. $\operatorname{cov}\left(\varepsilon(x+h)-\varepsilon(x),\hat{Z}(x+h)-\hat{Z}(x)\right)=2\gamma_{\hat{Z},\varepsilon}(h)=0$

iii) the covariance structure of the increments is symmetrical

$$\operatorname{cov}\left(\varepsilon(y)-\varepsilon(u_o),\hat{Z}(x)-\hat{Z}(u_o)\right)=\operatorname{cov}\left(\varepsilon(x)-\varepsilon(u_o),\hat{Z}(y)-\hat{Z}(u_o)\right)\quad\forall x,y,u_o$$

Remarks:

ii) is weaker than requiring residuals and predictions to be uncorrelated.
iii) is required in order for (5.29) and (5.39) to hold.

Under these assumptions, the co-kriging result (5.39) shows that the double kriging estimate can be obtained by kriging residuals and predictions separately i.e.

$$\sum_{l\in s_2}\lambda_l\gamma_\varepsilon(x_l-x_k)+\tau_1=\bar{\gamma}_\varepsilon(x_k,V_o)\quad\text{for}\quad k\in s_2 \qquad (5.42)$$
$$\sum_{l\in s_2}\lambda_l=1$$
$$\sum_{l\in s_1}\mu_l\gamma_{\hat{Z}}(x_l-x_k)+\tau_2=\bar{\gamma}_{\hat{Z}}(x_k,V_o)\quad\text{for}\quad k\in s_1$$
$$\sum_{l\in s_1}\mu_l=1$$

The mean square error is then found to be

$$\mathrm{E}\left\{Z^*(V_o)-Z(V_o)\right\}^2=\sum_{i\in s_2}\lambda_i\bar{\gamma}_\varepsilon(x_i,V_o)-\gamma_\varepsilon(V_o,V_O)+\tau_1 \qquad (5.43)$$
$$+\sum_{l\in s_1}\mu_l\bar{\gamma}_{\hat{Z}}(x_l,V_o)-\gamma_{\hat{Z}}(V_o,V_O)+\tau_2$$

Double kriging is therefore remarkably simple, it is obtained by the following steps:

1) Perform an ordinary kriging of the predictions.
2) Perform an ordinary kriging of the residuals.
3) Add up the point estimates and the expected mean square errors.

Once again we emphasize that double-kriging is equivalent to co-kriging observations and predictions according to (5.35) or to co-kriging predictions and residuals in the sense of (5.39); this is true essentially under the condition that predictions and residuals have uncorrelated increments, a natural assumption, particularly if an external prediction model is used, which, and this is of the utmost importance, does not have to be "true" (i.e. $E\varepsilon(x)=0$) for the procedure to be valid (simply because the model equation (3.6) actually defines the residual process). If one really believes in the model, then simple kriging instead of ordinary kriging could be used for the residuals, but this seems generally to be an over-optimistic view and, besides, unnecessary. Finally, it is easily seen that double-kriging yields an exact interpolator.

It is worth noting that the design-based regression estimate (linear or not) for double sampling can be rewritten in the double-kriging form (5.41) by setting $\lambda_i = n_2^{-1}$, $\mu_i = n_1^{-1}$; if the two variograms are pure nugget effects the mean square error corresponds formally to the design-based variance (D. Mandallaz, 1991, 1992). In this sense, double kriging is the natural generalization of the design-based regression estimates used in double-sampling schemes.

The situations in which co-kriging can be simplified or even reduced to separate kriging procedures has received much attention (G. Matheron, 1965; A.G. Journel, C. Huijbregts, 1978; J. Rivoirard, 1989).

Kriging of residuals has been used in many applications (for a review see S. P. Neumann, E. A. Jacobson, 1984), primarily to take non-stationarity into account. The approach presented here differs in as much as one wants to get better estimates by using auxiliary information, summarized in the prediction

process $\hat{Z}(x)$; that is, techniques originally developed for the non-stationary case can be used, even if the prediction and the residual processes are stationary or at least intrinsic; if they are not, one can, at least in principle, use non-stationary geostatistics to both components of (5.40), e.g. intrinsic kriging of order k (see G. Matheron, 1973; Delfiner, 1976; P. Chauvet, 1991; section 5.6 below). Also, it is more in line with design-based sampling theory.

So far, we assumed that the model was external, that is given beforehand, and not fitted on the basis of the present inventory data. If the model itself has to be constructed from the inventory data (i.e. with $\{z(x_i), A(x_i), i \in s_2\}$), or only the unknown parameter β to be estimated, several problems occur:

1) Any estimate $\hat{\beta}$ of β cannot be optimal, since this would require the knowledge of the true underlying variogram of the residual process.

2) Any estimation of the true underlying variogram $\gamma_\varepsilon(h)$ on the basis of the fitted residuals is biased. The bias is generally small for small lags h, but can be serious for large lags (G. Matheron, 1970;N. Cressie, 1991; M. Armstrong, 1984; J.P. Chilès, 1977). In the next section, we shall present a new model-based method, which eliminates these difficulties.

If the auxiliary information is reduced to a set of polynomials in the co-ordinates, it is possible to bypass the above circularity by using the theory of intrinsic functions of order k (G. Matheron, 1973). Unfortunately, polynomials do not really convey the idea of auxiliary information as needed for forest inventory. However, we shall now see how the IRF-k ideas can be adapted to our problem.

5.6 External Drifts and Universal Kriging.

In this chapter we consider the following decomposition

$$Z(x) = m(x) + \varepsilon(x)$$

where $m(x)$ is a deterministic function and $\varepsilon(x)$ is a stochastic process with zero expectation and finite variance (not necessarily stationary). If $m(x)$ is known everywhere, the estimation of

$$z(V_o) = \frac{1}{\lambda(V_o)} \int_{V_o} z(x)dx$$

is straightforward, namely

$$z^*(V_o) = \frac{1}{\lambda(V_o)} \int_{V_o} m(x)dx + e^*(V_o)$$

where $e^*(V_o)$ is the simple kriging estimate based on the true residuals $e(x) = z(x) - m(x)$. However, $m(x)$ is rarely, if ever, completely known and has to be modelled. To pursue further, we shall postulate a linear relationship, i.e. set

$$m(x) = F(x)'\beta \quad \text{with} \quad F(x), \beta \in R^p$$

Note: Vectors are always understood as column vectors and the upper index t stands for the transposition operator of vectors and matrices.

If the model is external and correct (that is β is known and $E\,Z(x) = m(x)$) and $F(x)$ is known everywhere (for instance when the auxiliary information rests upon thematic maps), then the previous remarks apply ($\int m(x)dx$ is known). If these conditions are not met, we are back to the circular problem mentioned in the previous chapter. However, if $F(x)$ is a polynomial of degree k in the coordinates (possibly with locally varying coefficients), the techniques of intrinsic kriging of order k can be applied (G. Matheron, 1973; Galli et al, 1987, G. Christakos, 1992). This approach is perfectly justified to model non-stationary behaviour with either :

(i) an obvious structure (e.g. a strong linear trend in one direction).

(ii) an unclear structure without an intuitive physical background.

Under case (ii) the , so called **internal**, drift can be approximately and purely pragmatically modelled by local polynomials. Unfortunately, this does not really address the

problem of combined forest inventory, primarily because the auxiliary information has generally a strong influence on $m(x) = \mathrm{E}\,Z(x)$, with a direct physical meaning (e.g. stand structure), which we shall call an **external** drift and which cannot be adequately modelled by polynomials ($m(x)$ being for instance discontinous at stand boundaries). The main idea of intrinsic kriging of order k is to consider only linear combinations of the data which filter out the polynomial trends, and this can be heuristically extended to any $F(x)$ as we shall see.Before going into the mathematics, let us emphasize a few points of more philosophical nature.

In the decomposition (5.43) **"drift+residual"** the process $Z(x)$ is by definition non-stationary, whereas a structural analysis performed on a realization $z(x)$ (and also on $e(x), \hat{z}(x)$) may well not reject the assumption of stationarity. Why, then ,use techniques specially designed to cope with non-stationarity, like universal or intrinsic kriging of order k ? Not for the sake of sophistication, but to take auxiliary information into account. This situation may seem rather confusing, as also are the dichotomies "stationarity/non-stationarity" and "drift / residuals", concepts which are highly dependent on the scale of the problem at hand and of the availability of auxiliary information (for a review see A. G.Journel, M. E. Rossi, 1990; D. M Myers, 1989). To formulate this issue in forestry terms: a forest may well look stationary if one does not have a stand map of it, but not if one has, and likewise if one looks at smaller or larger parts of it. From a pragmatic point of view the correct approach is the one giving the best estimates; since the true value of $z(V_o)$ is nearly always unknown (an exception being the case study of chapter 7), it is recommendable to use several techniques for comparison.

We define the following column vectors and simplify the notation by setting $n = n_2$ (formally $n_1 = \infty$, since the auxiliary information must be known everywhere).

(5.44)

$$z = \left(z(x_1), z(x_2), \ldots\ldots z(x_n)\right)^t \in R^n, x_i \in s_2$$

$$e = \left(e(x_1), e(x_2), \ldots\ldots e(x_n)\right)^t \in R^n, x_i \in s_2$$

(all vectors are defined as column vectors, and the vector e is the realization of the corresponding random vector ε).

The (n,p) matrix F is defined according to:

(5.45)

i-th row of $F = F(x_i) \in R^p$

We shall of course assume that the number of observations is much larger than the number of parameters, i.e. $n > p$.

The restriction of the model at the data points x_i reads

(5.46)

$$z = F\beta + e$$

z is observable but not e since β is unknown. We assume furthermore that

(5.47)

$$E\varepsilon(x) = 0, E\varepsilon\varepsilon^t = \Sigma_\varepsilon, \Sigma_{\varepsilon i,j} = \mathrm{cov}(\varepsilon(x_i), \varepsilon(x_j))$$

We define for future use

(5.48)

$$\sigma_o^2 = \mathrm{var}\left(\frac{1}{\lambda(V_o)}\int_{V_o}\varepsilon(x)dx\right) \in R^1$$

$$\sigma_{n,o}^2 = \mathrm{cov}\left(\varepsilon(x_i), \frac{1}{\lambda(V_o)}\int_{V_o}\varepsilon(x)dx\right) \in R^n, x_i \in s_2$$

$$F_o = \frac{1}{\lambda(V_o)}\int_{V_o} F(x)dx \in R^p$$

The universal kriging estimate of

$$z(V_o) = \frac{1}{\lambda(V_o)}\int_{V_o} z(x)dx$$

is the best linear unbiased estimate in the class

$$z^*(V_o) = \lambda^t z, \lambda \in R^n$$

which implies the condition

$$F^t\lambda = F_o$$

Unbiasedness can always be ensured if the following compatibility condition holds :

$$span(F(x); x \in s_2) = span(F(x); x \in V) \subset R^p$$

i.e. the data points span the entire space of auxiliary information.

The mean square error is easily found to be :

$$E\left(Z^*(V_o) - Z(V_o)\right)^2 = \lambda^t\Sigma_\varepsilon\lambda + \sigma_o^2 - 2\lambda^t\sigma_{n,o}^2$$

Minimizing the above expression under the unbiasedness constraint by the Lagrange's technique yields the universal kriging equations :

$$\begin{pmatrix} \Sigma_\varepsilon & F \\ F^t & 0 \end{pmatrix} \begin{pmatrix} \lambda \\ \mu \end{pmatrix} = \begin{pmatrix} \sigma^2_{n,o} \\ F_o \end{pmatrix} \qquad \textbf{(5.49)}$$

where $\mu \in R^p$ is the vector of Lagrange's multipliers and 0 is the (p,p) null matrix.

The mean square error is then equal to

$$\mathrm{E}\left(Z^*(V_o) - Z(V_o)\right)^2 = \sigma^2_o - \lambda^t \sigma^2_{n,o} - F_o^t \mu \qquad \textbf{(5.50)}$$

It must be emphasized that the kriging equations (5.49-5.50) are very general and in particular they do not require any assumptions of stationarity. Note that universal kriging gives also an exact interpolator.

The main drawback of the universal kriging equations (5.49-5.50) is, of course, that $\Sigma_\varepsilon, \sigma^2_{n,o}\ \sigma^2_o$ are unknown, which leads again to the circularity problem mentionned before (N. Cressie, 1991); by assumption F_o is known since the auxiliary information is known in every point, otherwise it must be estimated, which induces further difficulties we shall not deal with here. Furthermore, there is a fundamental difficulty, which is often overlooked, by postulating the model (in vector form)

$$Z = F\beta + \varepsilon$$

Indeed, one can add a zero mean random vector $\tilde{\beta}$ to β and redefine the residuals, i.e. set

$$Z = F(\beta + \tilde{\beta}) + \tilde{\varepsilon}$$

The two models are obviously indistinguishable. When $F(x)$ is a polynomial this leads to the equivalence class of intrinsic random functions of order k (IRF-k for short) as defined by G. Matheron (1973). We shall see that the same kind of indetermination holds for external drifts, but that it is irrelevant for estimation purposes.

In analogy with the IRF-k theory, we consider only linear combinations of the observations which filter out the unknown drift, i.e.

$$\forall \beta \quad \lambda' z = \lambda' F\beta + \lambda' e = \lambda' e \text{ for all } \lambda \text{ such that } \lambda' F = 0 \tag{5.51}$$

Since a unique realization is available the above authorized linear combinations of the data are, from an operational point of view, the only meaningful quantities as they are independent of the true but unknown underlying drift. In other words, λ must be orthogonal to the subspace of R^n generated by the columns of F. It is worth noting that the same idea is the starting point, in the gaussian case, of the so-called restricted maximum likelihood procedures (R. Christensen, 1990b; D.L. Zimmermann, 1989), which is presented in the mathematical appendix. To characterize this subspace we use the singular value decomposition of the matrix F (G.H. Golub, C.F. van Loan, 1983) given by :

$$F = UDV^t \tag{5.52}$$

where U and V are orthogonal (n,n) and (p,p) matrices and D is an (n,p) "diagonal" matrix, satisfying :
$D_{ij} = 0$ for $i \neq j, D_{ii} = \sigma_i$ for $i \leq p, D_{ij} = 0$ *for* $i > p$ and
$\sigma_1 \geq \sigma_2 \geq .... \geq \sigma_q > \sigma_{q+1} = ... = \sigma_p = 0$, $rank(F) = q \leq p$.
Partitioning the matrix U into its first q and last $n-q$ column vectors, $U = \left(U_q \middle| U_{n-q}\right)$, one can determine the projection operator $P^{\perp}$ onto the subspace orthogonal to $Range(F) = F(R^p)$ (i.e. the residual-space) according to

$$P^{\perp} = U_{n-q} U_{n-q}^t \ .$$

Note that the projection operator onto the model-space $Range(F)$ is $P = I - P^{\perp} = U_q U_q^t$ (see G.H. Golub, C.F. van Loan, 1983, p.21) and that $rank(P^{\perp}) = n - q$, $rank(P) = q$.
We calculate the projection operator via the singular value decomposition because of numerical efficiency and generality (non-full rank models with $q < p$ are very useful with qualitative data) and not via the computing intensive "hat-matrix", i.e. $P^{\perp} = I_n - F(F^t F)^{-1} F^t$, as commonly done.
The vector of the observed residual is defined by $r = P^{\perp} z$, and the corresponding random variable by

$$R = P^{\perp} Z = P^{\perp}(F\beta + \varepsilon) = P^{\perp}\varepsilon \quad \text{as} \quad P^{\perp} F = 0 \ .$$

Note that the so defined residuals are not the optimal ones as this would require the knowledge of the true covariance matrix.

The authorized linear combinations λ are then of the form $\lambda = P^{\perp}\mu$ for some $\mu \in R^n$ (recall that $P^{\perp}$ is idempotent and symmetric). One then has

$\lambda' Z = (P^{\perp}\mu)' Z = \mu' P^{\perp} Z = \mu' P^{\perp}\varepsilon = \mu' R$ and therefore $\mathrm{E}\,\lambda' Z = \mathrm{E}\,\mu' R = 0$

For instance, the individual empirical residuals are authorized linear combinations, indeed it suffices to choose $\mu_l, l = 1,2....n$ as the canonical basis of R^n and to set $\lambda_l = P^{\perp}\mu_l$.

More generally, one could take any basis μ_l of R^n. In all instances the variance of an authorized linear combination is given by:

$$\mathrm{var}(\lambda_l' Z) = \mathrm{E}(\lambda_l' Z)^2 = \mathrm{var}(\mu_l' P^{\perp}\varepsilon) = \mu_l' P^{\perp}\Sigma_{\varepsilon} P^{\perp}\mu_l = \lambda_l' \Sigma_{\varepsilon}\lambda_l$$

Hence, a sensible strategy for fitting a parametric model to Σ is to minimize the expression:

$$\sum_{l=1}^{n}\left((\lambda_l' z)^2 - \lambda_l' \Sigma_{\varepsilon}\lambda_l\right)^2 \qquad \textbf{(5.53)}$$

Alternatively, since $\lambda_l' Z = \mu_l' P^{\perp}\varepsilon = \mu_l' R$ one also has

$$\mathrm{var}(\lambda_l' Z) = \mu_l' \mathrm{E}\, RR' \mu_l = \mu_l' P^{\perp}\Sigma_{\varepsilon} P^{\perp}\mu_l$$

where Σ_{ε} denotes the covariance of the true residual vector. Therefore, minimizing (5.53) is equivalent to minimizing

$$\sum_{l=1}^{n}\left(\mu_l'\left(rr' - P^{\perp}\Sigma_{\varepsilon} P^{\perp}\right)\mu_l\right)^2 \qquad \textbf{(5.54)}$$

The question is now to choose a family of vectors μ_l for which (5.54) is to be minimized. It seems natural to impose a normalizing constraint on the vector norms and not to privilege any direction, which leads to using orthonormal bases only.

Furthermore, a sensible approach to estimate the unknown covariance matrix Σ_{ε} is to minimize the maximum of (5.54), i.e. to use a minimax strategy.

Let us define $\Delta = rr' - P^{\perp}\Sigma_{\varepsilon} P^{\perp}$ and the Frobenius norm of any matrix A by $\|A\| = \left(\sum_{i,j} a_{ij}^2\right)^{1/2} = \sqrt{Tr(A'A)}$.

Since μ_l is an orthonormal basis there exists an orthogonal matrix B such that $\mu_l = Be_l$, where e_l is the canonical basis. Set $\tilde{\Delta} = B^t \Delta B$, then one has:

$$\sum_{l=1}^{n} \left(\mu_l{}^t \Delta \mu_l\right)^2 = \sum_{l=1}^{n} \left(e_l{}^t \tilde{\Delta} e_l\right)^2 = \sum_{l=1}^{n} \tilde{\Delta}_{ll}{}^2 = \sum_{l,k=1}^{n} \tilde{\Delta}_{lk}{}^2 - \sum_{l \neq k}^{n} \tilde{\Delta}_{lk}{}^2 \leq \left\|\tilde{\Delta}\right\|^2 = \|\Delta\|^2 = Trace\, \Delta^2 = \sum_{l=1}^{n} \alpha_l{}^2$$

where we have used the invariance of the Frobenius norm under orthogonal transformations (G. H. Golub, van Loan, 1983, p. 15) and where $\alpha_1^2 \geq \alpha_2^2 \geq .... \geq \alpha_n^2$ are the squared eigenvalues of the matrix Δ (which is symmetric but not necessarily positive definite). It is now clear that the upper bound is actually achieved when the orthonormal basis μ_l coincides with the orthogonal basis consisting of the eigenvectors of the matrix Δ. For computational purposes it is useful to note that with $Q = P^{\perp}$,

$$\left\|rr^t - Q\Sigma Q\right\| = Trace\left((rr^t - Q\Sigma Q)(rr^t - Q\Sigma Q)\right) = Trace\left((rr^t - \Sigma Q)(rr^t - \Sigma Q)\right) = \left\|rr^t - \Sigma Q\right\| = \left\|rr^t - Q\Sigma\right\|$$

(because of the invariance of the trace under cyclic permutation and $Qr = r$, $Q^2 = Q$).

Hence we have proved the following important result:

(5.55)

The minimax least square estimate $\hat{\Sigma}_\varepsilon$ of the residual covariance matrix Σ_ε minimizes the Frobenius norm

$$\left\|rr^t - P^{\perp}\Sigma P^{\perp}\right\| = \left\|rr^t - P^{\perp}\Sigma\right\|$$

It is interesting to note that the Frobenius norm was also proposed by P. K. Kitanidis (1985) in the context of quadratic estimation of components of covariance; in this approach one postulates a decomposition of the form :

(5.56)

$$\Sigma_\varepsilon = \sum_{i=1}^{K} \sigma_i^2 V_i$$

with known matrices V_i; the coefficients can then be explicitely estimated via quadratic forms (see also R.J. Marshall and K.V. Mardia, 1985). The least square approach (5.55) is, of course, far more general, but requires numerical minimization, because the parametric model for the covariance matrix is usually non-linear in the parameters. This can be prohibitive if the dimension of the parameter space is large. However, if the model is adequate, the residual can be expected to have a simple covariance structure (i.e. spherical, circular, exponential, gaussian), depending only on sill, range and possibly a

nugget effect. Anisotropy or nested structures could indicate that not all the relevant auxiliary variables have been taken into account. To be specific, let us assume that $\Sigma(\vartheta)=\sigma^2 K(\vartheta)$, where $K(\vartheta)$ denotes a correlation matrix depending on the unknown parameter $\vartheta \in R^k$ and the unknown variance σ^2.

It is shown in the appendix that, under fairly general conditions on the correlation matrix for a set of sample points, the following algorithm yields, in the gaussian case, asymptotically consistent estimates $\hat{\vartheta}, \hat{\sigma}^2$:

(5.57)

Step 1

Find, with respect to ϑ, the absolute minimum of

$$\left\| rr^t - \sigma^2(\vartheta) P^{\perp} K(\vartheta) \right\|$$

or, equivalently, the absolute maximum of

$$LS := \left(r^t K(\vartheta) r\right) \sigma^2(\vartheta)$$

where

(5.58)

$$\sigma^2(\vartheta) = \frac{r^t K(\vartheta) r}{Tr(P^{\perp} K(\vartheta))^2}$$

Let $\hat{\vartheta}$ denote the unique absolute extremum

Step 2

Set

$$\hat{\sigma}^2 = \sigma^2(\hat{\vartheta})$$

Remarks:

(1) In the simple stationary case, the only "external drift" is the constant "1" and the minimax least square estimation amounts essentially to perform non-linear weighted least squares on the empirical covariances at all lags, the weights being proportional to the number of pairs (in practice one usually retains the lower lags of the empirical variogram in order to have enough pairs and a better estimate near the origin). Therefore, the above procedure is a mathematically sound generalization of the empirical fitting techniques commonly used.

(2) The gaussian assumption can be removed with technical conditions on the 4th moments, hence the least square procedure is more general than the gaussian maximum likelihood (though the later is often used in non-gaussian cases). The correlation matrix must not necessarily be stationary, though it will have to be in applications.

(3) The convergence holds for domain asymptotics, i.e. one considers a fixed grid and let the domain tend to infinity. It may not hold for infill asymptotics, i.e. when the number of sample points in a finite domain tends to infinity. Fortunately, this is not really relevant for forest inventory.

(4) In the stationary case, convergence holds not only for all covariances with finite ranges but also for some displaying long term dependencies.

(5) In finite samples, the convergence conditions imply that the correlation range must be substantially smaller than the dimension of the domain containing the sample points used for the estimation. If, a posteriori, this is not the case the estimation procedure can be unreliable. If it is, the kriging estimates for the entire domain will tend to be very close to the classical sampling theory estimates. In this sense, geostatistical estimation is primarily relevant to obtain unbiased and efficient local estimates. This important point is rarely mentioned, probably because formal convergence proofs were not available.

(6) The restricted maximum likelihood estimates are more efficient than the least square estimates. From a numerical point of view, the least square estimates are far less computer intensive under straightforward grid search techniques since, in contrast to maximum likelihood, they do not require inversions of the, sometimes very large, correlation matrices. The efficiency ratio tends to 1 for increasing nugget effect (see 7 below) or decreasing

correlation ranges. Formulae for the asymptotic covariances matrices of the estimates are given in the mathematical appendix, as well as a discussion in terms of the spectral density of the underlying residual process.

(7) To take a nugget effect into account simply set

$$\Sigma_\varepsilon = \sigma_1^2 K_1(\vartheta) + \sigma_2^2 I = \sigma^2\left((1-\lambda)K_1(\vartheta) + \lambda I\right), \sigma^2 = \sigma_1^2 + \sigma_2^2, \lambda = \frac{\sigma_2^2}{\sigma^2}$$

where λ is the proportion of the variance due to the nugget effect.

(8) In practice one should look at the empirical variogram of the residuals to determine the domain over which the optimization should be performed. The least square estimate could be used to perform a one step iteration with maximum likelihood in order to get shorter confidence intervals. In any case, the estimation process yields a unique extremum as long as the correlation matrices differ from the unity (i.e. a range shorter than the minimum distance between two sample points cannot be estimated !)

The interested reader will find formal or elaborated heuristic proofs of the above facts in the mathematical appendix.

The least square approach immediately suggests the use of other matrix norms to improve the robustness of the estimates. However, the euclidian Frobenius norm is intuitively more appealing and enjoys the important minimax interpretation.

From a numerical point of view, the required computing time can be a problem with very large matrices. One could then split the domain into subdomains and pool the estimates, provided they are coherent; if not, stratified universal kriging should be used instead.

The software BLUEPACK uses a related estimation procedure for estimating the generalized covariances in the IRF-k (k<=2) context. However, it directly works with a set of authorized linear combinations (i.e. with equation 5.53 instead of 5.55) based on moving neighbourhoods of moderate size. Moreover, it is purely empirical and convergence criteria are not available. Moving neighborhoods can be problematic with truly external drifts for two reasons; first because generalized covariances

are only legitimate with polynomial drifts and secondly because the compatibility conditions may not hold in small neighbourhoods, a problem rarely occuring with polynomials. In our experience, using generalized covariances with non-polynomial drifts can lead to completely misleading results.

We now briefly consider hypothesis testing for the residual covariance. Let Σ_0 be the true but unknown covariance matrix and Σ_1 the test covariance matrix.

Let ψ_l be the unit eigenvectors of the matrix $P^\perp\Sigma_1 P^\perp$ (symmetric and positive semi-definite) corresponding to the non-zero eigenvalues θ_l $l=1,2,....n-q=rank(P^\perp)=rank(P^\perp\Sigma_1 P^\perp)$. Under the null hypothesis $\Sigma_1=\Sigma_0$ the random variables $\psi_l^t R$ are uncorrelated since

$$\mathrm{cov}\left(\psi_l^t R,\psi_l^t R\right)=E\left(\psi_l^t P^\perp \varepsilon\varepsilon^t P^\perp \psi_k\right)=\psi_l^t P^\perp \Sigma_0 P^\perp \psi_k=\theta_k\delta_{k,l}$$

The linear combination $\psi_l^t r$ of the empirical residuals have the obvious non-parametric variance estimate $\left(\psi_l^t r\right)^2$ and as parametric estimate the eigenvalue . In the gaussian case, and if the model is correct, the random variable

(5.59)

$$X^2=\sum_{l=1}^{n-q}\frac{\left(\psi_l^t r\right)^2}{\theta_l}$$

is therefore distributed as a chi-square on $n-q$ degrees of freedom. Now, one can rewrite (5.59) as:

$$X^2=\sum_{l=1}^{n-q}\frac{\psi_l^t r r^t \psi_l}{\theta_l}=Trace\left(rr^t\sum_{l=1}^{n-q}\frac{\psi_l\psi_l^t}{\theta_l}\right)=r^t\sum_{l=1}^{n-q}\frac{\psi_l\psi_l^t}{\theta_l}r$$

The matrix $\sum_{l=1}^{n-q}\frac{\psi_l\psi_l^t}{\theta_l}$ is precisely the Moore-Penrose generalized inverse of $P^\perp\Sigma_1 P^\perp$ (see G.H. Golub, C.F. van Loan, 1983, p.139), denoted by $\left(P^\perp\Sigma_1 P^\perp\right)^+$. Hence one can also write:

$$X^2=r^t\left(P^\perp\Sigma_1 P^\perp\right)^+ r$$

Under the null-hypohesis, this random variable follows a chi-square distribution on $n-q$ degrees of freedom. This is a generalization of a well-known result on quadratic form (see C. R. Rao, 1967, p.493). The expected value is easily found to be

$$E\,X^2=Trace\left(P^\perp\Sigma_0 P^\perp\left(P^\perp\Sigma_1 P^\perp\right)^+\right).$$

Under the null hypothesis $P^\perp\Sigma_0 P^\perp\left(P^\perp\Sigma_1 P^\perp\right)^+$ is the orthogonal projection onto $Range\left(P^\perp\Sigma_0 P^\perp\right)$ (see G.H. Golub, C.F. van Loan,

1983, p.139) so that $EX^2 = n-q$ also holds in the non-gaussian case.

Let us emphasize that X^2 is not a goodness-of-fit test, i.e. it would be wrong to perform this test with an estimate of the covariance matrix and to infer from a non-significant value that the point estimate is compatible with the data (within the model) or that the parametric model is correct. As a matter of fact, the chi-square so obtained will be exactly $n-q$ at the restricted maximum likelihood estimate, no matter what model is fitted (see the mathematical appendix). On the other hand, assuming the model structure to be essentially correct (e.g. a spherical covariance), then one can easily determine an interval of ranges compatible with the data for a given sill.

We now go back to the fundamental indetermination of the model. The observable empirical residual random vector R is linked to the true but unobservable residual vector ε by the relation $R = P^{\perp}\varepsilon$. This equation for ε is consistent as $R \in P^{\perp}(R^n)$; the general solution is given by

$$\tilde{\varepsilon} = \left(P^{\perp}\right)^{+} R + \left(I_n - \left(P^{\perp}\right)^{+} P^{\perp}\right)\omega \qquad (5.60)$$

where $\omega \in R^n$ is arbitrary and A^+ denotes the Moore-Penrose generalized inverse of any matrix A; furthermore A^+A is the projection operator onto $Range(A^t)$ (G.H Golub, C.F. van Loan, 1983); now, $R = P^{\perp}Z$, $\left(I_n - \left(P^{\perp}\right)^{+} P^{\perp}\right)$ is the projection operator onto $Range(F)$ and projectors are symmetric, so that one finally obtains

$$\tilde{\varepsilon} = R + F\tilde{\beta} \qquad (5.61)$$

where $\tilde{\beta}$ is an arbitrary random vector with zero expectation (R has by definition of ε, and $\tilde{\varepsilon}$, as the corresponding model must also be correct).

In otherwords, for a given observed empirical residual vector r, there exist infinitely many compatible realizations $\tilde{e}$ of the underlying true but unobservable residual process. For a given version ε_1, the process defined by

$$\varepsilon_2 = \varepsilon_1 + F\tilde{\beta}$$

where $\tilde{\beta} \in R^p$ is a random vector with zero expectation and a given covariance matrix Σ_β, is equivalent from the point of view of statistical inference; note that since ω was arbi-

trary, one can assume without loss of generality that $\tilde{\beta}$ is independent of ε_1. Thus, the fundamental indetermination suggested earlier, purely heuristically, is in fact the only one. From (5.61) one easily obtains the following relations between the various covariances used in universal kriging:

(5.62)

$$\Sigma_2 = \Sigma_1 + F\Sigma_{\tilde{\beta}}F'$$
$$\sigma^2_{n,0,2} = \sigma^2_{n,o,1} + F\Sigma_{\tilde{\beta}}F_o'$$
$$\sigma^2_{0,2} = \sigma^2_{o,1} + F_o\Sigma_{\tilde{\beta}}F_o'$$

where $\Sigma_{\tilde{\beta}} = E\tilde{\beta}\tilde{\beta}'$ and the indices 1,2 refer to the versions. Let us denote by λ_1,μ_1 and λ_2,μ_2 the solutions of the kriging system (5.49) with respect to versions 1 and 2. Using equation (5.61), it is easily verified that the solutions and the expected mean square errors are equal, so that the indetermination is irrelevant for kriging. Likewise, the Frobenius (euclidian) norm $\|rr' - P^{\perp}\Sigma P^{\perp}\|$, used for estimating the covariance matrix, is independent of the version chosen, since $P^{\perp}F = 0$.

This result is in complete analogy with the IRF-k theory. The difference being that for arbitrary drifts we have to postulate the existence of a stationary residual process (other, non-stationary processes, would give the same observable quantities, but they are equivalent with respect to kriging), whereas the IRF-k theory actually establishes, under the crucial assumption of polynomial drifts, the existence of a stochastic process with stationary generalized increments and gives the general properties of the underlying structure (generalized covariances). It can be shown that in this case IRF-k kriging is equivalent to an appropriate universal kriging model (R. Christensen, 1990a).

The disadvantages of universal kriging, as compared with double kriging, are that it requires an exhaustive knowledge of the auxiliary information (in practice thematic maps with accurate determination of surface areas) and that the model structure must be correct (zero mean residual). In such a case, one could still get the least square or restricted maximum likelihood estimate of the residual covariance matrix and then use one of the following three methods:

(1) Perform universal kriging at the points $x \in s_1 - s_2$ followed by kriging with measurement errors according to the rule: $\sigma^2(x)$=estimated kriging variance when $x \in s_1 - s_2$, and $\sigma^2(x) = 0$ when $x \in s_2$.

(2) Perform universal kriging at all points $x \in s_1$ followed by double kriging. Note that kriging the residuals yields zero for the point estimate (since universal kriging is an exact interpolator), but not for the mean square error.

(3) Recalculate predictions and residuals with generalized weighted least squares based on the estimated covariance matrix and then perform double kriging.

The second procedure has the advantage that it is simpler than the third (and in most instances numerically very close, if the validity assumptions for convergence are fulfilled), whereas the first underestimates the mean square error as we have seen.

The main advantage of universal kriging as presented here is that it gives a mathematically correct treatment for internal models, bypassing the drift/residual circularity problem, and validity conditions.

The only remaining difficulty (assuming drifts and covariance have been adequately modelled) is to assess the impact of using an estimate of the covariance, instead of the true covariance, on the point estimates and the kriging variance. Analytically, this appears to be an extremely difficult problem; preliminary results indicate that the strength of the spatial correlation (the stronger the better) and the condition index (ratio of the largest to the smallest eigenvalue) of the covariance matrix (the smaller the better) play an important role (P. Diamond, M. Armstrong, 1984; J.J. Warnes, 1986; D. Posa, 1989; M. L. Stein, M.S. Handcock, 1989; D. L. Zimmermann, N. Cressie, 1992).

5.7 Estimation of Ratios

In many applications, one has to estimate quantities of the form:

(5.63)

$$r = r(V_o) = \frac{\int_{V_o} z_1(x)dx}{\int_{V_o} z_2(x)dx} = \frac{z_1(V_o)}{z_2(V_o)}$$

e.g. percentages of trees with some characteristics, or mean timber volume per tree, etc. Note that in most circumstances, it would be wrong to consider quantities like:

$$\frac{1}{\lambda(V_o)}\int_{V_o}\frac{z_1(x)}{z_2(x)}dx \neq r(V_o)$$

as the function $\frac{z_1(x)}{z_2(x)}$ is generally not even additive.

The technique we propose to estimate $r(V_o)$, and particularly its mean square error, differs from the standard procedure as presented in (A.G. Journel, C. Huijbregts, 1978). Instead, it rests upon a geostatistical reformulation of Fieller's theorem (D.R. Cox, 1967), which is more in line with the design-based approach.

For each $\rho \in R^1$ we define the following random variable:

(5.64)

$$\theta(\rho) = \frac{1}{\lambda(V_o)}\int_{V_o}\left(z_1(x) - \rho z_2(x)\right)dx$$

Note that from the definitions, one has :

(5.65)

$$\theta(r(V_o)) \equiv 0$$

Let $\theta^*(\rho)$ be an estimate of $\theta(\rho)$; this estimate can be obtained by any of the techniques described so far, for instance double kriging. For each given ρ, one needs to determine the variograms of the "predictions" $\hat{z}_1(x) - \rho\hat{z}_2(x)$ and of the "residuals" $\varepsilon_1(x) - \rho\varepsilon_2(x)$. This can be done either by direct model fitting of the empirical variograms of these two new processes or by using pre-existing models of variograms and cross-variograms, and the relation $\gamma_{U-\rho V}(h) = \gamma_U(h) + \rho^2\gamma_V(h) - 2\rho\gamma_{U,V}(h)$ for any two processes U, V. Obviously, the first technique is more time consuming, but also more robust and instructive, since it models the relevant variogram directly.

The estimate $r^* = r^*(V_o)$ of the ratio $r(V_o)$ is defined, because of (5.65), as the solution of the equation

$$\theta^*(\rho) = 0, \text{ i.e. } \theta^*(r^*) = 0 \tag{5.66}$$

The procedure is iterative, namely:

$$r_1^* = \frac{z_1^*(V_o)}{z_2^*(V_o)}$$
$$r_{n+1}^* = r_n^* + \frac{\theta^*(r_n^*)}{z_2^*(V_o)} \tag{5.67}$$

where $z_1^*(V_o), z_2^*(V_o)$ are the double kriging estimates of the true values $z_1(V_o), z_2(V_o)$. The scheme (5.67) is based on the following argument:

Set $r_{n+1}^* = r_n^* + \varepsilon_n$ in (5.64) to obtain $\theta(r_{n+1}^*) = \theta(r_n^*) - \varepsilon_n z_2(V_o)$. Because of (5.66), this suggests immediately to set:

$$0 = \theta^*(r_{n+1}^*) \cong \theta(r_n^*) - \varepsilon_n z_2^*(V_o)$$

which is precisely the iteration scheme (5.67).

We assume convergence and define $\lim_n r_n^* = r^*$.

Let $\theta^*(\rho) = \theta(\rho) + e(\rho)$, where $e(\rho)$ is the estimation error satisfying $\mathrm{E}e(\rho) = 0$ and $\mathrm{E}e^2(\rho) = \mathrm{MSE}(\theta^*(\rho))$. Hence, one can write $0 = \theta^*(r^*) = \theta(r^*) + e(r^*)$ and therefore

$$r^*(V_o) - r(V_o) = \frac{e(r^*(V_o))}{z_2(V_o)}$$

so that $r^*(V_o)$ is asymptotically unbiased, with the mean square error approximately given by:

$$\mathrm{E}\left(r^*(V_o) - r(V_o)\right)^2 \cong \frac{\mathrm{MSE}(\theta^*(r^*(V_o))}{\left(z_2^*(V_o)\right)^2} \tag{5.68}$$

in perfect analogy with the design-based approach (D. Mandallaz, 1991). In practice it appears that 0, 1, 2 iterations suffice. It is worth noting that the above procedure also reduces the bias of the first iteration estimate, which is simply the standard estimate. The minor and obvious modifications required when using the other estimation techniques (e.g. ordinary kriging, mixed kriging, universal kriging) are left to the reader.

5.8 Numerical Aspects

(1) Numerical integration

The boundary of the domain of interest V_o must be approximately defined by a set of polygons. It appears that point estimates are not too sensitive with respect to the accuracy of such polygons,whereas the variance can be substantially inflated if large zones of non-forest area are not explicitly declared as such. It is therefore important that the polygons match the overall shape of the "forest/non-forest" zones, the error with respect to surface area being less important. Once the polygons are defined, it is necessary to define a grid to perform the numerical integrations required by the quantities $\bar{\gamma}(x,V_o)$, $\bar{\gamma}(V_o,V_o)$. The discretization mesh should be fine enough to stabilize the numerical values, also under translations of the grid. Comparisons of different kriging methods are valid only with a common underlying grid satisfying the above conditions for each method. In any case,the numerical uncertainty should be much smaller than the kriging error.

(2) Kriging

For numerical reasons, it is preferable to keep kriging neighbourhoods under 300 points. For large areas it may therefore be necessary to work with roughly equally large sub-domains with less than 300 points. Each subdomain $V_i, i = 1,2...L$ is kriged with its interior data points only (according to any of the techniques described in the sections 5.1-5.7). The point estimates $z^*(V_i)$ and the mean square errors MSE_i are then combined into a global estimate for

$$V_o = \bigcup_{i=1}^{L} V_i$$

according to

$$z^*(V_o) = \sum_{i=1}^{L} \frac{\lambda(V_i)}{\lambda(V_o)} z^*(V_i) \quad , \quad \mathrm{MSE}_o \equiv \sum_{i=1}^{L} \left(\frac{\lambda(V_i)}{\lambda(V_o)} \right)^2 \mathrm{MSE}_i$$

the overall point estimate is obviously also unbiased. The formula for the overall mean square error neglects the correlation between neighbouring points in two adjacent subdomains.

Nevertheless, it is an excellent approximation if the individual kriging estimates are based on more than 50 points.

For the construction of kriging maps, one can use gliding neighbourhoods instead of a unique neighbourhood. With external drifts, one must choose neighbourhoods large enough to ensure the compatibility condition of universal kriging. In our opinion, if one is willing to accept the underlying stationarity hypothesis or cannot reject it, there are no stringent reasons for using moving neighbourhoods, except the size of the kriging matrix and the numerical stability of its inversion. Besides, fitting variograms or covariances is usually done globally. In any case, it is wise to investigate the impact of the kriging neighbourhoods on the point and error estimates, as it appears difficult to give general guidelines valid a priori.

(3) Software

The calculations required for the case study presented in chapter 7 were performed either with the software SAS (SAS Institute, 1987) on an IBM 3090 for the "classical" statistical parts and the linear algebra calculations, or with the software BLUEPACK (BlUEPACK, 1990), on a VAX 9000-420 for the geostatistical parts.

6 The Design-based Approach.

For easier reference, we briefly outline the main results of the design-based approach as needed to understand the case study of chapter 7. Details can be found in (D. Mandallaz, 1991).

The prediction process $\hat{Z}(x)$ and the residual process $\varepsilon(x) = Z(x) - \hat{Z}(x)$ are observed at points arranged in clusters. Each cluster is defined by a set of M points x_l according to:

$$x_l = x + e_l, e_l \in R^2, l = 1,2...M \qquad (6.1)$$

The random point x is uniformly distributed in a domain $A \supset V$, such that $\Pr(x_l \in V) \neq 0, \forall l$. The number of points of a cluster falling into the forest area is therefore also a random variable $M(x) = \sum_{l=1}^{M} I_V(x_l)$. To simplify the notation, we redefine the indices so that the points of the cluster lying in the forest area are $x_l, l = 1,2...M(x)$. We now define the mean of the processes at the cluster level according to:

$$z(x) = \frac{\sum_{l=1}^{M(x)} z(x_l)}{M(x)}, \hat{z}(x) = \frac{\sum_{l=1}^{M(x)} \hat{z}(x_l)}{M(x)}, e(x) = \frac{\sum_{l=1}^{M(x)} e(x_l)}{M(x)} \qquad (6.2)$$

(The realization of $\varepsilon(.)$ is denoted by $e(.)$).

The design-based regression estimate is given by

$$\hat{z}_{reg} = \frac{\sum_{x \in s_1} M(x)\hat{z}(x)}{\sum_{x \in s_1} M(x)} + \frac{\sum_{x \in s_2} M(x)e(x)}{\sum_{x \in s2} M(x)} \qquad (6.3)$$

which is simply the sum of the overall mean of predictions and residuals ignoring the cluster structure; it obviously belongs to the class of double kriging estimates, the weights being equal to constants for the predictions and the residuals. The reason for writing the regression estimate in the form (6.3) is to highlight the role of the weights $M(x)$, particularly in the variance, as shown below. The large sample s_1 consists of n_1 clusters, whose centres x are independently uniformly distributed in A, whereas the small sample $s_2 \subset s_1$ consists of n_2

clusters chosen in the large sample according to equal probability sampling without replacement. The regression estimate (6.3) is design-unbiased for external models, otherwise only asymptotically. Note that in contrast to sections 5.1-5.7, n_1, n_2 are the number of clusters and not the total number of points. The design-based variance can be estimated according to

$$\hat{\text{var}}\,\hat{z}_{reg} = \left(1-\frac{n_2}{n_1}\right)\frac{1}{n_2}\frac{1}{n_2-1}\sum_{x\in s_2}\left(\frac{M(x)}{\bar{M}_2}\right)^2\left(e(x)-\bar{e}_2\right)^2 + \frac{1}{n_1}\frac{1}{n_2-1}\sum_{x\in s_2}\left(\frac{M(x)}{\bar{M}_2}\right)^2\left(z(x)-\bar{z}_2\right)^2 \qquad \textbf{(6.4)}$$

where

$$\bar{z}_2 = \frac{\sum_{x\in s_2} M(x)z(x)}{\sum_{x\in s_2} M(x)} \quad ,\bar{e}_2 = \frac{\sum_{x\in s_2} M(x)e(x)}{\sum_{x\in s_2} M(x)} \quad ,\bar{M}_2 = \frac{\sum_{x\in s_2} M(x)}{n_2}$$

When estimating domains $V_o \subset V$, the above formulae remain valid after restriction to $s_i \cap V_o$. Similar formulae are available for ratios.

It is obvious but worth noting that one-phase or two-phase simple random sampling are special cases with

$$n_2 = n_1 \quad \text{or} \quad M(x) \equiv 1$$

If the centres of the cluster are lying on the nodes of a systematic grid (with random start and possibly also random orientation), the regression estimate is still asymptotically design-unbiased but the variance formula (6.3) is no longer valid (the entire sample can be considered as a single large cluster and no estimate of variance can be based on a single realization, which is precisely the dilemma mentionned in the introduction). However, in practice one nearly always assumes (as we shall do in the case study of chapter 7) that systematic samples can be regarded as random samples, so that (6.3) and (6.4) can be used; this postulate corresponds in some sense to the intrinsic or stationarity hypothesis in geostatistics.

7. Case Study.

7.1 Generalities

The major objective of this case study is to illustrate the theory and give a first empirical validation of the techniques described in chapters 5 and 6. The material of this case study has also been used for an extensive evaluation of various design-base techniques, as well as model-based and model-dependent ones (D. Mandallaz, 1991).

7.2 Material

The data for this case study rests upon an intensive inventory carried out in parts of the Zürichberg Forest belonging to the city and the Canton of Zürich. This forest is typical of the Swiss Plateau Forest, though its recreational purpose and its accessibility are above the average.

In what follows, the regeneration areas do not belong, by definition, to the forest area.

The inventoried forest covers 217.92 ha, of which 17.07 ha served for a full census with accurate determination of the tree co-ordinates. Aerial infrared photographs are also available (scales 1:9000 and 1:3000) as well as a stand map and other thematic maps.

Fig. 1 displays the stand map, together with the polygon defining the small area with full census and the location of the plots of the terrestrial inventory. Enclaves of non-forest or regeneration areas are blank. The stand map is constructed by interpretation of aerial photographs with control on the ground. Data management tasks (digitization, updates, overlay, drawing, etc.) were perfomed with the G.I.S. Arcinfo on a VAX Computer. The stand map is based on 3 qualitative variables, namely:

(1) Developmental stage

This stand attribute is defined by the dominant diameter d_{dom}, the average of the 100 trees per ha with the biggest diameter at breast height. Though the definition is quantitative it must be emphasized that the assessment is done by "expert judgement".

In the present case we have 4 categories:

'**3**' pole stage	$12 \leq d_{dom} < 20$ cm
'**4**' young timber tree	$20 \leq d_{dom} < 35$ cm
'**5**' middle age timber tree	$35 \leq d_{dom} < 50$ cm
'**6**' old timber tree	$50 \leq d_{dom}$.

(2) Degree of mixture

In the present case this attribute has 2 categories:

'**1**' predominantly conifers
'**2**' predominantly broadleaves.

(3) Crown closure

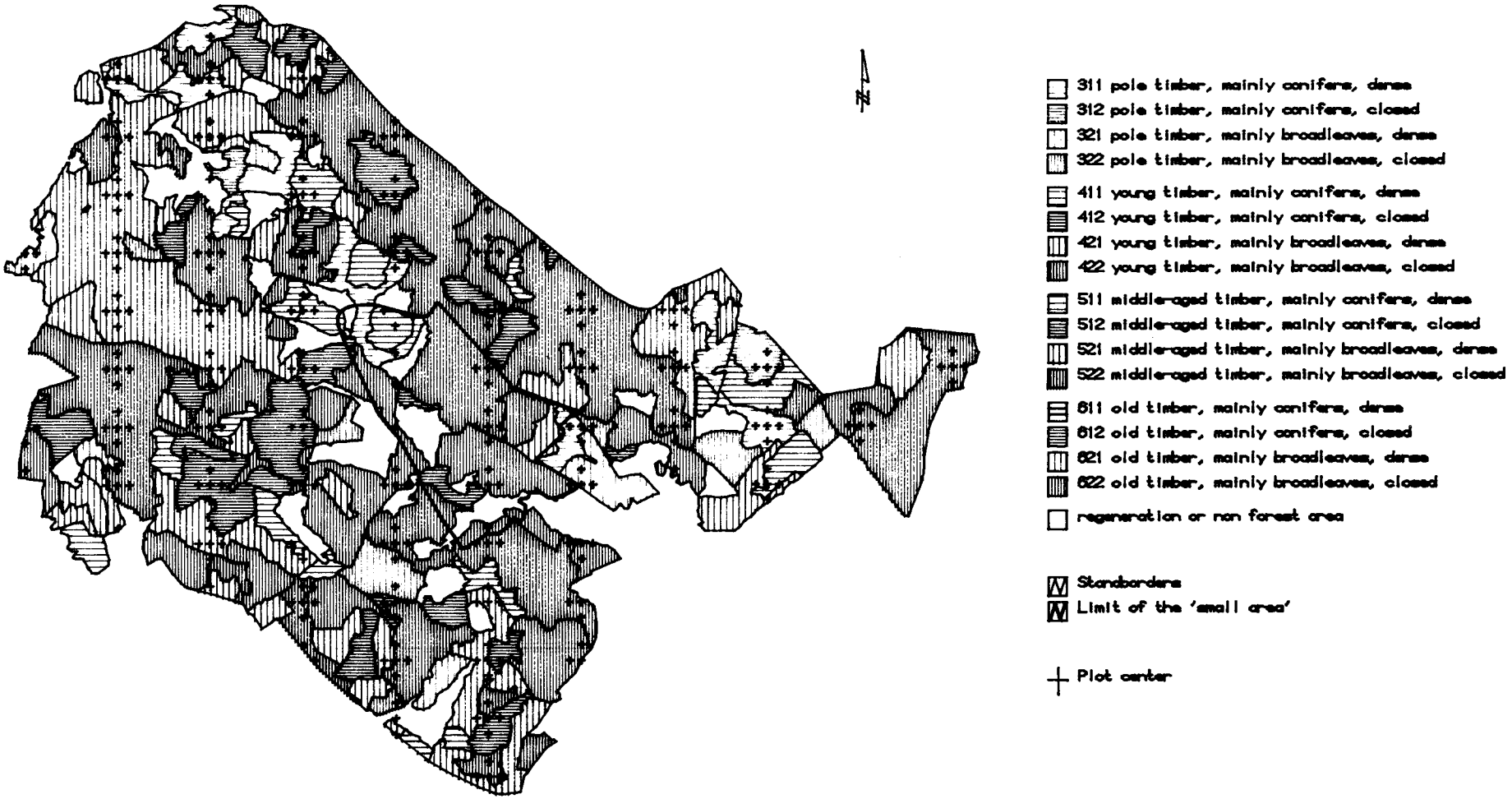
This stand attribute is based on the canopy density, defined as the proportion of total ground surface of the stand to the ground surface covered by the trees crowns.

In the present case we have only 2 categories:

'**1**' dense	canopy density > 0.9
'**2**' closed	0.6 ≤ canopy density < 0.9

The 16 possible stand structures actually all occur and are given in Fig. 1. Fig 2 displays the small domain with its 4784 trees. Table 1 gives the absolute and relative surface areas according to the different categories of the 3 stand attributes.

Fig. 1: Stand Map of the Zürichberg Forest



Seite Leer /
Blank leaf

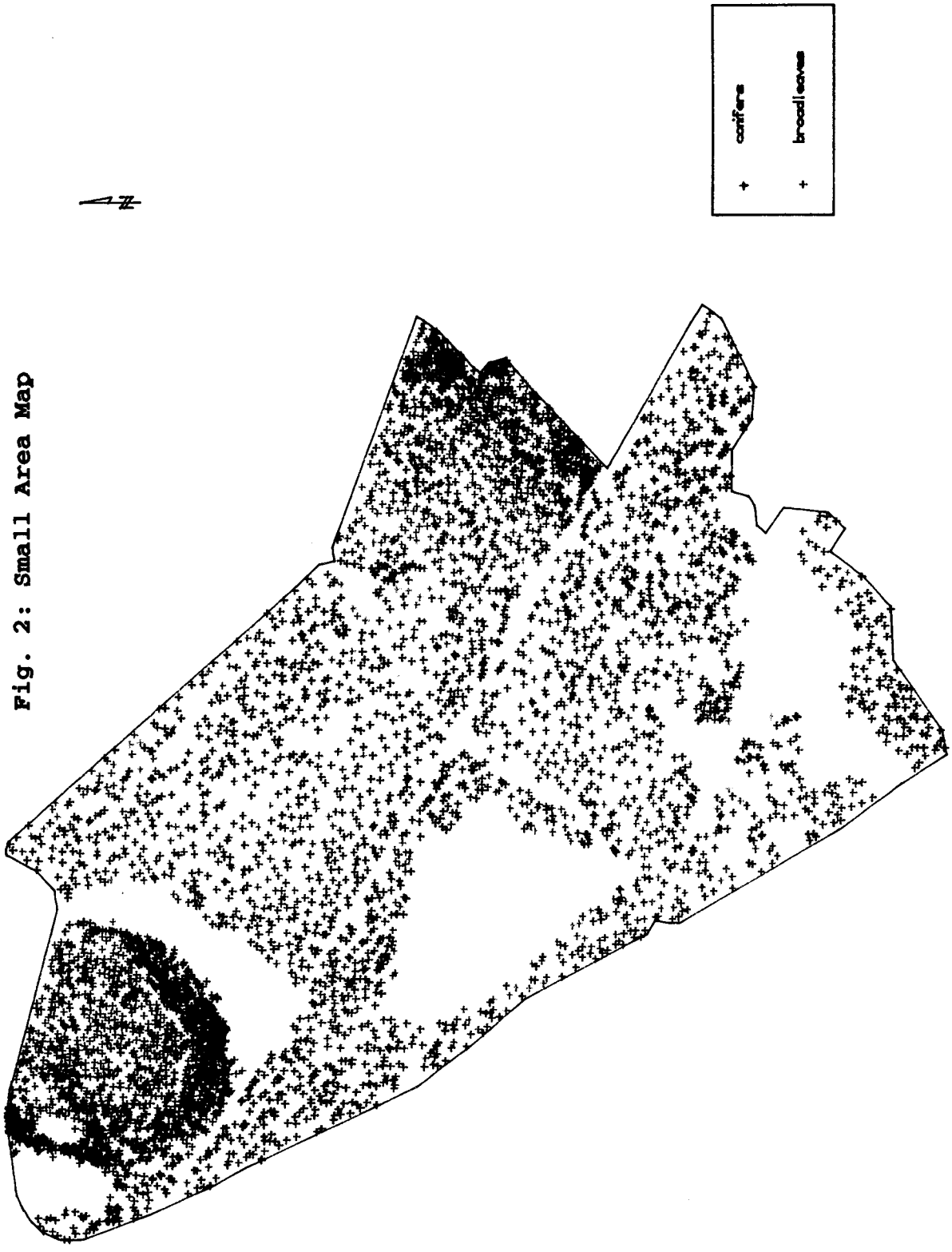


Fig. 2: Small Area Map

Seite Leer /
Blank leaf

Table 1: Absolute and relative surface areas
(without regeneration areas)

Variables	Entire Domain		Small Area	
	ha	%	ha	%
Develop. Stage				
3	24.93	11.4	1.27	7.5
4	25.87	11.9	2.99	17.5
5	134.73	61.8	7.32	42.9
6	32.38	14.9	5.49	32.1
Total	217.91	100.0	17.07	100.0
Degr. Mixture				
1	50.83	23.3	3.56	20.9
2	167.08	76.7	13.51	79.1
Total	217.91	100.0	17.07	100.0
Crown Closure				
1	84.71	38.9	3.27	19.1
2	133.20	61.1	13.80	80.9
Total	217.91	100.0	17.07	100.0

7.3 Inventory Methods

i) Sampling Scheme for Auxiliary Information

The sampling procedure rests upon a 5 points cluster: from the central point 2 points are taken 30 m in the W-E directions and another 2 points 40 m away in the N-S directions.

The first 1st phase procedure sets the central cluster point on a 120 m (W-E) by 75 m (N-S) systematic rectangular grid, which yields a nominal density of 5.56 points per ha, or 1 point per 0.18 ha (note that the clusters partially overlap in the N-S direction). The Geographical Information System provides for each point the stand map information.

The second 1st phase procedure does not actually draw sample points but gives instead the true means of the

auxiliary variables via the exact surface areas of the different stands; it is used only for universal kriging with external drifts.

ii) Sampling Scheme for the Terresterial Inventory

The 2nd phase sampling is based on a 5 plots cluster with the same geometrical structure as above. The plots have a horizontal surface area of 300 m^2 (9.77 m radius). Diameter at breast height (DBH in cm), species, crownclass, state of health and other qualitative variables are recorded on each tree in the plot whose diameter is above 12 cm. This sampling scheme sets the central plot on a 1:4 parallel subgrid of the 1st phase grid, i.e. on a 240 m (W-E) by 150 m (N-S) systematic rectangular grid with a nominal density of 1.39 plot per ha (1 plot per 0.72 ha).

7.4 Inventory Data

Table 2 displays the absolute and relative frequencies of the tree species drawn.

Table 2: Tree Data

Species	**Entire Domain** (samples) Frequencies		**Small Area** (census) Frequencies	
	absolute	relative	absolute	relative
Norway Spruce	807	28.1	1874	39.2
Silver Fire	76	2.6	147	3.1
Larch	185	6.6	119	2.4
other coniferous	76	2.6	24	0.5
Total coniferous	1144	39.9	2164	45.2
Beech	1164	40.6	1617	33.8
Norway maple	17	0.6	245	5.1
Sycamore maple	106	3.7	214	4.5
Ash	180	6.2	244	5.1
Elm	50	1.7	173	3.6
other broadleaves	209	7.3	127	2.7
Total broadleaves	1726	60.1	2620	54.8
T o t a l	2870	100.0	4784	100.0

Table 3 below gives the essential features of the actually observed sizes (at the cluster and plot levels), for the points or plots falling into the forest area. A point is "in", if its nominal co-ordinates are 'in' with respect to the relevant polygons of the stand map.

Likewise for a plot according to the nominal co-ordinates of its centre. For all but one plot no boundary problems ocurred; because of the inherent inaccuracies in the stand map no adjustment was performed

Table 3: Observed Sample Sizes

Samples	Entire Domain				Small Area			
	n_c	n_p	$\overline{M}$	V	n_c	n_p	$\overline{M}$	V
1st phase	298	1203	4.04	1.92	29	92	3.17	2.37
2nd phase	73	298	4.08	1.99	8	19	2.38	2.27

Legend:
n_c: number of clusters with at least one point in the forest area
n_p: total number of points in the forest area
$\overline{M}$: mean number of points per cluster in the forest area
V : variance of the number of points per cluster in the forest area.

7.5 Prediction Model

The stand map information can be fully characterized by the following 9-dimensional vector

$X(\omega)=(X_i(\omega) \quad i=1,2,.....9)$, where

$X_1(\omega)\equiv 1$, intercept term
for $i=2,3,4,5$ one sets
$X_i(\omega)=1$ if ω lies in development stage $i+1$, 0 otherwise
$X_6(\omega)=1$ if ω lies in a coniferous stand , 0 otherwise
$X_7(\omega)=1$ if ω lies in a broadleaved stand , 0 otherwise
$X_8(\omega)=1$ if ω lies in a dense stand , 0 otherwise
$X_9(\omega)=1$ if ω lies in a normal stand , 0 otherwise

This model has rank 6, which requires the use of generalized inverses. Alternatively, one could work with a regular design of rank 6 by setting, for example
$X_1(\omega)\equiv 1$
$X_2(\omega)=1$ if ω lies in development stage 3 , 0 otherwise
$X_3(\omega)=1$ if ω lies in development stage 4 , 0 otherwise
$X_4(\omega)=1$ if ω lies in development stage 5 , 0 otherwise
$X_2(\omega)=X_3(\omega)=X_4(\omega)=-1$ if ω lies in development stage 6

$X_5(\omega) = 1$ if ω lies in a coniferous stand , -1 otherwise
$X_6(\omega) = 1$ if ω lies in a dense stand , -1 otherwise

The model is of the form

$$Y(\omega) = X'(\omega)\beta + \varepsilon(\omega)$$

i.e. the response variable (stem or basal area density) follows a simple analysis of variance type model, without interactions. This model needs, in the classical sense, 6 degrees of freedom, whereas simple stratification with respect to the stand map needs 16 parameters.

The standard least square estimates of the regression coefficients for the regular design were:

$(411.96^{*}, 291.62^{*}, 19.86, -126.27^{*}, 11.22, 32.54^{*})$ for stem density

$(30.44^{*}, -11.61^{*}, -0.18, 7.13^{*}, 4.25, 1.96^{*})$ for basal area.

The asteriks denoting a 5% level significant value according to the standard tests (i.e. assuming independence and normality). The magnitudes and signs of the coefficients reflect common practical knowledge on forest stands.

According to Shapiro-Wilks test (which assumes independence), neither the raw observations nor the residuals of the basal area depart significantly from normality; on the other hand, the raw observations of the stem density depart significantly from normality and, to a lesser degree, also from lognormality; similarly, the residuals on the original scale and on the log-scale (i.e. using the same model structure on log-stem density) depart significantly from the normal distribution, though, optically, less so than the raw observations (see also Fig. 3 and 4 in section 7.6).

The auxiliary information is assumed to be error free. This implies that the nominal co-ordinates of a plot or a point are the same as the actual co-ordinates (no location errors), and that the stand map polygons are exact. Furthermore, the auxiliary information for plots at stand boundaries is determined by the plot centre only. In such cases, the ***0-1*** indicator variables could be allowed to take fractional values, a technique which is not presented here as it did not significantly improve the results.

To illustrate the estimation technique for ratios, we also consider the percentage of non-healthy trees (a tree being healthy if its apparent foliage loss is less or equal 10 %). To this end, we use a slightly different model. First, a logistic model with the same explanatory variables as above is fitted to the observed percentages in the terrestrial plots. The predicted number of non-healthy trees in any point is then obtained by multiplying the predicted number of stems by the predicted percentage at this point; this procedure is better than a direct fitting of the number of non-healthy trees. For details on the state of health and the logistic model see D. Mandallaz et al (1986).

The multiple coefficients of determination r^2 were 0.5 for stem density, 0.2 for basal area and 0.15 for the percentage of non-healthy trees, in agreement with previous experiences.

7.6 Variography

The figures 3 and 4 display the histograms of the stem and basal area densities observed in the 298 terrestrial plots.

Fig. 3: Histogram of the observed stem density.

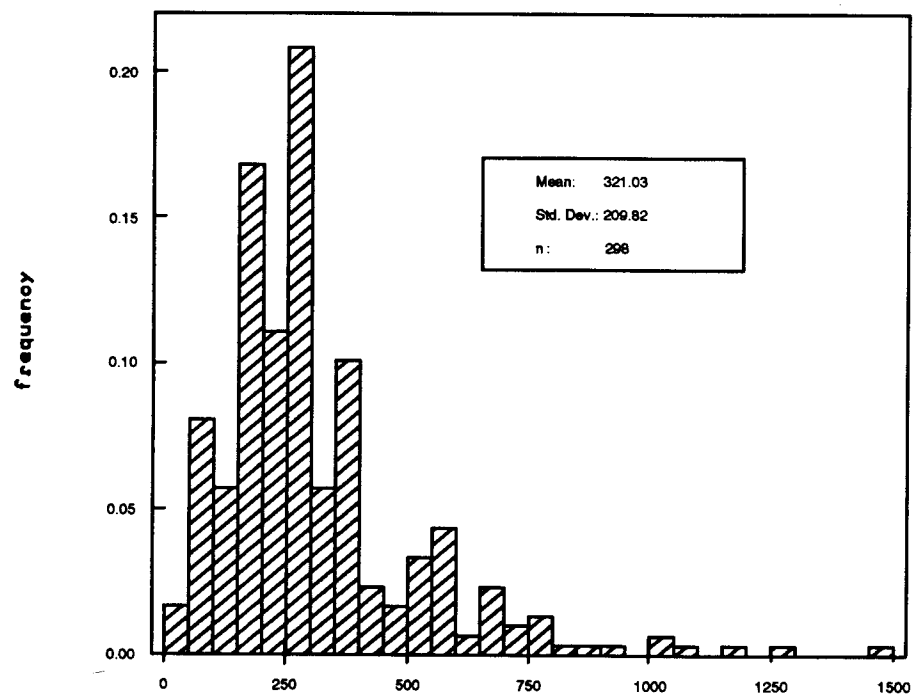
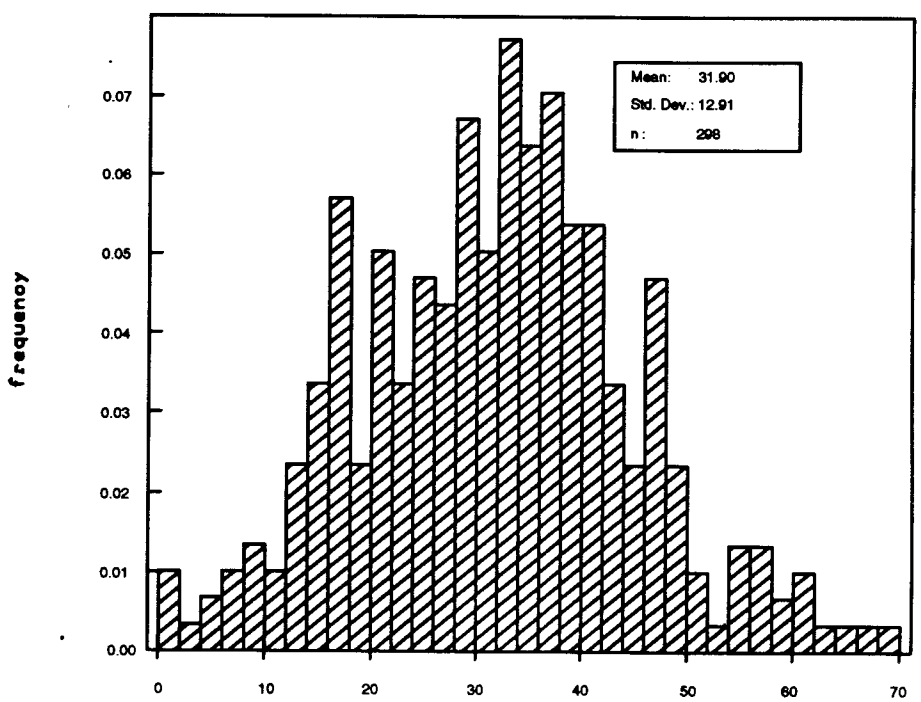
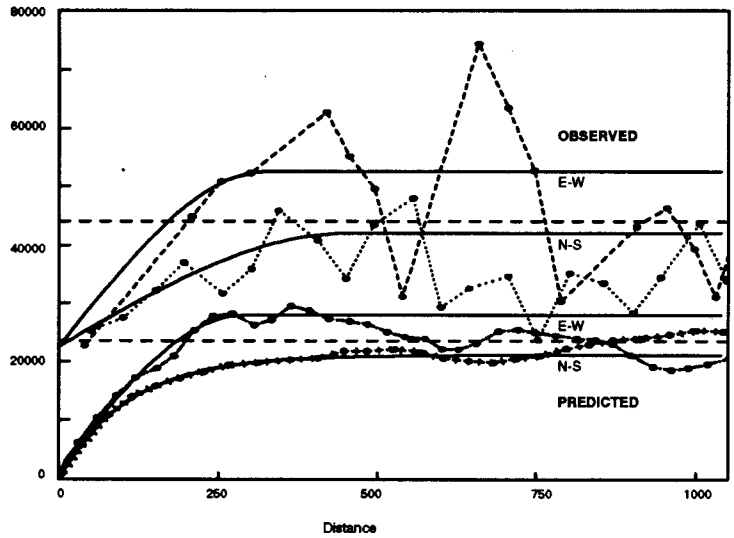


Fig. 4: Histogram of the observed basal area density (m^2/ha).



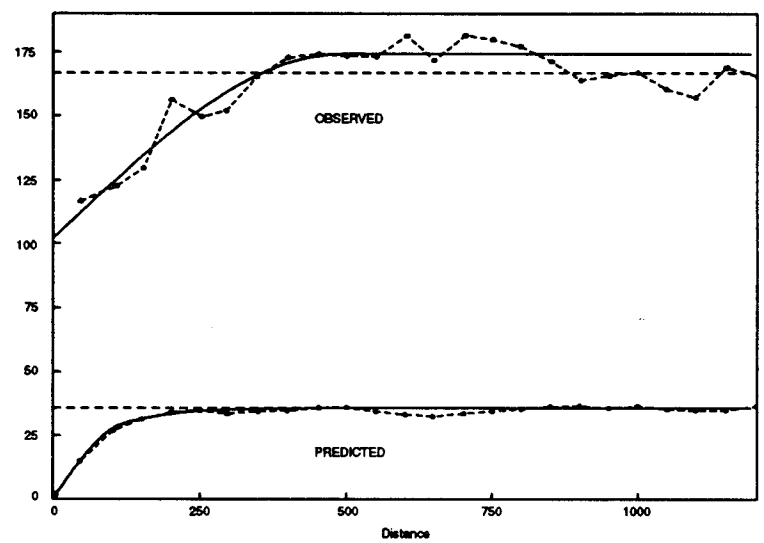
A selection of the variograms required for the various estimation problems is given in the figures 5 to 9; the fitted variograms are the sum of elementary variograms in the sense of (4.18). The interactive "fitting by eye" technique of BLUEPACK was used throughout.

Fig. 5: Variograms of the observed and predicted stem densities



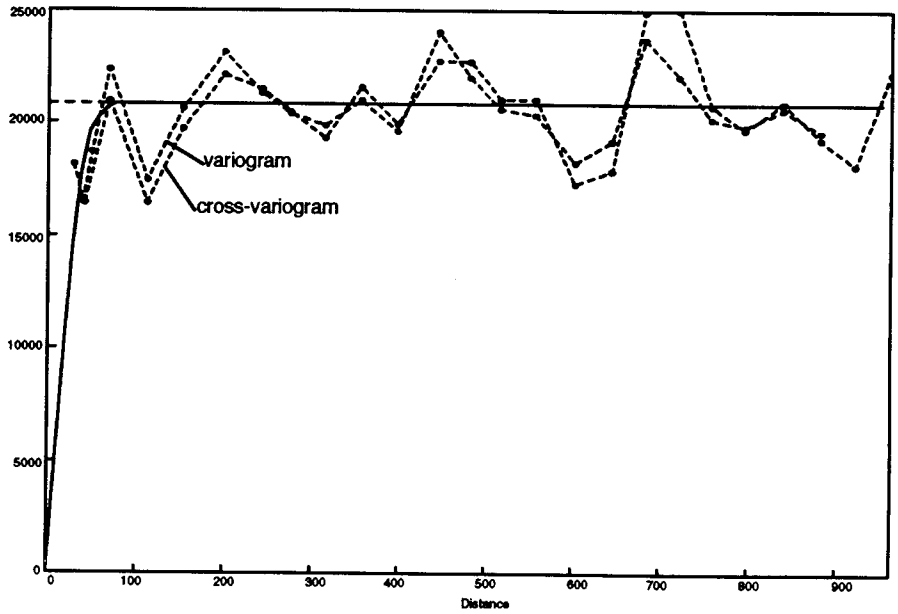
OBSERVED:					PREDICTED				
Nugget:	22500								
	Sill	**Range**	**Anisotropy**			**Sill**	**Range**	**Anisotropy**	
			X	**Y**				**X**	**Y**
Spherical:	19500	325.00	1.00	0.72	**Spherical:**	26000	290.00	1.00	0.00
Spherical:	10500	325.00	1.00	0.00	**Exponential:**	21080	110.00	0.00	1.00
					Spherical:	1900	0.10	1.00	0.00

Fig. 6: Variograms of the observed and predicted basal area densities.



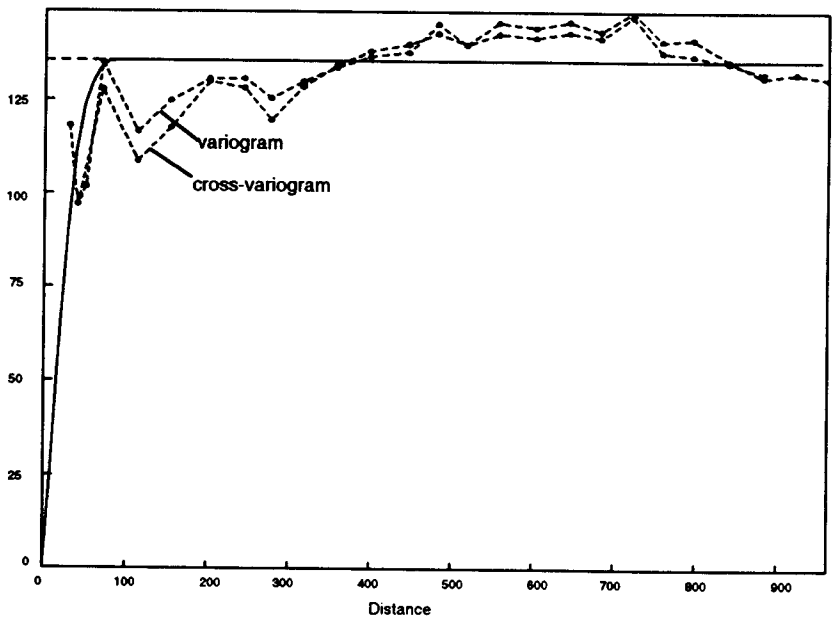
OBSERVED:		PREDICTED:			
Nugget:	102.00				
	Sill	**Range**		**Sill**	**Range**
Spherical:	72.00	490.00	**Spherical**	22.50	125.00
			Spherical	10.00	275.00
			Spherical	3.03	475.00

Fig. 7: Variogram of residuals and cross-variogram observation-residual for the stem density.



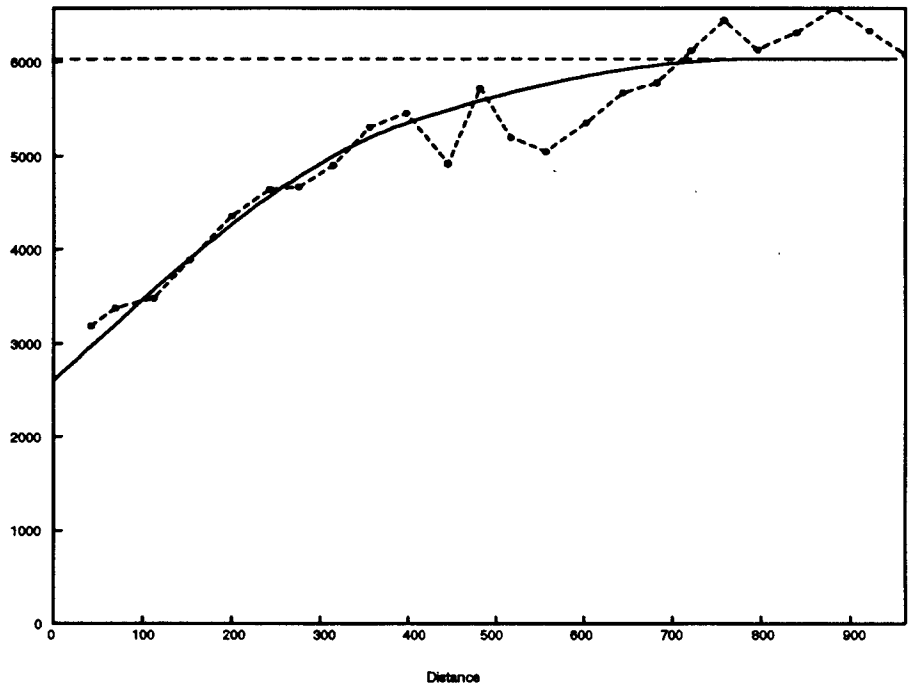
	Sill	Range
Spherical:	15000	50.00
Spherical:	5821	80.00

Fig. 8: Variogram of residuals and cross-variogram observations-residuals for the basal area density.



	Sill	Range
Spherical:	80.00	50.00
Spherical:	55.40	80.00

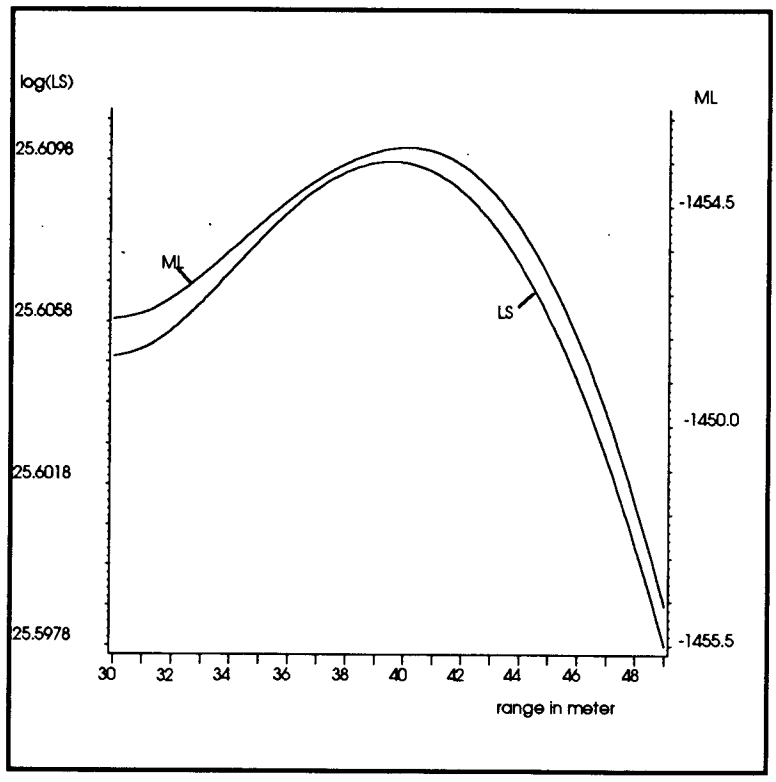
Fig. 9: Auxiliary variogram for the estimation of the percentage of non-healthy trees in the entire domain by ordinary kriging.



		Sill	Range
Nugget:	2600		
Spherical:		1250	400.00
Spherical:		2176	800.00

Fig. 10 and 11 below display the results obtained by the least square **(LS)** and restricted maximum likelihood **(ML)** techniques, as required for the estimation of the residual covariances in universal kriging (details are given in sections 9.2-9.3). The model used is a simple isotropic spherical covariance without nugget effect. Both techniques yield shorter ranges than the direct fitting of the empirical variograms given in Fig. 7 and 8. Similarly, Fig. 12 and 13 displays the results obtained by the **LS** and **ML** techniques for the observed basal area; the model used is a spherical isotropic covariance with nugget effect.

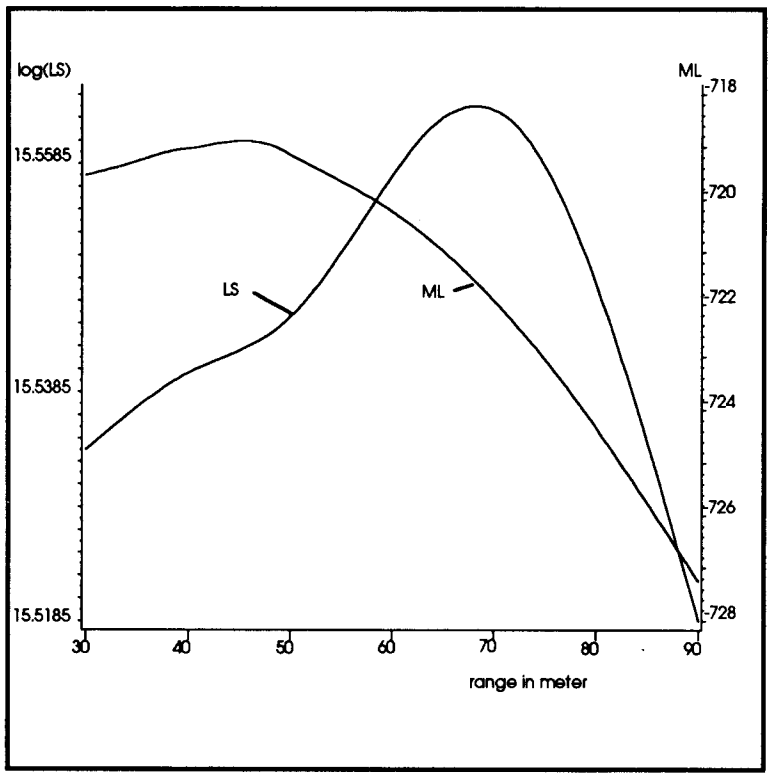
Fig. 10: Least square and maximum likelihood curves of the isotropic spherical model for the residual stem density



Point estimates (standard error)

	LS	ML
Range:	39.8m (5.9m)	40.2m (5.8m)
Sill:	21267 (1766)	21280 (1767)

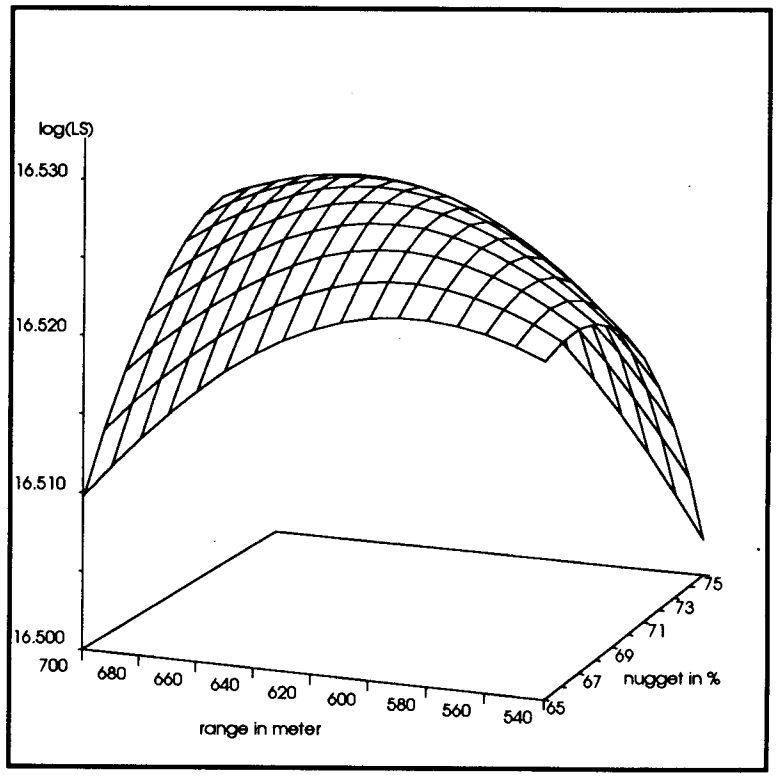
Fig. 11: Least square and maximum likelihood curves of the isotropic spherical model for the residual basal area



Point estimates (standard errors)

	LS	ML
Range:	68.2m (6.6m)	45.5m (5.1m)
Sill:	131.5 (12.2)	139.1 (11.6)

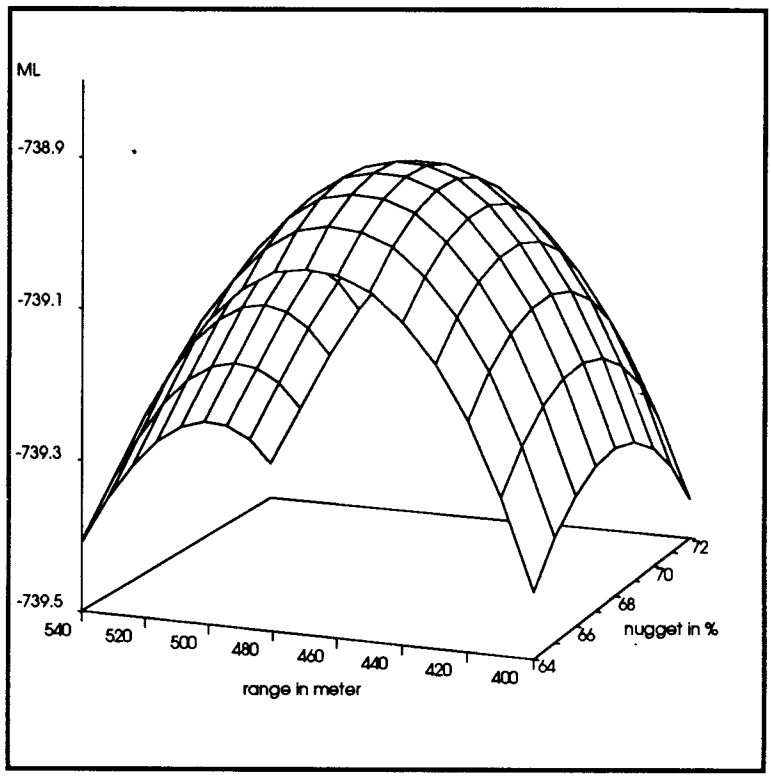
Fig. 12: Least square surface for the isotropic spherical model with nugget effect of the observed basal area



Point estimate (standard error)

Nugget effect:	70.3%	(10.5%)
Range:	633m	(196m)
Sill:	170.9	(20.7)

Fig. 13: Likelihood surface of the isotropic spherical model with nugget effect for the observed basal area



Point estimates (standard errors)

Nugget effect:	**67.6%**	**(8.2%)**
Range:	**466m**	**(77m)**
Sill:	**168.5**	**(17.8)**

Discussion:

1. Fig.3 may suggest a log-transformation for the stem density and, consequently, to apply log-normal kriging for estimation (see N. Cressie, 1991, p. 135) by ordinary kriging (generalization to double, mixed and universal kriging being obvious); whereas this technique is straightforward for punctual estimation it is computationally prohibitive for domain estimation (it requires in principle the integration of the backtransformed punctual estimates). Furthermore, as already mentionned in section 7.5, the log-normal distribution is not adequate either. For double and universal

kriging, the contribution of the residual to the mean square error is predominant: as we have seen, the residuals are closer to the normal than the raw observations, though, again neither the original scale nor the log-scale are really satisfactory. For these reasons, log-kriging was not performed; section 7.7 shows that the improvement, if at all, would have been irrelevant practically.

2.

The variograms of the predictions are continuous at the origin, which reflects the regular stand structure, and have ranges between 250m and 500m corresponding roughly to the average stand dimensions; stem density displays a slight anisotropy north-south / east-west. As a first approximation, the variograms of the observed densities are the sums of the prediction variograms and pure nugget effects.

3.

The empirical variogram of the observed basal area density gives a range $\cong 490m$ and a nugget effect of $\cong 60\%$, in good agreement with the **ML** estimates (466m); the **LS** range is much larger (633m), though not significantly different because of its substantial standard error (196m) (see Fig. 6, 12, 13).

4.

The variograms and cross-variograms of the residuals of the stem density show that the validity conditions (5.37) for double kriging are fulfilled. The empirical variogram strongly suggests a short range, which was set somewhat arbitrarily at 80m (i.e. the largest distance between 2 plots in the same cluster). The **LS** technique yields a spherical variogram with range 39.8m and sill 21'267, in perfect agreement with **ML** estimates; this is somewhat surprising as the distribution of the residuals is closer to a lognormal than a normal.

5.

The cross-variogram for the residuals of the basal area show that the validity conditions (5.37) for double kriging are also fulfilled. At first sight, the empirical variogram could suggest either a short range, or a range around 500m. In analogy with the stem density and because the prediction variogram itself has a range of 475m, the range was set at 80m. The **LS** range (68m) is larger than the **ML** range (45m), but the

difference is only borderline significant, which is also somewhat surprising since here the distribution of the residuals can be assumed to be normal. It is worth noting that, in this case, interactive "fitting by eye" used alone could be misleading.

6. The **ML** standard errors are smaller than the **LS** standard errors, particularly for the observed basal area range, in agreement with the theoretical results of section 9.3.

7.

It is interesting to note that under the assumption of a pure homogeneous Poisson forest, the correlation range with circular sampling plot of $300m^2$ would be just under 20m. Filtering out the inhomogeneities, approximatively taken into account by the stand map, should therefore lead to similar ranges if the forest departs mildly from the Poisson hypothesis. As the Poisson structure is generally an acceptable first approximation, this provides a further indirect qualitative check on the adequacy of the **LS** and **ML** estimates of the residual stem density range, whose 95% confidence limit is roughly (28m,52m).

8.

The auxiliary variogram for estimating the percentage of non healthy trees (Fig. 9) has a range of 800 m, well above the stand dimensions, which could reflect the long range dependence of forest damage.

7.7 Results

The following abbreviations have been used throughout:

O.K	**O**rdinary **K**riging
M.K	**M**ixed **K**riging
D.K	**D**ouble **K**riging
E.K	**K**riging with uncorrelated Measurement **E**rrors
E.D	Universal Kriging with **E**xternal **D**rifts.
D.B	**D**esign-**B**ased.

For illustration purposes, the kriging map of the stem density is given in Fig. 14; Fig.15 displays the corresponding error map.

Tables 4 and 5 below display the results for the various estimation techniques; the simple empirical variograms given in Fig. 5-9. were used throughout.

Fig. 14: Double Kriging map for stem density
(squares of 100m by 100m)

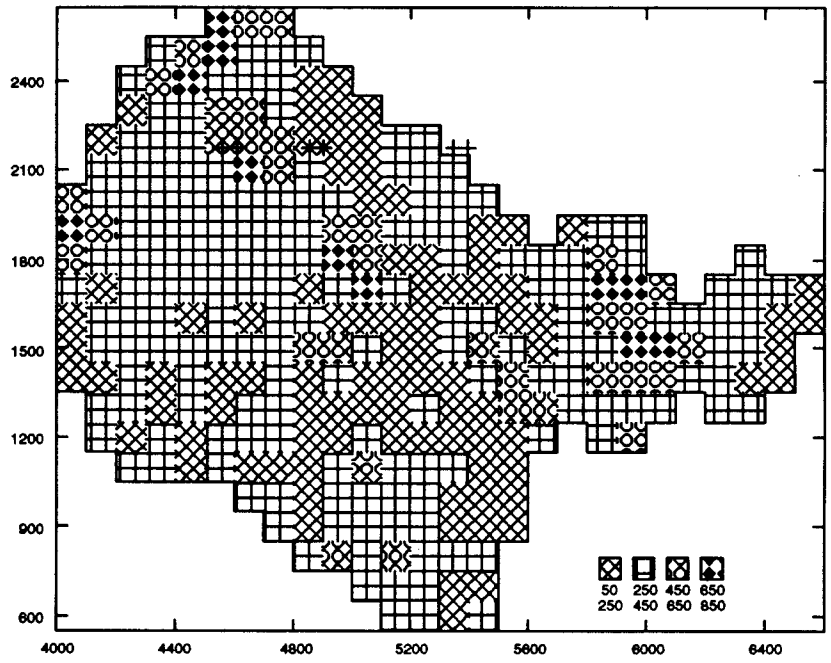


Fig. 15: Double Kriging error map for stem density.
(squares of 100m by 100m)

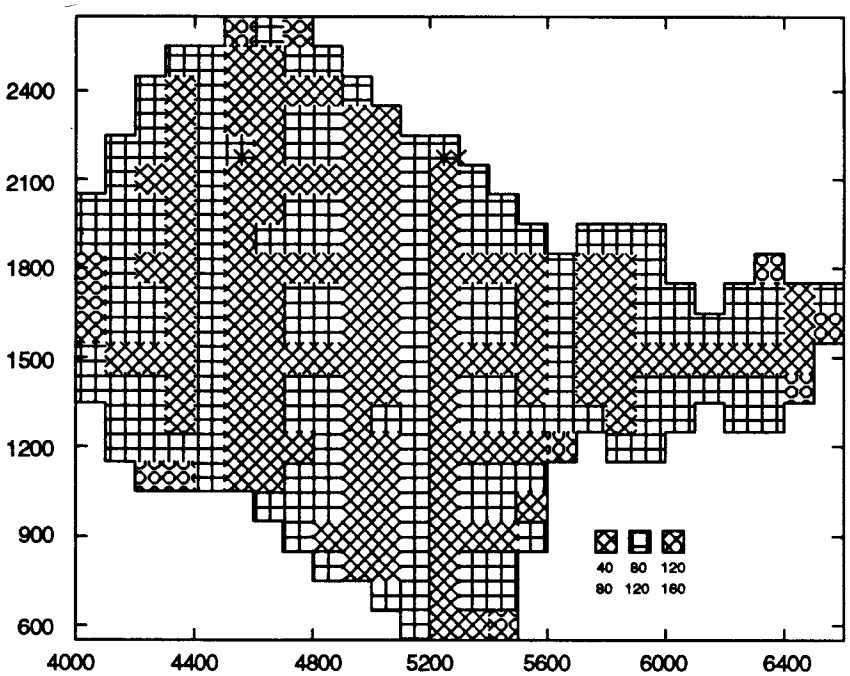


Table 4: Entire Domain
(218 ha)
Point Estimates
Kriging/standard error ( )

VARIABLES	OK	MK	EK	DK	ED	DB
Stem Density	324.76 (15.14)	326.15 (7.11)	327.33 (6.45)	325.84 (11.15)	327.47 (8.65)	325.08 (12.12)
Basal Area Density (m^2/ha)	31.85 (0.67)	31.29 (0.13)	31.44 (0.41)	31.39 (0.71)	31.48 (0.71)	31.23 (0.89)
Percentage of non-healthy tree	18.74 (1.07)	18.58 (0.23)	not done	18.83 (1.20)	to complex	18.90 (1.60)

Table 5: Small Area
(17 ha)
Point Estimates
Kriging/standard error ( )

VARIABLES	True value	OK	MK	EK	DK	ED	DB
Stem Density	**280.23**	294.35 (40.56)	282.10 (21.57)	288.31 (20.38)	281.67 (26.91)	290.58 (17.04)	257.14 (48.59)
Basal Area Density (m^2/ha)	**29.60**	27.68 (2.08)	30.28 (0.40)	29.00 (1.26)	29.46 (1.39)	29.28 (1.42)	24.00 (3.70)
Percentage of non-healthy tree	**24.20**	24.89 (4.07)	21.65 (0.93)	not done	22.13 (3.80)	to complex	26.40 (7.30)

Discussion

The most important facts are:

1. The point estimates for the entire domain are all very close to each other.

2. For the small area the empirical biases of the geostatistical techniques are much smaller than the empirical biases of the design-based approach, though both are within the respective standard errors.

3. The kriging errors are generally smaller than the design-based error, particularly for the small area. This statement must be somewhat weakened as the design-based technique is known to overestimate, generally, the error under systematic grids. Furthermore, universal kriging relies on the exact knowledge of the spatial means of the auxiliary variables and should be compared, therefore, to the corresponding design-based technique (D. Mandallaz, 1991), which gives slightly smaller errors than the **DB**-errors of tables 4 and 5, without changing the overall conclusions.

4. As expected on theoretical grounds, the mixed kriging errors are too small (especially for the small area), likewise but to a lesser degree, for kriging with measurement errors.

5. Universal kriging with external drifts yields, in most instances, a smaller error than double kriging, as expected on theoretical grounds.

6. Modified double kriging (in which the predictions are obtained by universal kriging, see section 5.6) gives essentially the same results as double kriging.

7. As compared to the design-based approach, double kriging yields, for the entire domain and the small area, variance reductions of 15% and 33% for the stem density, 36% and 85% for the basal area, respectively.

8. Whereas the design-based technique assigns constant weights to all the predictions and to all the residuals, the double kriging weights display, as expected, a very different behaviour: e.g. for the stem density in the small area, the 192 predictions weights (see point 4 below of the paragraph on numerical methods) ranged from -1.64% to 2.64%, roughly 20% being negative, whereas the 298 residual weights ranged from 0.06% to 1.53% and were far more stable, with a bulk around 0.3%, in agreement with the short range of the residual correlation, close to a nugget effect.

9. For the estimation of the percentage of non-healthy trees, ordinary and double kriging perform best and are essentially equivalent.

10. Double and Universal Kriging with the **LS** and **ML** give practically the same results as the empirical variograms. The same holds also for the Ordinary Kriging of the basal area, even with the substantially, but not significantly, larger **LS** range. In this sense, kriging results appear to be reasonably robust with respect to the choice of the variogram, in agreement with general experience. However, point 11 below shows that the above statement is not to be trusted blindly.

11. The model-dependent technique, which is essentially equivalent to Universal Kriging with a pure residual nugget effect (D. Mandallaz, 1991), yields point estimates very close to **ED** in tables 4-5, but it substantially underestimates the error for the small area (by 44% for the stem density and 27% for the basal area). Therefore, in Double and Universal Kriging, significantly to small ranges can yield misleading error estimates, in contrast to significantly to large ranges.

12. The kriging map for stem density (Fig. 14) displays obvious similarities with the stand map (Fig. 1); the kriging error map (Fig. 15) displays larger errors at the boundaries, as expected since the point estimates are based on less observations than in the central parts.

Further information on the numerical methods:

1. Prediction kriging for the entire domain was performed after dividing the domain in 8 roughly equal sectors, each kriged in a unique neighbourhood, with subsequent combinations of the kriging point estimates and errors.

2. Ordinary kriging was performed in a unique neighbourhood.

3. To check the accuracy of the combination technique, mentionned before, ordinary kriging was also performed with 4 roughly equal sectors. The numerical difference was well within the statistical error, so that the technique could be deemed reliable.

4. For the small area, the kriging of the predictions was performed with a unique neighbourhood of 192 points, out of which 92 were inside.

5. For the kriging maps, the predictions were kriged in moving 8 points neighbourhoods.

6. Kriging of the residual was always performed in a unique neighbourhood with all 298 points and with the variograms obtained by the least square technique of section 5.6; moving neighbourhoods led to poorer results, particularly in the small area where the terrestrial plots were, by chance, well below average.

7. The numerical integrations were all performed with the same 25m by 25m grid.

8. Several other tuning options were tried, without much impact on the results. The advantage of the choice presented here is that it is constant for all procedures.

9. For estimating ratios, it was found that one iteration suffices to determine the auxiliary variogram (see section 5.7).

10. The accuracy of the polygons defining the forest area does not have too much impact on the results as long as the main geographical features are captured; the kriging error is more sensitive than the point estimates. For instance, considerably simplifying the polygon for the entire domain from 2490 to 28 vertices leads to relative differences of 3.5% for the surface area, 0.8% for the ordinary kriging point estimate of the stem density, and 3.2% for its error, i.e. all the differences are well within the statistical uncertainty.

7.8 Validation.

Unbiasedness and coverage probability

To validate the estimation procedures the small area was divided in squares of 1ha and 0.25ha, for which point estimates and 95% confidence intervals (assuming normality) were calcu-lated. For squares overlapping the non-forest area surface, areas adjustments were done. For space reasons, the details are given for ordinary and double kriging only. The calculations were performed with the empirical variograms.

Figures 14-21 below display the scatter plots predicted versus observed values, for the stem density and the basal area. Circles indicate that the true value lies in the confidence intervals, whereas the stars indicate that it does not. The most important facts are:

1) Both methods give empirically unbiased point estimates.

2) For the stem density, double kriging yields an actual coverage rate close to 95 % (if the squares were independent the difference between actual and nominal coverage rates would not be significant) and performs better than ordinary kriging.

3) For basal area, neither ordinary kriging nor double kriging are really matching the nominal coverage rate with the 1ha squares; with the .25ha squares double kriging performs rather well and better than ordinary kriging.

3) The correlation between observed and predicted values is, in all but one case, higher for the 1ha than the .25ha squares, as intuitively expected.

Fig. 22-25 display the empirical distribution functions, over the squares, of the true and predicted values. They reveal that double kriging performs indeed better than ordinary kriging for the stem density; for the basal area both techniques have difficulties to cope with the large values.

Fig. 16: Double Kriging for stem density, 1ha squares.
($r^2 = 0.89$)

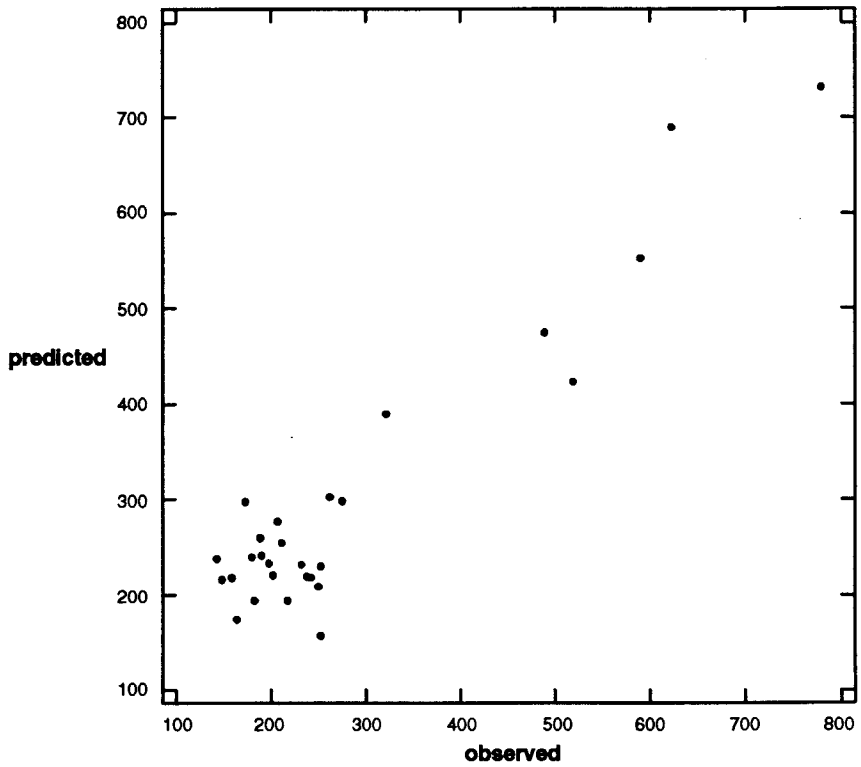


Fig. 17: Ordinary Kriging for stem density, 1ha squares.
($r^2 = 0.56$).

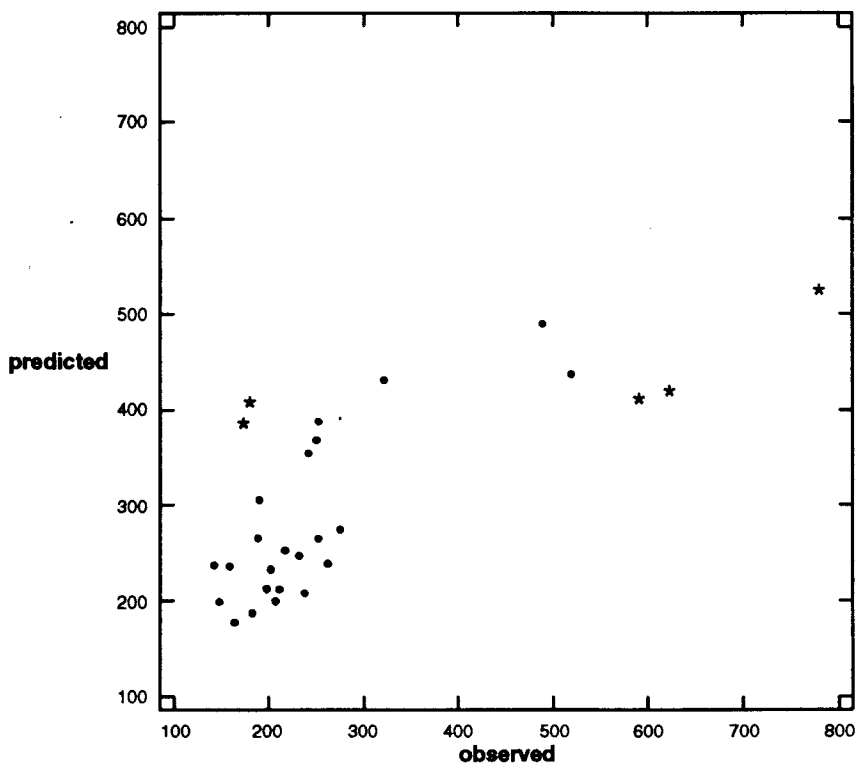


Fig. 18: Double Kriging for basal area, 1ha squares.
($r^2 = 0.12$)

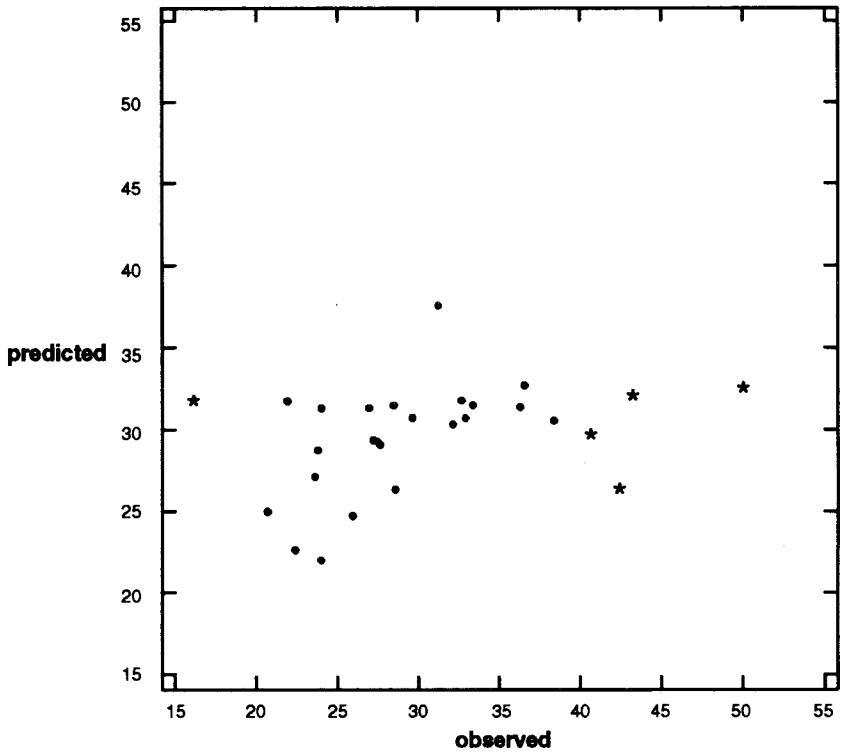


Fig. 19: Ordinary Kriging for basal area, 1ha squares.
($r^2 = 0.39$)

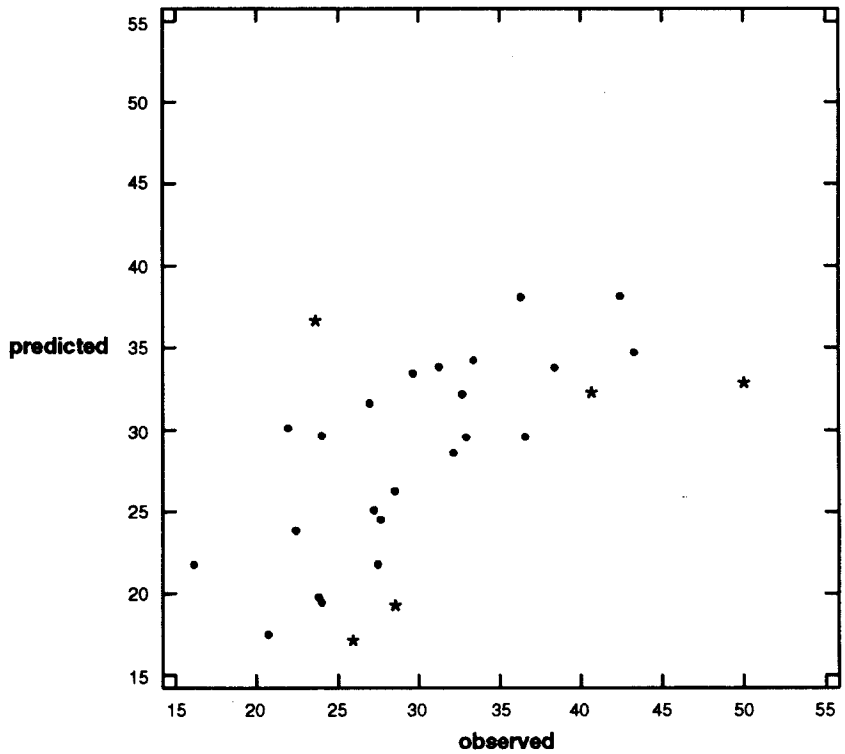


Fig. 20: Double Kriging for stem density, .25 ha squares.
($r^2 = 0.57$)

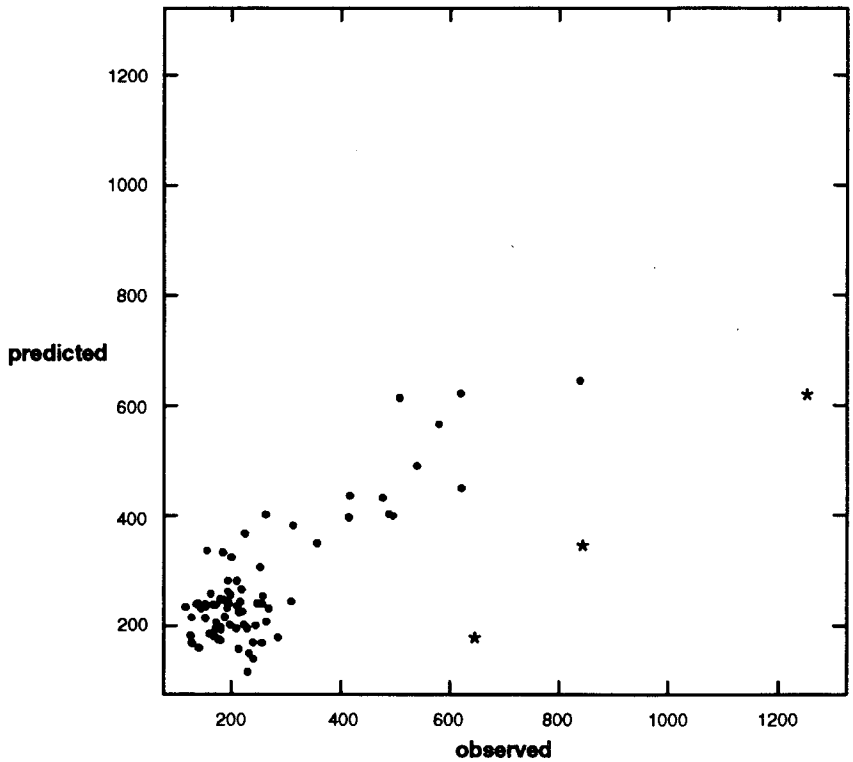


Fig. 21: Ordinary Kriging for stem density, .25ha squares.
($r^2 = 0.42$)

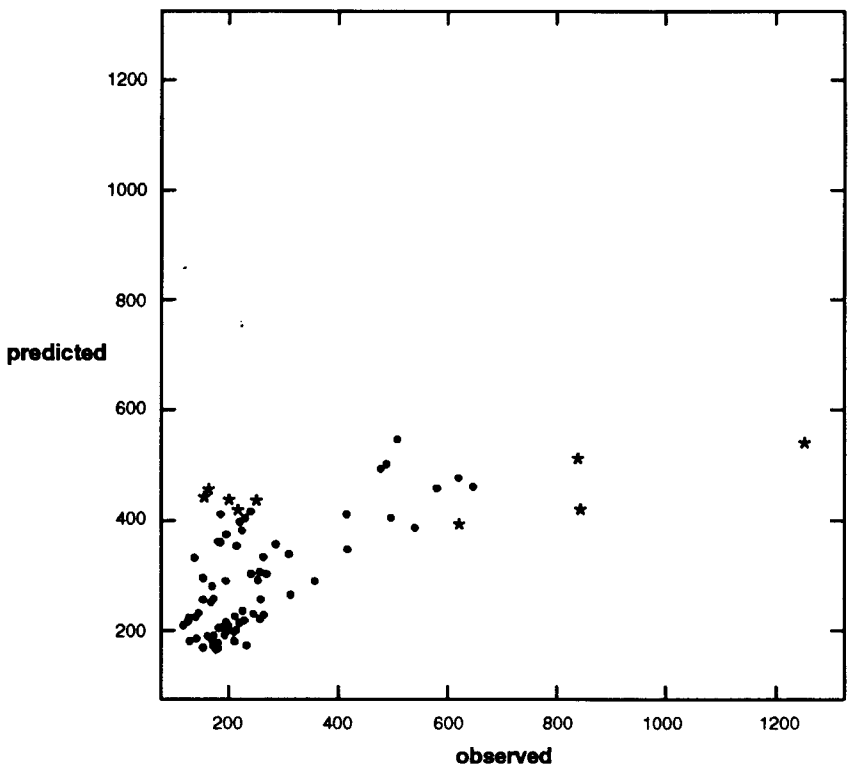


Fig. 22: Double Kriging for basal area, .25 ha squares.
($r^2 = 0.15$)

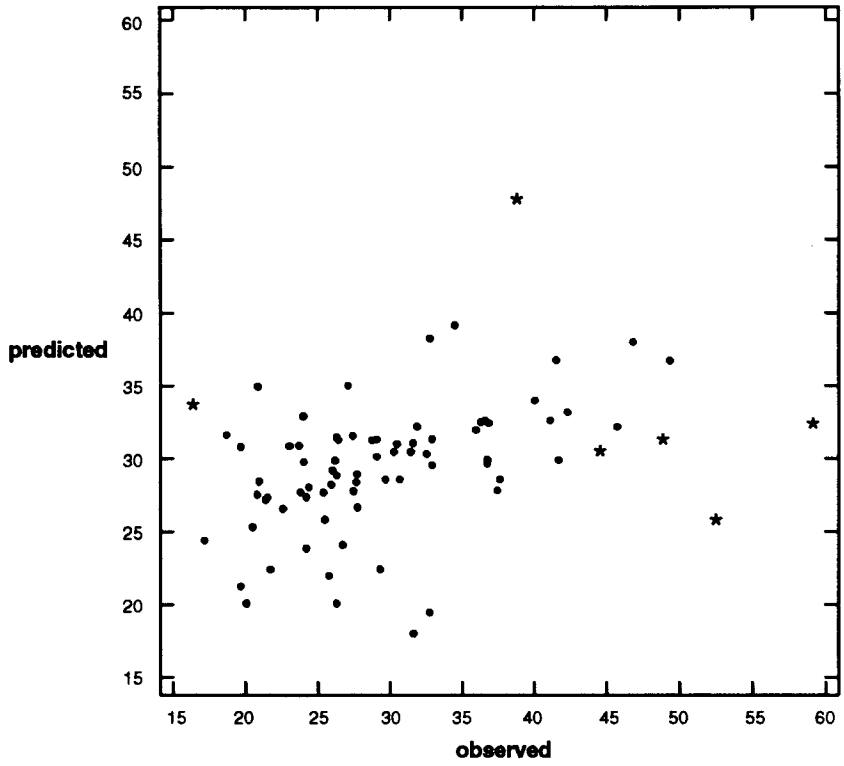


Fig. 23: Ordinary Kriging for basal area, .25 ha squares.
($r^2 = 0.35$)

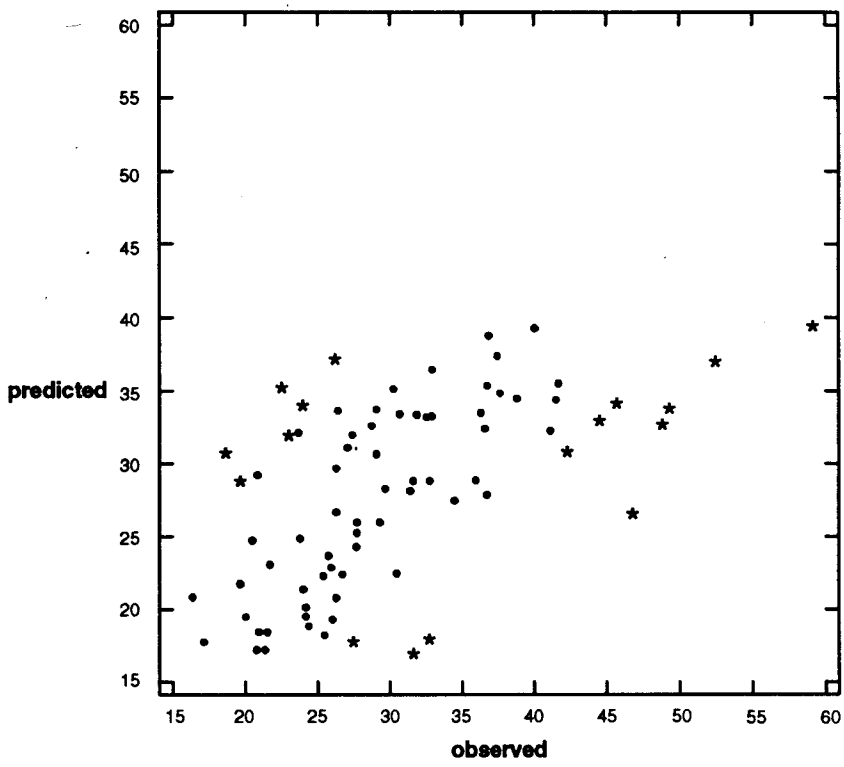


Fig. 24: Empirical distribution for 1ha squares, stem density

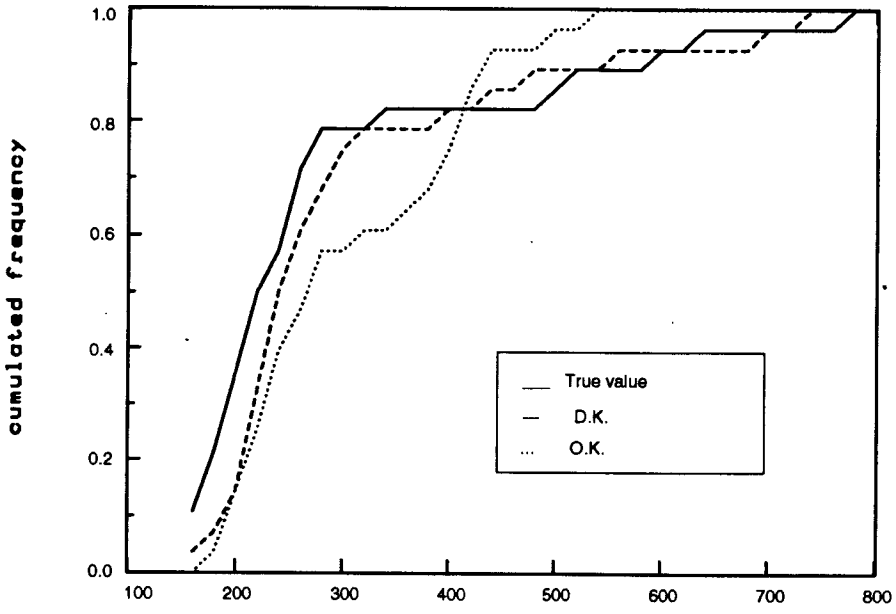


Fig. 25: Empirical distribution for .25ha squares, stem density

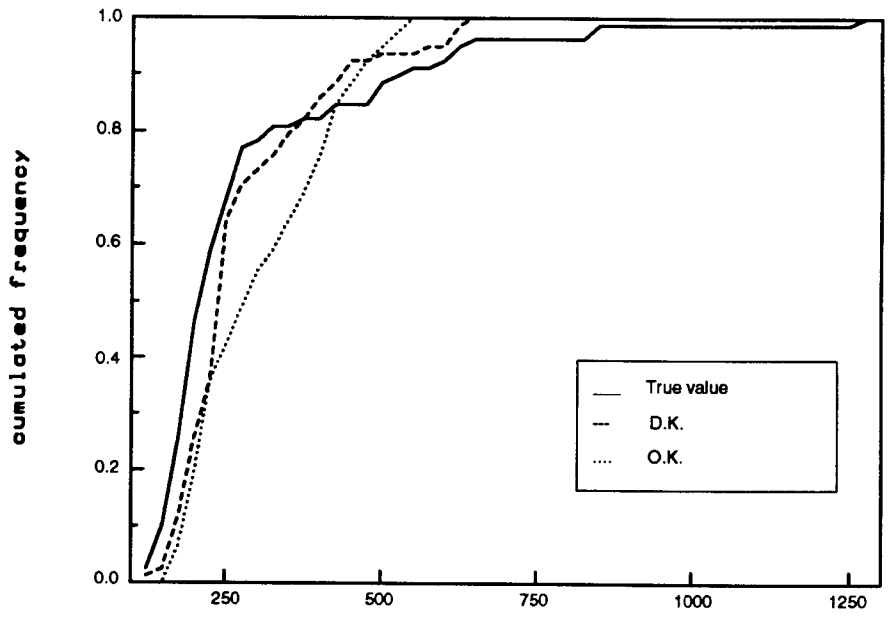


Fig. 26: Empirical distribution for 1ha squares,
Basal area density (m^2/ha)

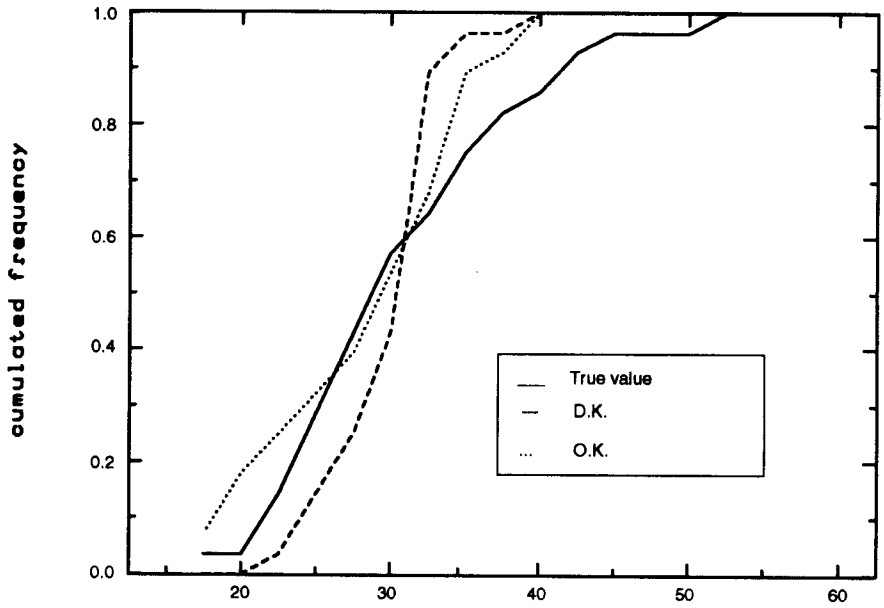
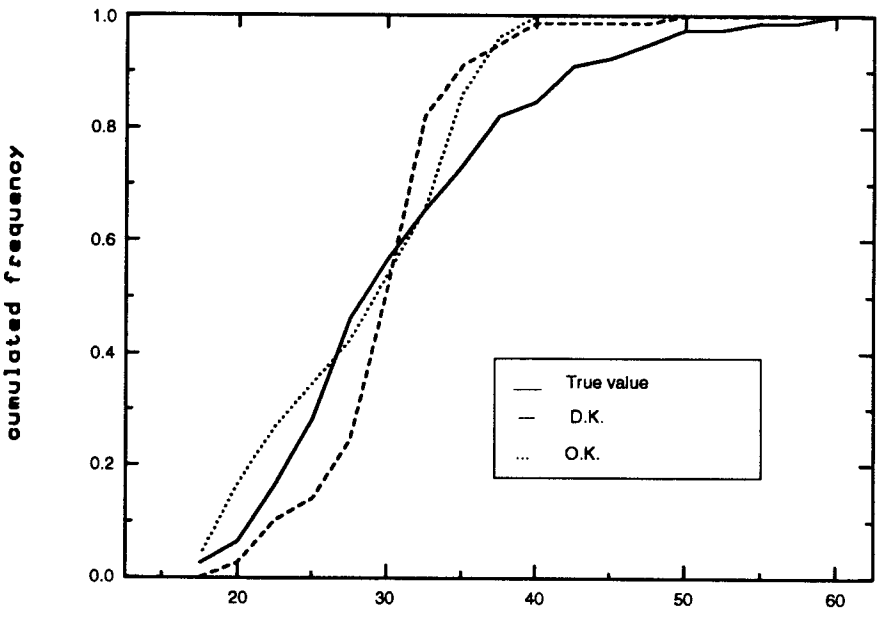


Fig. 27: Empirical distribution for 25ha squares,
Basal area density (m^2/ha)



The previous findings show that the point estimates are satisfactory, but that apparently the confidence intervals may be a source of concern, especially for the basal area. To get further insight, we compare the estimated mean square errors with the corresponding empirical values and remove the striking outliers (though there were no obvious reasons for their occurence). More precisely we consider the following two criteria:

$$\text{Chi-square} = \sum_{i=1}^{n} \frac{(z_i - z_i^*)^2}{\hat{\sigma}_i^2} \quad , \quad \text{Index} = \frac{\frac{1}{n}\sum_{i=1}^{n}(z_i - z_i^*)^2}{\frac{1}{n}\sum_{i=1}^{n}\hat{\sigma}_i^2}$$

If the squares were independent, and under a gaussian model, the first criterion will approximately follow a chi-square distribution on n degrees of freedom, hence its name. Both criteria indicate how well the model-based errors agree with their empirical counterparts. Tables 6 and 7 below display the results.

Table 6: Goodness-of-fit with all the observations.

Variables	Squares	Methods	Sample size	Chi-square	Index	r^2
Stem Density	100x100 1ha	DK	28	14	0.54	0.89
		OK	28	42	1.61	0.56
	50x50 .25ha	DK	78	123	1.48	0.57
		OK	78	203	2.34	0.42
Basal Area Density	100x100 1ha	DK	28	61	2.07	0.12
		OK	28	67	2.34	0.39
	50x50 .25ha	DK	78	98	1.20	0.15
		OK	78	197	2.57	0.35

Table 7: Goodness-of-fit without outliers.

Variables	Squares	Methods	Sample size	Chi-square	Index	r^2
Stem Density	100x100	DK	28	14	0.54	0.89
	1ha	OK	27	31	1.34	0.46
	50x50	DK	75	37	0.54	0.50
	.25ha	OK	74	79	1.20	0.42
Basal Area Density	100x100	DK	25	28	1.04	0.27
	1ha	OK	26	39	1.52	0.53
	50x50	DK	76	76	0.89	0.20
	.25ha	OK	71	108	1.56	0.40

One observes that:

1. Double kriging gives, on average, better error estimates than ordinary kriging; however, without adjustment for outlying observations, both techniques underestimate the error.

2. The goodness-of-fit criteria are very sensitive to a few outliers.

3. After removing the outliers, double kriging overestimates the error, whereas ordinary kriging underestimates it. Hence, double kriging does not only give smaller errors, but it is also more reliable.

4. It was also found that mixed kriging and kriging with measurement errors were empirically unbiased; both underestimate the error, sometimes severely and in any case more so than **O.K.** and **D.K.** (even after removing the outliers), particularly mixed kriging.

5. It can be expected on theoretical grounds that universal kriging with external drifts performs as well as double kriging, but the calculations were not carried out because of the enormous amount of work required.

8 Conclusions.

The general conclusion of this work is that geostatistics offers a natural theoretical framework to the estimation problems of forest inventory; it is also more germane to these problems than classical sampling theory as it takes the fundamental spatial aspects into account, which sampling theory essentially ignores.

The primary objective of this work was to adapt geostatistical techniques to combined forest inventory. This can be done in several ways: mixed kriging, kriging with measurement errors, co-kriging, double kriging and universal kriging; the latter only if the first phase is exhaustive (i.e. when thematic maps of the auxiliary informations are available).

The best procedure, in terms of simplicity, efficiency, reliability and mathematical coherence is double kriging, which is a straightforward generalization of the classical design-based regression techniques in double sampling; this procedure is very simple when the prediction model is external, i.e. known prior to the inventory: perform ordinary kriging of the predictions and residuals and add up the point estimates and kriging variances; this requires the covariances or variograms of predictions (available in large numbers) and residuals. Universal kriging is the limit case of double kriging when the first sampling phase is exhaustive. Mixed-kriging and kriging with measurement errors are simpler ad-hoc procedures which also yield unbiased point estimates; however, they tend to underestimate the kriging error. Mixed-kriging rests upon the variogram of the predictions only, which can be an advantage if the sample size of the second phase sample is small.

If the prediction model has to be estimated with the inventory data, the inference of the residual covariance is a difficult mathematical problem. This estimation can be performed either by a least square procedure, which is a mathematically correct version of the empirical "fitting by eye " technique, or by a restricted maximum likelihood procedure. Both methods filter out the drifts and are therefore also applicable with external models (and in the stationary case). It turns out that the two techniques yield consistent estimates of the parameters

of the residual covariance under essentially the same conditions, in particular, if the range of the correlation is small with respect to the dimension of the data field; if this condition is not met, consistency may or may not hold. The least square estimate is easier to compute but less efficient than the restricted maximum likelihood estimate, particularly under increasingly stronger correlation. The consistency conditions and first empirical evidence tend to indicate that kriging techniques are primarily relevant for local estimation.

Though all proposed geostatistical techniques yield theoretically and empirically unbiased estimates (as shown by the case study), double kriging is more reliable, with respect to the estimated error, than mixed kriging and kriging with measurement errors. Universal kriging performs even better than double kriging but is of course more expensive and is not always available.

The case study, at the level of the forest enterprise (220ha, with 300 terrestrial plots and a stand map), suggests that double kriging and universal kriging, together with an adequate estimation procedure of the residual covariance, are reliable procedures for the estimation of stem and basal area densities, as well as for estimation of ratios (such as the percentage of non-healthy trees). For very small areas (.25ha-1ha, stem and basal area densities only), they yield excellent point estimates and acceptable error estimates, whereas design-based techniques are either not even available or of little practical value. For small areas (20ha), they perform much better than the classical design-based techniques, with respect to bias and error (up to 85% variance reduction for basal area). For larger areas, all the point estimates get closer and the variance reduction is smaller, though still interesting (36% for basal area).

From a practical point of view, it must be emphasized that the quality of the prediction model and the cost ratios are key factors in combined inventory, in both the classical and geostatistical contexts. For optimization, one can rely on the classical techniques as a first approximation, since the criteria for optimization are generally defined at the global level, whereas geostatistics is primarily relevant for local estima-

tion. Since geostatistical techniques rely first of all on good estimates of the spatial correlation, particularly at short distances, it is recommended to use cluster sampling techniques, at least partially.

Future work will apply the techniques developed in this work at the regional and national levels (with the hope that a kind of scale invariance will hold for the conclusions) and extend them to continous inventory (i.e. to the estimation of growth), particularly under sampling with partial replacement. With respect to long term perspectives, a large advantage of geostatistics over sampling theory is that the sampling design is irrelevant for the calculations (but not for efficiency):in the design-based approach the calculations tend to become inextricable with more than two sampling occasions.

There is also a need to know more about the finite sample properties of the estimates of the spatial covariance, particularly when its range is not negligible with respect to the dimension of the data field, and to assess the impact thereof on the kriging estimates. This will require extensive simulations and numerical work.

9 Mathematical Appendix

This chapter gives a detailed discussion of the restricted maximum likelihood procedure for estimating the residual covariance matrix in universal kriging. The results are not new as such, but recent and difficult enough to have been mostly ignored by the applied literature. The proofs given are interesting because of their relative simplicity and in this sense also new (the literature deals generally with the full maximum likelihood approach). Moreover, this chapter presents a rigorous treatment of the least square procedure for estimating the residual covariance and compares it to the restricted maximum likelihood approach. These results are new. Though of a theoretical nature, the developments below are not without practical, sometimes even far reaching consequences, especially with respect to the underlying assumptions. These findings have been intuitively outlined in section 5.6.

Before going into the proper statistical discussion, it is necessary, for easier reference, to list or prove several technical facts. This is done in section 9.1.

9.1 Preliminaries

N.B. Vectors are understood throughout as column vectors and the upper index t stands for the transposition operator of vectors and matrices.

Lemma **(9.1)**

Let A, B be two (n,n) symmetrical matrices and $Tr(\ )$ denote the trace operator, then one has

$$(Tr(AB))^2 \leq Tr(A^2)Tr(B^2)$$

with equality if and only if $\exists \mu \in R \quad A = \mu B$.

This is essentially the Cauchy-Schwartz inequality. For a proof see (J.R. Magnus, H. Neudecker, 1988, p.201).

Lemma **(9.2)**

Let A, X be positive semi-definite (n,n) matrices, then

$$(\det A \det X)^{\frac{1}{n}} \leq \frac{1}{n} Tr(AX)$$

with equality if and only if $A=0$ or $\exists \mu \in R^{+} \quad X=\mu A^{-1}$. For a proof see (J.R. Magnus, H. Neudecker, 1988, p.226).

Lemma **(9.3)**

The Frobenius (euclidian) norm of any matrix C is defined by:

$$\|C\| = \left(Tr(C^t C)\right)^{1/2} = \left(\sum_{i,j} C^2{}_{i,j}\right)^{1/2}$$

Then for any quadratic (n,n) matrix C one has

$$Tr\left(C^2\right) \leq Tr\left(C^t C\right) = \|C\|^2$$

with equality if and only if C is symmetrical, i.e. $C^t = C$.

Proof:

By definition $Tr\left(C^2\right) = \sum_{i=1}^{n} C_{i,i}^2 + \sum_{k \neq i} C_{i,k} C_{k,i}$.

set $\Delta = Tr\left(C^t C\right) - Tr\left(C^2\right) = \sum_{k \neq i} C_{i,k}^2 - \sum_{k \neq i} C_{i,k} C_{k,i}$, which can be rewritten as: $\sum_{k<i} C_{i,k}^2 - \sum_{k<i} C_{i,k} C_{k,i} + \sum_{k>i} C_{i,k}^2 - \sum_{k>i} C_{i,k} C_{k,i}$, let $\Delta_{i,k} = C_{i,k} - C_{k,i}$, then one gets

$$\Delta = \sum_{k<i} C_{i,k} \Delta_{i,k} + \sum_{k>i} C_{i,k} \Delta_{i,k} = \sum_{k<i} C_{i,k} \Delta_{i,k} + \sum_{k<i} C_{i,k} (-\Delta_{i,k}) = \sum_{k<i} (C_{i,k} - C_{k,i}) \Delta_{i,k} = \sum_{k<i} \Delta_{i,k}^2 \geq 0,$$

with equality if and only if $\Delta_{i,k} \equiv 0$, which completes the proof.

Lemma **(9.4)**

Let A_1, A_2 be symmetrical (n,n) matrices with eigenvalues

$$\lambda_1(A_k) \leq \lambda_2(A_k) \leq \ldots\ldots\ldots \lambda_n(A_k), k = 1,2$$

then one has the following upper bound for the norm of the, not necessarily symmetrical, matrix product:

$$\|A_1 A_2\|^2 \leq \left(\sum_{i=1}^{n} \lambda_i^2(A_1)\right) \lambda_n^2(A_2)$$

Proof:

Diagonalizing the matrices one can write $A_k = Q_k \Lambda_k Q_k^t, k = 1,2$, where the Λ_k are diagonal and the Q_k are orthogonal; then,

because of the invariance of the norm under left/ right orthogonal transformations and the definition of the norm one has
$\|A_1A_2\|^2 = \|Q_1\Lambda_1Q_1^tQ_2\Lambda_2Q_2^t\|^2 = \|\Lambda_1Q_1^tQ_2\Lambda_2\|^2 = Tr\left(\Lambda_2Q_2^tQ_1\Lambda_1^2Q_1^tQ_2\Lambda_2\right)$
$= Tr\left(Q_2^tQ_1\Lambda_1^2Q_1^tQ_2\Lambda_2^2\right) = Tr\left(\Lambda_1^2(Q_2^tQ_1)^t\Lambda_2^2Q_2^tQ_1\right)$, because of the invariance of the trace under cyclic permutation. With the orthogonal matrix $R = Q_2^tQ_1$ this last expression can be written as $Tr\left(\Lambda_1^2R^t\Lambda_2^2R\right)$. Let $d_{i,j} = \left(R^t\Lambda_2^2R\right)_{i,j} = r_i^t\Lambda_2^2r_i$, where the r_i are the orthonormal column vectors of R. It is then straightforward to verify that $Tr\left(\Lambda_1^2R^t\Lambda_2^2R\right) = \sum_{i=1}^{n}\lambda_i^2(A_1)d_{i,i}$, and therefore the result since the quadratic form defining the $d_{i,i}$ takes its maximum at the largest eigenvalue.

Lemma **(9.5)**

Let A_1, A_2 be symmetrical matrices with eigenvalues as in Lemma 4 then one has the following inequality:

$$Tr\left(A_1A_2\right)^2 \le \left(\sum_{i=1}^{n}\lambda_i^2(A_1)\right)\lambda_n^2(A_2)$$

The proof follows at once from the Lemmas 3 and 4. This result will play a key role in the convergence proof.

Lemma **(9.6)**

(Poincaré separation theorem):

Let A be a real (n,n) symmetrical matrix with eigenvalues $\lambda_1 \le \lambda_2 \ldots\ldots \le \lambda_n$ and G a semi-orthogonal (n,k) matrix, i.e. $G^tG = I_k$. Let $\mu_1 \le \mu_2 \le \ldots\ldots \mu_k$ denote the non-zero eigenvalues of G^tAG, then

$$\lambda_i \le \mu_i \le \lambda_{n-k+i} \quad i = 1,2\ldots k$$

Furthermore, if M is (n,n) symmetrical idempotent matrix of rank k, then the non-zero eigenvalues $\mu_1 \le \mu_2 \le \ldots\ldots \mu_k$ of the matrix MAM satisfy the relations

$$\lambda_i \le \mu_i \le \lambda_{n-k+i} \quad i = 1,2\ldots k$$

Proof: see (J.R. Magnus, H. Neudecker, 1988, p.209-210).

Lemma **(9.7)**

If Y is a random vector following a multivariate normal distribution $N(\mu,\Sigma)$ and A,B are two matrices such that $A\mu=0, B\mu=0$, then one has the following expression for the covariance of two quadratic forms:

$$E\left(Y'AY\right)\left(Y'BY\right)=Tr(A\Sigma)Tr(B\Sigma)+2Tr(A\Sigma B\Sigma)$$

Proof:
This is a straightforward generalization of a result given in (J.R. Magnus, H. Neudecker, 1988, p.251) after using the relation $4\mathrm{cov}(X,Y)=\mathrm{var}(X+Y)-\mathrm{var}(X-Y)$.

Lemma **(9.8)**

Let Y be $N(\mu,\Sigma)$.
For the quadratic form $n^{-1}Y'BY$ with $B\mu=0$ to converge in mean square, and hence in probabitity, towards its expected value, i.e. $\lim_{n\to\infty} E\left(n^{-1}\left(Y'BY-Tr(B\Sigma)\right)\right)^2=0$, it is sufficient that the following condition holds:

$$\lim_{n\to\infty}\left(n^{-2}Tr(B\Sigma)^2\right)=0$$

Proof:
Expanding the square and lemma (9.7) yield the result after some simple algebraic manipulations.

Lemma (9.8) will play a key role as it allows the use of a law of large numbers with correlated random variables, provided the correlation is not "too strong". It is heuristically clear that the lemma will hold without the assumption of normality but with further conditions on the 4th moments.

Lemma **(9.9)**

Let C be a symmetrical (n,n) matrix of rank $n-r$ with non-zero eigenvalues $\lambda_1\le\lambda_2.....\le\lambda_{n-r}$. A generalized inverse of C is any

matrix C^- satisfying $CC^-C=C$. Then there exits a $(n,n-r)$ matrix A such that

$$C^- = A\left(A^tCA\right)^{-1}A^t$$

in particular $\left(A^tCA\right)$ is a regular $(n-r,n-r)$ matrix.

Proof:

There exists an orthogonal matrix R such that

$$C=R\begin{pmatrix}\lambda_1 & 0 & \dots & 0\\ 0 & . & \dots & 0\\ . & . & \lambda_{n-r} & 0\\ 0 & 0 & \dots & 0\end{pmatrix}R^t \quad , \text{ let } A=R\begin{pmatrix}1 & . & 0\\ 0 & . & 0\\ 0 & 0 & 1\\ . & . & 0\\ 0 & 0 & 0\end{pmatrix}$$

(so that dimensions match), it is easily verified that

$$A^tCA=\begin{pmatrix}\lambda_1 & 0 & .... & 0\\ 0 & \lambda_2 & .... & 0\\ . & . & .... & 0\\ 0 & \dots & 0 & \lambda_{n-r}\end{pmatrix}$$

and that $C^- = A\left(A^tCA\right)^{-1}A^t$ indeed satifies $CC^-C=C$.

This lemma simply amounts to reparameterizing a linear system with dependent equations.

Lemma **(9.10)**

Consider the differentiable mapping $\rho\in R^p\rightarrow V(\rho)$, where $V(\rho)$ is a symmetrical (n,n) matrix, then

$$\frac{\partial}{\partial\rho_i}\log\det V(\rho)=Tr\left(V^{-1}(\rho)\frac{\partial}{\partial\rho_i}V(\rho)\right)$$

$$\frac{\partial}{\partial\rho}V^{-1}(\rho)=-V^{-1}(\rho)\frac{\partial}{\partial\rho}V(\rho)V^{-1}(\rho)$$

See (J.R. Magnus, H. Neudecker, 1988, p.150) for a proof and the exact regularity conditions.

9.2 Restricted Maximum Likelihood Estimate of the Residual Covariance

In this section, we investigate the properties of the restricted maximum likelihood estimate (for short **REML**) of the residual covariance matrix; this estimate is simpler and has a better finite sample behavior than the full maximum likelihood estimate (which simultanously estimates the regression coefficients and the covariance). The idea is to consider only linear combinations of the data which filter out the drift.

The residual vector $r = P^{\perp}Y$ has a singular multivariate normal distribution with zero expectation and covariance matrix $\overline{\Sigma} = P^{\perp}\Sigma P^{\perp}$, where Σ is the regular covariance matrix of a multivariate normal random vector Y. By assumption, this random vector is the restriction of an underlying gaussian stochastic process in the plane at a finite set of sample points where the process is actually observed. According to the external drift model, described in section 5.6, one has $EY = \mu = F\beta$ and $P^{\perp}\mu = 0$ since by construction of the projection operator $P^{\perp}F = 0$. Except for the more intricate discussion of infill asymptotics this underlying process is never used explicitely. We assume that $\Sigma = \Sigma(\theta)$ with $\theta = (\sigma^2, \rho)' \in R^k$. The variance σ^2 is assumed to be constant and ρ is a $k-1$ dimensional parameter describing the structure of the correlation (range, anisotropies etc.). We therefore have the model $\Sigma(\theta) = \sigma^2 K(\rho)$, where $K(\rho)$ is a correlation matrix. The underlying stochastic process is not assumed, at this stage, to be necessarily stationary (even if the variance is constant).

The covariance matrix of the residual vector q is $\overline{\Sigma}(\theta)$ and has rank $n-q = \dim(Range(P^{\perp}))$; hence, there exists a $(n, n-q)$ full column rank matrix A such that $A'r = A'P^{\perp}Y$ has a regular multivariate normal distribution with zero expectation and a regular $(n-q, n-q)$ covariance matrix $\tilde{\Sigma}(\theta) := A'\overline{\Sigma}(\theta)A$. It is often useful to take A to consist of the $n-q$ eigenvectors of $P^{\perp}$ to the eigenvalue 1 (see also C.R. Dietrich, M.R. Osborne, 1991, for alternative algorithms details useful in special cases).

Following R. Christensen (1990b), the likelihood of the random

vector $A^t r$ is, up to an irrelevant multiplicative constant, independent of the choice of the matrix A, and is given by:

(9.11)

$$f(A^t r;\theta) = (2\pi)^{-n/2}\left(\det \tilde{\Sigma}(\theta)\right)^{-1/2} \exp\left\{-\frac{r^t A\tilde{\Sigma}^{-1}(\theta)A^t r}{2}\right\}$$

The REML equations can be obtained via lemma (9.10) and are

(9.12)

$$Tr\left(\tilde{\Sigma}^{-1}(\theta)\frac{\partial}{\partial\theta}\tilde{\Sigma}(\theta)\right) = r^t A\tilde{\Sigma}^{-1}(\theta)\frac{\partial}{\partial\theta}\tilde{\Sigma}(\theta)\tilde{\Sigma}^{-1}(\theta)A^t r$$

Specializing this to the case $\theta = (\sigma^2,\rho)^t$ the partial derivative with respect to the variance yields the following relation at the extremum:

(9.13)

$$\sigma^2(\rho) = \frac{Y^t P^{\perp} A\tilde{K}(\rho)^{-1}A^t P^{\perp} Y}{n-q} = \frac{Tr\left(\tilde{K}(\rho)^{-1}A^t P^{\perp} YY^t P^{\perp} A\right)}{n-q} \quad , \text{ with}$$

$$\tilde{K}(\rho) = A^t P^{\perp} K(\rho) P^{\perp} A =: A^t \overline{K}(\rho) A$$

Note that $E_{\rho_o}\sigma^2(\rho) = \sigma^2\dfrac{Tr\left(\tilde{K}(\rho)^{-1}\tilde{K}(\rho_o)\right)}{n-q}$

Substituting the expression for $\sigma^2(\rho)$ in the logarithm of the restricted likelihood (9.11) one obtains the so called restricted profile log-likelihood for ρ, which reads, up to irrelevant additive constants:

(9.14)

$$\log \tilde{f}(A^t r;\rho) = -\frac{n-q}{2}\log\sigma^2(\rho) - \frac{1}{2}\log\det\tilde{K}(\rho) =: ML$$

Therefore, after elimination of the variance, finding the REML estimate of the range parameter ρ is equivalent to minimizing the function $\sigma^2(\rho)^{n-q}\det\tilde{K}(\rho)$. The matrix $\tilde{K}(\rho)$ depends on the sampling design and not on the observations. To have consistency one should at least require $\sigma^2(\rho)$ to converge towards its expected value $E_{\rho_o}\sigma^2(\rho) = \sigma^2\dfrac{Tr\left(\tilde{K}(\rho)^{-1}\tilde{K}(\rho_o)\right)}{n-q}$. Some algebraic manipulations and lemma (9.8) show that this will be the case provided that $\lim\limits_{n\to\infty}\dfrac{1}{(n-q)^2}Tr\left(\tilde{K}(\rho)^{-1}\tilde{K}(\rho_o)\right)^2 \to 0$, ρ_o being the true value. The following theorem shows that this is essentially the case, if the true value lies in a known compact set and some technical regularity conditions are fulfilled.

Theorem (consistency of REML)

(9.15)

Let $U \subset R^{k-1}$ be a compact set whose interior contains the true value ρ_o of the correlation parameter. Consider an increasing sequence of sample points such that the following conditions are satisfied:

(i) $\forall \rho \in U \quad \lim_{n\to\infty} \frac{1}{(n-q)^2} Tr\left(\tilde{K}^{-1}(\rho)\tilde{K}(\rho_o)\right)^2 = 0$

(ii) The mapping $\rho \in U \to K(\rho) \in R^{n^2}$ is injective and continuous.

(iii) $\forall \rho_1, \rho_2 \in U \quad \left(\exists \mu > 0 \quad \tilde{K}(\rho_1) = \mu \tilde{K}(\rho_2)\right) \Rightarrow \rho_1 = \rho_2$

(iv) Let $B_j \quad j = 1,2.....K(M)$ a covering of $U \subset R^{k-1}$ with open spheres of radius M^{-1}. There exists a sequence $\varepsilon_M \geq 0, \lim_{M\to\infty} \varepsilon_M = 0$ such that:

$$\forall M \quad \lim_{n\to\infty} P_{\rho_o}\left\{\sup_{\rho\in B_j}\left|\log\frac{\sigma^2(\rho)}{E_{\rho_o}\sigma^2(\rho)}\right| > \varepsilon_M\right\} = 0 \quad \forall j = 1,2....K(M)$$

Let $\hat{\rho}_n \in U$ be the argument at the absolute minimum of the random function

$$f_n(\rho) := \log\sigma^2(\rho) + \frac{1}{n-q}\log\det\tilde{K}(\rho).$$

Then $\hat{\rho}_n$ converges in probability towards the true value ρ_o.

Condition (iv) means that $\sigma^2(\rho)$ converges towards its expected value not only pointwise (which is implied by the first condition), but also uniformly in ρ in arbitrary small spheres; the supremum can be taken over a dense countable subset and is therefore measurable.

Proof:

Define the (non-random) function

$$\phi_n(\rho) = \log\sigma^2 + \log Tr\frac{\tilde{K}^{-1}(\rho)\tilde{K}(\rho_o)}{n-q} + \frac{1}{n-q}\log\det\tilde{K}(\rho) = \log E_{\rho_o}\sigma^2(\rho) + \frac{1}{n-q}\log\det\tilde{K}(\rho).$$

We first show that $\phi_n(\rho_o) = \log\sigma^2 + \frac{1}{n-q}\log\det\tilde{K}(\rho_o)$ is the strict minimum of $\phi_n(\rho)$. This is equivalent to the inequality:

$$\frac{1}{n-q}\log\det\tilde{K}(\rho_o)<\frac{1}{n-q}\log\det\tilde{K}(\rho)+\log\frac{Tr(\tilde{K}^{-1}(\rho)\tilde{K}(\rho_o))}{n-q}\quad\text{for}\quad\rho\neq\rho_o$$

i.e.

$$\left(\det\tilde{K}^{-1}(\rho)\tilde{K}(\rho_o)\right)^{\frac{1}{n-q}}<Tr\frac{\tilde{K}^{-1}(\rho)\tilde{K}(\rho_o)}{n-q}.$$

By lemma(9.2) the latter is true unless $\tilde{K}(\rho)=\mu\tilde{K}(\rho_o)$ for some μ, which implies, by condition (iii), that $\rho=\rho_0$.

Next we prove the uniform convergence of $f_n(\rho)-\phi_n(\rho)=\log\frac{\sigma^2(\rho)}{E_{\rho_o}\sigma^2(\rho)}$,

i.e. $\forall\varepsilon>0\quad\alpha_n(\varepsilon)=P_{\rho_o}\left\{\sup_{\rho\in U}\left|f_n(\rho)-\phi_n(\rho)\right|>\varepsilon\right\}\to 0$ as $n\to\infty$. Take M so large that $\varepsilon_M<\varepsilon$ (by iv), then one has from the definitions:

$$\alpha_n(\varepsilon)\le\sum_{j=1}^{K(M)}P_{\rho_o}\left\{\sup_{\rho\in B_j}\left|\log\frac{\sigma^2(\rho)}{E_{\rho_o}\sigma^2(\rho)}\right|>\varepsilon_M\right\}$$

For any arbitrary small $\delta>0$, take n so large that by (iv) each summand in the above expression is less than $\frac{\delta}{K(M)}$, which implies $\alpha_n(\varepsilon)<\delta$ and the result. Therefore, we have shown that the random function $f_n(\rho)$ converges in probability and uniformly in ρ towards the non-random function $\phi_n(\rho)$ which has a unique absolute minimum at ρ_o .

We finally prove that the minimum $\hat{\rho}_n$ of $f_n(\rho)$ converges in probability towards ρ_o. Indeed, suppose that $\hat{\rho}_n$ did not converge in probability towards ρ_o. Then there exists a neighbourhood W of ρ_o such that $\hat{\rho}_n\notin W$ infinitely often.

Set $\gamma=\inf_{\rho\notin W}\left(\phi_n(\rho)-\phi_n(\rho_o)\right)>0$ (since $\rho_0\in W$ yields the minimum). Choose n so large that $\hat{\rho}_n\notin W$ and $\sup_{\rho\in U}\left|f_n(\rho)-\phi_n(\rho)\right|<\frac{\delta}{2}$. Then one has:

$$f_n(\rho_o)=\phi_n(\rho_o)+\left(f_n(\rho_o)-\phi_n(\rho_o)\right)<\phi_n(\rho_o)+\frac{\delta}{2}$$

$$f_n(\hat{\rho}_n)=\phi_n(\hat{\rho}_n)+\left(f_n(\hat{\rho}_n)-\phi_n(\hat{\rho}_n)\right)>\phi_n(\hat{\rho}_n)-\frac{\delta}{2}$$

and therefore also

$$f_n(\hat{\rho}_n)-f_n(\rho_o)>\phi_n(\hat{\rho}_n)-\phi_n(\rho_o)-\delta\ge 0\quad\text{as}\quad\hat{\rho}_n\notin W.$$

This implies at once $f_n(\hat{\rho}_n)>f_n(\rho_o)$, and a contradiction since $\hat{\rho}_n$ yields the absolute minimum of $f_n(.)$. Hence $\hat{\rho}_n$ converges in probability towards ρ_o, which completes the proof.

Remarks:

1) Condition (ii) implies in particular that the correlation matrices cannot be the identity for a subset of ρ values; in other words, the minimum distance between two sample points must be smaller than the correlation range (otherwise the likelihood is flat and any range below the minimum distance can be taken as the REML).

2) Condition (iii) is very technical but follows asymptotically from condition (ii) under reasonable regularity conditions. First recall, that in $\tilde{K}(\rho)=A'P^{\perp}K(\rho)P^{\perp}A$ the choice of A is arbitrary. In particular, we can choose A to consist of the $n-q$ orthonormal eigenvectors of the projector $P^{\perp}$ to the eigenvalue 1; thus $P^{\perp}A=A$ is a semi-orthogonal $(n,n-q)$ matrix. By lemma (9.6) one has the inequalities:

$$\lambda_i(\rho)\le\tilde{\lambda}_i(\rho)\le\lambda_{q+i}(\rho),i=1,2.....n-q$$

where the $\lambda_i(\rho)$ are the ordered eigenvalues of the original correlation matrices $K(\rho)$. Therefore $\tilde{K}(\rho_1)=\mu\tilde{K}(\rho_2)$ implies at once:

$$(n-q)^{-1}\sum_{i=1}^{n-q}\lambda_i(\rho_1)\le\mu(n-q)^{-1}\sum_{i=1}^{n-q}\lambda_i(\rho_2)\le(n-q)^{-1}\sum_{i=1}^{n-q}\lambda_{q+i}(\rho_1)$$

If the eigenvalues are bounded, letting $n\to\infty$ yields $\mu=1$ as the trace of the correlation matrix is n. If the asymptotic spectrum is a continuum then obviously $\tilde{\lambda}(\rho)\to\lambda(\rho)$; intuitively speaking, filtering out the drifts has asymptotically no effect on the eigenvalues and it suffices to consider processes with zero expectation when dealing with quantities depending only on the eigenvalues (we shall use this fact later for the calculation of the asymptotic relative efficiency). If the residual process is a stationary time series with bounded spectral density then $\tilde{\lambda}(\rho)\to\lambda(\rho)$; this follows from a famous theorem of Szegö on the asymptotic distribution of the eigenvalues of Toeplitz' forms (see U. Grenander, M. Rosenblatt, 1984, p. 104-105). For processes on the plane, we shall present later on a heuristic proof. In this sense , condition (ii) and (iii) will generally be asymptotically equivalent, at least for the standard stationary cases and for "non-pathological" sampling schemes (i.e. when the design matrix has a very particular

structure, depending on the drift, which, by chance, coincides with some pattern of the residual correlation).

3) X. Guyon (1993, theorem 3.3, p. 106) gives convergence proofs in a more general set-up, with a condition similar to (iv), but assuming that $\phi_n(\rho)-\phi_n(\rho_o)$ converges towards a so-called contrast function, which is not required here. From a practical point of view, it is difficult to think of situations where (i) holds, i.e. pointwise convergence, and (iv) not, i.e. uniform convergence on arbitrary small sets.

Using lemma (9.5) it is possible to give simple sufficient conditions for condition (i) to hold. Indeed, if $\tilde{\lambda}_i(\rho)$ denote the ordered eigenvalues of the matrices $\tilde{K}(\rho)$, then condition (i) holds if

$$\lim_{n\to\infty} \frac{1}{(n-q)^2}\left(\sum_{i=1}^{n} \frac{1}{\tilde{\lambda}_i^2(\rho)}\right)\tilde{\lambda}_n^2(\rho_o)=0$$

Using lemma (9.6) one obtains at once the following important result:

Theorem

(9.16)

If condition (i) implies (iv) and if conditions (ii), (iii) hold, then uniform boundedness, from below and above, of the eigenvalues of the correlation matrices, i.e.

$$\exists m_1, m_2 \quad 0 < m_1 \le \lambda_1(\rho) \le \lambda_n(\rho) \le m_2 < \infty \quad \forall \rho \in U$$

implies that the REML is consistent.

From a practical point of view, it can be expected that bounded eigenvalues suffices to insure consistency.

Remarks:

1) K.V Mardia and R.J Marshall (1984) obtained similar results for the full maximum likelihood estimate, but without the condition on the smallest eigenvalue (this is rather surprising since allowing the smallest eigenvalue to tend to zero would lead to singular covariance matrices; X. Guyon, 1982, has the

same condition for the smallest eigenvalue in terms of the spectral density). For a stationary process observed on a regular grid of the plane, the largest eigenvalue is uniformly bounded if the correlation is uniformly summable, i.e. if

$$\sum_{(l_1,l_2)\in Z^2} K((l_1,l_2),\rho) < M < \infty \quad , K((l_1,l_2),\rho) = corr((k_1,k_2),(k_1+l_1,k_2+l_2))$$

(the grid points are identified with their integer valued coordinates), see K.V. Mardia and R.J. Marshall (1984). From this follows that the REML is consistent for all stationary corre-lations with a uniformly bounded range if the domain tends to infinity and the grid mesh is fixed.

2) For a stationary time series, condition (i) can be formulated explicitly in term of the spectral density (M. Fox, M. J. Taqqu, 1987); this important but very difficult result shows that condition (i) can also hold for processes with a long range dependence (see F. R. Hampel, 1987, J. Beran, 1992). Using well-known relations between the behaviour of the spectral density and the distribution of the eigenvalues (U. Grenander, M. Rosenblatt, 1984, p. 104-105), we see that theorem (9.16) gives sufficient but not necessary conditions.

3) Unbounded eigenvalues can also occur with infill asymptotics, i.e when $n \to \infty$ in a finite domain (see B.D. Ripley, 1988). For time series this can be seen by using the aliasing relation between the spectral density of the underlying process in continuous time and the spectral density of the resulting time series under discrete systematic sampling (U. Grenander, M. Rosenblatt, 1984, p. 57), and the afore mentionned Szegö's theorem. It is therefore not quite clear when condition (i) will hold under infill asymptotics. Fortunately this case is not relevant for forest inventory (if we knew the forest in every point we do not need any statistics).

4) If the eigenvalues are uniformly bounded then, in particular, $n^{-2}\sum_{i=1}^{n}\lambda_i^2(\rho_o) = n^{-2}\|K(\rho_o)\|^2 \to 0$. For a regular grid this implies that the average correlation between all pairs tends to zero, i.e that the correlation range is small with respect to the dimension of the domain.

Using lemma (9.8) and equation (9.13) it is clear that, under the condition of theorem (9.15), one obtains a consistent estimate of the variance by setting:

$$\hat{\sigma}_n^2 = \sigma^2(\hat{\rho}_n)$$

In section (5.6) we considered the chi-square test

$$X^2 = Y^t P^\perp (P^\perp \Sigma(\theta_1) P^\perp)^+ P^\perp Y$$

based on the Moore-Penrose generalized inverse. Using lemma (9.9) this can be rewritten as

$$X^2 = Y^t P^\perp \tilde{A}(\tilde{A}^t P^\perp \Sigma(\theta_1) P^\perp \tilde{A})^{-1} \tilde{A} P^\perp Y$$

for some $\tilde{A}_{(n,n-q)}$. If we choose this matrix as $A_{(n,n-q)}$ and take $\theta_1 = (\sigma^2(\rho_1), \rho_1)$ we obtain, according to equation (9.13),

$$X^2 = \left(\sigma^2(\rho_1)\right)^{-1} Y^t P^\perp A \tilde{K}^{-1}(\rho_1) A^t P^\perp Y = n - q$$

Hence, if we take, in particular, as test matrix the REML estimate we always get exactly $n-q$. This shows why we cannot use this chi-square as a goodness-of-fit test.

To obtain asymptotic expressions for the asymptotic variances, the covariance matrix must satisfy further conditions, which are now outlined:

(9.17)

A.V 1 The mapping $R^{2n} \times (\sigma^2, \rho) \to \Sigma(\theta)$ is twice continuously differentiable.

A.V 2 $\lim_{n\to\infty} (n-q)^{-2} Tr\left(\tilde{\Sigma}^{-1}(\theta_o)\tilde{\Sigma}_{ij}(\theta_o)\right)^2 = 0 \quad , \quad \tilde{\Sigma}_{ij}(\theta) = \dfrac{\partial^2}{\partial\theta_i \partial\theta_j}\tilde{\Sigma}(\theta)$

A.V 3 $\lim_{n\to\infty} (n-q)^{-2} Tr\left(\tilde{\Sigma}^{-1}(\theta_o)\tilde{\Sigma}_i(\theta_o)\tilde{\Sigma}^{-1}(\theta_o)\tilde{\Sigma}_j(\theta_o)\right)^2 = 0, \quad \tilde{\Sigma}_i(\theta) = \dfrac{\partial}{\partial\theta_i}\tilde{\Sigma}(\theta)$

Remarks:

The assumption A.V 1 may not hold at a finite set of points, for instance with the spherical covariance. This can lead to multimodality of the likelihood (K. V. Mardia, A. J. Watkins, 1989). From a mathematical point of view one can always take, to any degree of accuracy, an infinitely differentiable regularization. In practice, such cases are potentially dangerous for iteration procedures based on derivatives, like Newton-Raphson.

Splitting the parameter vector into variance and correlation structure, conditions A.V 1 and A.V 2 hold mutatis mutandis for the correlation matrix $\tilde{K}(\rho_o)$ as well as, in obvious notation:

A.V 4 $\lim_{n\to\infty}(n-q)^{-2}Tr\left(\tilde{K}^{-1}(\rho_o)\tilde{K}_i(\rho_o)\right)^2=0$.

It is clear from lemmas (9.5) and (9.6) that conditions A.V. 2 and A.V 4 will hold when the absolute eigenvalues of the derivatives of the correlation matrix are bounded. Later on we shall give a heuristic argument showing that this also implies A.V. 3. K.V Mardia and R.J Marshall (1984) have shown that this is the case in the context of the full maximum likelihood. In any case, these conditions are rather technical and far less intuitive than the condition for consistency. In practice one can have reasonable hopes that the consistency conditions will suffice for everything.

We are now able to calculate the asymtotic variance by using standard arguments. It is easier to work with the complete restricted log likelihood function (see 9.11), i.e. with

$$g(Y,\theta):=-\frac{1}{2}\left\{\log\det\tilde{\Sigma}(\theta)+Y'P^{\perp}A\tilde{\Sigma}^{-1}(\theta)A'P^{\perp}Y\right\}$$

The gradiant g' is defined through $g_i'(\theta)=\frac{\partial}{\partial\theta_i}g(Y,\theta)$, and the Hessian matrix g'' through $g_{ij}''(\theta)=\frac{\partial^2}{\partial\theta_i\partial\theta_j}g(Y,\theta)$. Tedious but simple calculations based on lemma (9.7) yields:

(9.18)

$$E_{\theta_o}g_i'(\theta_o)=0, i=1,2....k$$

$$E_{\theta_o}g_i'(\theta_o)g_j'(\theta_o)=\frac{1}{2}Tr\left(\tilde{\Sigma}^{-1}(\theta_o)\tilde{\Sigma}_i(\theta_o)\tilde{\Sigma}^{-1}(\theta_o)\tilde{\Sigma}_j(\theta_o)\right)$$

$$E_{\theta_o}g''_{ij}(\theta_o)=-\frac{1}{2}Tr\left(\tilde{\Sigma}^{-1}(\theta_o)\tilde{\Sigma}_i(\theta_o)\tilde{\Sigma}^{-1}(\theta_o)\tilde{\Sigma}_j(\theta_o)\right)$$

Under the conditions A.V 1-A.V 3 and by repeated use of lemma (9.8) one obtains, after lenghty but simple manipulations of the trace operator, the convergence result

(9.19)

$\lim_{n\to\infty}n^{-1}\left(g''(\theta_o)-E_{\theta_o}g''(\theta_o)\right)=0$ in probability.

One has the Taylor expansion
$0=g'(\hat{\theta}_n)=g'(Y,\theta_o)+g''(Y,\tilde{\theta})(\hat{\theta}_n-\theta_o)$ where $\left|\tilde{\theta}-\theta_o\right|\le\left|\hat{\theta}_n-\theta_o\right|$

hence, because of (9.19) and the consistency theorem (9.15), we have, asymptotically, $\hat{\theta}_n - \theta_o = -\left(E_{\theta_o} g''(\theta_o)\right)^{-1} g'(Y,\theta_o)$. From (9.18) one obtains at once the asymptotic covariance matrix of the REML:

(9.20)

$$\Sigma_{\hat{\theta}_n} = \left(\frac{1}{2} Tr\left(\tilde{\Sigma}^{-1}(\theta_o)\tilde{\Sigma}_i(\theta_o)\tilde{\Sigma}^{-1}(\theta_o)\tilde{\Sigma}_j(\theta_o)\right)\right)^{-1}$$

It is worth noting that K.V. Mardia and R.J. Marshall (1984) obtained the same result for the full maximum likelihood but with $\Sigma(\theta_o)$ instead of $\tilde{\Sigma}(\theta_o)$. To get a consistent estimate of the asymptotic covariance matrix one replaces in (9.20) the true parameter θ_o by its estimate $\hat{\theta}_n$.

For the full maximum likelihood estimate K.V Mardia and R.J. Marshall (1984), based on results of B. T. Sweeting (1980), have proved asymtotic normality, so that the same holds for the REML because both estimates are asymtotically equivalent.

To get further insight into the asymptotic variance of the REML and in order to compare it later with the least square estimate, we shall restrict the discussion to stationary processes and express the results in terms of the spectral density (note that 9.20 does not require as such stationarity). We shall derive the main results with two different arguments and without bothering about the regularity conditions. To date, formal proofs are only available for stationary time series (M. Fox, M. J. Taqqu, 1987).

From now on, we restrict our attention to stationary processes with zero expectation (because, under regularity assumption, the estimation of the drift is asymptotically irrelevant for the calculation of the asymptotic variance) and to sampling schemes based on systematic rectangular grids. Note that this does not cover cluster sampling, a technique frequently used in forest inventory.

The sample points are identified with their integer coordinates with respect to the fundamental cell of the grid, i.e. we set:

$$x = (u,v) \in Z^2 \quad : u = 1,2....n_1 \, , v = 1,2...n_2 , n = n_1 n_2 \, .$$

The main idea of the first argument is to wrap the finite rectangular grid onto a torus by identifying $(n_1,v)=(1,v)\ \forall v$ and $(u,n_2)=(u,1)\ \forall u$, thus avoiding the boundary problems. If the grid mesh is constant and $n_1,n_2\to\infty$ the grid on the torus becomes "flatter" and, intuitively, can be viewed as a planar grid. This is a generalization of the time series technique identifying the line with a circle. In the spatial set-up this argument has been justified rigorously for a special class of markovian processes by P. A. Moran (1973) and used by many authors since (J. Besag, P. A. Moran, 1975; J. Besag, 1977; X. Guyon, 1982; K.V. Mardia, R.J. Marshall, 1984). Though intuitive and simple, this "wrapping on the torus" trick is not uncon-troversial in terms of the physical interpretation of the underlying process (M. Kendall, A. Stuart, J.K. Ord, 1983, p. 539). In any case, it has the advantage of simplifying the mathematics.

On the torus the covariance matrix is circular, i.e.

$$\operatorname{cov}\big((u_1,v_1);(u_2,v_2)\big)=c(u_1-u_2,v_1-v_2)=\operatorname{cov}\big((n_1-(u_1-u_2);(n_2-(v_1-v_2)\big)$$

It can be verified that this has the far reaching consequence that, within a model, all the covariance matrices commute, i.e.:

$$\Sigma(\theta_1)\Sigma(\theta_2)=\Sigma(\theta_2)\Sigma(\theta_1) \tag{9.21}$$

From a well-known result in linear algebra there exists then a fixed unitary matrix U ,$\overline{U}'U=I$, ($\overline{U}$ is the complex conjugate of U) which simultaneously diagonalizes all the covariance matrices, i.e.

$$\Sigma(\theta)=U^{-1}\Lambda(\theta)U\quad\forall\theta \tag{9.22}$$

The set of values of the spectral density, for the process on the torus, at the points $2\pi(\frac{i_1}{n_1},\frac{i_2}{n_2})\ \ i_k=1,2\ldots n_k\ \ ,k=1,2$ coincide with the set of eigenvalues in $\Lambda(\theta)$ (see for instance K.V. Mardia, R.J. Marshall, 1984). Since (9.22) implies at once

$$n^{-1}Tr\big(\Sigma^{-1}(\theta)\Sigma(\theta_o)\big)^2=n^{-1}Tr\big(\Lambda^{-1}(\theta)\Lambda(\theta_o)\big)^2=\frac{1}{n}\sum_{i=1}^{n}\frac{\lambda_i^2(\theta_o)}{\lambda_i^2(\theta)}$$

and therefore asymptotically the important result:

$$\lim_{n\to\infty}\frac{1}{n}Tr\big(\Sigma(\theta)^{-1}\Sigma(\theta_o)\big)^2=\frac{1}{4\pi^2}\int_{\Pi_2}\frac{f^2(\lambda,\theta_o)}{f^2(\lambda,\theta)}d\lambda \tag{9.23}$$

where $f(\lambda,\theta)=\sum_{l\in Z^2}\mathrm{cov}(l,\theta)\exp(il\lambda)$, $l\lambda=l_1\lambda_1+l_2\lambda_2$, is the spectral density of the original planar process and $\Pi_2=[0,2\pi]\times[0,2\pi]$. For time series this is precisely the famous result of M. Fox and M. J. Taqqu (1987). The consistency condition is therefore fulfilled for spectral densities satisfying $0<m_1<f(\lambda,\theta)<m_2<\infty \quad \forall\theta$.

We now assume that under further regularity conditions we can formally differentiate (9.22) with respect to θ, then one obtains the following relation:

(9.24)

$$\lim_{n\to\infty}\frac{1}{n}Tr\left(\Sigma^{-1}(\theta_o)\Sigma_i(\theta_o)\Sigma^{-1}(\theta_o)\Sigma_j(\theta_o)\right)=\frac{1}{4\pi^2}\int_{\Pi_2}\frac{f_i'(\lambda,\theta_0)f_j'(\lambda,\theta_0)}{f^2(\lambda,\theta_0)}d\lambda$$

where $f_i'(\lambda,\theta)=\frac{\partial f(\lambda,\theta)}{\partial\theta_i}$ $i=1,2...k$ are the partial derivatives of the bivariate spectral density. Using (9.20) we can therefore state the "theorem":

(9.25)

Under regularity conditions $\hat{\theta}_n-\theta_o$ is asymptotically multivariate normal with covariance matrix:

$$\Sigma_{\hat{\theta}}=\frac{2}{n}\left(\frac{1}{4\pi^2}\int_{\Pi_2}\frac{f_i'(\lambda,\theta_0)f_j'(\lambda,\theta_0)}{f^2(\lambda,\theta_0)}d\lambda\right)^{-1}=\frac{2}{n}\left(\frac{1}{4\pi^2}\int_{\Pi_2}\frac{\partial}{\partial\theta_i}\log f(\lambda,\theta_o)\frac{\partial}{\partial\theta_j}\log f(\lambda,\theta_o)d\lambda\right)^{-1}$$

For time series this is precisely the result given in (R. Fox, M.S. Taqqu, 1986, lemma on page 525), which is also valid for certain unbounded spectral densities resulting from long range dependence (like fractional brownian motion).

It can be expected that (9.25) will hold under mild regularity assumptions on the spectral density, not only for stationary processes but also after filtering out the drifts.

For autoregressive processes in the plane, (9.25) can be found in a famous paper of P. Whittle going back to 1954, whereas X. Guyon (1982) derives (9.25) in arbitrary dimensions for an estimate based on the modified spectrogram. The asymptotic expression (9.25) will be extremely useful to calculate the asymptotic relative efficiency of restricted maximum likelihood and least squares estimates of the covariance.

We now outline the second argument, which does not have the drawback, from a physical point of view, of distorting the covariance matrix by wrapping it on the torus; on the other hand it requires to work instead with an infinite sample from the onset. For stationary covariances matrices A, B on Z^2, the spectral theorem states that:

$$A_{p,q} = \frac{1}{4\pi^2} \int_{\Pi_2} \exp(-i(p-q)\lambda) f_A(\lambda) d\lambda \quad p,q \in Z^2$$

$$B_{q,r} = \frac{1}{4\pi^2} \int_{\Pi_2} \exp(-i(q-r)\lambda) f_B(\lambda) d\lambda \quad q,r \in Z^2$$

where $f_A(\lambda), f_B(\lambda)$ are the spectral densities defined now on $\Pi_2 = [-\pi,\pi] \times [-\pi,\pi]$ to comply with standard convention. Because $f_A(\lambda), f_B(\lambda)$ are even functions one can write the above Fourier coefficients as:

$$A_{p,q} = \frac{1}{4\pi^2} \int_{\Pi_2} \exp(-iq\lambda) f_{A,p}(\lambda) d\lambda \quad , f_{A,p}(\lambda) = \exp(ip\lambda) f_A(\lambda)$$

$$B_{q,r} = \frac{1}{4\pi^2} \int_{\Pi_2} \exp(-iq\lambda) f_{B,r}(\lambda) d\lambda \quad , f_{B,r,p}(\lambda) = \exp(ip\lambda) f_B(\lambda)$$

For spectral densities in $L^2(\Pi_2)$, i.e. square integrable, the infinite product matrix is well defined through

$$(AB)_{p,r} = \sum_{q \in Z^2} A_{p,q} B_{q,r} \quad p,r \in Z^2$$

Since the bivariate exponentials form a complete orthogonal system for $L^2(\Pi_2)$ the Parseval relation (see e.g. N. Wiener, 1958, p. 44) yields:

$$(AB)_{p,r} = \frac{1}{4\pi^2} \int_{\Pi_2} f_{A,p}(\lambda) \bar{f}_{B,r}(\lambda) d\lambda = \frac{1}{4\pi^2} \int_{\Pi_2} \exp(-i(p-q)\lambda) f_A(\lambda) f_B(\lambda) d\lambda$$

Thus, we have an Hilbert space isomorphism between the multiplication of doubly infinite stationary covariance matrices on Z^2 and the product of spectral densities. In particular, if $f_A^{-1}(\lambda)$ admits a Fourier expansion, then:

(9.26)

$$A^{-1}_{p,q} = \frac{1}{4\pi^2} \int_{\Pi_2} \frac{\exp(-i(p-q)\lambda)}{f_A(\lambda)} d\lambda$$

By induction on the number of matrices, and by tranlating the above relations into asymptotic results for finite matrices as

$n_1, n_2, n \to \infty$, one obtains in particular the important relation for the trace:

(9.27)

$$\lim_{n\to\infty} \frac{1}{n} Tr(AB)^k = \frac{1}{4\pi^2} \int_{\Pi_2} \left(f_A(\lambda) f_B(\lambda)\right)^k d\lambda \quad k = 1,2...$$

$$\lim_{n\to\infty} \frac{1}{n} Tr(ABCD) = \frac{1}{4\pi^2} \int_{\Pi_2} f_A(\lambda) f_B(\lambda) f_C(\lambda) f_D(\lambda) d\lambda$$

The first formula has been proved to hold for time series under fairly general conditions (F. Fox, M.S. Taqqu, 1987). The formulae (9.26-9.27) yield at once (9.24).

It is now clear that regularity conditions, essentially ensuring that $f, f^{-1}, f_j' \in L^2(\Pi_2)$ have Fourier coefficients generating $\Sigma, \Sigma^{-1}, \Sigma_j$, will imply (9.23) and (9.25).

It is worth noting that these important results have been obtained by two very different techniques.

From a pragmatic point of view, we can summarize the results of this chapter as follows:

1. If the estimated correlation range is small with respect to the dimension of the domain, the REML is reliable, asymptotic validity has been reached, and kriging is primarily relevant for local estimation.

2. If not, this might suggest that the drift model is not adequate and should, if possible, be revised. If the drift is deemed correct and the confidence interval yields range values comparable to the domain dimension, or even larger, then the asymptotic validity or the correlation model may be at fault. In any case caution is required.

3. Estimation of the covariance structure with small moving neighbourhoods can yield totally unreliable results.

9.3 Least Squares Estimate of the Residual Covariance

In this section, we investigate the properties of the least square estimate (for short LS) of the residual covariance matrix as defined in section (5.6); this estimate minimizes the maximum discrepancy between the empirical and model-dependent variances over a canonical set of authorized linear combinations filtering out the drift. It can also be viewed as the mathematically correct version of the widely used "fitting by eye" technique of the empirical variogram. It is computationally simpler than the REML as it does not require the inversion of the, sometimes very large, covariance matrices. We shall see that the LS is consistent under essentially the same conditions as the REML, but that it is less efficient, particularly under increasingly stronger correlation. These results appear to be new, even if intuitively expected and related to similar findings in slightly different contexts.

Most of the underlying concepts and the notation have been previously defined in section 9.2, so that only the main points will be given.

The LS estimate of the covariance matrix $\Sigma(\theta)$ minimizes over θ the expression:

$$\left\| rr' - P^{\perp}\Sigma(\theta)P^{\perp} \right\|^2 = \left\| P^{\perp}(YY' - \Sigma(\theta))P^{\perp} \right\|^2 = \left\| rr' - P^{\perp}\Sigma(\theta) \right\| \qquad (9.28)$$

(the last equality holds because of $P^{\perp}r = r, P^{\perp}P^{\perp} = P^{\perp}$).

Taking the partial derivatives with respect to $\theta = (\sigma^2, \rho)$ yields the "normal equations":

$$Tr\left(\Sigma(\theta)\frac{\partial}{\partial\theta}\overline{\Sigma}(\theta) \right) = Y'\frac{\partial}{\partial\theta}\overline{\Sigma}(\theta)Y, \quad \overline{\Sigma}(\theta) = P^{\perp}\Sigma(\theta)P^{\perp} \qquad (9.29)$$

Using $Tr\left(K(\rho)P^{\perp}K(\rho)P^{\perp}\right) = Tr\left(\overline{K}^2(\rho)\right)$, $\overline{K}(\rho) = P^{\perp}K(\rho)P^{\perp}$, the partial derivative with respect to σ^2 yields the relation at the extremum:

(9.30)

$$\sigma^2(\rho) = \frac{Y^t \overline{K}(\rho) Y}{Tr\left(\overline{K}^2(\rho)\right)}$$

which should be compared with (9.13) for the REML. Note that:

$$E_{\rho_o}\left(\sigma^2(\rho)\right) = \frac{Tr\left(\overline{K}(\rho)\overline{K}(\rho_o)\right)}{Tr\left(\overline{K}^2(\rho)\right)}$$

According to Lemma (9.8) $\sigma^2(\rho)$ will converge towards its expected value $E_{\rho_o}\left(\sigma^2(\rho)\right)$ if

$$\lim_{n\to\infty} \frac{1}{n^2} Tr\left(\overline{K}(\rho)\overline{K}(\rho_o)\right)^2 = 0$$

Substituting (9.30) into (9.28) yields after some algebraic manipulations:

(9.31)

$$\left\| rr^t - \sigma^2(\rho) P^{\perp} K(\rho) \right\|^2 = \left(rr^t\right)^2 - \frac{\left(Y^t \overline{K}(\rho) Y\right)^2}{Tr\left(\overline{K}^2(\rho)\right)}$$

Set

(9.32)

$$f_n(\rho) = 2\log\left(Y^t \overline{K}(\rho) Y\right) - \log Tr\left(\overline{K}^2(\rho)\right)$$
$$\phi_n(\rho) = 2\log Tr\left(\overline{K}(\rho)\overline{K}(\rho_o)\right) - \log Tr\left(\overline{K}^2(\rho)\right)$$

Minimizing (9.31) is equivalent to maximizing $f_n(\rho)$. Consider:

$$f_n(\rho) - \phi_n(\rho) = 2\log\frac{\sigma^2(\rho)}{E_{\rho_o}\sigma^2(\rho)}$$

This converge pointwise to zero in probability and also uniformly under the same condition as 9.15 (iv) (using the same arguments and notations as in the proof of 9.15). Furthermore, $\phi_n(\rho_o) = \log\left(Tr\left(\overline{K}^2(\rho_o)\right)\right)$ is the absolute maximum of $\phi_n(.)$; indeed,

$$\phi_n(\rho) \le \phi_n(\rho_o) \Leftrightarrow \left(Tr\, \overline{K}(\rho)\overline{K}(\rho_o) \le Tr\, \overline{K}^2(\rho)\overline{K}^2(\rho_o) \right)$$

which by lemma (9.1) is true, with equality if $\exists\mu \;\; \overline{K}(\rho) = \mu\overline{K}(\rho_o)$. Let $\hat{\rho}_n$ be the argument at the absolute maximum of $f_n(\rho)$, i.e. at the absolute minimum of (9.31) and hence the least square estimate. It is now clear that using exactly the same arguments (and notation) as in the proof of the consistency of the REML (9.15) we can state the following theorem:

Least squares consistency theorem

(9.33)

Let $U \subset R^{k-1}$ be a compact set whose interior contains the true value ρ_o of the correlation parameter. Consider an increasing sequence of sample points such that the following conditions are satisfied:

(i) $\forall \rho \in U \quad \lim_{n\to\infty} \frac{1}{n^2} Tr\left(\overline{K}(\rho)\overline{K}(\rho_o)\right)^2 = 0$

(ii) The mapping $\rho \in U \to K(\rho) \in R^{n^2}$ is injective and continuous.

(iii) $\forall \rho_1, \rho_2 \in U \quad \left(\exists \mu > 0 \quad \overline{K}(\rho_1) = \mu \overline{K}(\rho_2)\right) \Rightarrow \rho_1 = \rho_2$

(iv) Let $B_j \quad j = 1,2.....K(M)$ a covering of $U \subset R^{k-1}$ with spheres of radius M^{-1}. There exists a sequence $\varepsilon_M \geq 0, \lim_{M\to\infty} \varepsilon_M = 0$ with

$$\forall M \quad \lim_{n\to\infty} P_{\rho_o}\left\{\sup_{\rho \in B_j}\left|\log\frac{\sigma^2(\rho)}{E_{\rho_o}\sigma^2(\rho)}\right| > \varepsilon_M\right\} = 0 \quad \forall j = 1,2....K(M)$$

Let $\hat{\rho}_n \in U$ be the argument at the absolute maximum of the random function

$$f_n(\rho) = 2\log\left(Y^t \overline{K}(\rho) Y\right) - \log\left(Tr\left(\overline{K}^2(\rho)\right)\right).$$

Then $\hat{\rho}_n$ converges in probability towards the true value ρ_o.

Since $P^{\perp}$ is a symmetrical idempotent matrix of rank $n-q$, we can apply the second part of the Poincaré separation theorem (lemma 9.6) and proceed as for the REML to show that under regularity conditions, condition (ii) will imply asymptotically condition (iii) and that the minimum distance between the sample points must be smaller than the true range, in order to avoid a "flat" sum of squares over identity matrices. Again, it can be expected that condition (i) will generally imply condition (iv).

Using lemma (9.5) one gets at once the following result:

(9.34)

If condition (i) implies (iv) and if conditions (ii),(iii) hold, then the uniform boundedness of the eigenvalues from above, i.e.

$$\exists m_2 \quad 0 \leq \lambda_n(\rho) \leq m_2 < \infty \quad \forall \rho \in U$$

implies that the LS estimate is consistent.

The condition on the smallest eigenvalue is no longer necessary, but this is of purely academic interest. It is now clear that the remarks 1,3,4 following theorem (9.16) also apply; for time series the result of M. Fox and M. J. Taqqu (1987) also holds. In other words, restricted maximum likelihood and least squares are consistent under essentially the same conditions.

We are now able to calculate the asymptotic variance by using arguments similar to the REML. It is easier to work with

$$g(Y,\theta) := -2Y^t\overline{\Sigma}(\theta)Y + Tr\left(\overline{\Sigma}^2(\theta)\right) = \left\|P^{\perp}(YY^t - \Sigma(\theta))P^{\perp}\right\|^2 - Y^tP^{\perp}YY^tP^{\perp}Y,$$

the second equality resulting from simple algebra. Obviously the LS estimate yields the absolute minimum of this function. The gradiant g' is defined through $g_i'(\theta) = \frac{\partial}{\partial\theta_i}g(Y,\theta)$, and the Hessian matrix g'' through $g_{ij}''(\theta) = \frac{\partial^2}{\partial\theta_i\partial\theta_j}g(Y,\theta)$.

In perfect analogy with the REML we now assume that the following conditions hold:

A.V.L 1 The mapping $R^{2n}\times(\sigma^2,\rho)\to\Sigma(\theta)$ is twice continuously differentiable.

A.V.L 2 $\lim_{n\to\infty} n^{-2}Tr\left(\overline{\Sigma}(\theta_o)\overline{\Sigma}_{ij}(\theta_o)\right)^2 = 0$, with $\overline{\Sigma}_{ij}(\theta) = \frac{\partial}{\partial\theta_i\partial\theta_j}\overline{\Sigma}(\theta)$

A.V.L 3 $\lim_{n\to\infty} n^{-2}Tr\left(\overline{\Sigma}(\theta_o)\overline{\Sigma}_i(\theta_o)\overline{\Sigma}(\theta_o)\overline{\Sigma}_j(\theta_o)\right)^2 = 0$, with $\overline{\Sigma}_i(\theta) = \frac{\partial}{\partial\theta_i}\overline{\Sigma}(\theta)$

Tedious but simple calculations based on lemma (9.7) yields:

(9.35)

$$E_{\theta_o}\, g_i'(\theta_o) = 0, i = 1,2....k$$
$$E_{\theta_o} g_i'(\theta_o)g_j'(\theta_o) = 8Tr\left(\overline{\Sigma}(\theta_o)\overline{\Sigma}_i(\theta_o)\overline{\Sigma}(\theta_o)\overline{\Sigma}_j(\theta_o)\right)$$
$$E_{\theta_o} g''_{ij}(\theta_o) = 2Tr\left(\overline{\Sigma}_i(\theta_o)\overline{\Sigma}_j(\theta_o)\right)$$

Under the conditions A.V.L 1 - A.V.L 3 and by repeated use of lemma (9.8) one obtains, after lenghty but simple manipulations of the trace operator, the convergence result:

(9.36)

$\lim_{n\to\infty} n^{-1}\left(g''(\theta_o) - E_{\theta_o} g''(\theta_o)\right) = 0$ in probability.

One has the Taylor expansion:

$0 = g'(\hat{\theta}_n) = g'(Y,\theta_o) + g''(Y,\tilde{\theta})(\hat{\theta}_n - \theta_o)$ where $\left|\tilde{\theta} - \theta_o\right| \le \left|\hat{\theta}_n - \theta_o\right|$

Hence, because of (9.36) and the consistency theorem (9.33), we have, asymptotically, $\hat{\theta}_n - \theta_o = -\left(E_{\theta_o} g''(\theta_o)\right)^{-1} g'(Y,\theta_o)$. From (9.35) one obtains at once the asymptotic covariance matrix of the least square estimate:

(9.37)

$$\Sigma_{\hat{\theta}} = 2\left(Tr\overline{\Sigma}_i(\theta_o)\overline{\Sigma}_j(\theta_o)\right)^{-1} Tr\left(\overline{\Sigma}(\theta_o)\overline{\Sigma}_i(\theta_o)\overline{\Sigma}(\theta_o)\overline{\Sigma}_j(\theta_o)\right)\left(Tr\overline{\Sigma}_i(\theta_o)\overline{\Sigma}_j(\theta_o)\right)^{-1}$$

Substituting $\hat{\theta}_n$ for θ_o into (9.37) yields a consistent estimate of the asymptotic covariance.

Using (9.26) and (9.27), it is straightforward to rewrite the equation (9.37) in terms of the spectral density to obtain the "theorem":

(9.38)

Under regularity conditions $\hat{\theta}_n - \theta_0$ follows asymptotically a multivariate normal normal distribution with covariance matrix:

$$\Sigma_{\hat{\theta}} = \frac{2}{n}\left(\frac{1}{4\pi^2}\int_{\Pi_2} f_i'(\lambda,\theta_0) f_j'(\lambda,\theta_0) d\right)^{-1}\left(\frac{1}{4\pi^2}\int_{\Pi_2} f^2(\lambda,\theta_0) f_i'(\lambda,\theta_0) f_j'(\lambda,\theta_0) d\lambda\right)\left(\frac{1}{4\pi^2}\int_{\Pi_2} f_i'(\lambda,\theta_0) f_j'(\lambda,\theta_0) d\right)^{-1}$$

As for the REML, (9.38) can be expected to hold under weak regularity conditions on the spectral density, not only for stationary processes, but also after filtering out the drifts.

We now state the important result:

(9.39)

The restricted maximum likelihood estimate of the covariance structure is more efficient than the least squares estimate, i.e. the matrix

$$\Sigma_{\hat{\theta},LS} - \Sigma_{\hat{\theta},REML}$$

is positive definite.

Proof:
It is equivalent to show that $\Sigma^{-1}_{\hat{\theta},REML} - \Sigma^{-1}_{\hat{\theta},LS}$ is positive definite (J.R. Magnus, H. Neudecker, 1988, p.22). Using (9.25) and (9.38) this is equivalent to show this for the matrix

$$\int \frac{f_i' f_j'}{f^2} - \left(\int f_i' f_j'\right)\left(\int f^2 f_i' f_j'\right)^{-1}\left(\int f_i' f_j'\right)$$

where we have simplified the notation in an obvious way.
We now view the point λ as a uniform random point in Π_2 and consider the $2k$ dimensional random vector .

$$\Psi = \left(\frac{f_i(\lambda,\theta_o)}{f(\lambda,\theta_o)}, f(\lambda,\theta_o) f_i(\lambda,\theta_o)\right)^t, i = 1,2...k \quad .$$

Then, the matrix

$$E_\lambda \Psi\Psi^t = \frac{1}{4\pi^2}\begin{pmatrix} \int \frac{f_i' f_j'}{f^2} d\lambda & \int f_i' f_j' d\lambda \\ \int f_i' f_j' d\lambda & \int f^2 f_i' f_j' d\lambda \end{pmatrix} =: \begin{pmatrix} V & \Delta \\ \Delta & J \end{pmatrix}$$

is positive definite, hence the matrix $BE_\lambda \Psi\Psi^t B^t$ is also positive definite for any matrix B.
Choosing $B = \begin{pmatrix} I_k & -\Delta J^{-1} \\ 0 & J^{-1} \end{pmatrix}$

leads to $BE_\lambda \Psi\Psi^t B^t = \begin{pmatrix} V - \Delta J^{-1}\Delta & 0 \\ 0 & J^{-1} \end{pmatrix}$ positive definite and therefore also $V - \Delta J^{-1}\Delta$, which is precisely the result.
Note that this proof is absolutely similar to the proof of the multi-dimensional Cramer-Rao inequality.

H. R. Künsch (1980, p. 85) obtained the same result in a different context: he considered spatial models of the form $X_i = \sum_{j \neq i} a_{ij}(\theta) X_j + U_i$ (with correlated U_i) and the maximum likelihood as well as least squares estimates of θ.

If we define the relative efficiency as the ratio of the determinant of the asymptotic variances, we obtain by the previous results the following theorem:

Asymptotic relative efficiency

(9.40)

$$1 \le \frac{\det \Sigma_{\hat{\theta},LS}}{\det \Sigma_{\hat{\theta},REML}} = \frac{\det\left(\int_{\Pi_2} f^2(\lambda,\theta_o) f_i'(\lambda,\theta_o) f_j'(\lambda,\theta_o) d\lambda\right)\left(\int_{\Pi_2} f^{-2}(\lambda,\theta_o) f_i'(\lambda,\theta_o) f_j'(\lambda,\theta_o) d\lambda\right)}{\left(\det \int_{\Pi_2} f_i'(\lambda,\theta_o) f_j'(\lambda,\theta_o) d\lambda\right)^2}$$

This result allows us to get better insight into the dependence of the efficiency on parameters. For a model with a sill, range and nugget effect, the spectral density function can be written as :

$$f(\lambda,\theta) = \sigma^2((1-\alpha)g(\lambda,\rho)+\alpha) \quad ,0 \le \alpha \le 1, \frac{1}{4\pi^2}\int_{\Pi_2} g(\lambda,\rho)d\lambda = 1 \;\forall \rho$$

Even with this simple model it is extremely difficult to deal with the three parameters simultanously, so that we consider one parameter at a time, given that the other two as known. It is then easily verified that (9.40) yields the following properties:

$$1 \le e(\sigma^2|\rho,\alpha) = e_1(\rho,\alpha), \quad \lim_{\alpha\to 1} e_1(\rho,\alpha) = 1$$
$$1 \le e(\alpha|\sigma^2,\rho) = e_2(\rho,\alpha), \quad \lim_{\alpha\to 1} e_2(\rho,\alpha) = 1$$
$$1 \le e(\rho|\sigma^2,\alpha) = e_3(\rho,\alpha), \quad \lim_{\alpha\to 1} e_3(\rho,\alpha) = 1$$

The efficiency therefore does not depend on the sill and tends to one with increasing nugget effect.

To analyse further the efficiency with respect to the range parameter, we replace the product of the integrals by double integrals in (9.40), symmetrize the numerator and apply the mean value theorem since all integrands are positive. This then yields:

$$\exists \lambda_1,\lambda_2 \in \Pi_2 \quad e_3(\rho,\alpha) = \frac{1}{2}\left[\left(\frac{(1-\alpha)g(\lambda_1,\rho)+\alpha}{(1-\alpha)g(\lambda_2,\rho)+\alpha}\right)^2 + \left(\frac{(1-\alpha)g(\lambda_2,\rho)+\alpha}{(1-\alpha)g(\lambda_1,\rho)+\alpha}\right)^2\right]$$

Using the equivalent of Szegö's theorem (i.e spectral density on the torus gives the eigenvalues), we get from the previous equation the approximate upper bound for large samples:

(9.41)

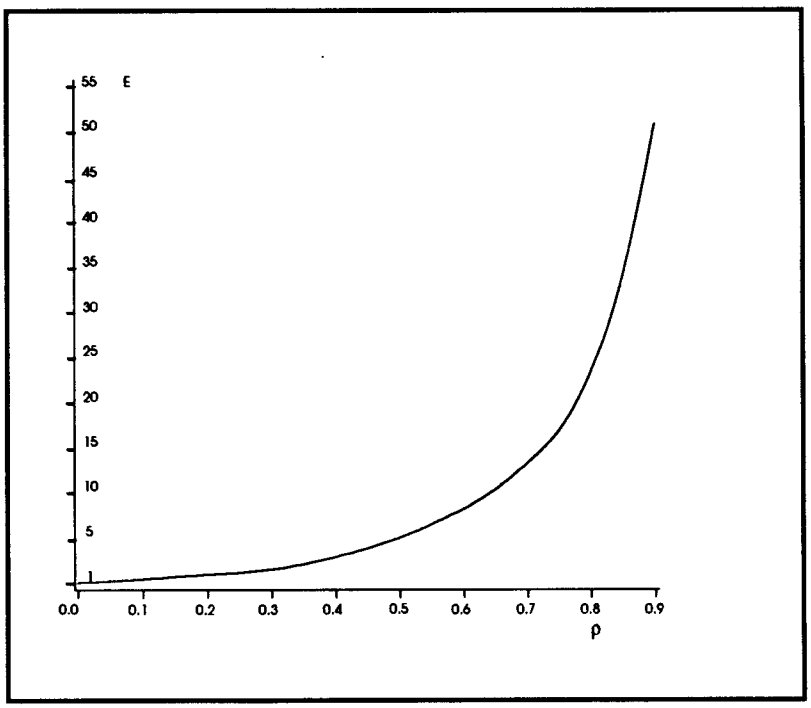
$$e_3(\rho,\alpha) < \frac{1}{2}\left[\left(\frac{(1-\alpha)\lambda_n(\rho)+\alpha}{(1-\alpha)\lambda_1(\rho)+\alpha}\right)^2 + 1\right]$$

This means: the stronger the correlation, the greater the advantage of maximum likelihood over least square; the larger the nugget effect, the smaller the difference.
To illustrate this effect, we briefly consider a time series with exponential covariance, i.e $c(k,\rho)=\rho^k$. The spectral density is found to be:

$$g(\lambda,\rho)=\frac{1-\rho^2}{1-2\rho\cos\lambda+\rho^2} \quad \lambda\in[-\pi,\pi]$$

Fig. 28 displays the relative efficiency as a function of the correlation ρ at unit lag in the absence of nugget effect. The curve is based on numerical integration. It is obvious that the least square estimate is loosing much efficiency for correlation over 0.5 and even dramatically so for correlation beyond 0.8. It can also be verified that the upper bound (9.41) is far too pessimistic for a correlation beyond 0.2 (e.g. for $\rho=0.6$ the bound is 128.5).

Fig. 28: Ratio E of least square versus maximum likelihood variance of the estimated correlation in markovian time series.



From a pragmatic point of view we can summarize the previous results as follows:

The least square estimate of the covariance (or variogram) is the mathematically correct version of the "fitting by eye" technique. It is consistent under essentially the same conditions as the restricted maximum likelihood, easier to compute, can be used in the non-gaussian case, but can suffer large efficiency losses in the presence of strong spatial correlation. If possible, a one step iteration with maximum likelihood should be performed using the least square estimate as a starting value. As for maximum likelihood, caution is required if the estimated correlation range is not small with respect to the dimension of the domain.

10 BIBLIOGRAPHY.

M. ARMSTRONG, (1984). Problems with universal kriging, Mathematical Geology, 16, pp. 101-108.

BLUEPACK, (1990). Manual, ENSM, Paris.

J. BERAN, (1992). A goodness-of-fit test for time series with long range dependence. J.R. Statist. Soc. B, 54, pp. 749-760.

J. BESAG, P. A. MORAN, (1975). On the estimation and testing of spatial interactions in gaussian lattice processes. Biometrika, 62, pp. 555-562.

J. BESAG, (1977). Errors in variables estimation for gaussian lattice schemes. J.R. Statist. Soc. B, 39, pp. 73-78.

J. BOUCHON, (1979). Structure des peuplements forestiers, Ann.Sci.Forest., 36, pp. 175-209.

P. CHAUVET, (1991). Vers les modèles non stationnaires. Aide-mémoire de géostatistique linéaire, Cahier de géostatistique, Fasc.2, Centre de Morphologie Mathématique , Fontainebleau, France.

J.P. CHILES, (1977). Géostatistique des phénomènes non stationnaires. Thèse de docteur-ingénieur, Université de Nancy I.

I. CLARK et al, 1989. MUCK: a novel approach to co-kriging. In Procedings of the conference on geostatistical , sensitivity, and uncertainty methods for ground-water flow and radionuclide transport modelling, B. E. Buxton, ed. Batelle Press, Columbus, OH, pp. 473-493.

G. CHRISTAKOS, (1992). Random field models in earth sciences. Academic Press, Inc. , New York.

R. CHRISTENSEN, (1990a). The equivalence of predictions from universal kriging and intrinsic random functions kriging, Mathematical Geology, 22, pp. 655-663.

R. CHRISTENSEN, (1990b). Linear models for multivariate, time series, and spatial data. Springer-Verlag, New York.

D.R. COX (1967). Fieller's theorem and a generalization, Biometrika, 54, pp. 567-572.

N. CRESSIE, (1990). The origin of kriging, Mathematical Geology, 22, pp. 239-252.

N. CRESSIE, (1991). Statistics for spatial data, John Wiley & Sons, Inc., New York.

P. DELFINER, (1976). Linear estimation of non stationnary phenomena, Advanced Geostatistics in the Mining Industry, eds M.Guaroscio et al:Dordrecht, pp. 49-68.

P. DIAMOND, M. ARMSTRONG, (1984). Robustness of variograms and conditionning of kriging matrices. Mathematical geology, 16, pp. 809-822.

C.R. DIETRICH, M.R. OSBORNE, (1991). Estimation of covariance parameters in kriging via restricted maximum likelihood. Mathematical geology, 23, pp. 119-135.

P. DUPLAT, G. PEYROTTE, (1981). Inventaire et accroissement, Annexe 4, Office National des Forêts, Section technique, Paris.

R. FOX, M.S. TAQQU, (1986). Large sample properties of parameter estimates for strongly dependent stationary time series. Ann. Stat.,14, pp. 517-532.

R. FOX, M.S. TAQQU, (1987). Central limit theorem for quadratic forms in random variables having long-range dependence. Probability theory and related fields, 74, pp. 213-240.

A. GALLI et al, (1987). Applied Geostatistics, Summer School 1987, ENSM, Paris.

N. GELFAND, Y. VILENKINE, (1964). Generalized random functions, Vol. 4, Academic Press, New York and London.

GIUDICELLI et al, (1972). Applications de la théorie des processus aléatoires à l'estimation de la précision d'un inventaire forestier par échantillonnage systématique. Ann.Sci.Forest., 29, pp. 267-293.

G.H. GOLUB, C.F. van LOAN, (1983). Matrix Computation,North Oxford Academic, Oxford.

U. GRENANDER, M. ROSENBLATT, (1984). Statistical analysis of stationary time series. Chelsea Publishing Company, New York.

J.J. de GRUIJTER, C.J.F. ter BRAAK, (1990). Model-free estimation from spatial samples:a reappraisal of classical sampling theory, Mathematical Geology, 22, pp. 407-415.

D. GUIBAL, (1973). L'estimation des Oukoumés du Gabon, note interne 333, Centre de Morphologie Mathématique, Fontainebleau, France.

I. GUIKMAN, A. SKOROHOD, (1980). Introduction à la théorie des processus aléatoires, Editions MIR, Moscou.

X. GUYON, (1982). Parameter estimation for a stationary process on a d-dimensional lattice, Biometrika, 69, pp. 95-105.

X. GUYON, (1993). Champs aléatoires sur un réseau, modélisation, statistique et applications, Mason, Paris.

F.R. HAMPEL, (1987). Data analysis and self-similar processes. Proc. 46th session Int. Statist. Inst., Tokyo, book 4, pp. 235-254.

F. HOUILLER, (1986). Echantillonnage et modélisation de la dynamique des peuplements forestiers, thèse de doctorat, Université Claude Bernard, Lyon I.

A. JOST, (1993). Geostatistische Analyse des Stichprobenfehlers systematischer Stichproben, PhD thesis, University of Freiburg in Breisgau.

A.G. JOURNEL, C. HUIJBREGTS, (1978). Mining Geostatistics, Academic Press, London.

A.G. JOURNEL, M.E. ROSSI, (1990). When do we need a trend model in kriging, Mathematical Geology, 21, pp. 715-739.

E. KAUFMANN, (1992). Tree volume estimation and sample tree selection in the Swiss NFI. IUFRO S4.02 proceedings of Ilvessalo symposium on national forest inventory, Finnish Forest Research Institute, University of Helsinki, pp. 185-194.

M. KENDALL, A. STUART, J.K. ORD, (1983). The advanced theory of statistics, Charles Griffin & Co Ltd, London.

P.K. KITANIDIS, (1985). Minimum variance unbiased quadratic estimation of covariances of regionalized variables. Mathematical Geology, 17, pp. 192-208.

H.R. KUENSCH, (1980). Reellwertige Zufallsfelder auf einem Gitter: Interpolationsprobleme, Variationsprinzip und statistische Analyse. ETH Zürich PhD thesis nr. 6648.

J.R. MAGNUS, H. NEUDECKER, (1988). Matrix differential calculus with applications in statistics and econometrics, John Wiley & Sons, New York.

D. MANDALLAZ et al, (1986). Dépérissement des forêts:essai d'analyse des dépendances. Ann.Sci.For., 43(4), pp. 441-458.

D. MANDALLAZ, (1991). A unified approach to sampling theory for forest inventory based on infinite population and superpopulation models. PhD Thesis no 9378, ETH Zürich, Chair of forest management and planning.

D. MANDALLAZ, (1992). Optimization of general double sampling schemes in infinite populations:an application to forest inventory. Proceedings IUFROS 4.11 Conference, London.

P. MARBEAU, (1976). Géostatistique forestière:état actuel et développements nouveaux pour l'aménagement en forêt tropicale thèse, Centre de Morphologie Mathématique, ENSM, Paris.

K.V. MARDIA, R.J. MARSHALL (1984). Maximum likelihood estimation of models for residual covariance in spatial regression, Biometrika, 71, pp. 135-148.

K.V. MARDIA, A.J. WATKINS, (1989). On multimodality of the likelihood in the spatial linear model, Biometrika, 76, pp. 289-295.

R.J. MARSHALL, K.V. MARDIA, (1985). Minimum norm quadratic estimation of components of covariance. Mathematical Geology, 17, pp. 517-525.

B. MATERN, (1960). Spatial variation, report of the Forest Research Institute of Sweden, 49, pp. 1-144.

G. MATHERON, (1962). Traité de géostatistique appliquée, tome I, mémoire du BRGM, no 14, Editions Technip, Paris.

G. MATHERON, (1963). Traité de géostatistique appliquée, tome II, mémoire du BRGM, no 24, Editions BRGM, Paris.

G. MATHERON, (1965). La théorie des variables regionalisées et ses applications, Masson, Paris.

G. MATHERON, (1970). La théorie des variables regionalisées et ses applications, Cahiers du centre de morphologie mathématique no 5, Fontainebleau, France.

G. MATHERON, (1973). The intrinsic random functions and their applications, Advances in Applied Probability, 5, pp. 439-468.

P.A. MORAN, (1973). A gaussian markovian process on a square lattice. J. Appl. Prob., 19, pp. 54-62.

D.E. MYERS, (1982). Matrix formulation of co-kriging, Mathematical Geology, 14, pp. 49-257.

D.E. MYERS, (1989). To be or not to be...stationary? That is the question, Mathematical Geology, 21, pp. 347-362.

S.P. NEUMANN, E.N. JACOBSON, (1984). Analysis of non intrinsic variability by residual kriging with applications to regional ground water levels, Mathematical Geology, 16, pp. 499-521.

D. POSA, (1989). Conditioning of the stationary kriging matrices for some well-known covariance models. Mathematical Geology, 21, pp. 755-765.

H. RAMIREZ-MALDONADO, (1988). On the relevance of geostatistical theory and methods to forest inventory problems, PhD Thesis, University of Georgia.

C.R. RAO, (1967). Linear statistical inference and its applications, John Wiley & Sons, Inc., New York.

B.D. RIPLEY, (1988). Statistical inference for spatial processes, Cambridge University Press, Cambridge.

J. RIVOIRARD, (1984). Le comportement des poids de krigeage, thèse de docteur-ingénieur, ENSM, Paris.

J. RIVOIRARD, (1989). Introduction au krigeage disjonctif et à la géostatistique non linéaire, Centre de Géostatistique, bibliogr. C-139, Fontainebleau, France.

SAS, (1987). SAS user guide,SAS Institute, Cary, N. C.

M.L. STEIN, (1987). Minimum norm quadratic estimation of spatial variograms, JASA, 82, pp. 765-772.

M.L STEIN, M.S. HANDCOCK, (1989). Some asymptotic properties of kriging when the covariance is miss-specified. Mathematical Geology, 21, pp. 171-190.

D. STOYAN, W.S. KENDALL, J. MECKE, (1987). Stochastic geometry and its applications, John Wiley & Sons, Inc., New York.

T. J. SWEETING, (1980). Uniform asymtotic normality of the maximum likelihood estimator. Ann. Stat., pp. 1375-1380.

E. TOMPPO, (1986). Models and methods for analysing spatial pattern of trees, PhD Thesis no 138, The Finnish Forest Research Institute, Helsinky.

de VRIES, (1986). Sampling theory for forest inventory, Springer Verlag, Berlin.

J.J. WARNES, (1986). A sensitivity analysis for universal kriging. Mathematical Geology, 18, pp. 653-676.

P. WHITTLE, (1954). On stationary processes in the plane, Biometrika, 41, pp.450-462.

N. WIENER, (1958). The Fourier integral and certain of its applications, Dover Publications, Inc., New York.

D.L. ZIMMERMANN, N. CRESSIE, (1992). Means squared prediction error in the spatial linear model with estimated covariance parameters. Ann. Inst. Statist. Math., 44, pp. 27-43.

D.L. ZIMMERMANN, (1989). Computationally efficient restricted maximum estimation of generalized covariance functions. Mathematical Geology, 21, pp. 655-672.

Seite Leer /
Blank leaf