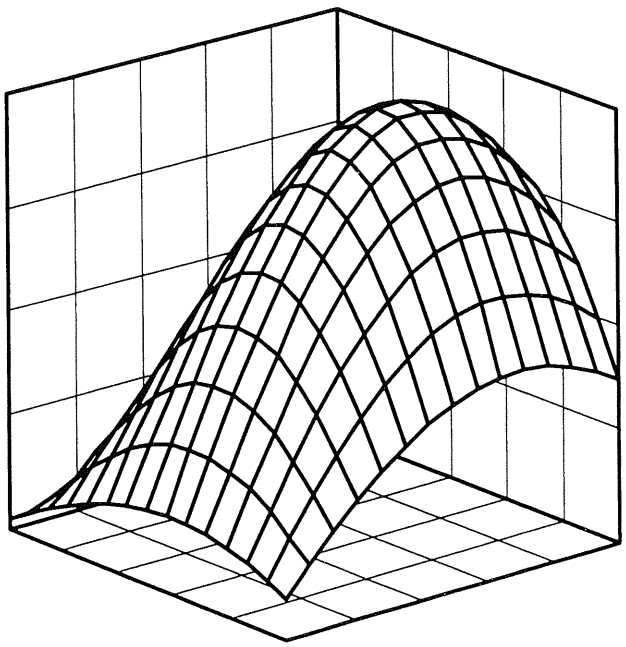
Data in Action



PRODUCTION
ECOLOGY

ab-dlo

Quantitative Approaches in Systems Analysis

The Quantitative Approaches in Systems Analysis series provides a platform for publication and documentation of simulation models, optimization programs, Geographic Information Systems (GIS), expert systems, data bases, and utilities for the quantitative analysis of agricultural and environmental systems. The series enables staff members, students and visitors of AB-DLO and PE to publish, beyond the constraints of refereed journal articles, updates of models, extensive data sets used for validation and background material to journal articles. The QASA series thus primarily serves to support peer reviewed articles published elsewhere. The inclusion of listings of programs in an appendix is encouraged.

All manuscript are reviewed by an editorial board comprising one AB-DLO and one PE staff member. The editorial board may consult external reviewers. The review process includes assessing the following: relevance of the topic to the series, overall scientific soundness, clear structure and presentation, and completeness of the presented material(s). The editorial board evaluates manuscripts on language and lay-out matters in a general sense. However, the sole responsibility for the contents of the reports, the use of correct language and lay-out rests with the authors. Manuscripts or suggestions should be submitted to the editorial board. Reports of the series are available on request.

Quantitative Approaches in Systems Analysis are issued by the DLO Research Institute for Agrobiology and Soil Fertility (AB-DLO) and The C.T. de Wit Graduate School for Production Ecology (PE).
AB-DLO, with locations in Wageningen and Haren, carries out research into plant physiology, soil science and agro-ecology with the aim of improving the quality of soils and agricultural produce and of furthering sustainable production systems.
The 'Production Ecology' Graduate School explores options for crop production systems associated with sustainable land use and natural resource management; its activities comprise research on crop production and protection, soil management, and cropping and farming systems.

Address for ordering copies of volumes in the series:
Secretariat
TPE-WAU
Bornsesteeg 47
NL-6708 PD Wageningen
Phone: (+) 31 317.482141
Fax: (+) 31 317.484892
E-mail: office@sec.tpe.wau.nl

Addresses of editorial board (for submitting manuscripts):
H.F.M. ten Berge
AB-DLO
P.O. Box 14
NL-6700 AA Wageningen
Phone: (+) 31 317.475951
Fax: (+) 31 317.423110
E-mail: h.f.m.tenberge@ab.dlo.nl

M.K. van Ittersum
TPE-WAU
Bornsesteeg 47
NL-6708 PD Wageningen
Phone: (+) 31 317.482382
Fax: (+) 31 317.484892
E-mail: martin.vanittersum@staff.tpe.wau.nl

Data in Action

Proceedings of a seminar series
1996/1997

A. Stein, F.W.T. Penning de Vries & J.W. Schut

Quantitative Approaches
in Systems Analysis No. 12
June 1997

CIP-DATA KONINKLIJKE BIBLIOTHEEK, DEN HAAG

Stein, A., F.W.T. Penning de Vries & J.W. Schut

Data in action - proceedings of a seminar series 1996-1997
Stein - Wageningen : Agricultural University Dept. of Soil Science and Geology;
Wageningen; The C.T. de Wit Graduate School for Production Ecology. -
(Quantitative approaches in systems analysis ; no. 12)
NUGI 835
Subject headings: data and models, statistics for agricultural data,
data and information systems, data and scale.

Guidelines 'Quantitative Approaches in Systems Analysis'

Manuscripts or suggestions should be submitted to the editorial board (H.F.M. ten Berge, AB-DLO, or M.K. van Ittersum, TPE-WAU). The final version of the manuscripts should be delivered to the editors camera-ready for reproduction. The submission letter should indicate the scope and aim of the manuscript (e.g. to support scientific publications in journals, program manual, educational purposes). The costs of printing and mailing are borne by the authors.

The English language is preferred. Authors are responsible for correct language and lay-out. Overall guidelines for the format of the texts, figures and graphs can be obtained from the publication editor at AB-DLO, or from the PE office:

H. Terburg
AB-DLO

P.O. Box 14
NL-6700 AA Wageningen
Phone: (+) 31 317.475723
Fax: (+) 31 317.423110
E-mail: h.terburg@ab.dlo.nl

Th.H. Jetten
Secretariat C.T. de Wit Graduate School
for Production Ecology
Lawickse Allee 13
NL-6701 AN Wageningen
Phone: (+) 31 317.485116
Fax: (+) 31 317.484855
E-mail: theo.jetten@beleid.spp.wau.nl

Preface

Alfred Stein[1], Frits Penning de Vries[2] and Johan Schut[3]

1. Dept. of Soil Science and Geology, Wageningen Agricultural University, PO Box 37, 6700 AA Wageningen, The Netherlands

2. DLO Research Institute for Agrobiology and Soil Fertility (AB-DLO) P.O. Box 14, 6700 AA, Wageningen, The Netherlands

3. Department of Plant Breeding, Wageningen Agricultural University, P.O.Box 386, 6700 AJ Wageningen, the Netherlands

email: alfred.stein@bodlan.beng.wau.nl

Data are crucial to derive valid conclusions in various applications. Data are needed to feed and develop models, and to validate models. Data are needed for a quantitative statistical approach, and data are needed in information systems to reliably display spatial units. Every specific datum has its own quality: how reliably was it measured, is it representative for a larger area of land, is it constant in time, what does it represent, and what factors have determined its values? Data are becoming increasingly available by new digital techniques, the electronic superhighway and the laws which enforce open access of data collected with public money. But are these the data that we need? And if so, how can these be accessed and handled efficiently?

In this publication the written papers of the PE-seminars 'Data in Action' are collected. The series was a follow-up of the previous series 'Models in Action' (Stein, Penning de Vries and Schotman, 1996). This time we emphasize the role of data for agricultural purposes. Again we aimed at a volume which well extends beyond the disciplinary context. We focused on statistics, information science, models and scale, although we realize that these concepts act intermixedly.

At this place we like to thank the CT de Wit Research School of Production Ecology for their financial and organizational support.

ref. Stein A, Penning de Vries FWT, Schotman P, 1996, Models in Action. Proceedings of a seminar series 1995/1996. Quantitative Approaches in Systems Analysis No. 6, June 1996.

Table of Contents

1. Data and models

No modelling without experimentation

Data are used to develop models, to feed models, to validate models and to apply models. In particular the use of dynamic models has already become widespread. Despite this development, the basic data are the cornerstone for developing, validating, running and extrapolating models. In the first seminar we focus on the interaction between data and models.

1.1 Modelling crop-weed interactions and weed population dynamics: the need to combine modelling and experimentation[1]

Martin J. Kropff[1,2], Jacco Wallinga[2] and Bert (L.A.P.) Lotz[2]

1. Wageningen Agricultural University, Department of Theoretical Production Ecology P.O. Box 430, 6700 AK Wageningen, The Netherlands
2. DLO Research Institute for Agrobiology and Soil Fertility (AB-DLO) P.O. Box 14, 6700 AA, Wageningen, The Netherlands
E-mail martin.kropff@staff.tpe.wau.nl

For the development of weed management systems with minimum herbicide use quantitative understanding of weed population dynamics and crop-weed interactions is required. Models, that integrate the available quantitative knowledge, can be used to design preventive measures, to develop long-term and short-term strategies for weed management, to assist in decision making to determine if, when, where and how weeds should be controlled and to identify new opportunities for weed control. Eco-physiological simulation models for crop-weed competition simulate growth and production of species in mixtures, based on eco-physiological processes in plants and their response to the environment. Such models helped to improve insight into the crop-weed system and can be used for various purposes like the development of simple predictive yield-loss models, threshold levels or the design of competitive plant types. The collection and generation of data was essential for the development and evaluation of these models. For strategic weed management decisions, preventive measures and the identification of new opportunities for weed control, quantitative insight into the dynamics and spatial patterns of weed populations is also required. The complexity of the process and the long-term character of weed population dynamics make the use of models necessary. Different modelling approaches have been developed and are described briefly. The development and application of these models has been limited because of data availability.

1. Introduction

Weed management has been one of the key issues in most agricultural systems, especially before herbicides became available. The use and application of herbicides was one of the main factors enabling intensification of agriculture in developed countries in the past decades. More recently, the availability of herbicides has been coupled to intensification of agriculture in developing countries as well. However, increased concern about environmental side effects of herbicides, the development of herbicide resistance in weeds and the necessity to reduce cost of inputs have resulted in greater pressure on farmers to reduce the use of herbicides. This has led to the development of new strategies for weed management.

The strategy to improve weed management systems based on increased precision with respect to weed management consists of three components (Kropff 1996):
(1) reduce weed effects through adapted crop management (Prevention),

---

[1]This paper is based on earlier publications (Kropff and Van Laar, 1993; Kropff, 1996 and Kropff, Wallinga and Lotz, 1996)

(2) improve decision-making with respect to weed control (Decision making),
(3) improve control technology (Control)

The first component (*prevention*) involves any aspect of management that favours the crop relative to the weeds. This includes the development of competitive crop varieties with minimum trade-off between competitiveness and yield potential of the cultivar. Systems analysis and simulation are indispensable tools for the study of such complex interrelationships and may help bridge the gap between knowledge at the process level and management at the field level.

The second component (*decision making*) consists of strategic (long term) decisions, tactic decisions (for a season) and operational decisions in the field. That requires long-term and short-term strategies for weed management, to assist in decision making to determine if, when, where and how weeds should be controlled

The third component (*control*) deals with the development/improvement of weed control technology and is strongly related to precision technology. Three ways to control weeds can be distinguished: biological, mechanical and chemical weed control. There are many ways in which control technology can be improved ranging from precision mechanical weed management tools to precision herbicide treatments

To improve weed management systems, based on this strategy with the three components, quantitative insight into both crop-weed interactions and the dynamics of weed populations in space and in time is required. Because of the complexity of the processes and the long term aspects in population dynamics, models are required to obtain such quantitative insight and to make the knowledge operational.

This paper reviews the state of the art with respect to modelling crop-weed interactions and weed population dynamics and discusses the role of data collection and generation.

2. Modelling Crop-Weed Interactions And Weed Population Dynamics

2.1. Modelling crop-weed interactions

Weeds affect crop yield as a result of competition between the crop and weeds for growth-limiting resources, i.e. light, water and nutrients. The quantitative understanding of crop-weed interactions was reviewed in detail by Kropff & Van Laar (1993). Two types of models for crop-weed interactions have been developed: (i) descriptive regression models with a few parameters that can be determined by fitting the model to observed data, and (ii) expanded eco-physiological crop growth models that simulate competition for the growth-determining (light) and -limiting (water and nutrients) resources between species.

One of the first systematic approaches to study competitive phenomena, developed by De Wit (1960), used an experimental design (the replacement series in which the mixing ratio varies, but total density remains constant) with a model to analyze the results. However, only in the early 1980s an approach was developed to describe competition over a range of population densities with varying mixture ratios and at a range of total densities (reviews by Spitters, 1989; Kropff and Van Laar, 1993). These descriptive regression models are based on the same principle as the approach of De Wit (1960): the non-linear hyperbolic relationship between yield and plant density.

These descriptive regression models have a static character -- a description is given of competition effects at a certain moment.

It has been demonstrated that competition effects can be simulated by expanded eco-physiological crop growth models that simulate competition for the growth-determining (light) and -limiting (water and nutrients) resources (review Kropff and Van Laar, 1993).

On the basis of insights obtained with the eco-physiological models, an alternative regression model was developed to predict yield loss using observations of relative leaf area or cover of weeds shortly after crop emergence (Kropff & Spitters, 1991). This is an example of a study where a model was used to generate hypotheses and theoretical data to facilitate evaluation of a simple model before starting laborious experiments to test the model.

2.1.1 Descriptive models for crop weed-interactions

The most widely used regression model to describe effects of competition at a certain moment is the hyperbolic yield-loss weed density model (Cousens 1985):

$$Y_L = \frac{a N_W}{1 + \frac{a}{m} N_W} \qquad (1)$$

where Y_L gives the yield loss, N_W is the weed density, a describes the yield loss caused by adding the first weed per m^2 and m the maximum yield loss. These hyperbolic yield-density equations fit well to data from experiments where only the weed density is varied (Kropff et al. 1984, Cousens 1985). However, the parameters a and m may vary strongly among experiments due to the effect of other factors on competition processes (Kropff and Van Laar 1993). Because both weed density and the period between crop and weed emergence strongly determine the competitive relationship between crop and weeds (Kropff *et al.* 1984, Cousens *et al.* 1987, Kropff *et al.* 1992), more robust prediction of yield loss on the basis of early observations should be based on these two factors.

The eco-physiological model INTERCOM (Kropff and Van Laar, 1993) was used to analyse the possibility to predict yield loss early in the season, based on observations of the relative leaf area index of the weeds. A very close relationship was predicted between relative leaf area index of the weeds and yield loss for a wide range of weed densities and periods between crop and weed emergence. Based on these results, an alternative approach for early prediction of crop losses by weed competition was introduced by Kropff and Spitters (1991). A simple descriptive regression model for early prediction of crop losses by weed competition was introduced by Kropff & Spitters (1991) and extended by Kropff *et al.* (1995). This model relates yield loss to relative weed leaf area (L_W expressed as weed leaf area /crop+weed leaf area) shortly after crop emergence, using the 'relative damage coefficient' q as the main model parameter next to the maximum yield loss m:

$$Y_L = \frac{q L_W}{1 + \left(\frac{q}{m} - 1\right) L_W} \qquad (2)$$

Because leaf area accounts for density and age of the annual weeds, this regression model accounts for the effect of weed density and the effect of the time of weed emergence (Kropff & Spitters 1991). An example is given in Figure 1 which shows the relation between yield loss in sugar beet and the parameters density and relative leaf area of *Chenopodium album* L. based on data from Kropff *et al.* (1995). Lotz *et al.* (1996) evaluated the approach over a wide geographic region and found that the descriptive value of the model is good, but that the current predictive ability is still insufficient for precision weed management.

In conclusion, the relative leaf area-yield loss regression model accounts for the effect of weed densities, different flushes of weeds and the period between crop and weed emergence. However, the effect of other factors, such as transplanting shock or severe water stress, is not accounted for, because the regression models do not account for underlying processes. Clearly,

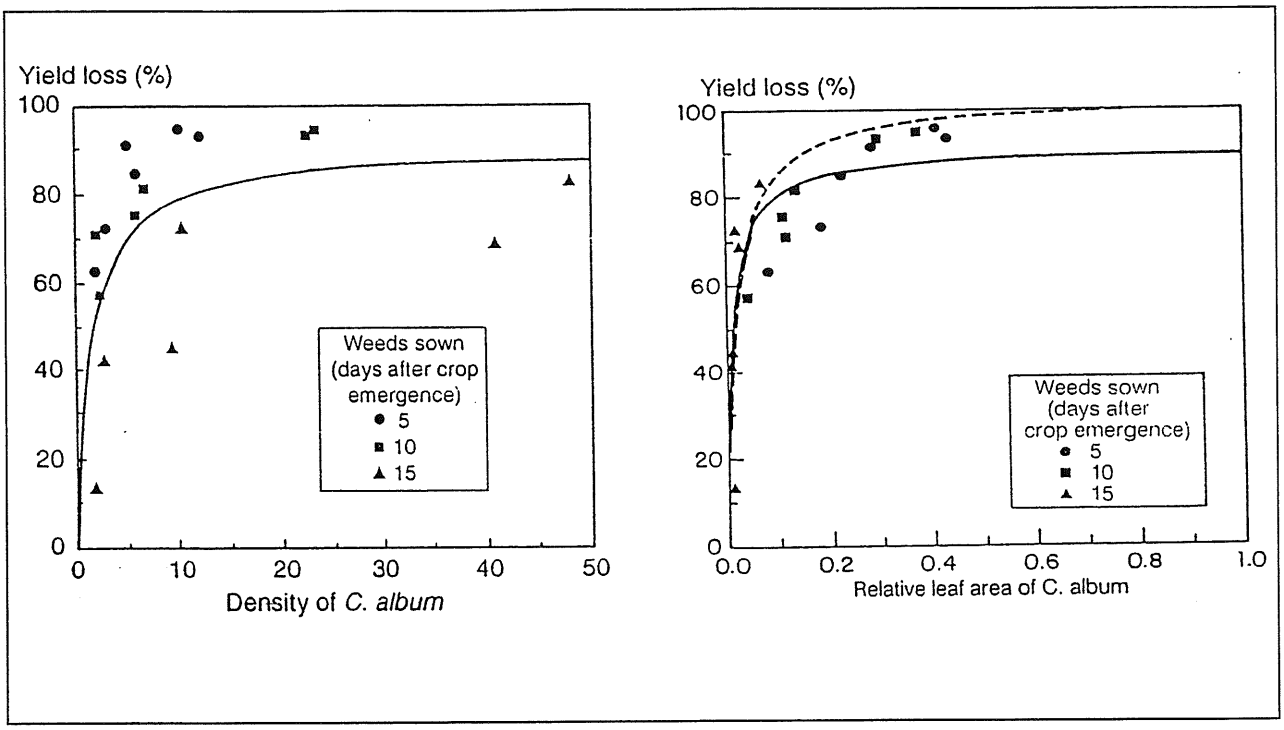


Figure 1. Yield loss in sugarbeet related to the density of *Chenopodium album* L. (A) and relative leaf area of *Chenopodium album* L. (B) for different dates of weed emergence: circles 1, squares 5 and triangles 12 days after the crop. Redrawn after Kropff *et al.* 1995.

the interaction of models and data has been crucial in developing the approaches, but will also be essential in further developing models that can be used for practical application.

2.1.2 Eco-physiological models for crop weed-interactions

Competition is a dynamic process that can be understood from the distribution of the growth-determining (light) or -limiting (water and nutrients) resources over the competing species and the efficiency with which each species uses the resources. Eco-physiological models that simulate these processes provide insight into competition effects observed in (field) experiments and may aid in seeking ways to manipulate competitive relations, such as those between crop and weeds by determining the most important factors in crop-weed competition.

The eco-physiological model INTERCOM described by Kropff and Van Laar (1993) consist of coupled crop growth models equal to the number of competing species. Under favorable growth conditions, light is the main factor determining the growth rate of the crop and its associated weeds. From the leaf area index (LAI) of the species, the vertical distribution of their leaf area and their light extinction properties, the light profile within the canopy is calculated. It is assumed that the horizontal distribution of leaves is homogeneous. Based on the species characteristics for the photosynthetic light response of single leaves, the vertical photosynthesis profile of each species in the mixed canopy is obtained. Integration over the height of the canopy and over the day gives the daily assimilation rate for each species. After subtracting the respiration requirements for maintenance, the net daily growth rate in kg dry matter/ha per day is obtained using the conversion factor. The dry matter produced is partitioned among various plant organs, using partitioning introduced as a function of the phenological development stage of the species. Phenological development rate is tracked in the model as a function of ambient daily average temperature.

Height growth rate is calculated as a function of temperature. However, when competition results in rates of daily dry matter growth are insufficient to allow further height growth, height growth rate is reduced.

To account for drought stress effects, a simple water balance for a free draining soil profile is attached to the model, tracking the available amount of soil moisture during the growing season. Transpiration and growth rates of each species are reduced when available soil moisture drops below a certain level. Competition for water is thus closely linked to aboveground competition for light because transpiration is driven by the absorbed amount of radiation and the vapor pressure deficit inside the canopy. Direct competition for water as a result of differences in rooting density is not accounted for.

For the parameterisation of these processes, many controlled environment studies and field experiments have been conducted. Kropff and Van Laar (1993) described the most important species specific data that are required for the model and also how these parameters can be estimated or experimentally determined. Experiment-specific input requirements of the eco-physiological model include geographical latitude, standard daily weather data, soil physical properties, dates of crop and weed emergence, and weed density. A detailed description of the eco-physiological simulation model is given by Kropff & Spitters (1992).

Most parts of the eco-physiological model have been evaluated and tested for monoculture crops (Table 1). The eco-physiological competition model has been tested with data from competition experiments with maize (*Zea mays* L.), yellow mustard (*Sinapis arvensis* L.) and barnyard grass (*Echinochloa crus-galli* L.) (Kropff *et al.*, 1984; Spitters, 1984; Spitters & Aerts, 1983). Additional validation of the eco-physiological model was performed using data from critical period experiments with the same species (Weaver *et al.*, 1992). This model was evaluated subsequently (Kropff *et al.*, 1992) using data with tomato (*Lycopersicon esculentum* L.)-pigweed (*Amaranthus retroflexus* L.) and tomato-eastern black nightshade (*S. americana*)

Table 1. Observed and simulated yields of weed-free crops, in Harrow (Canada) and Wageningen (Netherlands) (Kropff and Van Laar, 1993)

Crop	Site	Year	Observed yield** (kg ha^{-1})	Simulated yield (kg ha^{-1})
Tomato (seeded)	Harrow	1984	3172 ± 222	3009
Tomato (seeded)	Harrow	1985	2704 ± 260	3290
Tomato (transpl.)	Harrow	1984	2736 ± 164	2990
Tomato (transpl.)	Harrow	1985	4189 ± 330	4312
Maize	Wageningen	1982	13110 ± 1940	13901
Maize	Wageningen	1983	8440 ± 210	8459
Sugar beet	Wageningen	1984	14900 ± 1397	14870
Sugar beet	Wageningen	1985	23100 ± 1233	20644
Sugar beet	Wageningen	1986	20400 ± 687	20450
Rice	Los Banos	1992	7068 ± 169	7002

* Yields of tomato, maize, sugar beet and rice represent fruit, grain, root and panicle dry weight, respectively.

** Means ± standard errors.

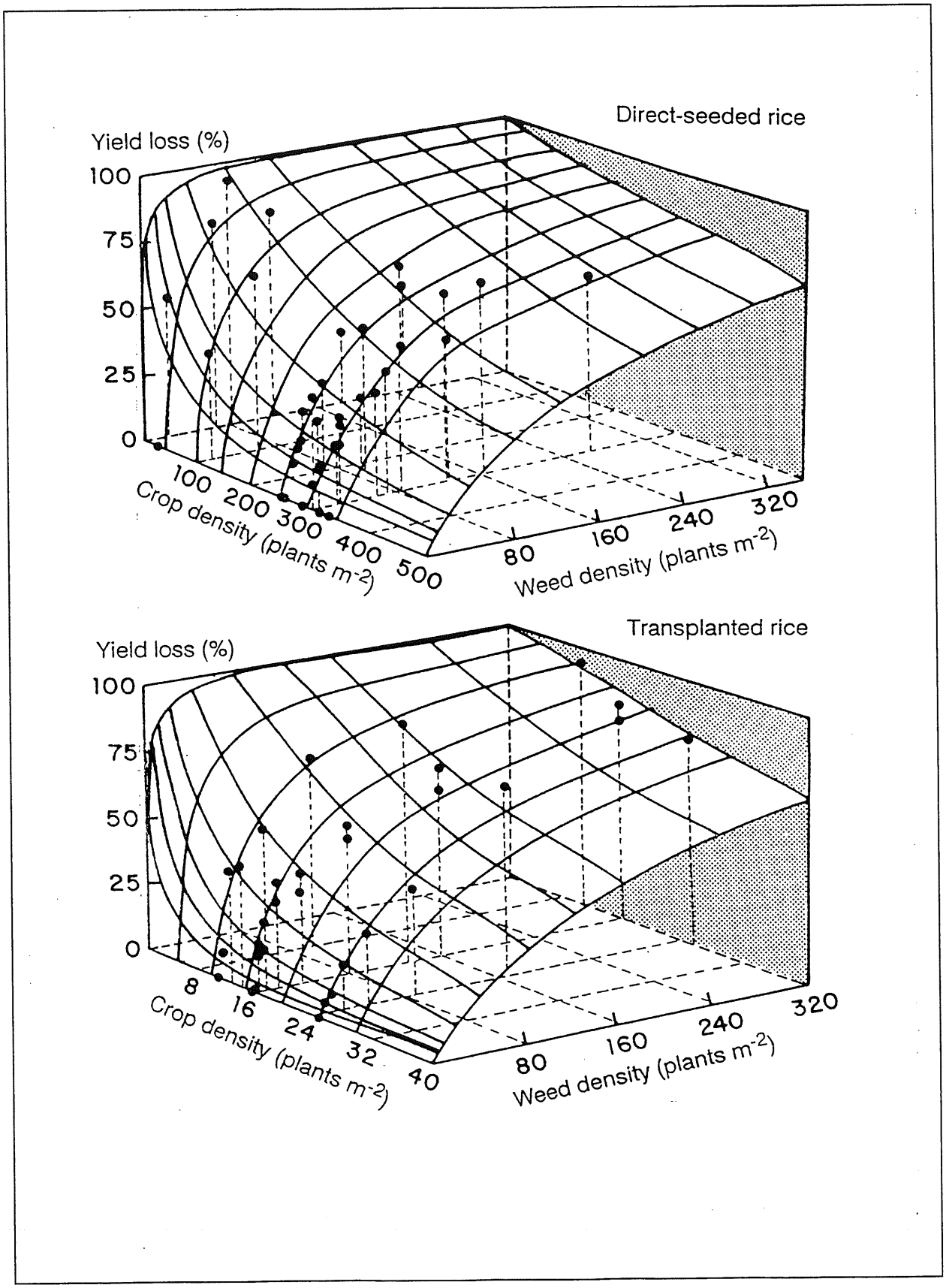


Figure 2. Simulated (using the model INTERCOM) and observed yield loss in rice by competition with *Echinochloa crus-galli* L. for direct seeded and transplanted rice for different data sets. Redrawn after Kropff & Van Laar (1993).

(Weaver *et al.*, 1987) in Canada. Yield loss-weed density responses, effects of transplanting vs. direct seeding and effects of weed-free or weed-infested periods (critical period experiments) were simulated accurately. In figure 2 the model INTERCOM was also evaluated for rice and *E. crus-galli* (Kropff and Van Laar, 1993).

Although further validation of the eco-physiological models using independent data sets may be required, the results of different studies indicate that interplant competition can be well understood from the underlying physiological processes.

2.2 Modelling weed population dynamics

Models have been developed to integrate the knowledge on life-cycle processes. Figure 3 shows the life cycle of annual weeds. The main processes involved are germination and emergence of seedlings from seeds in the seed bank in the soil, establishment and growth of the weed plants, seed production, seed shedding and seed mortality in the soil. Competition plays a major role in establishment and growth and therefore strongly affects the population dynamics of weeds. Besides natural processes, man has a major impact on the spread of weeds at all different scales. The different mechanisms of dispersal have been discussed in detail by Cousens & Mortimer (1995), who concluded that apart from wind dispersal few quantitative studies have been conducted on these mechanisms. Because most weed seeds remain very close to the plant (Harper 1977), weed patterns in fields do not change dramatically in time (Wilson & Brain 1991) which may be a basis for precision agricultural practices.

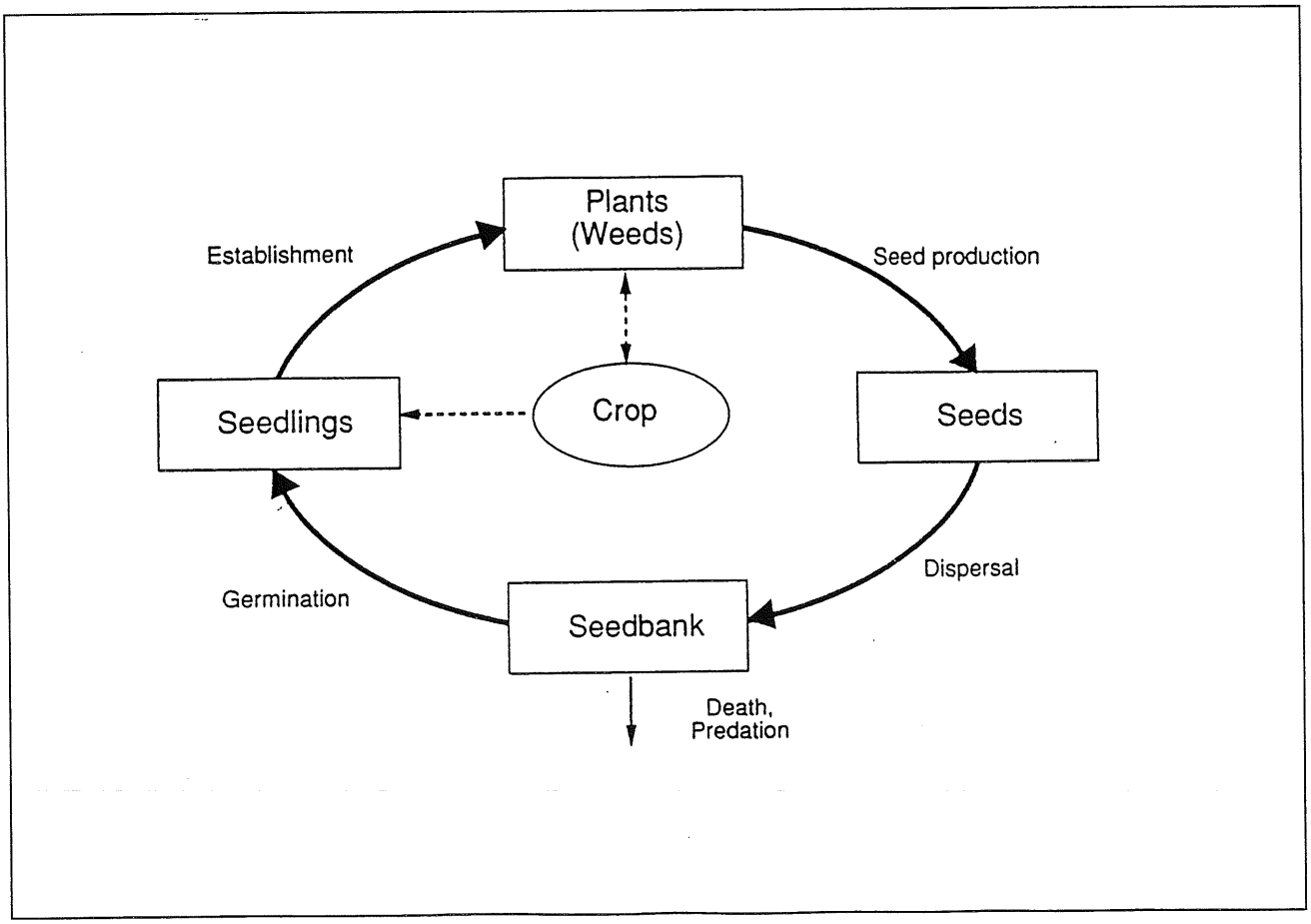


Figure 3. Schematic diagram for the life cycle of a weed in competition with crops in annual cycles.

Comprehensive models that are based on physiological principles are only available for parts of the life cycle: (as discussed) plant growth and competition (Kropff & Van Laar 1993) and germination and emergence (L.M.Vleeshouwers, in prep.). In contrast, processes like seed shedding, seed dispersal and predation of seeds are poorly understood. The most detailed models that encompass the whole life-cycle have been developed for species like *Avena fatua* L. (Cousens *et al.* 1986), *Alopecurus myosuroides* Huds. (Doyle *et al.* 1986) and *Galium aparine* L. (Van der Weide & Van Groenendael 1990). The basic structure of most models was described by Spitters (1989).

Not all models are aimed at understanding and integrating detailed knowledge. Another objective is to predict future weed infestations. Models for forecasting need to be robust, and they generally exhibit a better predictive capability when they contain only a few parameters, even if there is complete understanding of underlying processes. The various complex processes in the life-cycle are then blended into a few lumped parameters like a germination rate, a reproduction rate and a mortality rate. Forecasting future infestations is bound up with very large error margins, irrespective of our understanding of weed population biology, since some key factors like future weather conditions are unknown.

Apart from the level of detail at which the life-cycle is studied, three different modelling approaches to integrate individuals into a population can be distinguished (Durrett & Levin 1994, Kropff *et al.* 1996): (i) the density based models, (ii) the density based models that take spatial gradients in density into account and (iii) the individual based models which also account for spatial processes.

Of the modelling approaches, individual based models are the most comprehensive, but as a result of their complexity they quickly run into computing problems and the complexity is not always required. The density based model can be very useful to roughly explore options for long term weed management strategies. The individual based models can be very helpful to identify opportunities for site specific weed management (Kropff *et al.* 1996).

The availability of data, especially on soil bound processes has been a major setback for the development and evaluation of population dynamics models.

3. Modelling and Weed Management

As mentioned in the introduction, three aspects can be distinguished in the strategy to improve weed management systems (Kropff 1996):
(1) reduce weed effects through adapted crop management (Prevention),
(2) improve decision-making with respect to weed control (Decision making),
(3) improve control technology (biological, mechanical and chemical) (Control)
Modeling approaches that have been developed can be used for all three aspects.

The first component of the strategy to improve weed management involves any aspect of management that favors the crop relative to the weeds. The eco-physiological simulation model INTERCOM for interplant competition was used to identify traits that determine the competitive ability of a crop. The most important traits were: rapid early leaf area -, tiller - and height development, and more horizontally oriented leaves in early growth stages (vertical ones later on because of yield potential) (Kropff & Van Laar 1993). In experiments, rice varieties that differed in these traits were evaluated with respect to their competitive ability versus a standard purple-coloured variety. The variety with all required traits, Mahsuri, reduced the growth of the purple variety so much, that all purple rice plants had died before the final harvest (M.J.Kropff, unpublished results). Detailed studies on trade-offs between different traits are underway (L. Bastiaans *et al.* in prep.).

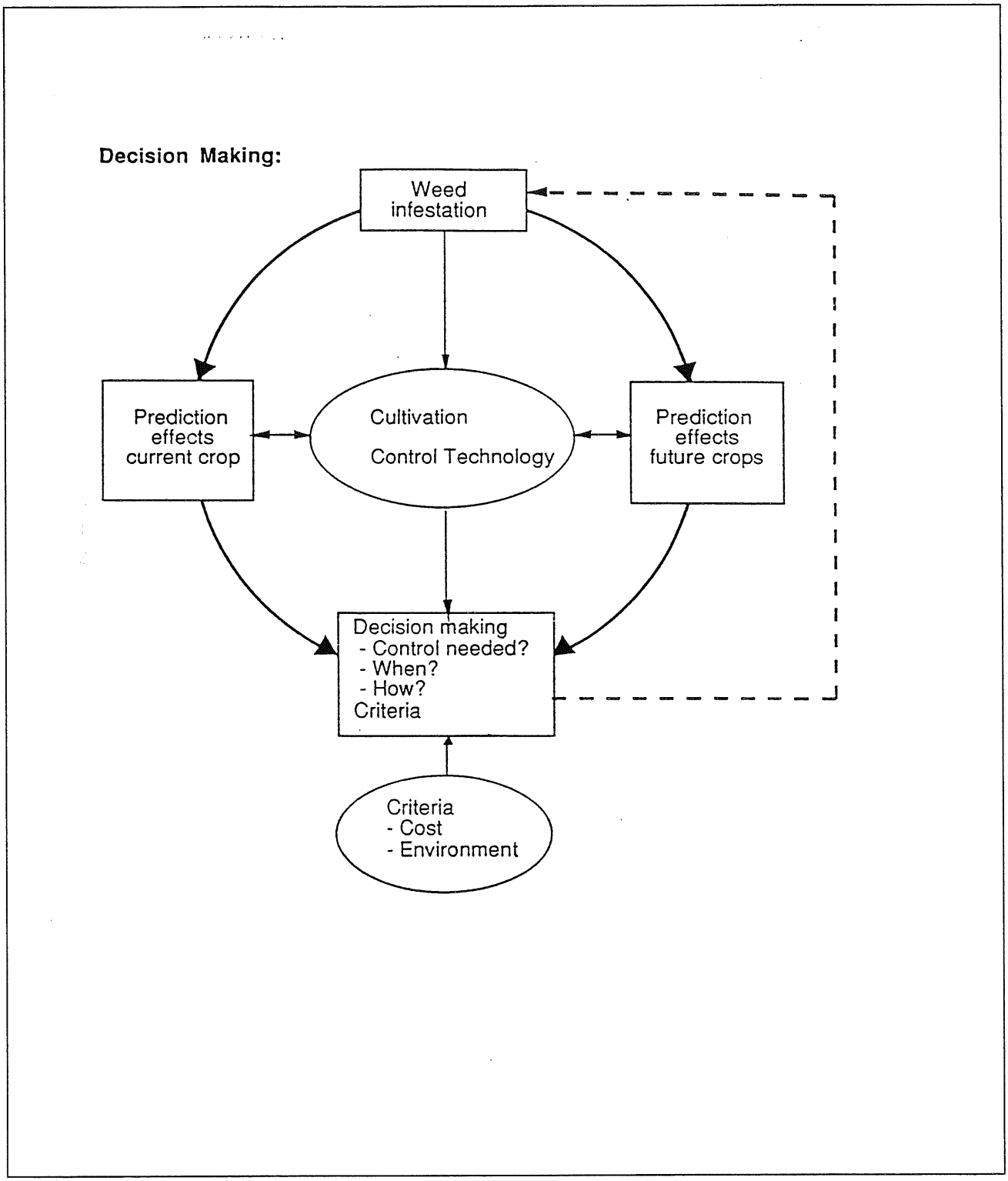


Figure 4. Schematic representation of the decision making process in weed management.

Other examples of preventive measures are stale seed beds, specifically designed crop rotations (like the inclusion of grass, which is very competitive, in the rotation), intercropping or relay cropping etc.

The second component of the strategy is the improvement of the decision making process which consists of strategic (long term) decisions, tactic decisions (for a season) and operational decisions in the field. Here, precision in time and space is required. It involves long-term and

short-term strategies for weed management, to assist in decision making to determine if, when, where and how weeds should be controlled.

The decision-making process for tactical and operational decisions in a weed management system based on post-emergence observations is illustrated in Figure 4. To allow rational decision making, the severity of weed infestation shortly after crop emergence should be estimated. Criteria must be defined (i.e., the objectives and planning horizon of the farmer) to enable economic decision making.

The third component is to improve control technology. Here the models can be very helpful as well. For example, to evaluate the impact of sublethal dosages of herbicides, and the risk involved, or the analysis of the impact of biological control agents.

4. Conclusions

Options to improve weed management systems with a minimum herbicide use exist in all components of the strategy: prevention, decision making and control. Quantitative insight in weed population dynamics and crop weed interactions is essential for that purpose. Such quantitative insight is summarized and made operational in models. From the experience obtained it can be concluded that model development and experimentation at different levels of complexity is needed for successful progress in the development of weed management programs with minimum herbicide use.

5. References

Auld BA, Coote BG 1980 A model of a spreading plant population. Oikos 34:287-292

Cousens R 1985 An empirical model relating crop yield to weed and crop density and a statistical comparison with other models. J of Agric. Sci. 105:513-521

Cousens R, Mortimer AM 1995 Dynamics of Weed Populations, Cambridge: Cambridge University Press

Cousens R, Doyle CJ, Wilson BJ, Cussans GW 1986 Modelling the economics of controlling *Avena fatua* in winter wheat. Pesticide Science 17:1-12

Cousens R, Brain P, O'Donovan JT, O'Sullivan A 1987 The use of biologically realistic equations to describe the effects of weed density and relative time of emergence of crop yield. Weed Sci. 35:20-725

Doyle CJ, Cousens R, Moss SR 1986 A model of the economics of controlling *Alopecurus myosuroides* Huds. in winter wheat. Crop Protection 5:143-150

Durrett R, Levin SA 1994 The importance of being discrete (and spatial). Theoretical Population Biology 46:363-394

Harper JL 1977 Population biology of plants, London: Academic Press

Kropff MJ 1996 Strategic balancing. Inaugural address, Wageningen Agricultural University.

Kropff MJ, Spitters CJT 1991 A simple model for crop loss by weed competition on basis of early observation on relative leaf area of the weeds. Weed Research 31:97-105

Kropff MJ, Van Laar HH (eds) 1993 Modelling Crop-Weed Interactions, Wallingford: CAB International

Kropff MJ, Vossen FJH, Spitters CJT, De Groot W 1984 Competition between a maize crop and a natural population of *Echinochloa crus-galli* L.). Neth J Agric Sci 32:324-327

Kropff MJ, Weaver SE, Smits MA 1992 Use of eco-physiological models for crop-weed interference: Relations amongst weed density, relative time of weed emergence, relative leaf area, and yield loss. Weed Research 40:296-301

Kropff MJ, Lotz LAP, Weaver SE, Bos HJ, Wallinga J, Migo T 1995 A two parameter model for prediction of yield loss by weed competition from early estimates of relative leaf area of the weeds. Annals of Applied Biology 126:329-346

Kropff MJ, Wallinga J, Lotz LAP 1996 Weed population dynamics. In: Brown H, Cussans GW, Devine MD, Duke SO, Fernandez-Quintanilla C, Helweg A, Labrada RE, Landes M, Kudsk P, Streibig JC (eds) Second International Weed Control Congress Copenhagen, EWRS. p 3-14.

Lotz LAP, Christensen S, Cloutier D, Fernandez Quintanilla C, Légère A, Lemieux C, Lutman PJW, Pardo Iglesias A, Salonen J, Sattin M, Stigliani L, Tei F (1996) Prediction of the competitive effects of weeds on crop yields based on the relative leaf area of weeds. Weed Research 36:93-101

Spitters CJT 1984 A simple simulation model for crop weed competition. 7th International Symposium on Weed Biology, Ecology and Systematics. COLUMA-EWRS, Paris, p 355-366

Spitters CJT 1989 Weeds: population dynamics, germination and competition. In Rabbinge R, Ward SA, Van Laar HH (eds), Simulation and systems management in crop protection, Simulation Monographs, Pudoc, Wageningen p 182-216

Spitters CJT, Aerts R 1983 Simulation of competition for light and water in crop weed associations. Asp of Appl Biol 4:467-484

Van der Weide RY, Van Groenendael JM 1990 How useful are population dynamical models: an example from *Galium aparine* L. Zeitschrift für Pflanzenkrankheiten und Pflanzenschutz Sonderheft XII:147-155

Weaver SE, Smits N, Tan CS 1987 Estimating yield losses of tomato (*Lycopersicon esculentum*) caused by nightshade (*Solanum spp*) interference. Weed Science 35:163-168

Weaver SE, Kropff MJ, Groeneveld RMW 1992 Use of eco-physilogical models for crop-weed interference: The critical period of weed interference. Weed Research 40:302-307

Wilson BJ, Brain P 1991 Long-term stability of distribution of *Alopecurus myosuroides* Huds. within cereal fields. Weed Research 31:367-373

Wit, CT de 1960 *On Competition.* Agricultural Research Reports 66.8, Pudoc, Wageningen

1.2 Development of a morphogenetic model from field and lab data: modeling the seed number per node on a pea stem

Marie-Helene Jeuffroy

Institut National de la Recherche Agronomique, Unité d'Agronomie,
78 850 Thiverval-Grignon, France.
E-mail: jeuffroy@bcgn.grignon.inra.fr

1. Introduction

Crop models can be used for many purposes (Boote et al., 1996). They can be an aid in interpreting experimental results, they can be research tools in order to understand the interactions of the numerous factors in crop growth and production, they can assist farmers in decision making. For these different aims, models must give a realistic representation of the complex reality. Therefore, the model must not only be compared to data from the field, to evaluate its ability to give a good account of the reality, but it must also be build in taking into account the results from experimental data, so that the descriptions of the natural phenomena (model structure and parameters) will be derived from observation rather than on speculation (Monteith, 1996).

In this paper, several relationships existing between a model (its building and its use) and data will be illustrated in taking as an example the model I realized to simulate the seed number on each reproductive node of a pea stem (Jeuffroy, 1994).

A pea plant is generally composed of a main stem, and can bear some branches, the number of which being highly variable. On each stem, from the base to the top, can be firstly observed some vegetative nodes, bearing only a composed leaf. Then, from the first flowering node, which is about the fifteenth node on the cultivar Solara, each following node bears a composed leaf, and one or several flowers, which will give pods containing seeds, if they do not abort. Flowers appear successively on the different nodes, and the development and growth of the pods and seeds is sequential : thus various organs at different ages can exist on a same stem at a given date : (1) vegetative organs in development (as the apex continues to develop during the reproductive period), (2) young flowers (on the upper nodes), (3) young pods below and (4) old pods with filling seeds on the first reproductive nodes. Many reproductive organs (flowers, pods and seeds) generally abort on a pea stem. The proportion of aborted organs is not constant among the different nodes, resulting in a heterogeneous distribution of seed numbers among the different nodes, which is called the profile of seed number. As development is sequential, the seed number on each node is not fixed at the same time. Thus the seed formation on the successive nodes is not realized in the same environmental conditions. For example, a limiting factor occurring early will reduce the seed number on the first nodes ; if it occurs later, it will only have an effect on the last nodes of the stem. Moreover, it can be assumed that the seed formation on the last nodes depends on what happened on the first nodes. Thus in many situations, the total seed number per stem can be understood only if we understand the seed formation on each node.

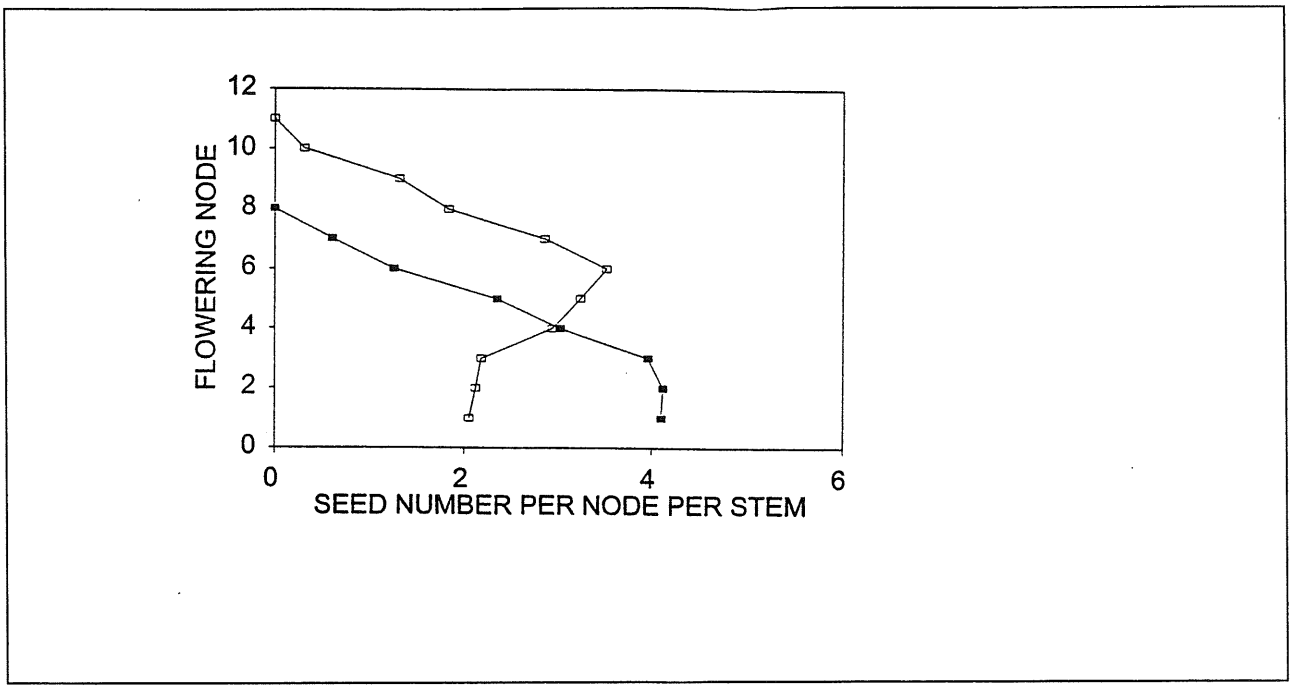


Figure 1. The two main forms of profiles of seed numbers per node observed in the fields.

In farmers' fields, various profiles are observed. They have mainly two typical forms (Figure 1), even when no limiting factor is observed. Some of them are rather high, with a high number of flowering nodes, and generally bear few seeds on the first reproductive nodes; the others are low, with few flowering nodes, and a high number of seeds on the first reproductive nodes. In the two cases, the seed number decreases regularly on the upper nodes. The total seed number per stem is linked to the form of the profile, and the highest seed numbers per stem are generally obtained on profiles rather low. In order to understand the high yield variability observed in farmers, which is mainly linked to the variability of seed numbers (Doré, 1992), it is thus essential to understand the diversity of these profiles.

In the literature, the main hypothesis to explain the profiles is nutritional : as there are many sinks on a stem, of various kinds and ages, there is a high competition for assimilates between them, and particularly between pods, explaining their different seed numbers. As the organs involved in this competition are numerous and from various nature, as the complexity of the competition is high, the only means to understand this competition seemed to make a model including assimilate partitioning. It is thus assumed that the variability of the profiles of seed numbers per node observed in the fields will only be understood with a model.

The model is a tool to understand the diversity observed in the fields.

2. The structure of the model

2.1 Seed and plant development

In the pod life, three main developmental stages can be defined : flowering, final stage in seed abortion and physiological maturity. Between flowering and final stage in seed abortion, cell divisions occur in the ovules (Ney et al., 1993). During this period, a whole pod or some of the ovules contained in it may abort, reducing the final seed number of the pod. After the final stage in

seed abortion the seeds cannot abort any more in the pod (Pigeaire et al., 1986) : the seed number is fixed in the pod. This stage corresponds also to the beginning of seed filling (Ney et al., 1993).

At the stem scale, the progression of these three main stages along the stem is linear in cumulative degree-days (Ney and Turc, 1993) : the duration between the occurrence of a stage on two consecutive nodes is constant, about 45 degree-days for the flowering, and about 40 degree-days for the final stage in seed abortion in cv. Solara. It is thus possible to know precisely the age of each pod of the stem, every day during the reproductive period, as soon as the date of beginning of flowering on the stem is recorded. Moreover the period of formation of seed number can be defined precisely between flowering of the first flower (FLO1) and the final stage in seed abortion of the last reproductive node (FSSA). This will be the period concerned for the model.

2.2 The model framework (Figure 2)

The original assumption of the model is that the seed number of a pod depends on the pod growth dynamics before its final stage in seed abortion (Figure 2). This was demonstrated for individual pods from all nodes, on various situations in the field (Jeuffroy and Chabanet, 1994). Then, the growth dynamics of each pod depends on the whole stem growth dynamics and on the pattern of assimilate partitioning among all sinks of the stem along time. Assimilate partitioning can then depend on the nature and age of the various sinks in competition, which are mainly developmental variables.

The inputs of the model are thus the characteristics of plant development and mean daily temperature, in order to simulate pod age at each node and the end of the vegetative development, and the mean growth rate per stem, in order to estimate the amount of available assimilate at each daily step.

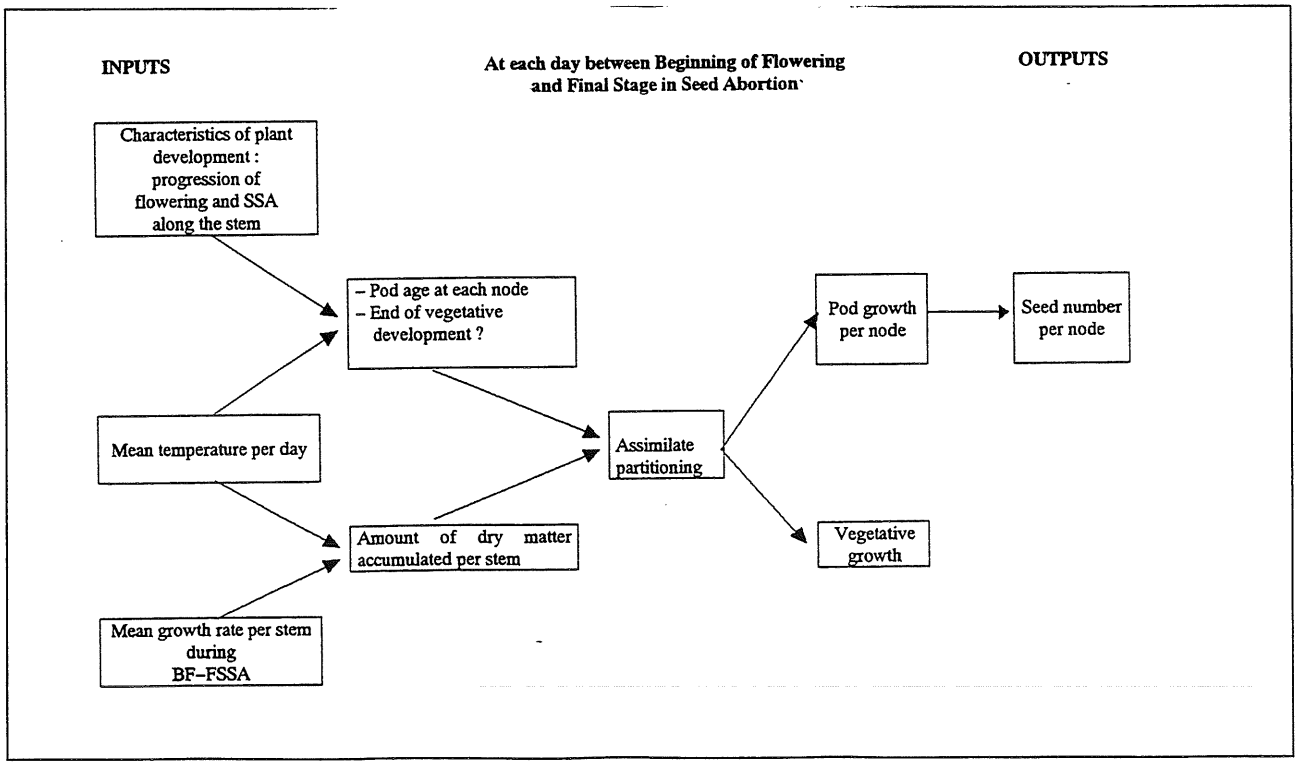


Figure 2. Structure of the model (according to hypothesis of plant functioning). BF = beginning of flowering : FSSA = final stage in seed abortion.

The whole model is thus composed of three main modules, (1) the estimation of seed number per pod according to early pod growth, (2) assimilate partitioning between sinks, and (3) determinism of the end of flowering (fixing the final number of flowering nodes). In this paper, I will focus on the module of assimilate partitioning.

3. The module of assimilate partitioning

3.1 The hypothesis of assimilate partitioning

Assimilate partitioning is the weak point of the majority of the available crop models (Whisler et al., 1986). Several analytical results exist in the literature in order to build this module. Yet, they are sometimes in contradiction or insufficient. Finally some questions still arise from literature. The first one concerns the possibility of internodal translocation of assimilate during the period of seed formation. Some physiological studies showed that the majority of assimilates allocated to a pod came from the subtending leaf (Flinn and Pate, 1970) while other studies showed the possibility for a pod to receive assimilates even if its subtending leaf did not produce assimilate (Szynkier, 1974). If the possibility of assimilate transfer between nodes exists, the second question concerns the existence of a hierarchy in the carbon distribution among the various sinks, some of them having priority on others. Finally the rules of assimilate partitioning among sinks during the whole period of seed set must be determined. In order to answer these questions, which are the structural assumptions of the model, several experiments were realized.

Data are necessary to test the hypothesis of plant functioning, given the structure of the model

3.2 The experiment to test these hypothesis

A first experiment was realized in order to analyze the possibility of assimilate transfer among nodes, and the general determinants for assimilate partitioning. It was thus necessary to work on short periods and not only with biomass balance sheets, because of the high variability between plants. The experiment was then carried out in controlled environment, with ^{14}C-labeling. For the question of assimilate translocation, some individual nodes were shaded, the whole plant was fed $^{14}CO_2$, and the pods from the shaded nodes were analyzed (see Jeuffroy and Warembourg, 1991, for details). For the question of priority of some sinks, control plants and partially shaded plants were given $^{14}CO_2$ (Jeuffroy and Warembourg, 1991), the latter producing only one-fourth assimilates recovered form controls.

3.3 Some results

The possibility of assimilate translocation between nodes was demonstrated (Jeuffroy and Warembourg, 1991), for several nodes and at various stages.

For the question of hierarchy, the amount of assimilates recovered in the different vegetative organs from the shaded plants was very low compared to the control, but the proportion of ^{14}C recovered in one vegetative organ compared to the total amount in the plant was the same between shaded and control plants (Figure 3). Thus the proportion of assimilates allocated to one vegetative sink does not vary according to the total amount of available assimilate in the plant.

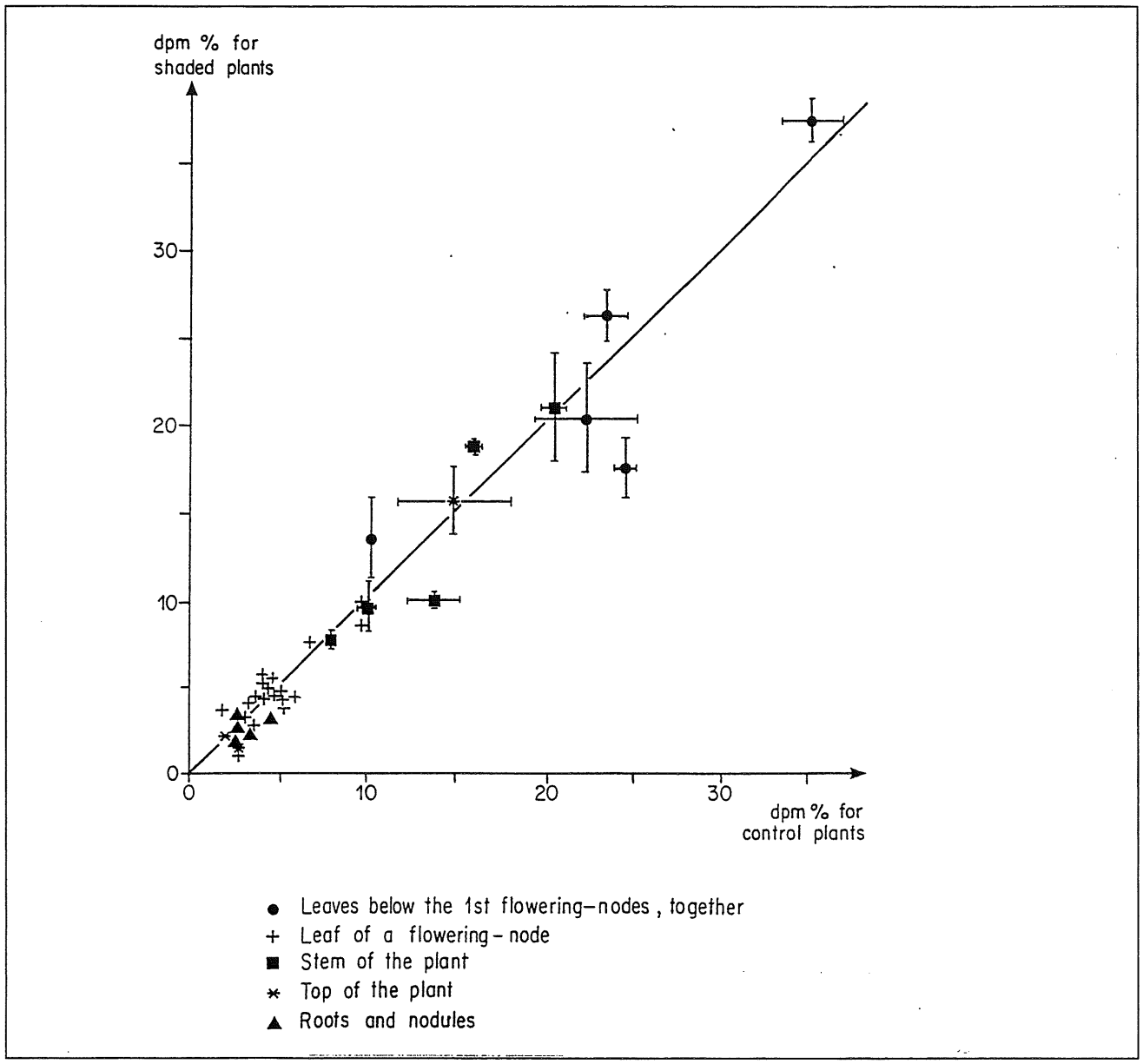


Figure 3. Relative distribution of ^{14}C-assimilates among vegetative organs (dpm % of total recovered in plant). Comparison between control and shaded plants. Bars indicate standard errors (only for the high % dpm). (From Jeuffroy and Warembourg, 1991).

On reproductive organs, the result was less obvious, as there was a high variability between pods (Figure 4). Nevertheless, it was not possible to show an effect of the total amount of assimilates produced on the proportion allocated to each pod. Looking at the pods more precisely, it was possible to link the proportion of ^{14}C allocated to one pod either to its relative dry matter (compared with all the pods) or to its relative seed number, according to the stage of the pod (before or after its final stage in seed abortion).

Thus, it was not possible to determine some priority of sinks over others, including in the competition between vegetative and reproductive sinks. The amount of assimilate allocated to each sink depends on its demand, proportional to the pod dry matter (for pods before final stage in seed abortion), proportional to the pod seed number (after this stage), and proportional to the difference between the final number of flowering nodes and the number already developed (for vegetative organs, Jeuffroy, 1991).

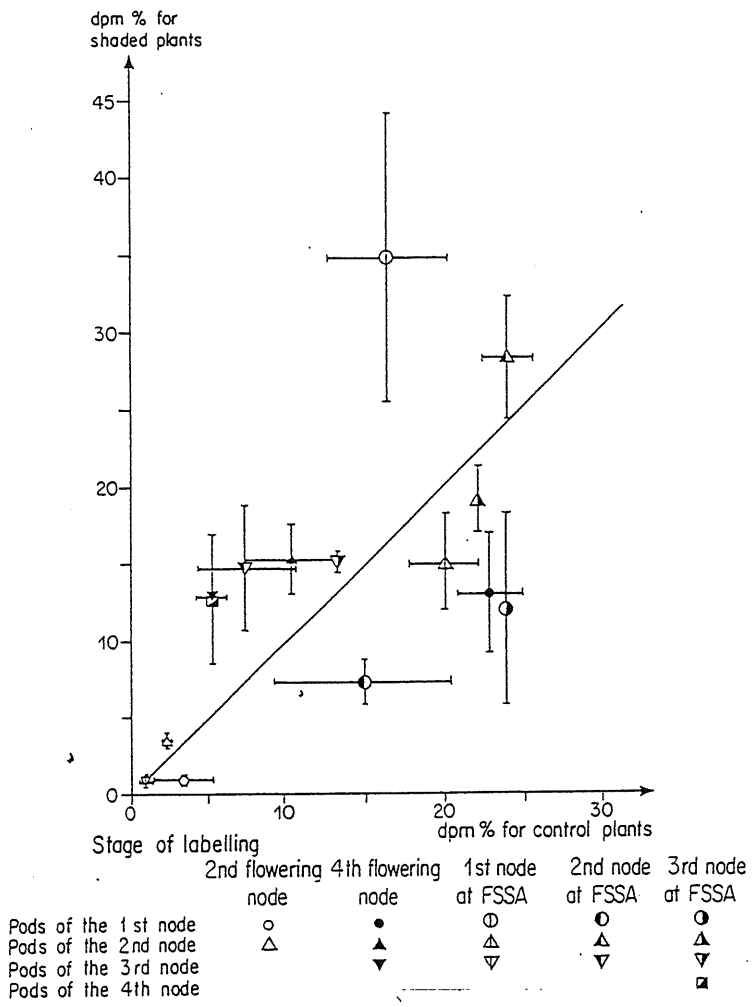


Figure 4. Relative distribution of ^{14}C-assimilates among pods (dpm % of total recovered in plant). Comparison between control and shaded plants. Bars indicate standard errors. (From Jeuffroy and Warembourg, 1991).

The calculations of the amounts of assimilates allocated to each sink at each step then required the estimation of the parameters for these demands.

The experiment is a source for the estimation of the model parameters

3.4 Estimation of the model parameters

This new step required another experiment, as it was better to estimate these parameters directly in the field, where the model had to be used. In this experiment, some particular conditions were necessary, in order to measure the potential growth of a pod, which was the definition for pod demand. On field grown plants, we cut off the apex after the development of 4 flowering nodes and let only one pod on each stem, in order to prevent competition on the stem. Then, we measured

individual pod dry matter, which was adjusted to an exponential model in the first part of the life (before final stage in seed abortion) and to a linear model in the second part (between this stage and physiological maturity). This curve gave us the parameters for pod demand (Jeuffroy and Devienne, 1995). The same principles were applied for the estimation of the vegetative demand parameters (Jeuffroy, 1991).

3.5. Evaluation of the assimilate partitioning module in the field

With these two experiments (labeling and parameter estimation), the rules for assimilate partitioning were determined. But, it was necessary to verify them in the conditions where the model had to be used, because the plants obtained in controlled environments in the first experiments were very different form those observed in the field. Thus, another experiment was realized in the field, different from that used for parameter estimation (Jeuffroy and Devienne, 1995).

Experimental data are necessary to evaluate the sub-models in the range of situations where the model must be used

Evaluating this module of assimilate partitioning in a large range of situations to be covered by the whole model, we got various sowing dates and densities during two years, inducing a large range of inputs of the model (Table 1). The mean growth rate per stem ranged from 5 to 14 mg/stem/degree-day, and the final number of flowering nodes from 4 to 10. Variability existed also in the developmental parameters. Here, observed and simulated values of dry matter per organ were compared. There was a good fit between observed and simulated vegetative dry matter increments, calculated between two successive sampling dates (Figure 5), and also for pod dry matter at each sampling date (Figure 6). The model gave also a good simulation of the growth dynamics of vegetative and reproductive organs (Figure 7).

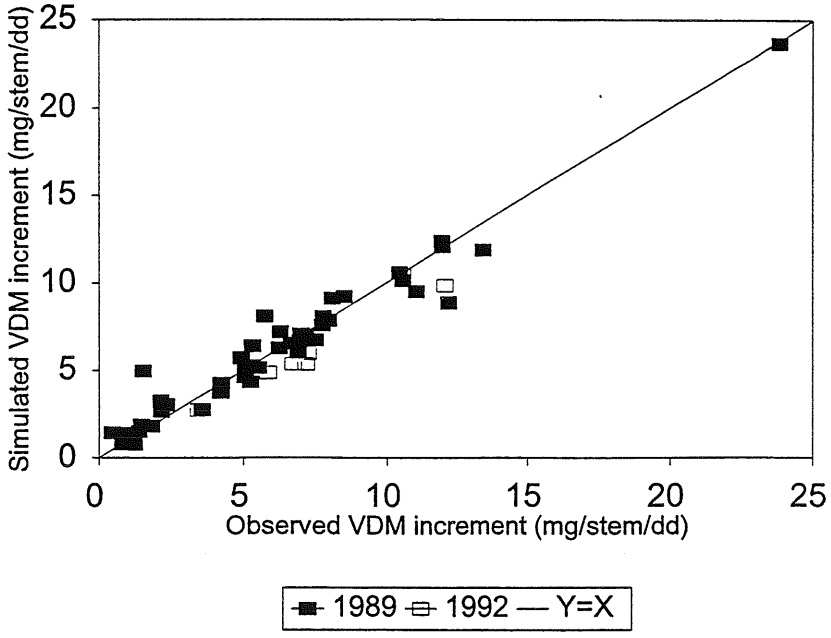


Figure 5. Relationship between observed and simulated vegetative dry matter increments (VDM), calculated between two successive sampling dates during the period of seed set, on the situations from Table 1.

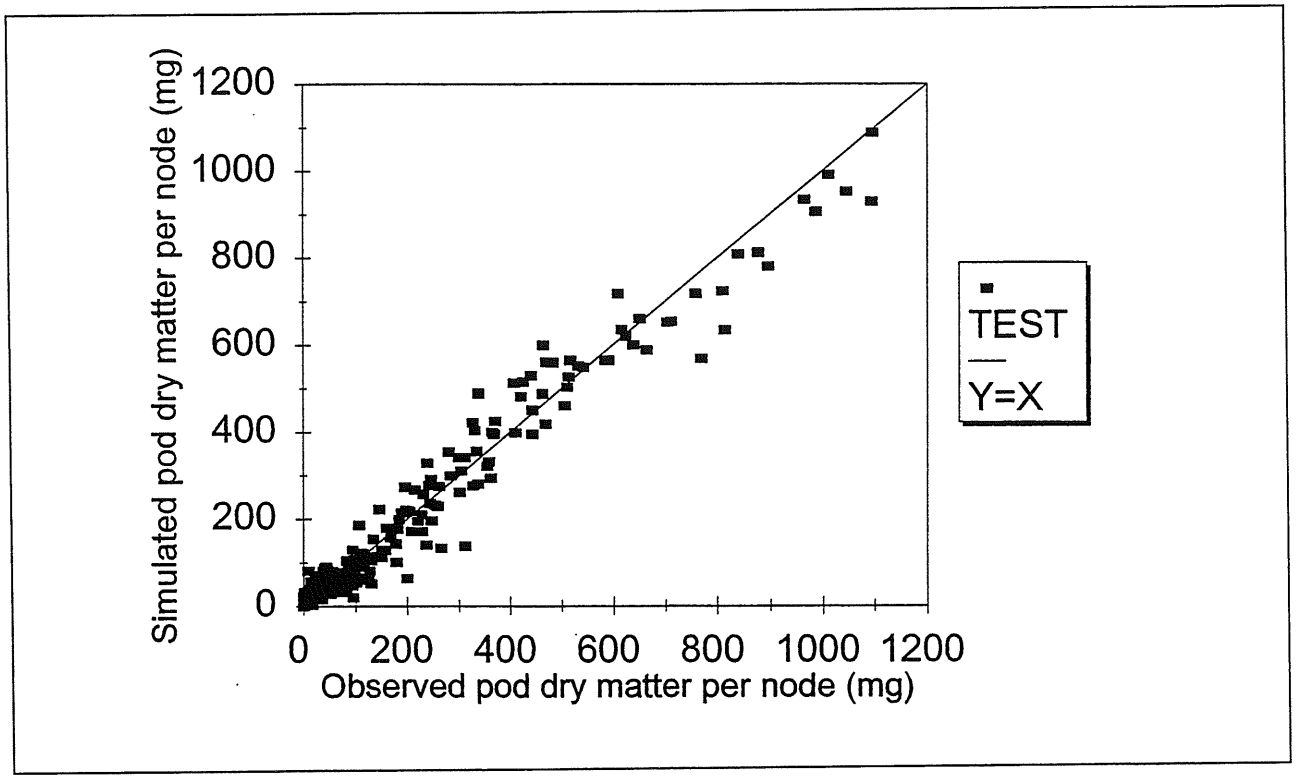


Figure 6. Relationship between observed and simulated pod dry matter per node, at each sampling date during the period of seed set, on the situations from Table 1.

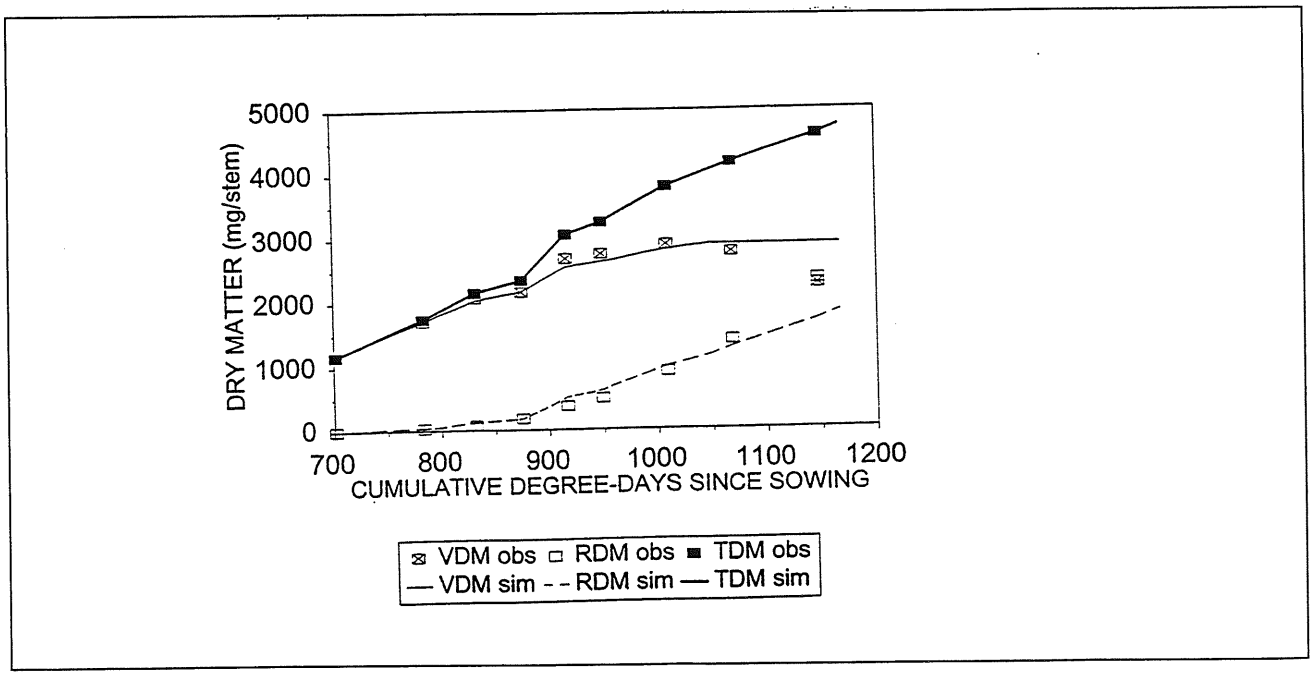


Figure 7. Simulated and observed growth dynamics of vegetative and reproductive parts of a pea plant. VDM = vegetative dry matter ; RDM = reproductive dry matter ; TDM = total dry matter; obs = observed ; sim = simulated.

Table 1. Characteristics of the different crops used to test the model. (From Jeuffroy and Devienne, 1995).

Year	Sowing date	Nr. of stems m^{-2}	Mean growth rate per stem (mg $stem^{-1}$ dd^{-1})	S_F (dd)	$1/R_F$ (dd)	r^2 a	S_S (dd)	$1/R_S$ (dd)	r^2 b	Final nr. of flowering nodes per stem
1989	14 -3	117	9.47	756	68.5	.99	1019	40.0	1.00	6
		156	8.19	727	79.5	.98	1022	35.0	c	5
		257	5.16	785	54.0	.99	1022	54.0	c	4
	31 -3	61	14.30	775	36.5	.98	943	41.0	.96	7
		107	8.87	773	35.5	.99	935	35.0	.94	6
		194	5.89	778	49.0	.99	948	39.0	c	5
	19-4	81	12.63	729	36.0	.91	908	68.0	.92	6
		141	7.80	705	46.0	.95	908	59.5	.90	5
		175	5.77	711	40.0	.95	910	35.0	c	5
1992	24 -2	67	12.98	860	37.5	.98	1068[d]	45.0[e]		10
		97	10.60	852	35.9	.98	1060[d]	45.0[e]		9
		159	7.56	841	42.3	.97	1049[d]	45.0[e]		9

S_F: date of flowering of the first node (in cumulative degree-days from sowing)
R_F: rate of progression of flowering (in node degree day^{-1})
S_S: date of final stage in seed abortion on the first node (in cumulative degree-days from sowing)
a: coefficient of correlation of the adjustement of the number of nodes which were flowered with cumulative degree-days since sowing
b: coefficient of correlation of the adjustement of the number of nodes which had passed FSSA with cumulative degree-days since sowing
c: only 2 observations were available
d: estimated from mean S_S-S_F of 1989

4. Analysis of the various profiles of seed numbers per node observed in the fields with the model

The initial aim for the building of the model was to understand the diversity of the profiles of seed numbers per node observed in the field. In order to test the ability of the model to give a good account of this range, an evaluation of the model in a large range of situations in the field was carried out. This step also aimed to determine the range of validity for the model.

The model must be evaluated in the range observed in the fields

The evaluation of the whole model was realized on the range of sowing dates and densities presented in Table 1, resulting in a large range of the inputs and outputs. Figure 8 shows the observed profiles of seed number for the situations tested : some of them are high with few seeds at the bottom, while others are lower with many seeds at the base.

The analysis can then be detailed on two situations particularly differing in number of flowering nodes and in mean growth rate, the 3rd (D3S1) and the 10th (S11) cases on Table 1. There is a good fit between the observed and the simulated profiles in the two cases (Figure 9).

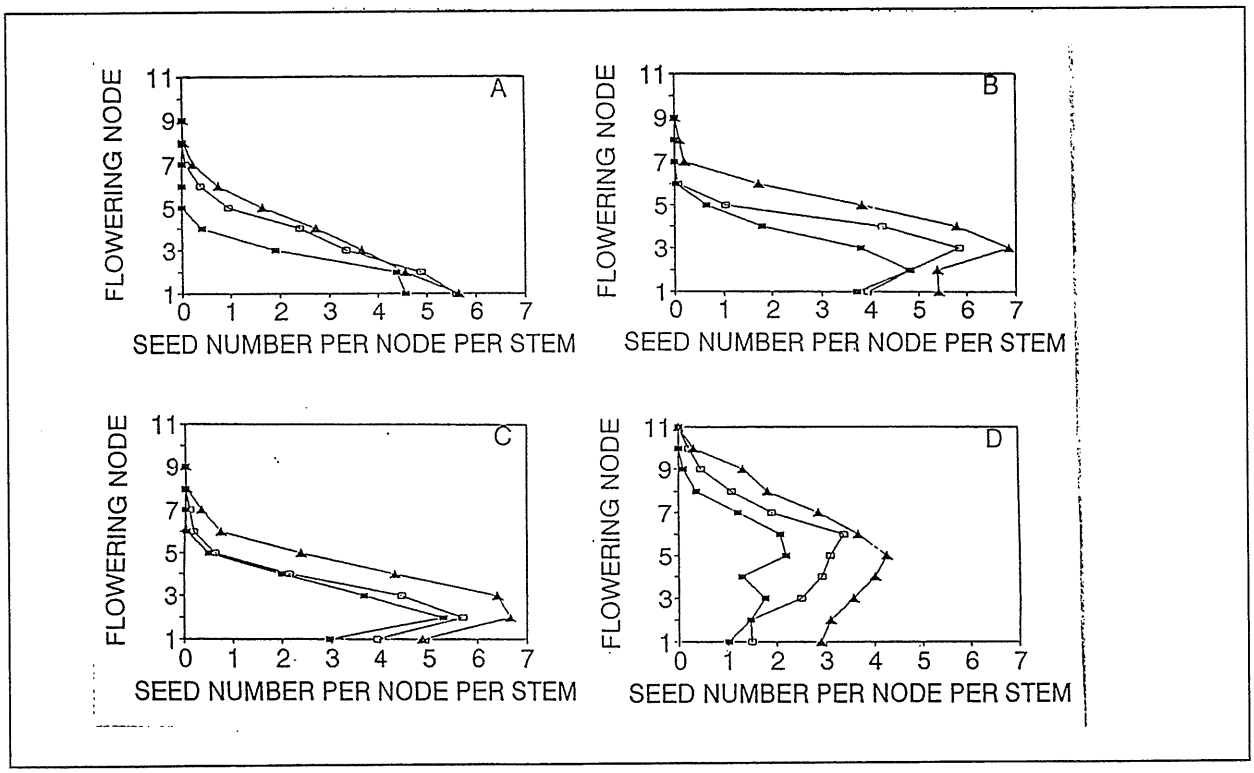


Figure 8. Observed profiles of seed numbers per node on the situations from Table 1. (From Jeuffroy and Devienne, 1995).
A: sowing 14 March 1989, 117 stems m^{-2} (▲), 156 stems m^{-2} (□), 257 stems m^{-2}(■)
B: sowing 31 March 1989, 61 stems m^{-2}(▲), 107 stems m^{-2}(□), 194 stems m^{-2}(■)
C: sowing 19 April 1989, 81 stems m^{-2} (▲), 141 stems m^{-2}(□), 175 stems m^{-2}(■)
D: sowing 24 February 1992, 67 stems m^{-2}(▲), 97 stems m^{-2}(□), 159 stems m^{-2}(■)

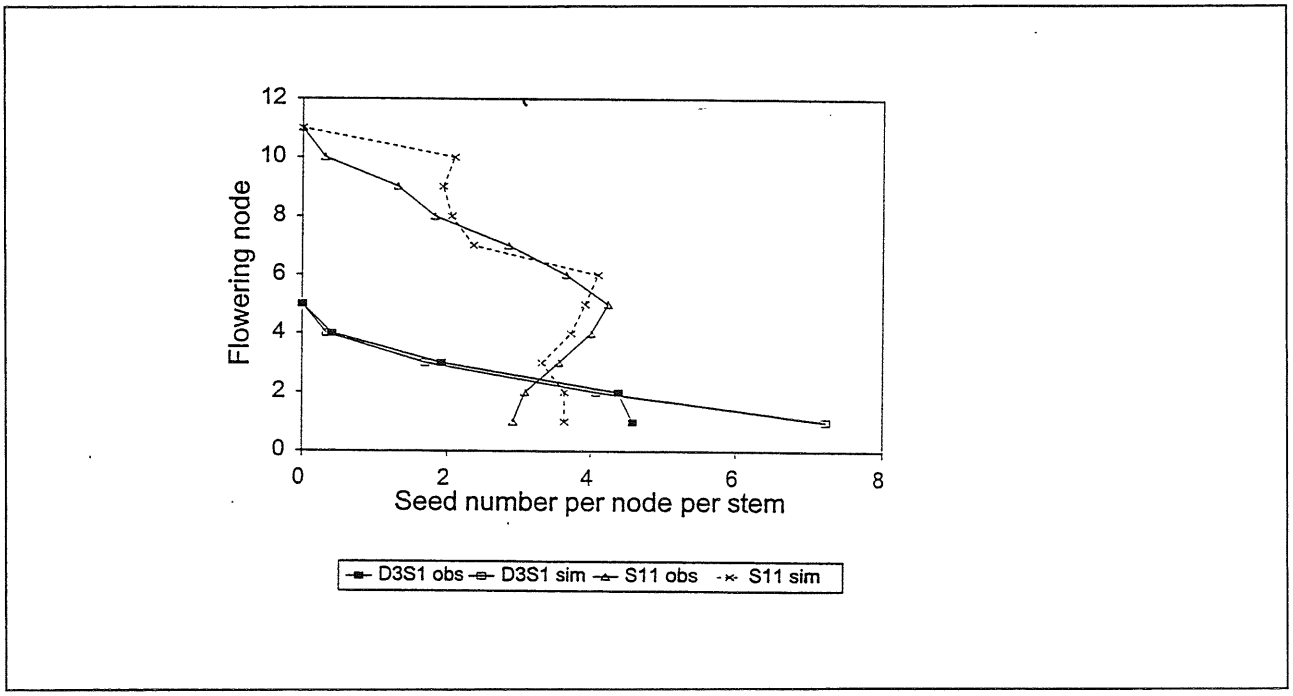


Figure 9. Comparison of the simulated and observed profiles of seed numbers per node on two situations from Table 1 (S11 and D3S1). S11 = sowing 24 February 1992, 67 stems m^{-2}, D3S1 = sowing 14 March 1989, 257 stems m^{-2}, obs = observed, sim = simulated.

The two different forms of profiles can be interpreted with the model. The higher number of reproductive nodes obtained on S11 (because of a better nitrogen nutrition status of the crop at the beginning of flowering, Jeuffroy and Sebillotte, 1997) means that the vegetative development and growth continued during a longer period on this treatment. In the competition between vegetative organs and the first growing pods, a very small amount of assimilate was allocated to the pods, in comparison to their demand. This resulted in a small number of seeds on these nodes, when compared to the same nodes on the other treatment D3S1, for which the competition between vegetative and reproductive organs was very short. Later in the cycle, the competition for assimilates among pods in the two situations resulted in a higher amount of assimilates allocated to the older pods, reducing the seed number on the upper last ones.

As the model gave also a good account of the observed profiles for the other situations tested , it can be concluded that it is a good tool to understand the diversity of the situations observed in the range of inputs tested. Yet there are some situations still not explained by the model. For example, the model generally overestimated the seed number on the last nodes, probably because it is necessary to consider the filling seeds as having a priority in assimilate distribution, as it was shown on soybean (Munier-Jolain, 1994). The lack of fit is then an help to improve the model. In the case of this model, some modules are being improved in our team, for example the estimation of the vegetative demand, and the prediction of the final number of reproductive nodes.

The lack of fit between observed and simulated situations helps to improve the model

5. Analyzing the effect of limiting factors with the model

The model already allows us to go further, and to give an account of the effect of limiting factors which are not included in the model. For example, in evaluating the model during the year 1991, the simulated profile was very different from the observed one, particularly on the nodes 3 and 4 (Figure 10). Looking at the climatic data, it was observed that several days with very high maximum temperature occurred during the period of seed set. And it is known in the literature that high temperatures during seed set cause seed abortion (Lambert and Linck, 1958 ; Karr et al;, 1959) and that the period of highest sensitivity for the pods is the second half of the period between flowering and FSSA (Jeuffroy et al., 1990). In comparing the days with high temperature to the period of seed formation on each node, it appeared that these days occurred during the period of sensitivity of the nodes 3 and 4 (Figure 11), and could explain the gap between the observed and the simulated profiles. The model then enables to quantify the yield loss due to the limiting factor.

The model is a tool to quantify the effects of limiting factors which are not included in the model.

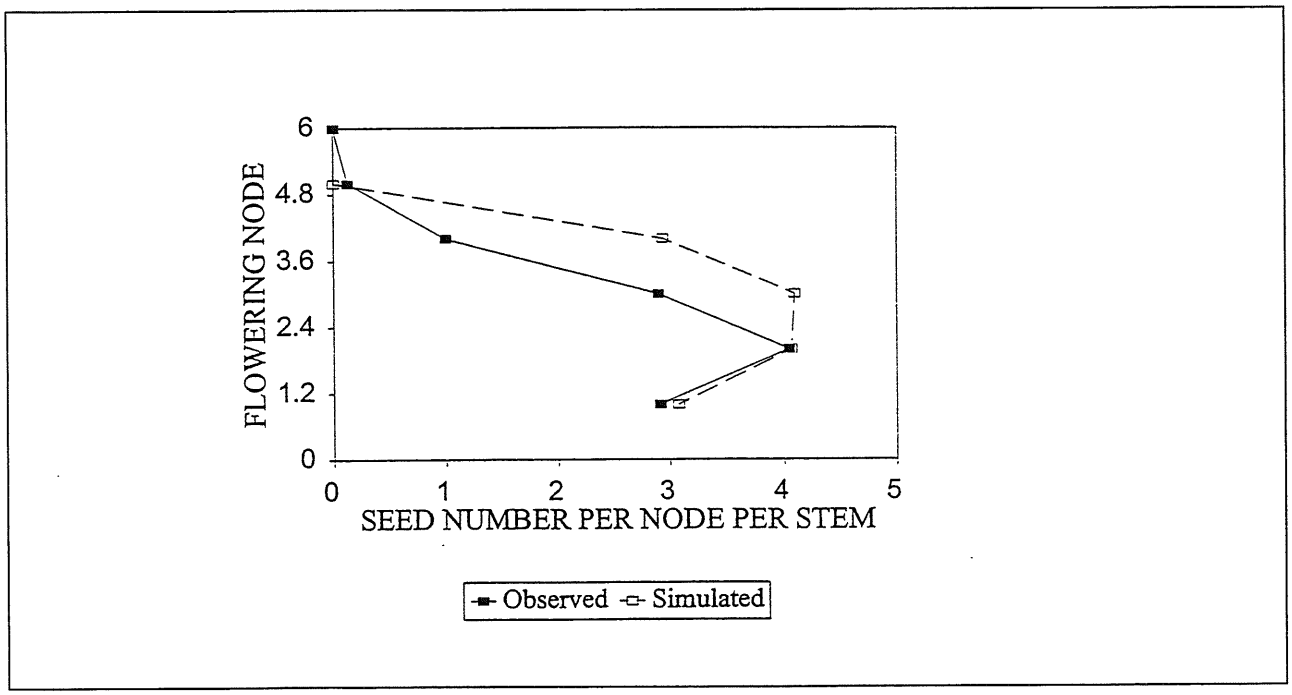


Figure 10. Comparison of the simulated and observed profiles of seed numbers per node on one situation in 1991.

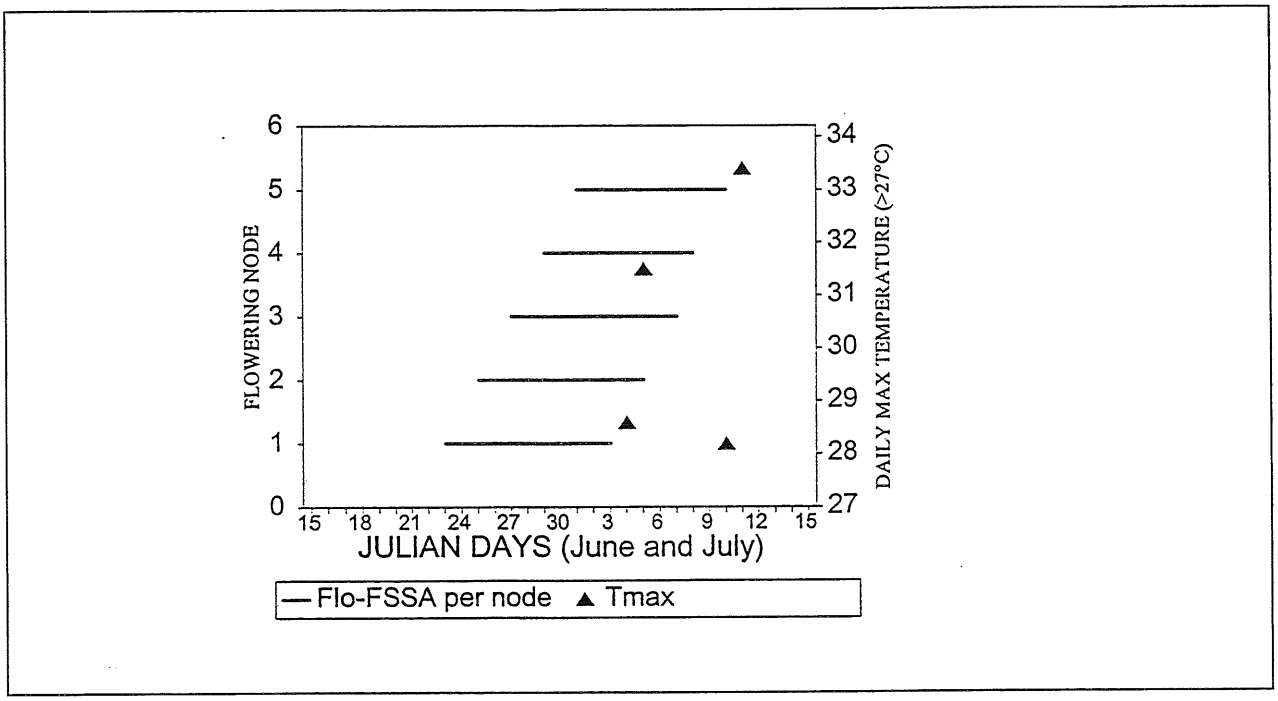


Figure 11. Comparison of the periods of seed formation on each node and the days with high maximum temperature.

6. Conclusion

In this description of building and use of a model for seed number per node on pea, several relationships between data and model were illustrated.

1. The model is a tool to understand the diversity observed in the fields.
2. Data are necessary to test the hypothesis of plant functioning, giving the model structure.
3. The experiment is a source for the estimation of the model parameters.
4. Experimental data are necessary to evaluate the sub-models in the range of situations where the model must be used.
5. The lack of fit between observed and simulated situations helps to improve the model.
6. The model gives an account of the effects of limiting factors which are not included in the model and quantifies their effects.

7. References

Boote KJ, Jones JW, Pickering NB 1996 Potential uses and limitations of crop models. Agron. J. 88: 704-716

Doré T 1992 Etude, par voie d'enquête, de la variabilité des rendements et des effets précédent du pois protéagineux de printemps (*Pisum* sativum L.).Thèse de Doctorat, INA Paris-Grignon.

Flinn AM, Pate JS 1970 A quantitative study of carbon transfer from pod and substending leaf to the ripening seeds of the field pea (*Pisum arvense* L.). J. Exp. Bot. 21: 71-82

Jeuffroy M-H 1994 Le nombre de graines par tige. Agrophysiologie du pois protéagineux ; applications à la production agricole. INRA, ITCF, UNIP Eds

Jeuffroy M-H 1991 Rôle de la vitesse de croissance, de la répartition des assimilats et de la nutrition azotée dans l'élaboration du nombre de graines du pois protéagineux de printemps (*Pisum sativum* L.). Thèse de Doctorat, Univ. Paris XI

Jeuffroy M-H, Chabanet C 1994 A model to predict seed number per pod from early pod growth rate in pea (*Pisum sativum* L.). J. Exp. Bot. 45: 709-715

Jeuffroy M-H, Devienne F 1995 A simulation model for assimilate partitioning between pods in pea (*Pisum sativum* L.) during the period of seed set ; validation in field conditions. Field Crops Res. 41: 79-89

Jeuffroy M-H, Duthion C, Meynard J-M, Pigeaire A 1990 Effect of a short period of high day temperatures during flowering on the seed number per pod of pea (*Pisum sativum* L.). Agronomie 2: 139-145

Jeuffroy M-H, Sebillotte M 1997 The end of flowering in pea : influence of plant nitrogen nutrition. Eur. J. Agron., in press

Jeuffroy M-H, Warembourg FR Carbon transfer and partitioning between vegetative and reproductive organs in *Pisum sativum* L. Plant Physiol. 97: 440-448

Karr EJ, Linck AJ, Swanson CA 1959 The effect of short periods of high temperature during day and night periods on pea yields. Amer. J. Bot. 46: 91-93

Lambert RG, Linck AJ 1958 Effects of high temperature on yield of peas. Plant Physiol. 33: 347-350

Monteith JL 1996 The quest for balance in crop modeling. Agron. J. 88: 695-697

Munier-Jolain NG 1994 Etude de la variabilité du poids individuel des graines du soja de type indéterminé (*Glycine max.* L. Merrill, cv Maple Arrow). Influence de l'apparition séquentielle des organes reproducteurs. Thèse de Doctorat, INRA Paris-Grignon

Ney B, Duthion C, Fontaine E 1993 Timing of reproductive abortions in relation to cell division, water content and growth of pea seeds. Crop Sci. 33: 267-270

Ney B, Turc O 1993 Heat-unit-based description of the reproductive development of pea. Crop Sci. 33: 510-514

Pigeaire A, Duthion C, Turc O 1986 Characterization of the final stage in seed abortion in indeterminate soybean, white lupin and pea. Agronomie 6: 371-378

Szynkier K 1974 The effect of removing of supply or sink organs on the distribution of assimilates in two varieties of garden pea, *Pisum sativum* L.. Acta Agric. Scand. 24: 7-12

Whisler FD, Acock B, Baker DN, Fye RE, Hodges HF, Lambert JR, Lemmon HE, Mc Kinion JM, Reddy VR 1986 Crop simulation models in agronomic systems. Adv. Agron.: 141-208

1.3 Quantification of Soil and Nutrient Losses by Wind Erosion in Niger, West Africa

Geert Sterk[1], Alfred Stein[2] and Ludger Herrmann[3]

1. Department of Irrigation and Soil & Water Conservation, Wageningen Agricultural University, Nieuwe Kanaal 11, 6709 PA Wageningen, The Netherlands.

Email: Geert Sterk@Users@TCT.WAU

2. Department of Soil Science and Geology, Wageningen Agricultural University, PO Box 37, 6700 AA Wageningen, The Netherlands.

3. Department of Soil Science (310), University of Hohenheim, 70593 Stuttgart, Germany.

Wind erosion is a major soil degradation processes in the African Sahel. Wind-blown material contains nutrients and therefore wind erosion causes a decline in soil fertility. This paper quantifies soil and nutrient losses by wind erosion. Soil particle transport was measured in the Sahelian zone of Niger with 21 sediment catchers in a plot of 40 by 60 m during four storms. During the two biggest storms, sediments were collected and analysed for total K, C, N, and P. The main mass of nutrients was transported just above the soil surface by saltation. The suspended nutrient mass fluxes were an order of magnitude lower than the saltation fluxes but were extended to greater heights. Therefore, suspension transports also significant quantities of nutrients. Mass budgets were calculated for the four nutrients and the following losses from the experimental plot during the two storms were estimated: 57.1 kg ha^{-1} K, 79.6 kg ha^{-1} C, 18.3 kg ha^{-1} N, and 6.1 kg ha^{-1} P. The observations of total mass transport were used for a geostatistical analysis. Storm based maps of mass transport were produced with kriging. The maps show a large spatial variation in particle mass transport and were used to estimate net soil losses from the experimental plot. In total, 45.9 ton ha^{-1} got lost during the season.

1. Introduction

Wind erosion is the removal of soil material (soil particles, nutrients and organic matter) by wind. Linked to wind erosion is sedimentation, which is the deposition of wind-blown material. Between erosion and sedimentation, the material is transported by saltation, creep and suspension (Bagnold, 1973). Saltating particles jump and bounce over the surface, reaching a maximum height of approximately 1 m, but the main particle mass moves just above the soil surface. Saltation transports soil particles with sizes roughly between 63 and 500 mm. When saltating particles fall to the soil surface they not only eject other saltating particles but also induce creep, the rolling and sliding of larger particles (>500 mm), and suspension, the raising and transport of dust particles (<63 mm). During a storm, creep can move particles over distances from a few centimetres to several metres, saltating particles travel from a few metres to a few hundred metres, and suspension transport ranges from several tens of metres to thousands of kilometres.

Wind-blown material contains nutrients and organic matter (Zobeck and Fryrear, 1986), and hence, wind erosion causes a decline in soil fertility. Long-term agricultural effects are decreasing crop yields and, in the worst cases, degraded land which cannot be used for agriculture anymore. The loss of nutrients is usually ascribed to suspension transport (e.g., Zobeck and Fryrear, 1986; Leys and McTainsh, 1994). Suspension selectively removes the finest soil particles that contain

disproportionately greater concentrations of plant nutrients (Young et al., 1985). However, little information exists on nutrient transport by saltation, which moves the main mass of soil particles (Chepil, 1945). Creep is considered not to tranport significant quantities of nutrients, since it transports mainly coarse sand which is poor in nutrients.

Quantification of wind erosion is difficult because of large spatial and temporal variability in particle mass fluxes (Wilson and Cooke, 1980). Soil erodibility is determined by several variables like soil texture, surface roughness and topography. These variables usually show spatial variation, resulting in spatial variation in soil erodibility and thus in wind-blown mass transport as well. Moreover, erosive storms differ in duration, wind speed, and wind direction, which causes a wide range of particle mass transport rates.

The objective of this study was to quantify soil and nutrient losses by wind erosion from detailed observations of soil particle transport.

2. Materials and Methods

A field experiment was conducted in the African Sahel, at the International Crops Research Institute for the Semi-Arid Tropics (ICRISAT) Sahelian Center (ISC). The site is located in South-West Niger, 45 km south of the capital, Niamey. The region is characterised by chemically and physically poor, sandy soils, and harsh climatic conditions. Wind erosion occurs mainly during the early rainy season (May - July) when short (10 - 30 min) wind storms often precede rainfall events. Normally, about ten of those events occur during the season.

2.1 Quantification of particle mass transport

Wind erosion was measured in a plot of 40 by 60 m on a sandy alfisol, during the rainy season of 1993. The measurements of horizontal particle mass transport were conducted with Modified Wilson and Cooke sediment catchers (Fig. 1; Sterk and Raats, 1996). Twenty-one catchers were regularly distributed over the plot in three rows of six, and with the three remaining catchers placed between two rows (Fig. 2). Four storms were sampled during the whole season.

The MWAC catcher has seven sediment traps attached at heights between 0.05 and 1.00 m (Fig. 1), which means that trapped materials were a mixture of saltation and suspension particles. Creep particles were not trapped. After a storm, the sediment in each trap was collected and weighed. Through the seven observations of horizontal particle mass flux per catcher, a model was fitted to describe a vertical profile (Sterk and Raats, 1996):

$$q(z) = a\left(\frac{z}{\alpha} + 1\right)^{-b} + c\exp\left(-\frac{z}{\beta}\right) \qquad (1)$$

where $q(z)$ is the horizontal particle mass flux (kg m^{-2} s^{-1}) at height z (m), and a, α, b, c, and β are regression coefficients. Integration of the vertical profile over height from z = 0 to 1 m and correction for the trapping efficiency of the MWAC catcher (= 0.49) resulted in a total particle mass transport rate (kg m^{-1} s^{-1}) at the point of sampling. When multiplied by the storm duration, a total particle mass transport value (kg m^{-1}) was obtained. This value represents the total mass of sediment below 1 m that passed a strip 1 m wide and perpendicular to the mean wind direction of the storm. It is assumed that the contribution of the sediment moving above 1 m to the total mass transport was negligible.

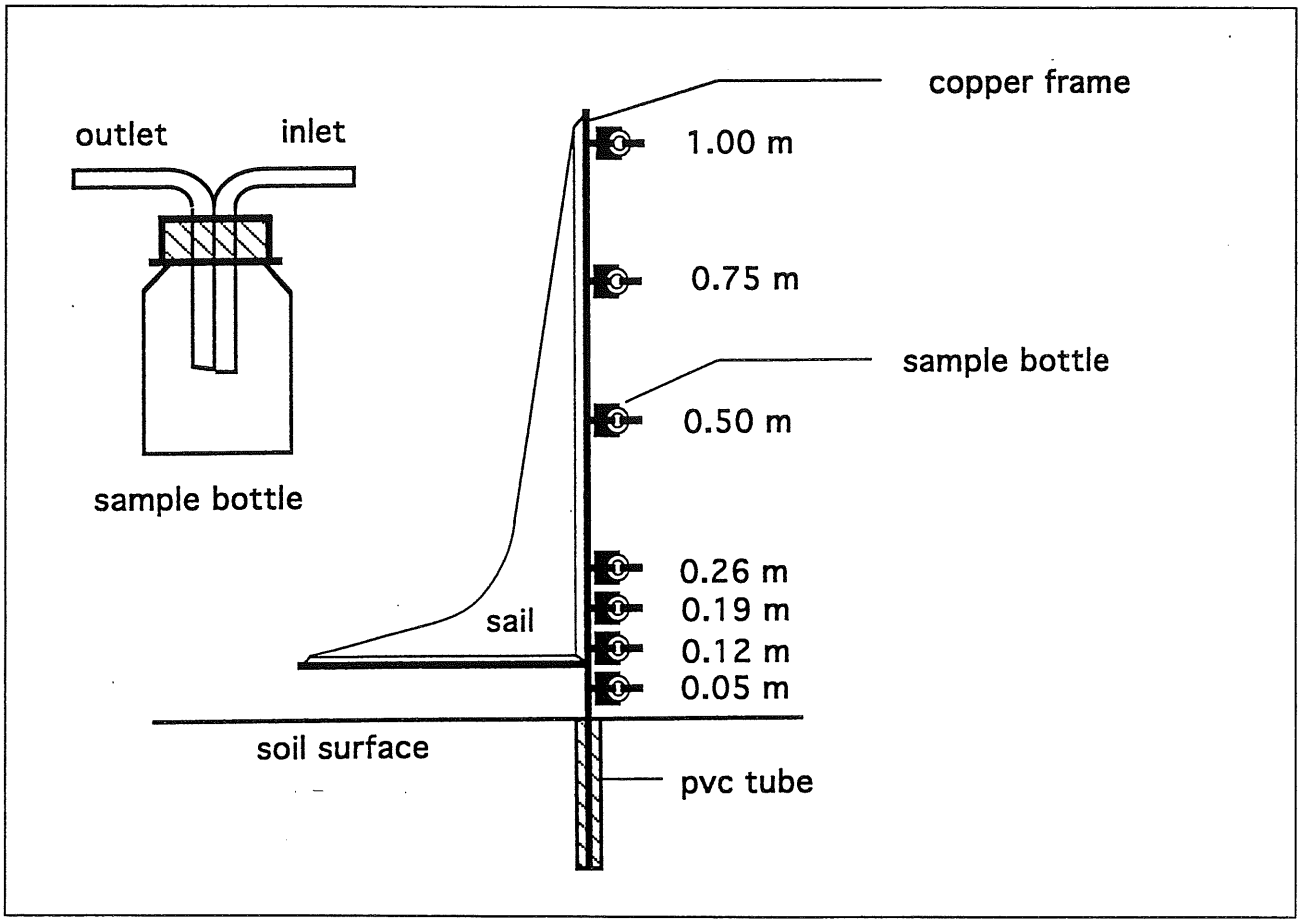


Figure 1. The Modified Wilson and Cooke sediment catcher.

2.2 Quantification of nutrient transport

From the 18 catchers in the three rows of six, the materials trapped at three heights (0.05, 0.26 and 0.50 m) and samples of topsoil material were collected. The samples were analysed with regard to total element contents of potassium (K), carbon (C), nitrogen (N), and phosphorus (P). Total K and P were measured with X-ray fluorescence, and total C and N were measured with gas chromatography (Sterk et al., 1996).

Through the nutrient contents of the wind-blown sediment a simple power function was fitted to describe the vertical profiles of K, C, N, and P contents (Zobeck and Fryrear, 1986):

$$T(z) = p z^q \tag{2}$$

where $T(z)$ is the total element content (mg kg^{-1}) at height z, and p and q are (positive) regression coefficients. In this model, $T = 0$ at $z = 0$ suggesting that creep is not transporting nutrients. Multiplication of eq. (1) and (2) yields an equation that describes the vertical profile of horizontal nutrient mass fluxes (Sterk et al., 1996):

$$f(z) = p z^q [a(\frac{z}{\alpha}+1)^{-b} + c\exp(-\frac{z}{\beta})] \tag{3}$$

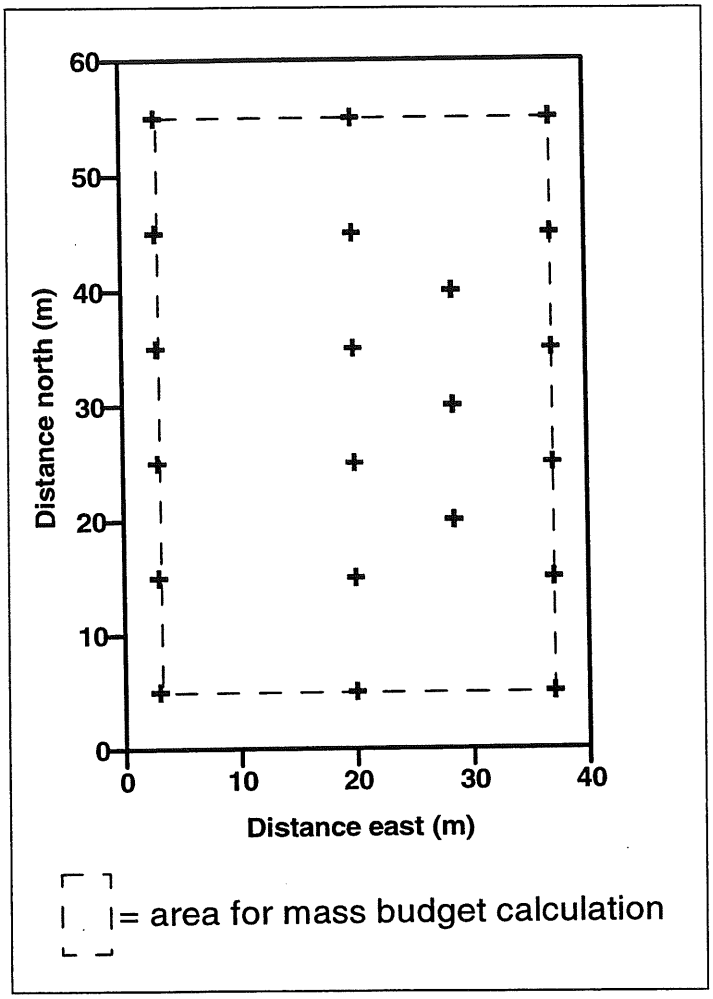


Figure 2. Experimental plot with positions of 21 MWAC sediment catchers.

where $f(z)$ is the horizontal mass flux (mg m^{-2} s^{-1}) of a certain element at height z. The nutrient mass flux profiles were numerically integrated over height from z = 0 to 1 m. The total nutrient mass transport rates (mg m^{-1} s^{-1}) were obtained after correction for the overall trapping efficiency of the MWAC catcher. Multiplying the total nutrient mass transport rates by the storm duration resulted in total nutrient mass transport values (mg m^{-1}).

2.3 Geostatistical analysis

The observations of total particle mass transport were used for a geostatistical analysis. Mass transport is denoted by the regionalised variable $X(s,t)$, where s and t indicate space and time, respectively. The observations for each location and storm are denoted by $x(s_{ij},t_j)$, with $j = 1,..,m$ the different storms, and $i = 1,..,n_j$ the different spatial locations. Measurements were done during $m = 4$ storms at $n_j = 21$ locations.

Spatial dependence between different observations was modelled with the variogram. Since 21 observations are not sufficient to reliably estimate a variogram for each storm (Webster and Oliver, 1992), the analysis was extended from the space domain into the space-time domain (Sterk and Stein, 1997).

The four storms were pooled to create one data set for variogram estimation. Two assumptions underlie this procedure: (i) the expectation exists and is independent of the position s but may depend upon time t, $E[X(s,t)] = m(t)$ and (ii) the variance of the difference between $X(s,t)$ and $X(s+h,t)$ is finite and is not dependent of s. It is defined as:

$$Var[X(s,t) - X(s+h,t)] = E[X(s,t) - X(s+h,t)]^2 = 2\gamma(h,t) \quad \forall_s \tag{4}$$

where $\gamma(h,t)$ is the variogram at time t, and h is the lag distance between any two points in the space domain. Moreover, it is assumed that the four storms are independent temporal replicates of wind-blown mass transport, with similar spatial correlation structures (Sterk and Stein, 1997). Hence, by standardizing the four storms a variogram can be estimated which is independent of time.

The storms were standardized to the mean mass transport value of all four storms:

$$y(s_{ij}, t_j) = \hat{\mu}_s \frac{x(s_{ij}, t_j)}{\hat{\mu}_j} \quad (5)$$

where $y(s_{ij},t_j)$ is the standardized mass transport at observation location i and for storm j, and $\hat{\mu}_s$ and $\hat{\mu}_j$ are the mean mass transport values of the standardized data set and the j-th storm, respectively. Obviously, the conditions (i) and (ii) are now fulfilled, i.e. $E\ [y(s_{ij},t_j)] = \hat{\mu}_s$, and $Var\ [Y(s,t) - Y(s+h,t)]$ is independent of time t.

Using the standardized observations, variogram values were estimated with (Sterk and Stein, 1997):

$$\hat{\gamma}_s(h) = \sum_{j=1}^{4} \frac{1}{2n_j(h)} \sum_{i=1}^{n_j(h)} [y(s_{ij}, t_j) - y(s_{ij} + h, t_j)] \quad (6)$$

Now each pair of standardized observations at times t_j with j = 1,..,4 contributes to estimate the variogram. Through the estimation variogram a spherical model was fitted:

$$g(h) = \begin{cases} 0 & for \quad h = 0 \\ C_0 + C \cdot \left(\frac{1}{2}(h/r) - \frac{3}{2}(h/r)^3\right) & for \quad 0 < h < r \\ C_0 + C & for \quad h > r \end{cases} \quad (7)$$

where the parameter C_0 is the nugget constant, C the sill parameter, and r the range parameter of the variogram model (Journel and Huijbregts, 1978). The parameters C_0, C, and r were estimated with a weighted nonlinear regression procedure.

The obtained variogram model was converted into four variogram models, one for each storm. The range parameter of the model is the same for all storms, but the nugget constant and the sill parameter are dependent on the storm's mean mass transport. New values were calculated using the mean mass transport values of the four storms ($\hat{\mu}_j$) and the mean mass transport value of the standardized data set ($\hat{\mu}_s$):

$$C_{0,j} = C_{0,s}\left(\frac{\hat{\mu}_j}{\hat{\mu}_s}\right)^2 \quad (8)$$

$$C_j = C_s\left(\frac{\hat{\mu}_j}{\hat{\mu}_s}\right)^2 \quad (9)$$

where j denotes the storm number and s the standardized data set.

Maps of wind-blown mass transport were produced with kriging. The program OKB2D from the Geostatistical Software Library (Deutsch and Journel, 1992) was applied. This program uses a two-dimensional ordinary kriging algorithm. At each time t_j (j = 1,..,4) it predicts a value of mass

transport (X) at a location (s_0) on the basis of a number (v) of surrounding neighbourhood points measured at the same storm. The predictor P for the value $X(s_0,t_j)$ is a linear combination of the v observations $x(s_1,t_j),.....,x(s_v,t_j)$:

$$P = \sum_{i=1}^{v} \lambda_i X(s_i, t_j) \tag{10}$$

The v weights l_i are calculated such that P is unbiased and that the variance of the prediction error is minimal. This procedure requires information about the variogram of the regionalised variable. A detailed description of the calculation procedure can be found in geostatistical handbooks (e.g., Journel and Huijbregts, 1978).

3. Results

3.1 Particle mass transport

A total of four storms occurred during the 1993 rainy season (Table 1). During the first storm, one catcher was not functioning, so, only 20 measurements of particle mass transport were made during that particular event. For each sediment catcher and storm, eq. (1) was fitted through the measured particle mass fluxes at seven heights. Figure 3 shows a fitted particle mass flux profile for one sediment catcher during the storm of 1 July 1993. The profile has a maximum at the soil surface and decreases sharply with increasing height. In general, the fitted profiles showed good agreement with the observations. Average deviations between measured and calculated mass fluxes increased from 0.1% at the lowest sampling level to 38.0% at the highest level. The larger deviations at the higher sampling levels were caused by bigger measurement errors due to the very small quantities trapped at those heights (Sterk and Raats, 1996).

In Table 1 the summary statistics of the total mass transport observations for the four storms are given. The mean values indicate that the storms had very different magnitudes. The second and third storms were classified as small storms, whereas the first and the fourth were large storms.

Table 1. Characteristics of four wind erosion events in SW Niger, 1993 rainy season.

			Total mass transport		
Storm date	Wind speed	Wind direction	Mean	CV	Range
	$m\ s^{-1}$		$kg\ m^{-1}$	%	$kg\ m^{-1}$
13 June	10.3	SE	102.7	35.9	24.0 - 213.6
27 June	7.6	S	15.5	33.4	7.2 - 26.0
30 June	8.9	SE	32.0	46.3	9.6 - 68.9
1 July	9.2	SSE	149.9	34.5	68.9 - 282.7

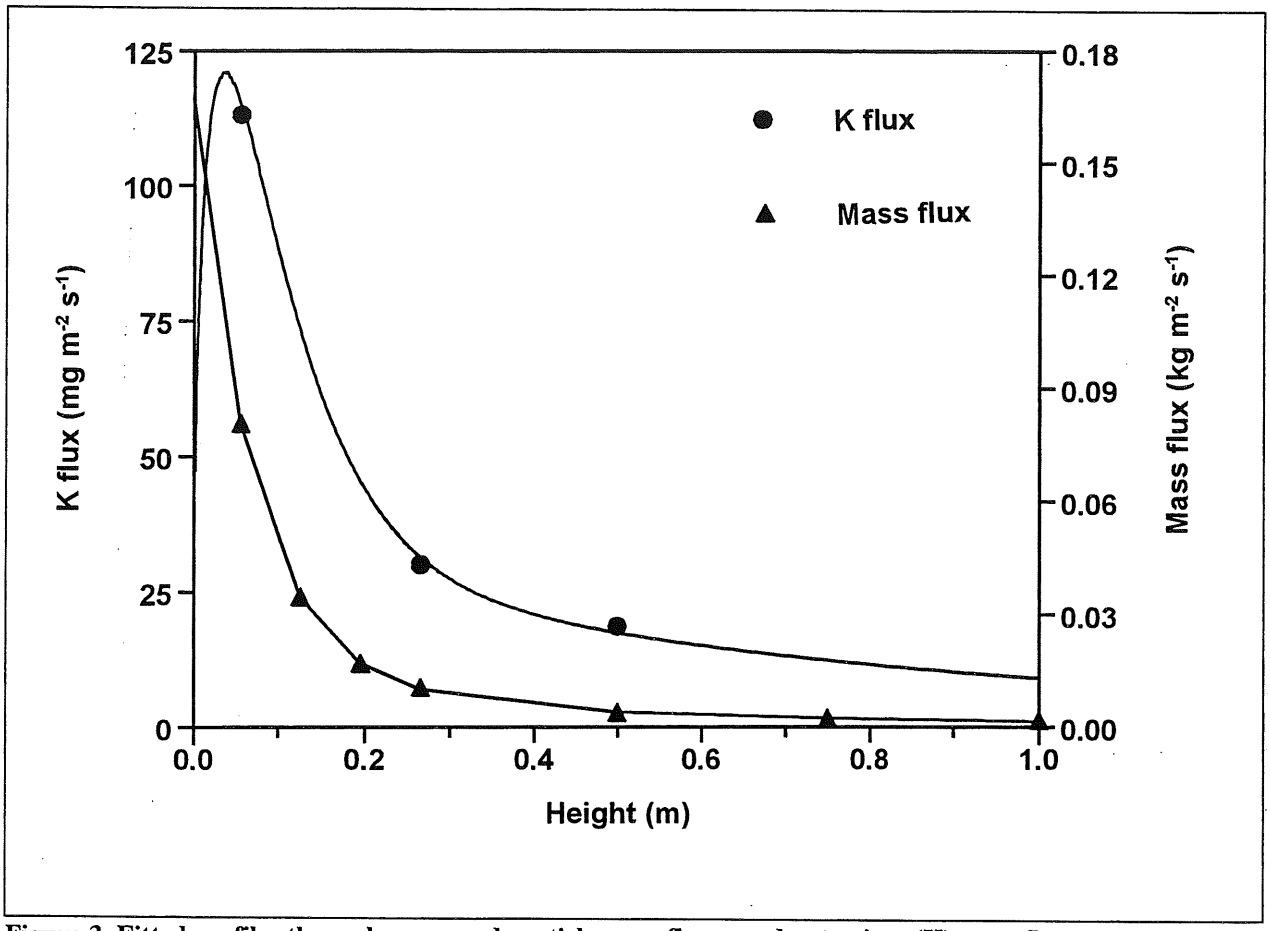


Figure 3. Fitted profiles through measured particle mass fluxes and potassium (K) mass fluxes.

Table 2. Average enrichment ratios of trapped erosion materials for two storms in SW Niger, 1993 rainy season.

		Enrichment ratio†			
Storm date	Height	K	C	N	P
	m				
13 June	0.05	1.18	1.33	0.83	0.98
	0.26	1.91	2.39	1.42	1.10
	0.50	2.06	3.02	2.08	1.14
1 July	0.05	1.27	1.47	0.83	0.91
	0.26	2.58	2.63	1.33	1.40
	0.50	4.15	4.74	2.25	1.98

† Ratio of the content of a certain element in the eolian material and the content of that element in the topsoil.

3.2 Nutrient mass transport

Chemical samples of wind-blown material were only taken during the first and fourth storms. The quantities obtained from the second and third storms were too small for chemical analysis. The measured nutrient contents of the wind-blown material showed an increase with height. This is illustrated by the average values of the enrichment ratio (Table 2), which is the ratio of the contentof a certain element in the eroded material to the content of that element in the topsoil. Equation (2) was fitted to all nutrient contents and the regression coefficients obtained were used in eq. (3) to describe the nutrient mass flux profiles of K, C, N and P. In Fig. 2, a fitted curve of the potassium mass flux is shown for the same catcher and storm as the particle mass flux profile. For C, N and P similar shaped curves were obtained. From the nutrient mass flux profiles the total nutrient mass transport values were determined and used to calculate nutrient losses from the experimental plot. Mass budgets were determined for both storms and all four elements by averaging the total nutrient mass transport values in the outermost rows of MWAC catchers (Fig. 3). Two rows contributed to input and two to output of sediment. The average wind direction of a storm determined the rows contributing to input and output, respectively. During both storms, wind-blown particles were entering the plot over the southern and eastern boundaries, and left the plot over the northern and western boundaries. The calculated nutrient losses from the experimental plot were: 57.1 kg ha^{-1} K, 79.6 kg ha^{-1} C, 18.3 kg ha^{-1} N, and 6.1 kg ha^{-1} P. These losses are equal to approximately 3% of the nutrient masses that were present in the top 0.10 m of the soil. The actual losses were even higher because (i) the nutrient-rich suspended material above 1 m was not included in the calculations, and (ii) the second and third storms caused also nutrient losses.

3.3 Mapping particle mass transport

All 83 observations of particle mass transport were standardized with eq. (5) to the reference value of 75 kg m^{-1}, equal to the mean of the four storms. The standardized variogram was estimated with eq. (6), and a spherical model (eq. (7)) was fitted through the variogram values. This model has a nugget variance ($C_{0,s}$) of 15.1 kg^2 m^{-2}, a sill value (C_s) of 1220 kg^2 m^{-2}, and a range (r) of 52.1 m. The variogram was converted with eq. (8) and (9) into four storm-specific variograms, which were used for kriging.

Table 3. Calculated soil losses from the experimental plot.

	Net soil loss	
Storm date	Linear interpolation	Kriging
	Mg ha^{-1}	Mg ha^{-1}
13 June	8.4	12.5
27 June	2.0	2.0
30 June	6.6	4.6
1 July	26.4	26.8
Total	43.4	45.9

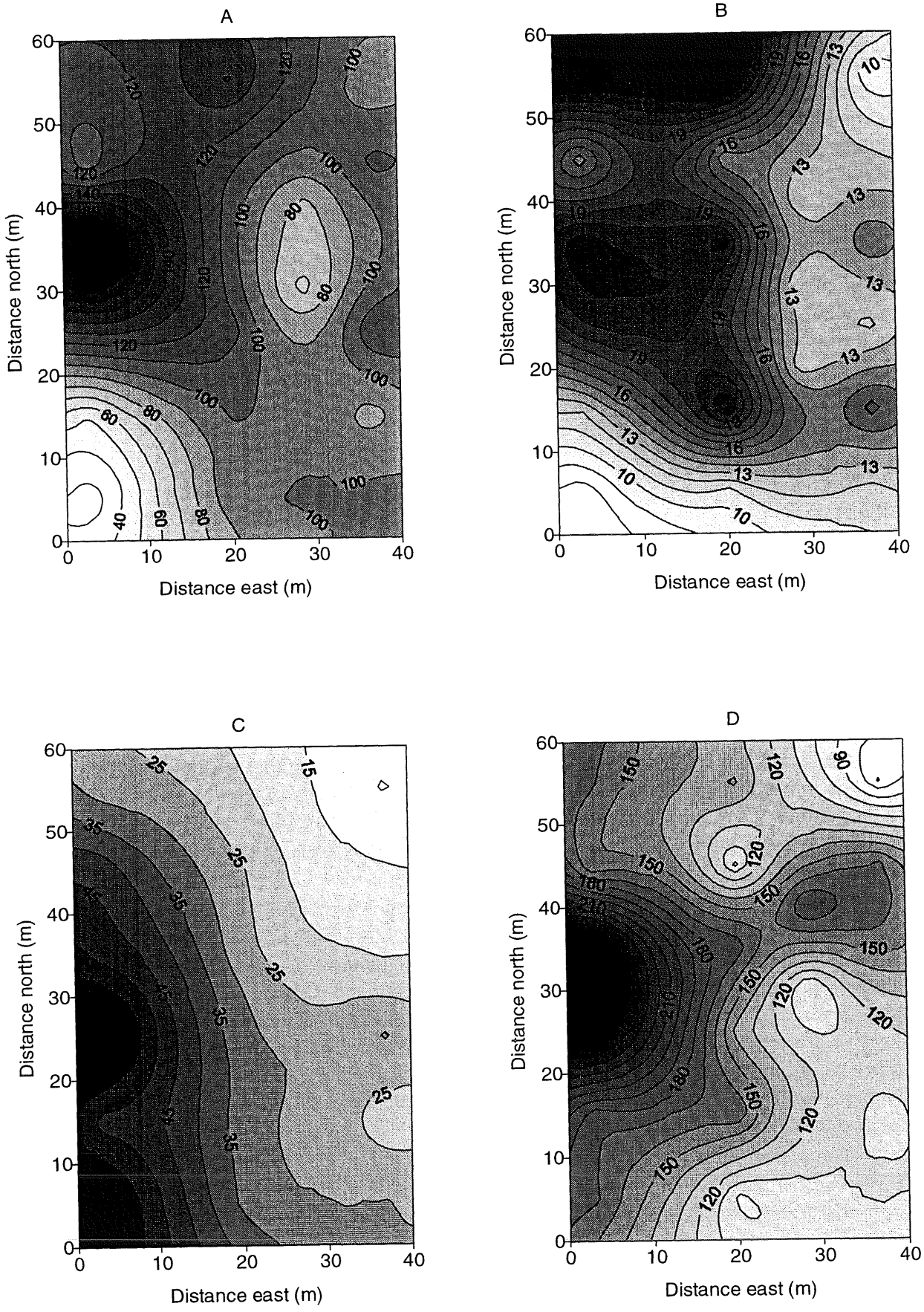


Figure 4. Maps of wind-blown mass transport (kg m^{-1}) produced by kriging. Storm dates are (A) 13 June, (B) 27 June, (C) 30 June, and (D) 1 July, 1993.

The produced maps (Fig. 4) provide the best linear unbiased predictions (BLUP) (Stein and Corsten, 1991) of total particle mass transport at each unsampled location at a given time and are therefore well suited for calculating net soil losses from the experimental plot. For all four storms mass budgets were determined by averaging the kriging predictions along the four boundaries. Two boundaries contributed to input and two to output of sediment, respectively. The best possible estimate of net soil loss is obtained when the boundaries coincide with the outermost rows of sediment catchers (Fig. 2), because the variance in prediction error is lowest along these rows. During the first, third, and fourth storms, the sediment moved into the plot across the southern and eastern boundaries and left the plot via the northern and western boundaries. During the second storm, only the southern and northern boundaries contributed to input and output of particle mass. The calculated soil losses from the kriging predictions were compared with soil losses calculated by averaging the mass transport values at the observation locations only (Table 3). This second procedure, which is similar to the nutrient mass budget calculations, is based on linear interpolation between the MWAC catchers in the outermost rows.

4. Discussion

The horizontal mass flux of wind-blown particles decreases sharply with height (Fig. 3). Just above the soil surface, saltation is the dominant mass transport mode, whereas suspension mass transport becomes dominant around the 1 m level. At 0.50 m, saltation and suspension are equally important (Sterk and Raats, 1996). The strong decrease of particle mass flux with height justifies the assumption that the mass of sediment moving above 1 m can be neglected.

In general, suspended dust is richer in nutrients than the coarser saltation material (Table 2), owing to a higher percentage of clay and silt particles. However, the K mass flux profile (Fig. 3) shows that the main mass of nutrients is transported in the height range where saltation is dominant. This was explained by the presence of saltation-size aggregates that contain clay and silt particles and thus nutrients (Sterk et al., 1996). The nutrient mass fluxes moved by suspension are an order of magnitude lower than the saltation mass fluxes but not insignificant. They are extended to a height of several hundred metres. Hence, suspension may also transport considerable amounts of nutrients.

The maps produced with kriging (Fig. 4) show the horizontal distribution of storm based mass transport. They clearly show that particle mass transport is characterized by a large spatial variability. Because of this variation, the traditional method of describing soil loss per unit area (Table 3) is not the best erosion indicator (Wilson and Cooke, 1980). It is more useful to distinguish erosion and deposition areas in the plot from downwind gradients in gray values. Positive gradients (from light to dark) are associated with erosion, negative gradients (from dark to light) with deposition, and zero gradients with transport only. Combining this information with maps of soil characteristics, surface roughness, crop characteristics and topography may lead to a better understanding of wind erosion processes.

The pattern of mass transport is not constant but changes from storm to storm. Prediction of those patterns with a wind erosion model would require very detailed input of soil characteristics, crop characteristics and wind field at many positions, which makes modelling at such a detailed level virtually impossible.

The total soil loss from the plot as determined from the kriging predictions is slightly (5.8%) higher than the total soil loss calculated with linear interpolation (Table 3). Comparing individual storms, however, shows that larger differences in net soil losses exist between the two calculation methods. Since kriging takes the spatial correlation structure of mass transport into account, it is assumed to provide better estimates of net soil losses from the experimental plot.

5. Conclusions

Wind erosion in the Sahelian zone of Niger causes significant losses of soil particles and nutrients. Saltation transports the main mass of both, but suspension may also transport significant quantities since the fluxes are extended to much greater heights than the saltation fluxes.

Saltation can only result in a local transport of soil particles and nutrients. Once picked up by the wind, saltating soil particles are moved in a downwind direction until they are deposited near some obstacle such as trees, bushes, fences, etc. Therefore, it leads to a local redistribution of soil particles and nutrients. The main sources are unprotected fields, and the main sinks are areas protected by vegetation (e.g. fallow land) or soil conservation measures. Redistribution within a field may also occur, which can be determined from the maps produced by kriging (Fig. 4).

The fine suspended dust which is raised from the field by a convective storm may be carried over long distances, resulting in a loss of fine soil particles and nutrients from the area and thus enhancing regional soil degradation. Apart from losses, there are also inputs of dust. During the dry season, fine suspended dust from the Sahara is transported towards the Sahel, where it partly settles. During the early rainy season, some of the dust raised by a convective storm is immediately deposited by rain wash-out. Whether there is a net loss or gain of dust due to suspension transport in the Sahel is unclear. Soil surfaces that are well protected against erosion can only benefit from dust deposits. However, most areas have inadequate protection against strong winds and probably lose more fine particles than they gain.

6. References

Bagnold RA 1973 The physics of blown sand and desert dunes 5th ed. London: Chapman and Hall.

Chepil WS 1945 Dynamics of wind erosion I. Nature of movement of soil by wind. Soil Sci. 60: 305-320

Deutsch CV, Journel AG 1992 GSLIB: Geostatistical software library and user's guide. New York: Oxford University Press

Journel AG, Huijbregts CJ 1978 Mining geostatistics. London: Academic Press

Leys J, McTainsh G 1994 Soil loss and nutrient decline by wind erosion - cause for concern. Aust. J. Soil Water Conserv. 7: 30-35

Stein, A, Corsten LCA 1991 Universal kriging and cokriging as a regression procedure. Biometrics 47: 575-587

Sterk G, Herrmann L, Bationo A 1996 Wind-blown nutrient transport and soil productivity changes in southwest Niger. Land Degradation Developm. 7: 325-335

Sterk G, Raats PAC 1996 Comparison of models describing the vertical distribution of wind eroded sediment. Soil Sci. Soc. Am. J. 60: 1914-1919

Sterk G, Stein A 1997 Mapping wind-blown mass transport by modeling variability in space and time. Soil Sci. Soc. Am. J. 61: 232-239

Webster R, Oliver MA 1992 Sample adequately to estimate variograms of soil properties. J. Soil Sci. 43: 177-192

Wilson SJ, Cooke RU 1980 Wind erosion. p. 217-251. *In* MJ Kirkby and RPC Morgan (ed) Soil Erosion. Chichester: John Wiley & Sons

Young RA, Olness AE, Mutchler CK, Moldenhauer WC 1985 Chemical and physical enrichments of sediment from cropland. *In* Erosion and Soil Productivity. ASAE Publication No. 8-85, p 107-116

Zobeck TM, Fryrear DW 1986 Chemical and physical characteristics of windblown sediment II. Chemical characteristics and total soil and nutrient discharge. Trans. ASAE 29: 1037-1041

2. Statistics for agricultural data

Maximum information from data

The contribution of statistics to data collection and data use is clearly visible when collecting representative data. Modern approaches include the use of data 'as they are collected', and data-driven statistical techniques are increasingly developed. Model parametrization methodology, comparable to optimization with least squares and the maximum likelihood criteria in statistics are important subjects in modelling efforts, as well as handling missing data.

2.1 Some Spatial Statistical Tools for Pattern Recognition

Andrew B. Lawson

Mathematical Sciences Division, University of Abertay Dundee, UK
Email: mctal@ zippy.dct.ac.uk

Modern analysis of patterns in spatial or mapped data relies increasingly on spatial statistical models of varying degrees of complexity. In this paper two specific modelling approaches are examined which can be applied to discrete spatial patterns ie. the distribution of objects or counts of objects in spatial regions. These two approaches are : the analysis of correlated random effects via empirical Bayes and full Bayes methods; and the analysis of clustering where cluster locations are directly modelled. Both approaches require the use of Markov chain Monte Carlo methods in their full Bayesian implementation.

1. Modern Pattern Recognition

1.1 Introduction

Pattern Recognition has developed within a wide variety of disciplines. Perceptual recognition problems are found in Psychology, while image processing (segmentation or higher level recognition) have been studied within Engineering, Physics and Statistics. Many of the tasks of image processing involve the recognition of structured patterns in arrays of data elements which have a spatial distribution. As such, the area of spatial statistics (Cressie, 1991) has contributed a considerable amount to the analysis of pattern recognition problems. In agricultural applications, there are many cases where mapped data, eg. from satellite imagery or ground-based surveys, are used and problems of a recognition nature arise. The remainder of this paper deals with two possible approaches to modelling mapped data, where discrete observation units are found eg. counts of events within regions, or locations of objects. These data types are frequently encountered in recognition problems and applications in agriculture should be clear.

1.2 Applications

The analysis of heterogeneity in discrete spatial data has received some increased attention in recent years (Clayton and Kaldor, 1987; Besag et al., 1991). Most recently, Breslow and Clayton (1993) have proposed a general framework for the analysis of generalised linear models with random effects which can be correlated or uncorrelated. In the work here described, we specify an approach to modelling discrete data of various kinds which allows the introduction of spatial correlation via a prior distribution for parameters of the model. This correlation is specified quite generally and leads to Geostatistics-liked estimators for parameters, when suitable prior covariance functions are specified. Here we examine situations where the canonical parameter of the observation distribution (and hence the data likelihood) can be described by a prior spatial Gaussian process and hence a realisation of the parameter will be multivariate normally distributed. In what follows we examine both Maximum aposteriori (MAP) estimators for the models considered and fully Bayesian analysis via Markov Chain Monte Carlo methods. The former are equivalent to Kriging estimators of standard Geostatistics.

Note the models discussed below have wide applications in an agricultural context, wherever discrete data (counts or locations) are studied in a spatial context. Some examples of agricultural applications are: land use map reconstruction; remote sensing of animal herds; soil science:object recognition/modelling in thin sections; veterinary science: disease mapping, pollution assessment and medical imaging.

2. General Random Effect Modelling Framework

In this section we propose a general framework for the incorporation of spatial prior structure in the analysis of observational data of discrete nature. The approach can be applied to Poisson, Binomial and also point process observational models and it will be shown that the normal observational model, which yields Kriging estimators is a special case of the approach advocated here. Define $x_i : i = 1(1)n$ to be the cartesian coordinates of locations in a planar region. A study window (A) is defined within which events are mapped. In the examples considered here, complete realisations are examined, and the events could consist of counts within regions contained wholly or partly within A or point events themselves. In the case of regions, we define n regions wholly or partly in A. For point events, the $\{x_i\}$ are point event locations, whereas for counts they are region 'centres'. We define the log-likelihood of n events as:

$$l(x|\eta) = \sum_{i=1}^{n} \ln f(x_i;\eta) \qquad (1)$$

where $f(.)$ is a probability density function and η is a p dimensional parameter vector ($p \leq n$). We assume that (1) is a conditional likelihood given η, where η has a prior distribution denoted by $\pi(\eta)$. In general, we assume that $\pi(\eta)$ includes the spatial correlation structure and trend components describing η. Hence η can be regarded as a hyperparameter in this case. The joint posterior distribution is given by:

$$P_0(x;\eta) \alpha \exp(l(x|\eta)).\pi(\eta) \qquad (2)$$

In general, it is often not possible to derive simple closed-form estimators for trend or covariance parameters in (2). Clayton and Kaldor (1987) suggested using a quadratic approximation to a poisson data likelihood and this led to the use of the EM algorithm. It is also possible to use Markov chain Monte Carlo methods such as the Gibbs Sampler or Metropolis-Hastings(M-H) (see, eg. Besag et al., 1991, Breslow and Clayton, 1987). This does not require the use of an approximation to the data likelihood.

In our work we demonstrate the use of the predictive distribution based on an asymptotic expansion (quadratic Taylor approximation) of the likelihood. This approximation yields maximum aposteriori (MAP) estimators with little computational effort. These estimators reduce to Kriging estimators when the likelihood is exactly normal. We also compare this approximation to a simple Metropolis-Hastings algorithm for exploring (2), which avoids some of the computational problems of the Gibbs Sampler applied to this problem. A novel feature of this work is the use of fully specified spatial covariances in the spatial prior model for the data. This is equivalent to the Bayesian interpretation of Kriging, where a spatial Gaussian prior distribution is assumed and a normal likelihood for the observational data is also employed (see eg. Cressie, 1991, p172). Lawson and

coworkers have applied these models to Poisson count data in epidemiological examples (Lawson, 1994, Lawson et al., 1996a).

2.1 The Quadratic Approximation

In this approach, we define first the saturated estimate of η based on (1), ie. the solution of

$$l(x|\eta)'_h = 0$$

where the prime denotes differentiation with respect to the subscript. This saturated estimate is denoted as $\tilde{\eta}$. The likelihood evaluated at $\tilde{\eta}$ is denoted as ψ. By adopting a second order Taylor expansion of $l(x|\eta)$ about $\tilde{\eta}$ it is possible to integrate out η from the posterior distribution. If a spatial Gaussian prior distribution is assumed with $\pi(\eta) \sim MVN(F\alpha, K)$, where F is an $n \times p$ design matrix, α is a $p \times 1$ vector of parameters, and K is a covariance matrix, then the predictive density of $\tilde{\eta}$ leads to generalised least squares estimates for α:

$$\hat{a} = \left(F' K_*^{-1} F\right)^{-1} F' K_*^{-1} \tilde{\eta} \tag{3}$$

where $K_* = K - (\Psi'')^{-1}$ and Ψ'' is the second derivative of Ψ *wrt* η. The parameter covariances, based conditionally on $\tilde{\eta}$ are:

$$\mathrm{cov}(\hat{a}) = \left(F' K_*^{-1} F\right)^{-1}. \tag{4}$$

This is just the regression of $\tilde{\eta}$ on F with covariance matrix K_*. This general result can be applied to a range of data likelihoods and table 1 displays a few examples of saturated estimators for a variety of discrete data models. The final entry in table 1 is included as non-stationary Poisson process models are of importance in the analysis of health risks related to putative sources of pollution hazard and spatial ecological modelling. Note that this approach can be generalised to the Generalised liner model. Define, in the usual notation, with canonical parameter η, the log-likelihood

Table 1 : Common discrete models and their saturated canonical parameter estimators and information

data type	m_i (count)	y_i (count)	x_i (location)
model	Poisson (λ_i)	Bin $(m_i P_i)$	Poisson process
η	$\lambda_i = e^{\eta_i}$	$P_i = \exp(\eta_i)/\{1+\exp(\eta_i)\}$	$\lambda(x_i) = e^{\eta(x_i)}$
$\tilde{\eta}$	$\ln(m_i)$	$\ln\{y_i/(m_i - y_i)\}$	$-\ln(A_i)$
$-\psi''$	$diag(m_i)$	$diag\left\{\frac{(m_i - y_i)y_i}{m_i}\right\}$	I_n

m_i : region total in Binomial example; A_i : ith Dirichlet tile area

$$
\begin{aligned}
l &= \left(y\theta - b(\theta)\right)/a(\phi) + c(y,\phi) \\
\text{and } V(\mu) &= (d\mu/d\theta).a(\phi), \\
\text{with } E(y) &= \mu = db/d\theta \text{ with } \eta = g(\mu). \\
\text{Then } \Psi' &= \left\{y - b'(\theta)\right\}/\left\{V(\mu).g'(\mu)\right\} \\
\text{and } E(-\Psi'') &= W.V(\mu) \\
\text{where } W &= \left(V(\mu)^{-1}.(d\mu/d\eta)\right)^2.
\end{aligned}
$$

The validity of the approximation will depend on the closeness of the quadratic approximation. For example, in the Gamma case, only certain combinations of parameters will be valid, and in the Poisson process case $\lambda(x_i) >> 0 \ \forall x \in A$.

2.2 Goodness of Fit and Residual Analysis

Under the full posterior distribution the MAP estimate of η, η^m say, is given by

$$\eta^m = R^{-1}T \tag{5}$$

where $R = K^{-1} - \Psi''$ and $T = K^{-1}Fa - \Psi''\tilde{\eta} + \Psi'$. Hence the MAP estimate can be directly evaluated by substitution of $\hat{\alpha}$ and estimated covariance parameters in K. In the assessment of goodness-of-fit it is possible to compare models by their posterior probability or its log. In addition, an ABIC criterion is also available (Ogata, 1991). Crude residuals can be computed as

$$\hat{e}_i = \tilde{\eta}_i - \eta_i^m$$

and their variance can also be estimated. Assessment of residual diagnostics can always be carried out by generating a residual envelope from samples of residuals from the fitted model and comparing the observed residuals with this envelope (Gelman et. al. (1995) provide examples of this approach).

2.3 Model Comparison, MCMC and data example

The above estimators use an approximation to the data likelihood and in some cases this may yield relatively poor estimates of parameters. To assess the appropriateness of the method we have compared the results to those obtained from posterior sampling using an Markov chain Monte Carlo method. The available methods of posterior sampling are reviewed in Gilks et al. (1996), Smith and Roberts (1993) and also in Gelman et al. (1995). Breslow and Clayton (1993) have applied a Gibbs Sampler to a disease mapping problem where the adjacencies of regions define a simple autocorrelation structure (Gaussian Intrinsic Autoregression). Our approach differs in two respects. First, we adopt a full spatial Gaussian prior distribution for η This allows full definition of *distance*-based dependence between regions which we believe to be more appropriate for disease correlation. Second, while it is possible to construct a Gibbs Sampler which includes spatial covariance and range parameters, it is more straightforward to adopt a Metropolis-Hastings (M-H) algorithm. The reason for this is that the acceptance probability for a proposed new parameter value is defined as:

$$\min\left\{1, \frac{P_0(x;\eta(\theta'))}{P_0(x;\eta(\theta))} \frac{q(\theta',\theta)}{q(\theta,\theta')}\right\}$$

where $q(,)$ is the proposal distribution for θ, and only the proposal parameter value $\{\theta'\}$ is compared to the current value $\{\theta\}$. This allows vector θ, as well as single θ updating (Besag and Green, 1993), a feature not usually directly available within a Gibbs Sampler. The other advantages are that the conditional distribution of θ^* (say) given other parameters is not required. This often requires rejection sampling, and, usually finding a conditional maximum aposteriori estimate of θ. Hence, M-H sampling has significant advantages in this application area.

We apply the above MAP estimators and MCMC method to an example of discrete data: Sudden Infant death incidence in North Carolina USA. In this example we compare the MAP results with modal estimates produced by a MCMC Metropolis-Hastings algorithm. We assume that the proposal distribution is symmetric and independent of previous θ values. In addition, we employ uniform indifference prior distributions for covariance and α parameters to allow comparison of the resulting estimates with those obtained by approximate MAP estimation.

Example: Sudden Infant Deaths in North Carolina

Cressie and Chan (1989) presented an analysis of counts of Sudden Infant Death (SID) in the 100 counties of North Carolina USA, for the period 1974-1978 (see also Cressie, 1991). It is thought that the counts of SID are related to deprivation gradients in the state. The original analysis addressed this issue, while here we provide an example of spatial modelling based on a constant state wide expected rate (2.06/1000 live births). Figure 1 displays the smoothed standardised mortality ratio (SMR) county-wise for the state of North Carolina. In this example the regions (counties) are irregular and they display considerable spatial structure. The SMRS appear to be high in the north-west, north-east and the south. For modelling purposes we have assumed a Poisson data likelihood with intensity:

Table 2. Results for the best subset ABIC model

parameter	MAP estimate	standard error	Modal M-H estimate	standard error[*2]
1	-1.0867	3.8419	-1.216	0.921
x	28.780	6.2697	20.572	8.915
y	17.677	6.6576	14.232	4.782
x^2	8.3156	8.8170	2.159	2.019
y^2	-26.603	4.9683	-24.782	5.293
σ^2	1.0008	0.3557[*1]	1.8647	0.0122
R	0.2008	0.0354[*1]	0.00151	0.00139
log(posterior)	-116.151		-104.2	
ABIC	242.30		218.4	

[*1]: s.e. estimated from REML likelihood curvature
[*2]: s.e. estimated from final 100 converged iterations

$\lambda_i = e_i \cdot \exp\{\eta_i\})$ where e_i is the expected count, m_i is the SIDS count for the *ith* county, and the saturated estimate is $\tilde{\eta}_i = \log(m_i / e_i)$. Table 2 displays the results for the best subset model for a set of 5 spatial variables. Figure 2a displays the MAP estimate of log (relative risk) for the best model fit. The M-H algorithm was checked for convergence using conventional diagnostic checks (see eg. Gelman et al., 1995) and convergence occurred within 5000 iterations.The results suggest that there is a linear (*x*-*y*) component and also a quadratic term in the SID surface $\left(y^2\right)$. The main difference between MAP and MCMC modal estimates here is the lack of spatial correlation found in the M-H result. Otherwise the two approaches give similar results. The residual surface for the MAP estimates (figure 2b) shows considerable unexplained structure in the north-east and north-west counties, and hence the model is not completely successful in accounting for the spatial structure over the whole study region. Note that Cressie and Chan (1989) also found considerable residual structure in such areas after fitting deprivation models.

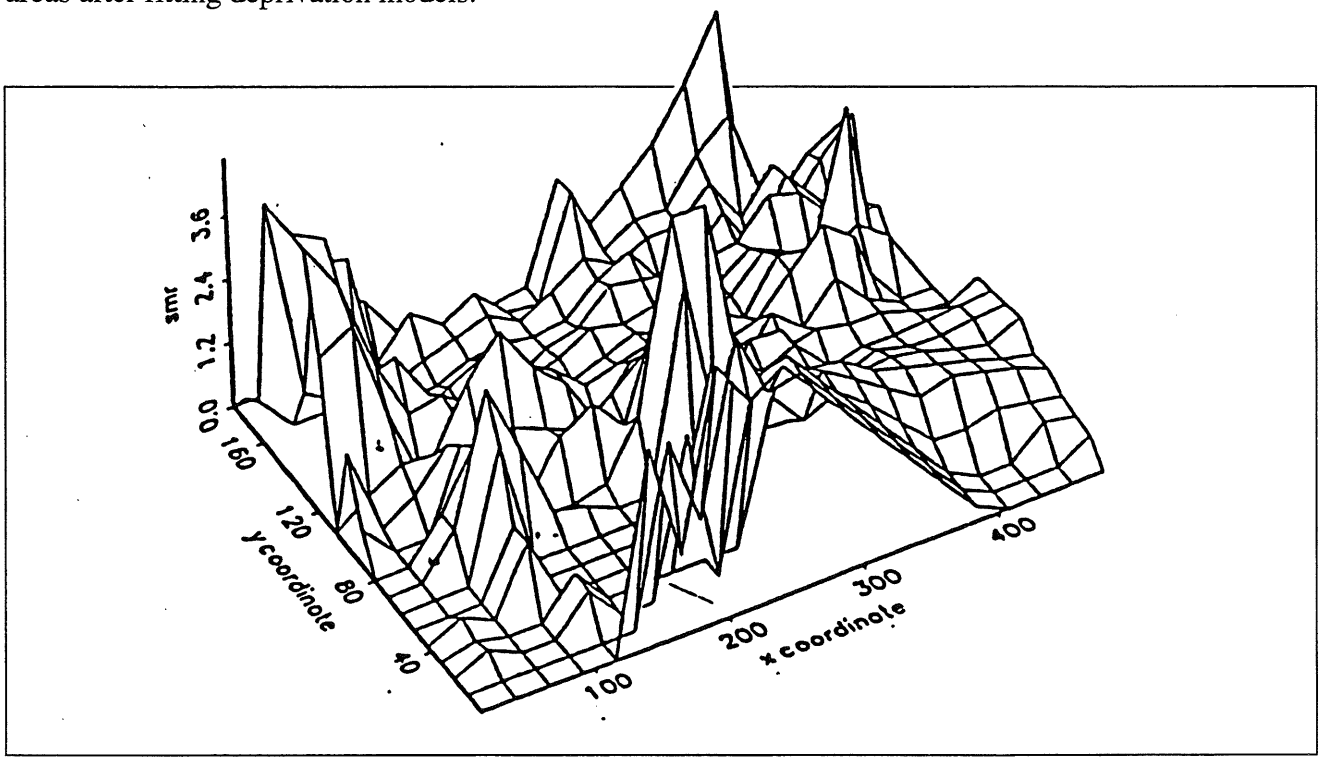


Figure 1 North Carolina: SMR surface for SIDS in the 100 counties

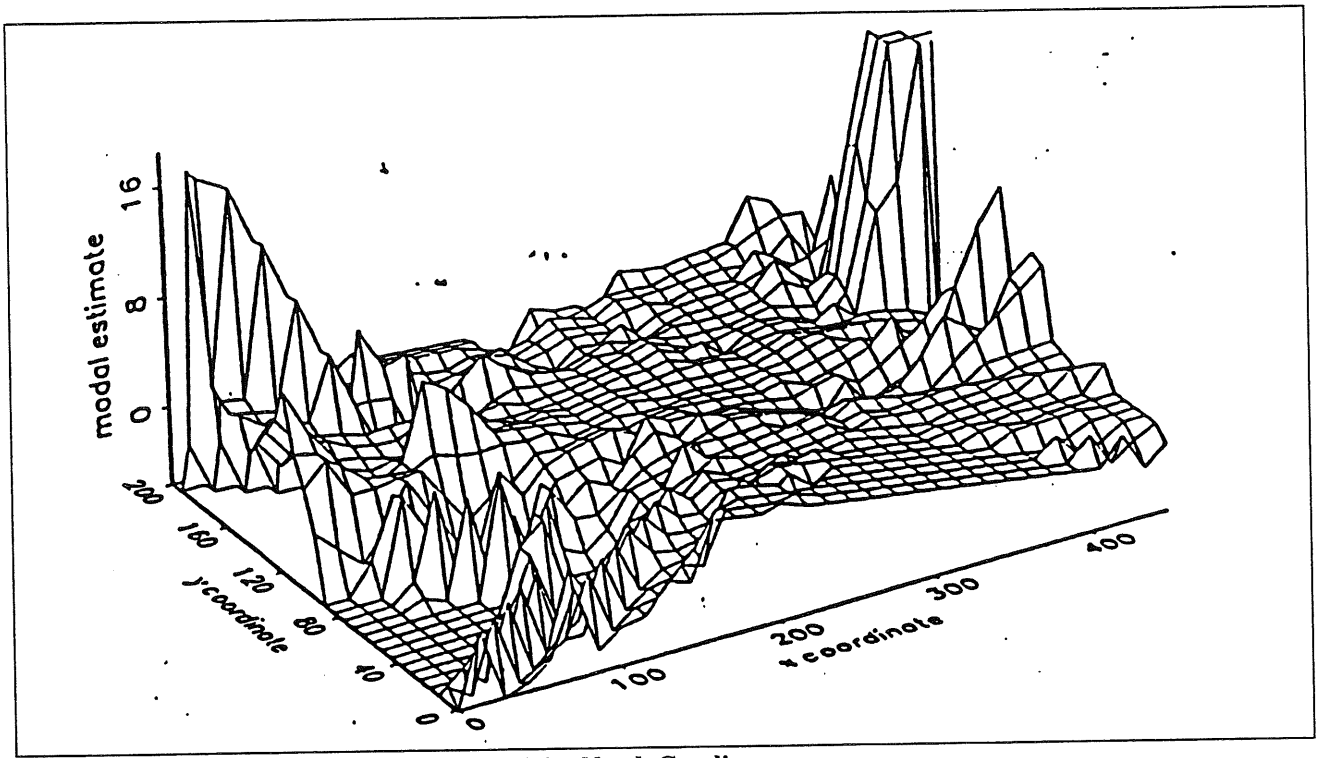


Figure 2a. MAP estimate of log(relative risk) for North Carolina

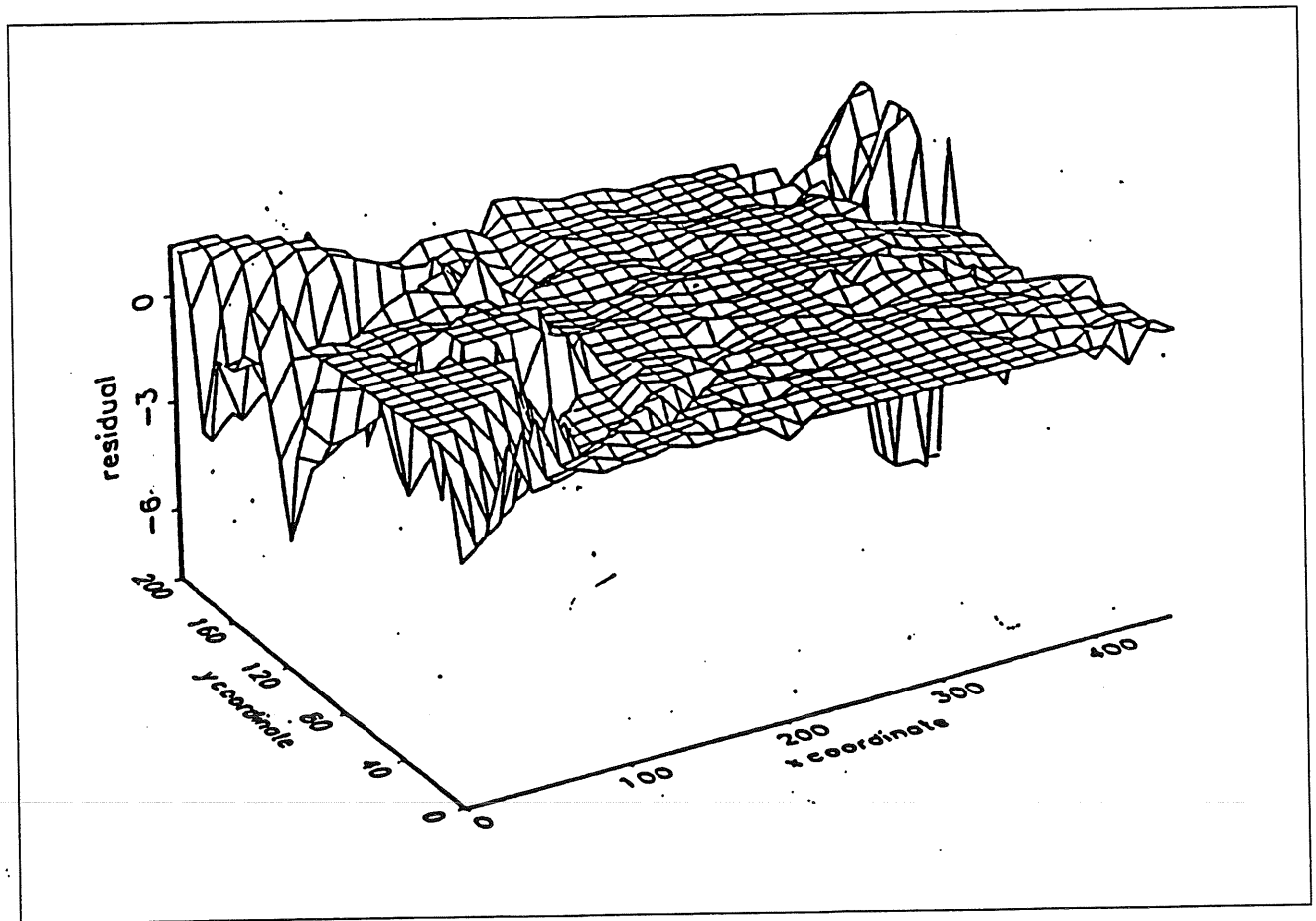


Figure 2b. Residual surface from MAP estimation for North Carolina

3. General Cluster Modelling Framework

The analysis of clustering in spatial point data has attracted increased interest in recent years. A growing interest in environmental issues both in the general public and the scientific community, has led to interest in clusters of disease related to environmental hazards eg. power stations, incinerators, electromagnetic fields or toxic waste dumping sites.

In this section, an approach to the analysis of clustering in small area health data is proposed, which can accommodate both of the above cases, via direct modelling of clustering within a more general model framework. The methods used are primarily Bayesian, as considerable use is made of MCMC methods. The methods have considerable generality and can be applied to both case event and counts of cases in arbitrarily-defined regions.

3.1 Model Development

The data **y** and cluster centres **x** are point patterns:

$$\mathbf{y} = \{y_1, \ldots\ldots\ldots y_m\},\ m > 0,\ y_i \in T$$
$$\mathbf{x} = \{x_1, \ldots\ldots\ldots x_n\},\ n \geq 0,\ x_i \in U,$$

where T, U are bounded open sets in $\Re^2$. By allowing U to differ from T, we allow the possibility of locating putative cluster centres outside the window of observation of the data (T). This makes some allowance for the edge effect where data could appear in the window but a centre lies outside ie. the boundary splits a cluster so that some part of the form is censored. The observed data $\{y\}$ are address locations of cases of disease, observed within T and a fixed time period. Diseases of interest could be leukaemias, which are thought to cluster weakly (Cuzick and Hills, 1991), or possibly, respiratory disease, such as respiratory cancer, larynx cancer or bronchitis, which could relate to one or more sources of health hazard (eg. incinerators, waste dump sites etc.). In either case, unobserved heterogeneity in the environment and/or population experiencing the disease events could lead to clustered disease incidence over the window T.

In what follows we represent this modelling approach by using the first order intensity of the process, with suitable parameterisation for particular applications:

$$\lambda(\mathbf{y}|\mathbf{x}) = g(\mathbf{y}).f(\mathbf{y},\mathbf{x},\theta). \qquad (6)$$

In all the applications we examine, $g(\mathbf{y})$ is considered to represent the background 'at-risk' process, and $f(.)$ is defined as a function of $\mathbf{y}$, $\mathbf{x}$ and a parameter vector θ. The exact form of $f(.)$ will be determined by the application. We assume a multiplicative link between the population background and $f(.)$ which implies that any spatial structure modelled in $f(.)$ will be directly modified by variations in $g(\mathbf{y})$. The alternative of a pure additive link (see, eg. Breslow and Day, 1987, p142), would imply that spatial structures modelled in $f(.)$ (eg. clusters) were of fixed size and hence unaffected by the population structure. This would appear to be inappropriate for spatial epidemiological data. In the application discussed here $g(\mathbf{y})$ is estimated non-parametrically from an external data set and analysis is made conditional on $\hat{g}(\mathbf{y})$. Issues relating to the estimation of $g(\mathbf{y})$ are discussed further by Lawson (1996c) and Lawson and Waller (1996).

Case Event Models

Where case event locations are to be modelled, it is possible to define a general point process model. Assume that, conditional on the x, the case events are independently distributed as a modulated heterogeneous Poisson Process with intensity given by (6). In the case of clustering, we condition on the realisation of a spatial stochastic process governing the $\boldsymbol{x}$ locations. Hence, for this case, we require a 'prior' spatial distribution to describe the $\boldsymbol{x}$ behaviour. In both cases, however, the general model of conditional independence of cases is assumed to have intensity:

$$\lambda(\mathbf{y}|\mathbf{x}) = g(\mathbf{y}).\left(1+\sum_{j=1}^{n_x} h(\mathbf{y}-x_j)\right).\prod_{i=1}^{n}(1+f(\mathbf{y}*\beta))$$

where the disease is known to cluster, but the analysis is also a function of y^* covariables (Lawson, 1996b).

Prior Distributions and Cluster Structure

Prior distributions must be provided for the components n_x, $\boldsymbol{x}$ and parameters in $h(y - x)$. Typically, the number of centres is assumed to have a Poisson (ρ) distribution, while $\boldsymbol{x}$ could follow a homogeneous Poisson process. The author and co-workers (Lawson, 1996a; Baddeley et al., 1996) discuss the theoretical justification for this in non-modulated cluster processes and Cox processes. Alternative specifications for the $\boldsymbol{x}$ prior distribution (eg. a Markov inhibition process) can be suggested. The cluster distribution function can take a variety of forms. Here we use

$$h(\mathbf{y}-\mathbf{x}) = \frac{\mu}{2\pi\kappa} e^{-[\|\mathbf{y}-\mathbf{x}\|]^2/2\kappa} \qquad (7)$$

a radially isotropic gaussian form with parameters μ and cluster variance κ. The possibility of allowing a flexible cluster shape, via density estimation of $h(.)$, may be attractive in situations where the exact form of clusters cannot be parameterised.

Applications in Count Modelling

Small area data is often available only as counts of cases within arbitrary regions (usually census tracts). The methods applied to case event data can be applied here also. Given the conditional independence assumption, then counts in disjoint regions are independent Poisson distributed with integrated intensity given by $\int_A \lambda(\mathbf{u}|\mathbf{x})d\mathbf{u}$, where A is an arbitrary region. It is therefore possible to recover the intensity function $\lambda(\mathbf{y}|\mathbf{x})$, suitably parameterised, based on count data.

3.2 Algorithm

The development of Markov Chain Monte Carlo (MCMC) methods and other iterative simulation tools (Tanner, 1991; Besag and Green, 1993), has allowed the implementation of algorithms which can explore posterior distributions of the spatial problems identified above.

Basic Point Event Algorithm

In the most general Bayesian formulation of the cluster model, we define the joint posterior distribution of $\{x,\theta\}$ as

$$P(\mathbf{x},\theta) \alpha\ L(\mathbf{y}|\mathbf{x}).p(\mathbf{x}).g(\theta) \tag{8}$$

where

$$L(\mathbf{y}|\mathbf{x}) = \left\{\prod_{i=1}^{m}\lambda(y_i|\mathbf{x})\right\}.\exp\left\{-\int_T \lambda(\mathbf{u}|\mathbf{x})d\mathbf{u}\right\} \tag{9}$$

$p(\boldsymbol{x})$ ≡ prior distribution for $\boldsymbol{x}$ (Markov inhibition or Uniform) and n_x (Poisson (ρ)), $g(\boldsymbol{\theta})$ ≡ prior distribution for cluster function parameters, and

$$\lambda(y_i|\mathbf{x}) = g(y_i).\left(1+\sum_{j=1}^{n_x} h(y_i - x_j)\right) \tag{10}$$

where there are n_x unknown centres. Note that the final fixed-foci term has been dropped for simplicity.

The derivation and properties of the following algorithm are discussed in Lawson (1996a) and Lawson et al. (1996b). The posterior distribution (8) could be explored by conventional iterative simulation methods, except for the cluster term, where a summation with a random upper limit occurs. This is essentially a mixture problem, and the parameters in this problem are best explored by a reversible jump Metropolis-Hastings (MH) sampler (Geyer and Mϕller, 1994; Green, 1995), involving a mixture kernel. Essentially the joint distribution of x and n_x must be explored during each iteration. This can be achieved by a spatial-birth-death-shift (SBDS) algorithm, where centres are added, deleted or shifted with given probability. A sequence of likelihood ratios can be specified for each case. In general, for a new configuration x', the posterior density ratio is, conditional on other parameters:

$$\frac{L(\mathbf{y}|\mathbf{x}')}{L(\mathbf{y}|\mathbf{x})}.\frac{p(\mathbf{x}')}{p(\mathbf{x})} \tag{11}$$

This ratio is evaluated for x' within the SBDS algorithm based on an MH criterion. A proposal configuration x' is accepted with probability

$$A(x,x') = \min\left\{1, \frac{P(x',\theta)}{P(x,\theta)}.\frac{q(x',x)}{q(x,x')}\right\} \tag{12}$$

where $q(x',x)$ is the proposal distribution for the new state. Often the proposal distribution for a point $\boldsymbol{u}$ is defined as a function of $h(\boldsymbol{y} - \boldsymbol{u})$ itself, (e.g. $\frac{1}{m}\sum_{i=1}^{m} h(y_i - \mathbf{u}),$) as simpler uniform proposals can lead to high rejection rates. We use Markov inhibition priors for $\boldsymbol{x}$, as the peaked nature of the likelihood surface can lead to multiple response, and it is important to propose spatially-separate new x values to avoid this problem. To this end, the Strauss prior can be used, and is defined for the proposed addition of a point $\boldsymbol{u}$ as

$$\frac{p(\mathbf{x}\cup\mathbf{u})}{p(\mathbf{x})}=\beta\gamma^{n_{R(u)}} \tag{13}$$

where β and $0<\gamma<1$ are parameters and $n_R(\mathbf{u})$ counts the number of x within R of $\boldsymbol{u}$. Similar ratios can be defined for deaths and shifts. The detailed specification of the acceptance ratios are found in Lawson (1995) and Lawson (1996a).

The parameters of the cluster distribution function, and other prior distributions can be treated conventionally. In most cases here, we assume that n_x has a Poisson (ρ) prior distribution. This parallels the assumptions which specify a Poisson Cluster Process in ordinary point process models (Diggle, 1983; Lawson, 1995). It is also possible to assume a prior distribution for ρ, and a Gamma distribution is often used. We have no strong prior reason to assume any other distribution than a uniform indifference prior on a suitable range (usually $\leq m$).

The cluster distribution parameters (μ,κ), based on model (7), are also assumed to have uniform indifference priors. The sampler steps used for ρ, μ and κ differ depending on whether a Gibbs or MH step is simple to implement. A Gibbs step is straight forward for ρ, whereas to implement a Gibbs step for κ or μ requires an optimisation step (to obtain ml estimates), and in these cases an MH step is used.

The count cluster modelling case

It is possible to extend this basic point event algorithm to the case where only counts of the case disease are observed within arbitrary regions. This application of the algorithm is of considerable importance given the ready availability of such data and level of interest in its analysis.

We assume that conditional on the $\boldsymbol{x}$, the process is a regionalised heterogeneous Poisson process governed by $\lambda(\mathbf{y}|\mathbf{x})$ and

$$E(n_i)=\int_{A_i}\lambda(\mathbf{u}|\mathbf{x})d\mathbf{u}\equiv\Lambda(A_i|\mathbf{x}) \tag{14}$$

where A_i denotes the i*th* region, and n_i the disease count in the region. As disjoint regions are independent under conditioning, then the $\{n_i\}$ are Poisson distributed with rates $\Lambda(A_i|\mathbf{x})$. Conditional on N, where $N=\sum_{i=1}^{p}n_i$, the likelihood for p regions is

$$L(\mathbf{n}|\mathbf{x},\theta)=\prod_{i=1}^{p}\left[\frac{\Lambda(A_i|\mathbf{x})}{\sum_{l=1}^{p}\Lambda(A_l|\mathbf{x})}\right]^{n_i} \tag{15}$$

Now in this case we do not observe the point case events but only know their region totals. However, for unknown foci locations ($\boldsymbol{x}$) we can use directly the basic point process algorithms and replace the likelihood ratios with those based on (15). Note that the use of (15) requires integration over arbitrary regions.

Define $\{z_{ij}\}$, $j=1(1)n_i$, the point locations of case events within the ith region. In each region, the conditional distribution of z given $\{n,\theta\}$ is given by:

$$(\mathbf{z}|\mathbf{n},\theta) \sim \frac{\lambda(\mathbf{z})}{\sum_{i=1}^{p}\int_{A_i}\lambda(\mathbf{u})d\mathbf{u}} \tag{16}$$

Hence, within the ith region, the joint distribution of $\{z_{ij}\}$ is

$$\frac{\prod_{j=1}^{n_i}\lambda(z_{ij})}{\left\{\int_{A_i}\lambda(\mathbf{u})d\mathbf{u}\right\}^{n_i}} \tag{17}$$

The important result of this algorithm is that the likelihood (or full posterior distribution) is now a function of the 'pseudo-data' (z) and hence point process modelling can be used via augmentation to model count data. Assuming that it is required to estimate cluster centres $\{x\}$ from the count data, then suitable parameterisation of $\lambda(z)$ with cluster terms and the inclusion of an inner M-H iteration for $\{\mathbf{x}, n_x\}$, prior to the θ step, provides a cluster version of this algorithm. We use the following finite element mesh numerical approximation to evaluate regional integrals:

$$\int_{A_i}\lambda(\mathbf{u})d\mathbf{u} = \sum_{j=1}^{l} T_j\lambda_j \tag{18}$$

where j denotes the jth mesh triangle for the ith region, and T_j is the triangle area (see eg. George, 1991). The intensity, λ_j, is evaluated at the triangle incentre.

Data Example: Respiratory Cancer in Central Scotland

In this example we examine the clustering tendency of respiratory cancer (ICD code: 162) in Falkirk, central Scotland, a town formerly associated with a variety of heavy industries during the early to mid 20th century. For the purposes of this example, respiratory cancer incidence in a subset of 26 contiguous Falkirk census enumeration districts (eds) has been recorded for a five year period 1978-1983. The total cancer count, expected count based on 16 age × sex strata and external (Scottish) rates for the period, and digitised ed boundaries are available for this example. Figure 3 displays the location map and outline ed map. Deprivation indices in this case were not available. The intention in the following analysis is to demonstrate the application of the count data algorithm to the estimation of cluster structure.

We have applied the count data algorithm, including augmentation of point events, with the following conditions. We initialise z with CSR events in $A_T = \sum_{i=1}^{p} A_i$. New values of z are rejection sampled from $\lambda(\mathbf{z}^l)$. M-H updates are used for μ, whereas a Gibbs step is straightforward for ρ (the cluster rate parameter) and κ. These steps are based on the augmented z likelihood, with E_i assumed constant across regions and $\lambda(z_{ij}) = E_i C(z_{ij})$, where $C(z_{ij})$ represents the cluster model terms. We have only included the unknown foci term for this example. A Markov (Strauss) prior has been included. Figure 4 displays the results of augmentation applied to this data set. Convergence occurred relatively quickly (< 200 iterations of main algorithm). There is some evidence that the number of centres lies in the range of 1 to 3, although the parent rate mode is 1.12. The posterior marginal distribution of centres is relatively uniform.

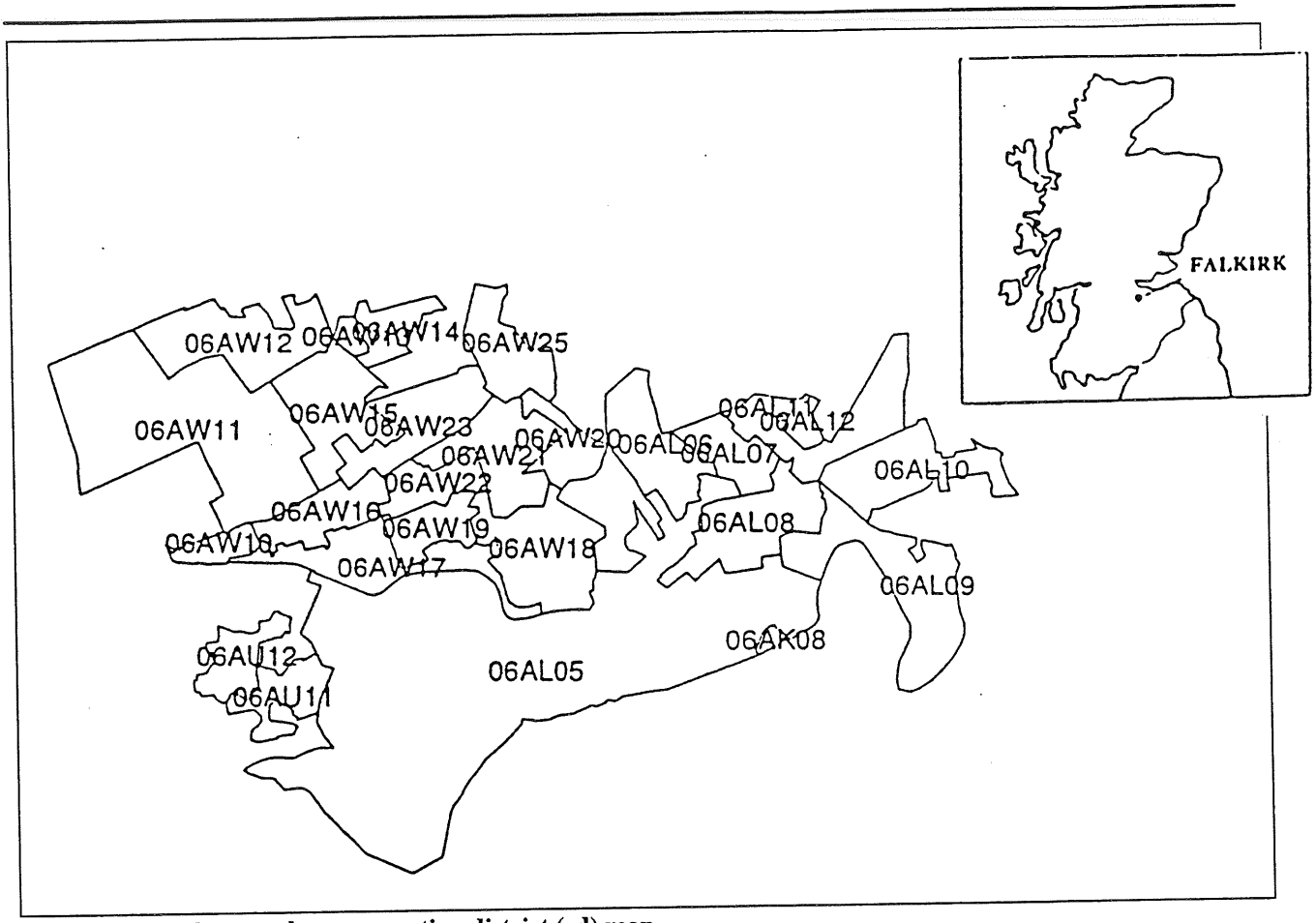


Figure 3. Falkirk example: enumeration district (ed) map.

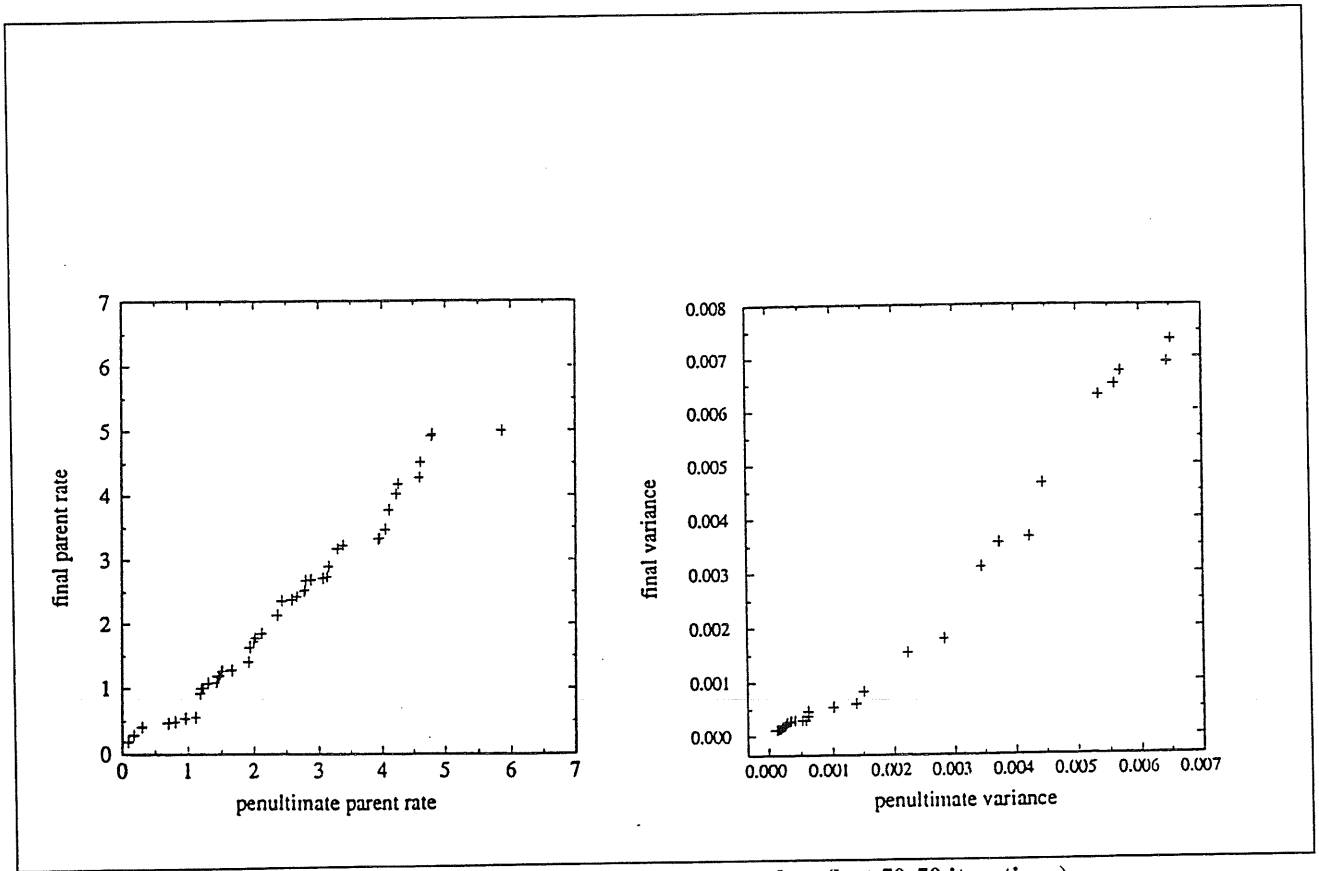


Figure 4a. Falkirk example: parent rate and cluster variance qq plots (last 50-50 iterations)

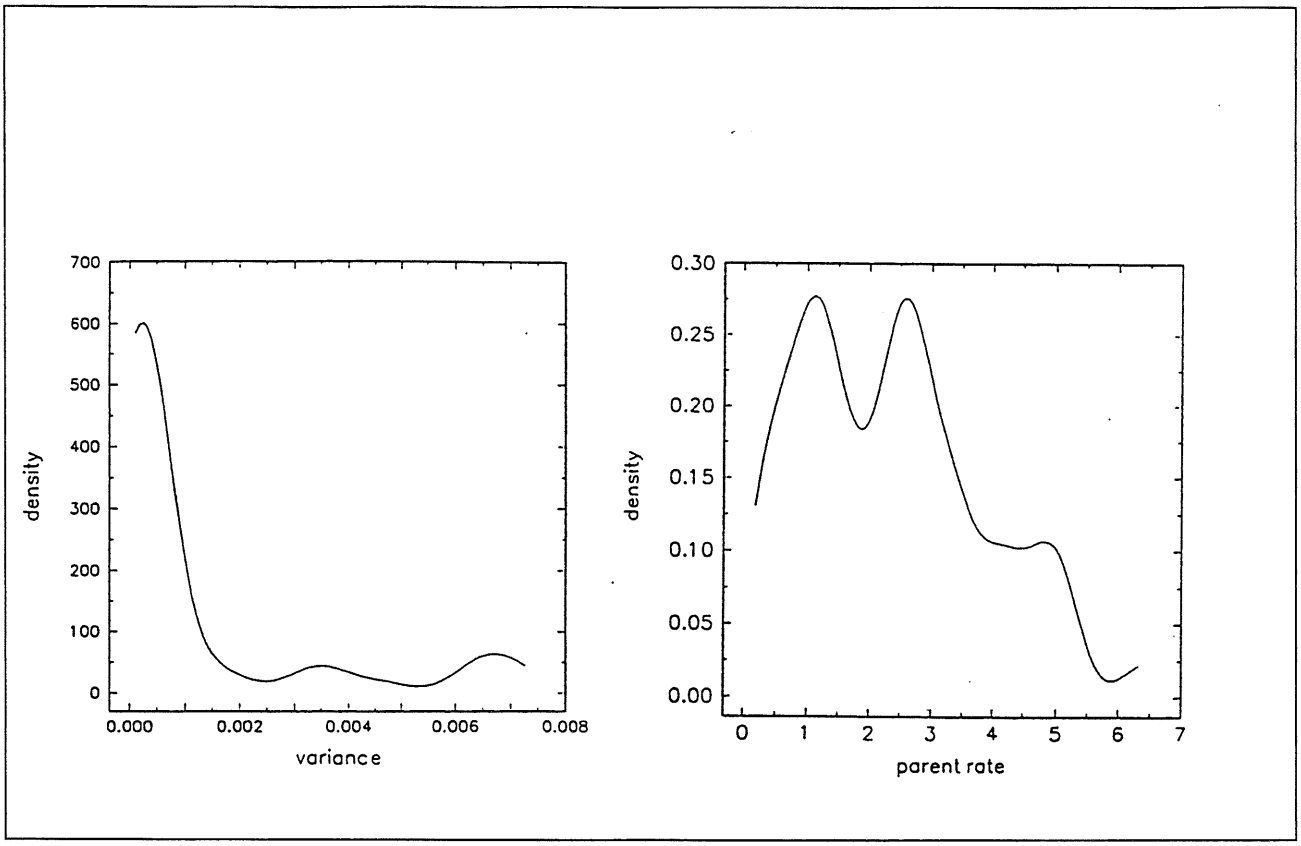


Figure 4b. Falkirk example: posterior marginal densities of parent rate and cluster variance.

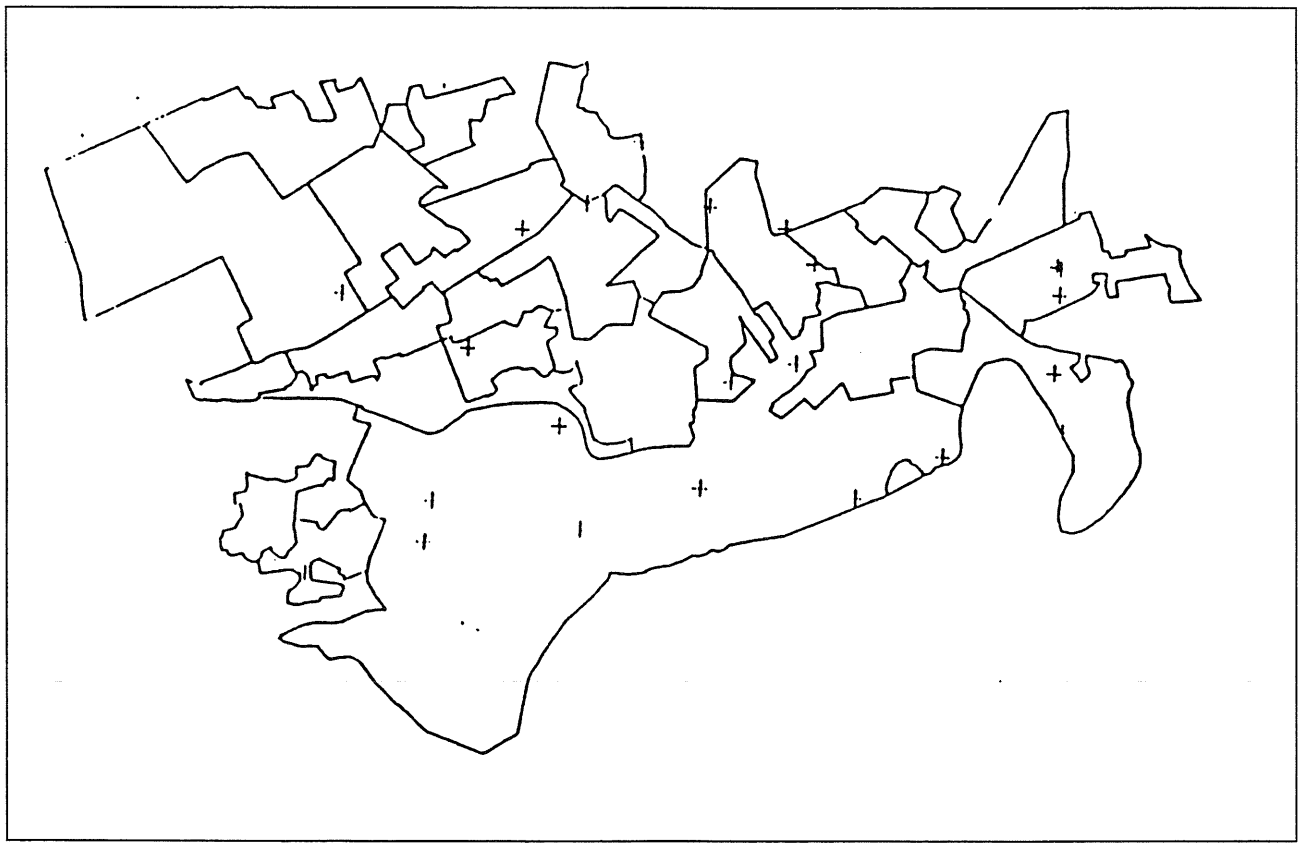


Figure 4c. Falkirk example: cluster centre posterior marginal distribution.

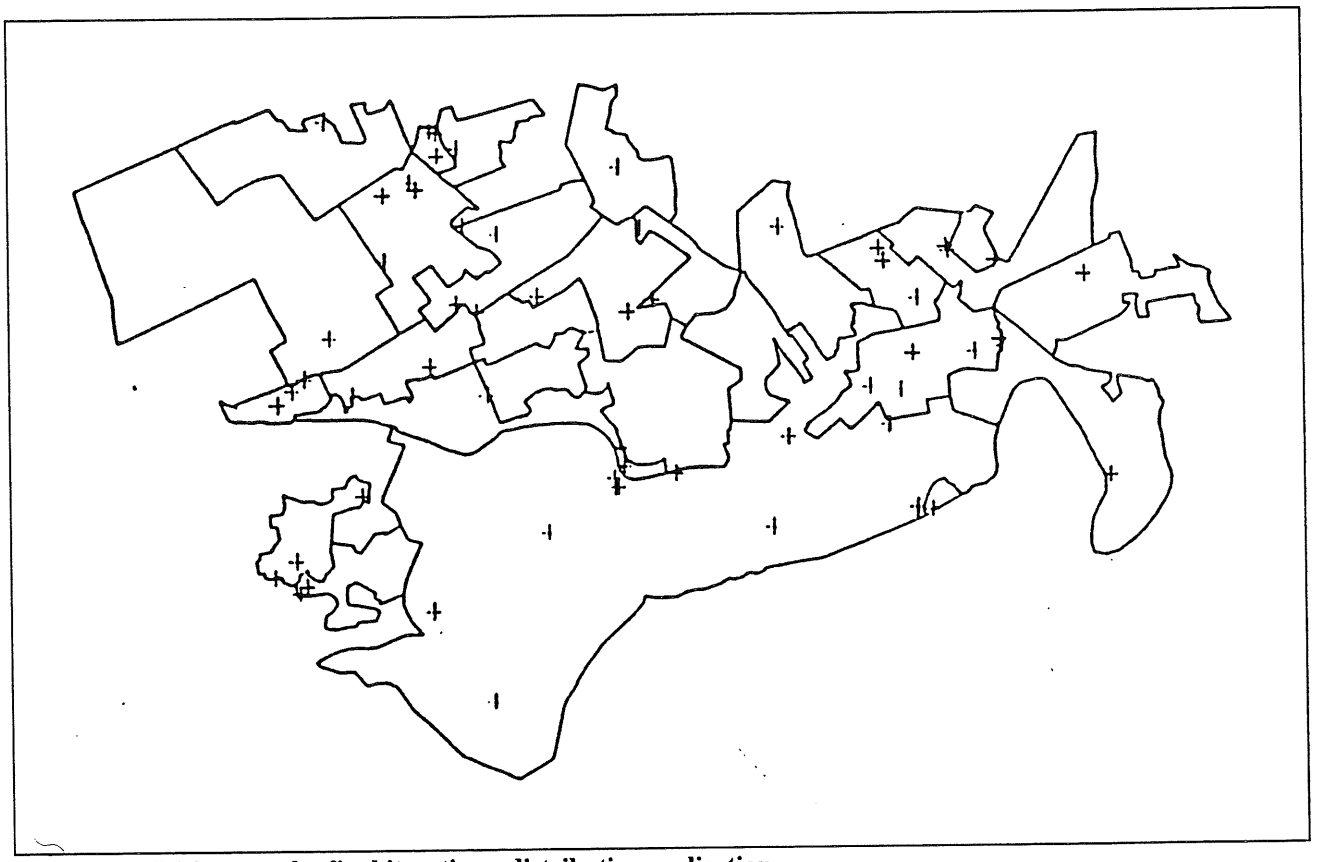


Figure 4d. Falkirk example: final iteration z distribution realisation

4 Conclusions and Further Work

The main conclusion of this work is that it demonstrates the flexibility of hybrid MCMC algorithms in the analysis of point process models with heterogeneous backgrounds. The main future development of this work lies in the modelling of general cluster situations with additional (non-specific) random effects (correlated or uncorrelated). The extension to space-time is straightforward but portends a wide application area.

5. Acknowledgements

I would like to acknowledge support of an European Union Biomed2 concerted action grant (contract nl: BMH4-CT96-0633) which has facilitated this work.

6. References

Baddeley AJ, Cheng HYW, Lawson AB, Van Lieshout MNM, Fisher NI 1996 Extrapolating and interpolating spatial patterns. Submitted for publication

Besag J, PJ Green 1993 Spatial statistics and Bayesian computation. Journal of the Royal Statistical Society Series B 55:25-37

Besag J, York J, Mollié A 1991 Bayesian image restoration with two applications in spatial statistics. Annals of the Institute of Statistical Mathematics 43:1-59

Breslow N, Clayton D 1993 Approximate inference in generalised linear mixed models. Jour. Amer. Statist. Assoc. 88:9-25

Breslow N, Day N 1987 Statistical Methods in Cancer Research, volume 2 : The design and analysis of Cohort Studies. Lyon: International Agency for Research on Cancer

Clayton DG, Kaldor J 1987 Empirical bayes estimates of age-standardised relative risks for use in disease mapping. Biometrics 43:671-691

Cressie N 1991 Statistics for spatial data. New York: John Wiley and Sons

Cressie N, Chan NH 1989 Spatial modelling of regional variables. Jour. Amer. Statist. Assoc. 84:393-401

Cuzick J, Hills M 1991 Clustering and clusters-summary. In G Draper (ed) Geographical epidemiology of childhood leukaemia and non-hodgkin lymphomas in Great Britain 1966-1983, pages 123-125. London: HMSO

Diggle P 1983 Statistical analysis of spatial point patterns. London: Academic Press

Gelman A, Carlin J, Stern H, Rubin DB Bayesian Data Analysis 1995 London: Chapman and Hall

George P 1991 Automatic Mesh Generation: applications to Finite Element Methods. New York: Wiley

Geyer C, Mϕller J 1994 Simulation procedures and likelihood inference for spatial point processes. Scandinavian Jour Statist. 21:84-88

Gilks WR, Richardson S, Spiegelhalter DJ (eds) Markov Chain Monte Carlo in Practice. London: Chapman and Hall

Green PJ 1995 Reversible jump mcmc computation and bayesian model determination. Biometrika 82:711-732

Lawson AB 1994 On using spatial gaussian priors to model heterogeneity in environmental epidemiology. The Statistician 43:69-76

Lawson AB 1995 Markov chain monte carlo methods for putative pollution source problems in environmental epidemiology. Statist. Med. 14:2473-2486

Lawson AB 1996a Markov chain monte carlo methods for spatial cluster processes. In Computer Science and Statistics 27:314-319

Lawson AB 1996b The analysis of putative sources of hazard in clustered small area data via mcmc methods. Statist. Med. 1996

Lawson AB 1996c Markov chain monte carlo methods for clustering in case event and count data in spatial epidemiology (submitted)

Lawson AB, Biggeri A, Lagazio C 1996a Modelling heterogeneity in discrete spatial data models via map and mcmc methods. In Forcina A, Marchetti G, Hatzinger R, Galmacci G (eds) Proceedings of the 11th International Workshop on Statistical Modelling. Graphos, Citta di Castello

Lawson AB, Van Lieshout MNM, Baddeley AJ 1996b Markov chain monte carlo inference for spatial cluster processes (submitted)

Lawson AB, Waller L 1996 A review of point pattern methods for spatial modelling of events around sources of pollution. Environmetrics 7:471-488

Ogata Y 1991 Goodness-of-fit of bayesian models by the monte carlo simulation. Annals of the Institute of Statistical Mathematics 43:25-32

Smith AFM, Roberts G 1993 Bayesian computation via the gibbs sampler and related markov chain monte carlo methods. Jour. Royal Statist. Soc. B 55:3-23

Tanner M 1991 Tools for Statistical Inference. Springer Series in Statistics. New York: Springer Verlag

2.2 Data use and Bayesian statistics for model calibration

Michiel J.W. Jansen

Centre for Biometry Wageningen, P.O. Box 100, 6700 AC Wageningen, The Netherlands
E-mail: m.j.w.jansen@cbw.dlo.nl

Bayesian statistics summarizes knowledge about a model's parameters in the form of a probability distribution. It provides a theoretically simple way to combine existing information about the parameters with information from new data. This feature makes Bayesian statistics particularly attractive for the combination of the diverse types of information required to set the parameters of a parameter-rich model. By way of example, the problem of calibrating parameter-rich models using system-specific observational data and general information from literature is considered from the Bayesian viewpoint. This viewpoint provides theoretical insights. Moreover, recent developments in Bayesian numerical methods hold a promise for actually performing calibration. It is noted that Bayesian methods provide an excellent starting point for uncertainty analysis and decision support.

1. Introduction

Suppose one has to adapt a model's parameters to a specific system. In general, mere observational (non-experimental) data about the system are insufficiently informative to set all parameters when the model has many of them. One often encounters the problem that the exogenous circumstances under which the data originated constitute only a small part of what the system may experience, so that different parameter settings may result in nearly equal fits to the observations, but to widely different predictions under new circumstances.

A major asset of system modelling is its capability to integrate information from diverse sources. Accordingly, it would be sensible and economical to remedy data shortage by scrutiny of the literature for additional information on the parameters. Literature data often originated under a broader range of exogenous circumstances, but they rarely pertain to the specific system for which the parameters have to be set. Most often, literature information alone is insufficient to parameterize a model for a specific situation (e.g. Kabat et al., 1995; Metselaar and Jansen, 1995a).

When both sources of information are insufficient by themselves, one obviously should try to combine them.

Bayesian statistics provides elegant ways to combine old (prior) information with new data, into accumulated (posterior) information. In Bayesian statistics, one summarizes knowledge about a parameter in the form of a probability distribution for that parameter. If the parameter is a vector, the distribution will be multivariate. Both prior and posterior are distributions. Probabilities that summarize one's knowledge are called 'epistemic' probabilities. Their status differs from frequentist probabilities, as occur for instance in coin tossing: such probabilities can be measured, whereas epistemic probabilities are more subjective. (e.g. Box & Tiao, 1973; Kroese et al. 1995.)

This paper discusses Bayesian methods for calibration and validation of a model. Moreover, it is briefly indicated that these methods connect quite well to uncertainty analysis and decision support.

Section 2 describes common problems with the classical analysis of observational data when the number of parameters to be estimated is large. We present the elements of Bayesian statistics in

Section 3. The next section treats the construction of a distribution describing the information about parameter values retrieved from literature. Section 5 discusses how parameter uncertainty is translated into prediction uncertainty. Bayesian calibration forms the subject of Section 6. It results in probabilistic predictions. Section 7 briefly discusses possibilities to validate such predictions. The consequences for decision support, of having probabilistic predictions, are touched upon in Section 8. The paper ends with a discussion in Section 9.

2. Problems with the classical analysis of observational data

A typical example of a classical statistical problem is the following. Let vector y denote observations of a system, while vector θ denotes the model-parameters to be estimated, and vector x the exogenous circumstances. Vector $f(x, \theta)$ denotes the corresponding model predictions. The whole is related by

$$y = f(x, \theta) + \varepsilon$$

where vector ε denotes the difference between observations and predictions, caused by measurement and model errors.

Typical regularity conditions for the above problem are the following. (i) Identifiability. When vector $\theta^{(1)}$ is unequal to $\theta^{(2)}$, vector $f(x, \theta^{(1)})$ is unequal to $f(x, \theta^{(2)})$. (ii) Differentiability: vector $f(x, \theta)$ of model predictions has finite partial derivatives with respect to θ, up to order 3. (iii) Presence of a randomness model: vector ε is random, with known probability density, $g(\varepsilon)$ say, that may involve a small number of additional parameters. There are quite a few more regularity conditions, but these three suffice for the present purposes.

Under the above luxurious circumstances it is often possible to show that, with a large set of data, an estimator of θ exists that is approximately normally distributed with mean θ, and with a covariance matrix that can also be estimated (e.g. Cox and Hinkley, 1974).

Now the problem with calibrating parameter-rich models is that very often none of the above regularity conditions is satisfied. (i) Identifiability. The model predictions can for instance be insensitive to changes of a sleeping parameter, belonging to a subprocess that is not activated under the prevailing exogenous circumstances. Anyhow, if the number of parameters is greater than the number of observations, the identifiability condition will not be met. (ii) Differentiability. Many parameter-rich models respond discontinuously to parameter changes, by which fact the differentiability condition is violated. (e.g. Metselaar and Jansen, 1995b.) (iii) Presence of a randomness model. The observations of a system can constitute quite disparate quantities, e.g. time series of leaf area, time series of weights of organs, waiting times for first occurrences. In such a case, it is implausible to assume independent identical normal errors or something equally simple. A more refined model is required that somehow reflects the different types of data. Duplicate measurements are quite welcome to enable the formulation of a model for the random effects.

The differentiability requirement is the least fundamental of the three conditions. Approximate normality of the estimator will be spoilt under non-differentiability. Moreover, many common tools for finding the estimator and assessing its accuracy require differentiability. But apart from that, non-differentiability can be overcome by brute computer force. In contrast, non-identifiability reflects a serious data shortage, which might be overcome, however, when the data are supplemented with additional information. Obviously, for all statistical approaches, a model for the randomness in the observations is essential.

It will be shown that in case of non-identifiability, the Bayesian approach allows distributional parameter estimation when the observational data are supplemented with sufficient prior information on the parameters. The only additional requirement is that one has a model for the randomness in the data. In theory, the Bayesian solution is simple and elegant, but the actual calculation of the estimator may pose formidable numerical problems.

It is not claimed that the Bayesian method provides the only way-out from the identifiability problems signalled. In particular, a set theoretic approach may constitute an attractive solution when one cannot formulate a plausible randomness model (Keesman and Van Straten, 1990).

3. Bayes' rule

The basics of Bayesian statistics are simple. Denote the prior density of parameter vector θ by $p(\theta)$. This means that the prior probability that θ lies in some domain D in parameter space is given by $\int_D p(\theta)\, d\theta$. The new data y come in through the likelihood function, say $l(\theta|y)$. Formal definitions of the likelihood are given in many statistics textbooks, we restrict ourselves to the remark that $l(\theta|y)$ is calculated from the data y, the model predictions at parameter value θ, and the stochastic model for the differences between predictions and observations. In Section 6 we will show how to calculate the likelihood function in a common calibration situation. The prior information, p, and the new data, y, are combined into a posterior density function, q say, by Bayes' rule:

$$q(\theta|y) = c\, p(\theta)\, l(\theta|y),$$

where c is a constant determined by the requirement that the integral of $q(\theta|y)$ over parameter space equals 1. The posterior probability that θ lies in a domain D is given by $\int_D q(\theta|y)\, d\theta$. Informative data should lead to a posterior distribution more peaked than the prior. When more new data become available, the posterior q can return in the role of prior: Bayes' rule describes a step in a learning process.

The problem with Bayesian statistics is that one always needs a prior, also at the first application of the rule, when one still does not know anything. Usually one starts with an uninformative prior, that is a flat distribution. Expert knowledge may be used in the construction of the first prior.

Bayes' rule provides an elegant, theoretically simple, answer to the calibration problem stated in the introduction. Before we come to that, however, we will first discuss how to construct priors that summarize the results from literature, and how to evaluate whether the literature prior suffices for predictions.

3.1 Constructing a prior from literature reports

By way of example, we consider the case of a single parameter θ that may differ between situations modelled. This natural variability of the parameter is modelled as randomness. Suppose the parameter has been measured or estimated in n situations that constitute an aselect sample from some population of situations. The parameter values are denoted by $\theta_1 \ldots \theta_n$; their measurements by $y_1 \ldots y_n$, with reported standard deviations $s_1 \ldots s_n$. Such a list may be the result of a literature search. The case of a parameter vector is more complicated, but can be treated similarly.

Parameters θ_i are independent with mean and standard deviation, say μ and τ, both unknown. They have been measured with error say ε_i, with different standard errors σ_i, which have been estimated by s_i. The accuracy, or number of degrees of freedom, of s_i is not known. In view of this

lack of information, it will be assumed that s_i is perfectly accurate, i.e. $s_i = \sigma_i$. The main problem is to predict θ_{new}, the parameter value in a randomly chosen new situation. A related problem is to estimate θ_i in a situation that has been measured. The problem may be solved in a classical and a Bayesian way. In the next example we shall briefly discuss a Bayesian solution, as described for instance in Gelman et al. (1995).

3.2 Example

A model parameter has been measured or estimated in nine situations. The measurements y and their standard errors are given in the next table.

y	8.8	9.0	9.1	9.7	9.8	10.1	10.2	10.7	11.6
s	1.8	1.0	1.2	1.3	1.1	2.0	1.0	1.2	1.6

Table 1. Measurements y_i of parameter θ_i in nine situations; with standard errors s_i (i=1...9).

The analysis reported below assumes that parameters θ_i, and measurement errors ε_i are normally distributed. More refined assumptions, for instance with distributions for positive parameters, can also be implemented; but we want to keep this example simple. Standard uninformative priors are used in the analysis.

Some quantiles of the posterior prediction of the parameter value (not its measurement) in a new situation, and in the ninth situation are tabulated below.

cumulative probability	0.025	0.250	0.500	0.750	0.975
θ_{new}	8.26	9.45	9.89	10.26	11.65
θ_9	8.78	9.64	9.99	10.41	11.84

Table 2. Quantiles of posterior distributions of parameter θ in a new situation, and in the ninth situation from the dataset

The variance of q_{new} equals 0.6816, which is less than the variance, 0.7961, of the measurements. Note that the 95% interval of q_9 is smaller than the interval that one may construct from information about the ninth situation only, namely $y_9 \pm 1.96\ s_9$.

If one knows for sure that the parameter studied is constant in the population of situations studied, that is t^2=0, one should use that knowledge. In that case one gets a posterior distribution for the parameter with variance 0.1737: the knowledge is rewarded with a much sharper estimate.

4. Uncertainty analysis

Uncertainty analysis estimates prediction uncertainty due to given uncertainty of model inputs: parameters, initial values, exogenous circumstances etcetera. Structural errors in the model, however, are beyond the scope of uncertainty analysis. Uncertainty analysis also tries to pinpoint inputs, or groups of inputs, that contribute most to prediction uncertainty. The results of such an analysis can help to set research priorities.

Very often uncertainty analysis proceeds from a probabilistic description of input uncertainties. So Bayesian parameter estimation forms an excellent starting point for such an uncertainty analysis.

Metselaar and Jansen (1995a) performed an uncertainty analysis for SUCROS-MAIZE, based on parameter uncertainty remaining after a thorough literature study. The resulting prediction uncertainty appeared to be too large for practical application of the model thus parameterized. Apparently, more information on the parameters is required.

5. Bayesian calibration

Suppose that analysis of literature data has lead to a distributional prediction of parameter values in a new situation. Very often it will appear that this distribution is too vague for sharp predictions. But suppose one also has independent observations of that particular situation; plus a stochastic model for the error in those measurements. Then one should try to combine the two forms of information in order to get sharper predictions. Obviously, there is no guarantee that the predictions will become sufficiently sharp, but it is worthwhile to try. Even more obviously, a sharp prediction need not be a good prediction, since your model and the additional assumptions may be erroneous: that aspect will be treated later, in the section on validation.

The posterior parameter distribution derived from the literature will now serve as prior distribution. The error model allows to calculate a likelihood function. In the case described in Section 2, with observation vector y modelled as $f(x,q) + e$, where e has probability density $g(e)$, the likelihood has the form $l(q|y) = g(y-f(x,q))$. Thus a model run is required for each calculation of l; usually this run takes nearly all the time required to calculate l. The prior information, p, and the new data, y, are combined into a posterior density function, q say, by Bayes' rule, $q(q|y) = c\ p(q)\ l(q|y)$, that was described in Section 3. Discontinuities in the likelihood, i.e. discontinuities in the model predictions, cause no problems in the formulation of the rule. Theoretically impossible parameter values can be excluded by zero prior values. The presence of prior information prevents identifiability problems. Uncertainty analysis with posterior probabilities can assess whether or not the posterior prediction uncertainty is acceptably small under specific new exogenous circumstances.

It will appear that the numerical implementation of the solution is far from simple, especially when the number of parameters is large. The rest of this section is devoted to the numerical implementation.

5.1 Grid method

Reilly (1976) and Aldenberg et al. (1995) implement Bayes' rule in a simple, intuitively appealing manner, namely by discretization of parameter space. Each of the say k parameters are restricted to a discrete set of say m values, which gives a grid of m^k points. Subsequently, the values of prior and likelihood are calculated at each grid point. The calculation of the likelihood involves m^k model runs. It will be obvious that the grid method will become prohibitively computer-intensive with more than a few parameters.

5.2 Monte Carlo methods

Monte Carlo methods provide alternatives to the grid method. Their aim is to produce a sample from the posterior distribution. Two typical instances will be discussed.

Rejection sampling

Rejection sampling, the simplest Monte Carlo method, works as follows. Let L denote an upper bound L of the likelihood function $l(\theta)$ over the part of parameter space Θ where the prior density is non-zero. One draws many times aselectly from the prior distribution. For each parameter vector drawn the likelihood is calculated. The parameter vector is accepted with probability $l(\theta)/L$. The procedure is most efficient when L when L is the maximum of the likelihood. The accepted parameter vectors constitute an independent random sample from the posterior distribution. The method is feasible if the average acceptance probability is sufficiently large, but might become very computer-intensive with small acceptance probabilities.

Markov Chain Monte Carlo

The Metropolis method is the original form of Metropolis-Hastings type Markov Chain Monte Carlo methods (e.g. Gelman et al., 1995).

The chain starts with an arbitrary parameter vector with non-zero prior probability and non-zero likelihood. Denote the latest parameter vector in the chain by θ, with prior $p(\theta)$ and likelihood $l(\theta)$. Then the next vector is obtained as follows. One draws a candidate point, say θ', by some random proposal generating mechanism: for instance $\theta'=\theta+\eta$, with η multivariate normal. The candidate draws the next place in the chain by lot: it is accepted when

$$[p(\theta')\, l(\theta')] / [p(\theta)\, l(\theta)] > u,$$

in which u is random with homogeneous distribution on interval (0, 1). Thus, higher posterior densities are always accepted, lower ones sometimes. If θ' is rejected, θ will succeed itself. The chain thus constructed will converge to a stationary chain with elements randomly, but not independently, drawn from the posterior distribution.

The performance of the Markov Chain Monte Carlo method depends critically on the proposal generating mechanism. Choosing an efficient proposal generating mechanism and monitoring the performance of the chain, constitute the main problems of the procedure.

5.3 Comparison of grid method and MC methods

An attractive feature of the above Monte Carlo methods is that inclusion of a parameter with little or no influence on the model outputs considered, will hardly or not affect the acceptance probability, so it will hardly or not lead to extra work. In an informal sense, the workload seems to be proportional to the information content of the new data. With the grid method, in contrast, each parameter multiplies the workload. Thus it seems worthwhile to try whether workload of Monte Carlo procedures grows less explosively with the number of parameters than the grid method.

A disadvantage of the Monte Carlo methods mentioned is that it is hard to assess the required computational effort in advance, without actually performing some sampling. But the effort might be evaluated in a sampling experiment of moderate size.

5.4 Example: an exercise with SUCROS

The acronym SUCROS stands for 'Simple and Universal CROp growth Simulation model' (Spitters et al. 1989). From the statistical point of view, however, the model is still quite complex: it has about 50 parameters. SUCROS is a dynamic model in discrete time, namely days. Daily radiation and minimum and maximum temperatures are exogenous input. The major state variables are the development stage, the leaf area index and the dry weights of some organs.

In the course of time, the crop passes through consecutive phases; the dynamical laws change discontinuously from phase to phase. By the distinct phases, the parameter-response-functions have jump-discontinuities (Metselaar and Jansen, 1995b). These give rise to complications during sensitivity analysis and calibration: Taylor approximations make no sense, and classical asymptotic statistical estimation theory does not apply.

The following, constructed, example provides a striking case of non-identifiability. Eight parameters were calibrated, about which prior information had been obtained from a literature search (Metselaar & Jansen, 1995a). The, artificial, system-specific data consist merely of a single measurement of total above-ground dry-weight of maize at harvest in Wageningen in 1985: 15 ton per ha, with normal measurement error having standard deviation 1. That measurement can be fitted with many different parameter settings, so classical calibration methods for more than one parameter are deemed to fail. A Bayesian posterior sample of size 1000 is constructed with the rejection method.

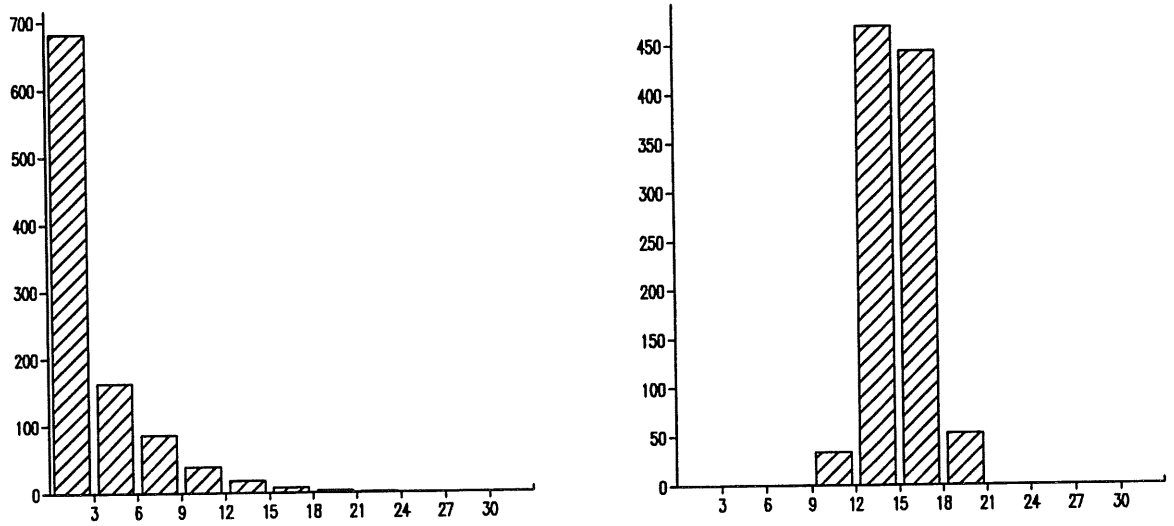


Figure 1 Histograms of samples of size 1000 of prior and posterior predictive distributions of total above-ground dry-weight (ton per ha) of maize in Wageningen in 1986; left: prior; right: posterior.

Figure 1 shows prior and posterior predictive samples of total above-ground dry weight at the same location under different weather circumstances, namely those of 1986. The flat prior is useless for prediction; its low mean reflects that the prior comprises variability of maize cultivars over the world, whereas the cultivars actually used in The Netherlands have been selected for suitability to the local climate. The posterior prediction is much sharper: apparently the measurement from the previous year adds valuable information, which can be recovered by the Bayesian calibration procedure.

6. Validation

Model validation is a multifaceted subject; see for instance the special issue of Advances in Water Resources (Hassanizadeh & Carrera, 1992). But this section will only touch upon the difference between validation of a point prediction and a probabilistic one, based on Bayesian calculus of parameter uncertainty.

The meaning of the word validation is a floating one. In this paper the term is used in the utilitarian sense of characterizing prediction error of a calibrated model in a specified population of

situations. With point predictions, the mean squared prediction error constitutes an obvious characterization of prediction error, the mean being taken over a random sample from the population. The mean squared prediction error incorporates input uncertainty, structural error, and error in the validation data measurements (e.g. Wallach & Goffinet, 1989). But how to express the quality of a probabilistic prediction? It is easy to make a probabilistic prediction so vague that it is never entirely wrong, in the sense that any outcome is predicted to be possible. Such vague predictions should get a low score. On the other hand, the score should discourage sharp predictions when the outcome is known to be very uncertain. The following measure for the quality of a probabilistic prediction is based on such considerations. Let $P(z)$ denote the cumulative distribution function of the prediction; let y denote the outcome, and let $U_y(z)$ denote the cumulative distribution of the outcome, that is, $U_y(z) = 0$ when $z < y$ and 1 when $z \geq y$. Then the quality of the prediction may be expressed by the loss function

$$L(P,y) = \int (U_y(z)-P(z))^2 \, dz$$

(see Figure 2). For a point prediction, the loss function is equal to the absolute value of the prediction error. Loss functions for probabilistic predictions are discussed in more detail in Kroese et al. (1995).

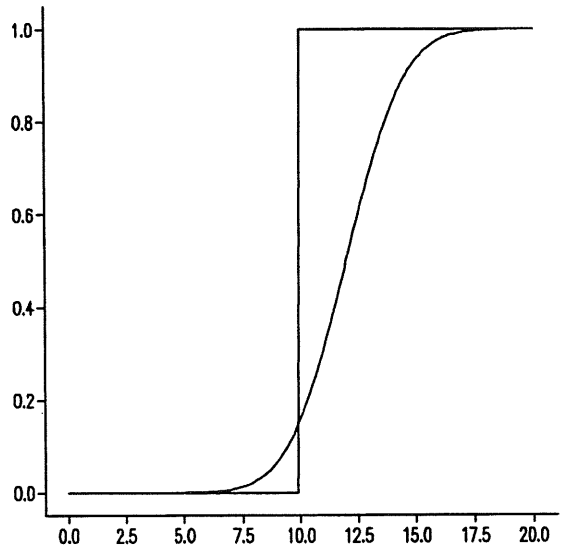


Figure 2. Loss function for probabilistic prediction. The smooth curve represents the cumulative distribution of the prediction. The step function depicts the cumulative distribution of the observation (which equals 10 in this graph). The loss is the integral of the squared difference between the two functions.

7. Decision support

Modelling is often used to support decisions. In theory, the model is used to predict a system's behaviour under various alternative actions, and the decision maker chooses the action leading to the most desirable predicted behaviour. In practice, however, matters are more complicated: the decision maker won't have absolute trust in the model and its parameterization, the model will seldom cover all relevant aspects, and it will often be hard to decide which outcome is most desirable. For such reasons, decisions are not calculated, but merely supported.

A Bayesian prediction takes the form of a probability distribution, usually multivariate, that represents uncertainty about the system's behaviour. Thus, one shortcoming of deterministic predictions for use in decision support is avoided. Different actions lead to different distributions, so the decision maker has to choose between distributions. There is a rich literature on Bayesian decision theory; see for instance Berger (1985).

7.1 Example

In the 18th century there was a heated discussion on the desirability of inoculation against smallpox. Inoculation was a method of immunization against smallpox by means of a slight artificial infection with human smallpox. Inoculation, however, was not altogether harmless: one might die from the artificial infection. (The less risky form of immunization with cow-pox was only discovered at the end of the century.) The risk of inoculation had to be compared with the permanent and considerable risk of natural infection. The physician and mathematician Daniel Bernoulli (1760) constructed a model that should enable comparison of the two risks. The comparison made below proceeds from a very individualistic point of view; for instance, the risk of contaminating others is neglected. Bernoulli was well aware that his model and its parameters were only approximate, but he stated that they conformed reasonably to the facts known. Nevertheless, Bernoulli was fiercely criticized because of the simplicity of his model and the uncertainty in its parameters (e.g. Bradley, 1971).

For the sake of an amusing example, and because parameter uncertainty was a major issue in the discussion, we ventured to translate statements made at the time about uncertainty of the model parameters into probability distributions. This allows to account for parameter uncertainty in statements about survival probability. We considered the case of a child that just reached the age of five and is still susceptible. Figure 3 shows the survival probabilities for the child when inoculated at age five and when never inoculated. It can for instance be seen that by inoculation the probability to reach the age of 20 rises from 72% to 79%. Only if one assigns gigantically more value to the next few months than to the rest of life, one might prefer not to inoculate. (Provided, of course, that one does not dispute the assumptions on which the calculations were based; and that one does not differentiate between death as a consequence of a deliberately taken risk, and death by natural infection.)

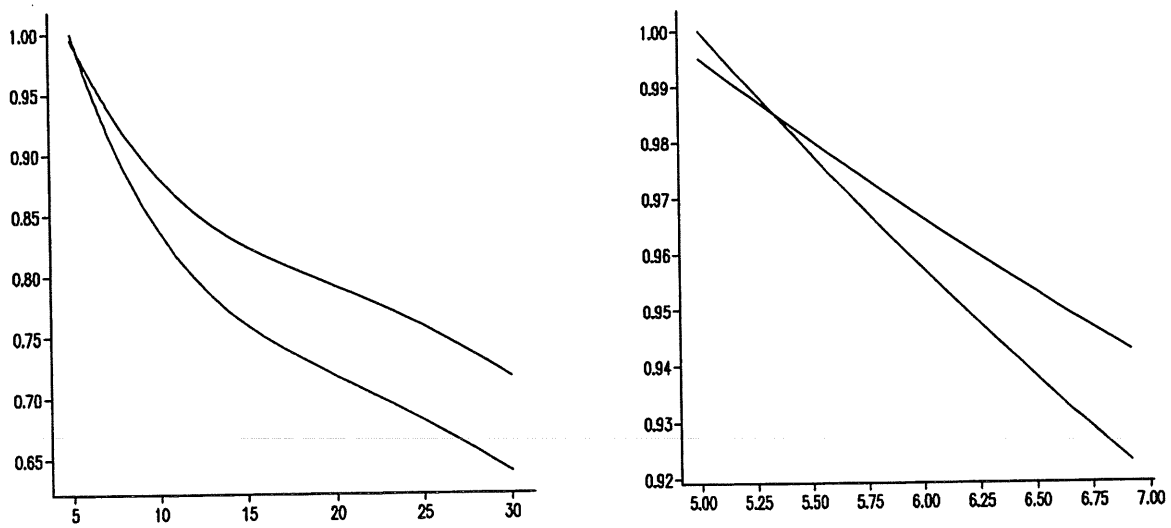


Figure 3. Survivor curves for a child of five that has not had smallpox, when inoculated (continuous) and when not inoculated (dotted). The curves represent the probability to live at least to some age. The detail-graph on the right clearly shows the risk to die from the inoculation.

8. Discussion

Bayesian methods are eminently suitable to update existing information about model parameters when new information becomes available. The information about parameters is represented by a probability distribution. The ensuing prediction uncertainty under specific exogenous circumstances can be evaluated by means of Monte Carlo simulation.

The paper discussed the possibilities for assessment of post-calibration uncertainty of a parameter-rich model as an integrated part of the calibration. The calibration was based on a prior parameter distribution and new observations. It was seen that this calibration problem has a simple theoretical solution, when one has a stochastic model for the difference between prediction and observations. Prior information prevents identifiability difficulties. The actual computation might become prohibitively time-consuming with increasing number of parameters.

From a theoretical point of view, the computer-intensive nature of Bayesian calibration is a mere technical problem. More fundamental problems may well arise that throw doubt on the applicability of the method, for instance through errors in the model structure, in the prior information, or in the stochastic model for the residuals. The relevance of the prior is sometimes controversial: it may for instance be questioned whether processes studied in laboratory experiments really are the same as those occurring outside. When calibrating parameter-rich models, one often fixes some moderately sensitive parameters in advance. The ensuing error will cause difficulties in the formulation of a plausible stochastic model for the residuals. In cases where the basic assumptions are very doubtful, the calibration procedure is no more strongly directed by such assumptions, which may lead to some embarrassment of choice.

Bayesian calibration offers prospects to enhance the usefulness of model predictions through an indication of their precision. It might sometimes happen, however, that the basic assumptions of the calibration are doubtful, or that the procedure is too computer-intensive. In such cases, prediction quality may still be evaluated by means of independent validation data. Anyhow, the mere fact that the basic assumptions are numerous, constitutes an argument to validate on independent data.

9 References

Aldenberg T, Janse JH, Kramer PRG 1995 Fitting the dynamical model PCLake to a multi-lake survey through Bayesian statistics. Ecological modelling 78: 83-99

Berger JO 1985 Statistical decision theory and Bayesian analysis (second edition). New York: Springer Verlag

Bernoulli D 1760 Essay d'une nouvelle analyse de la mortalité causée par la petite vérole et les avantages de l'inoculation pour la prévenir. Hist. Ac. Roy. Sci., 1-45 (published in 1766)

Box GEP, Tiao GC 1973 Bayesian inference in statistical analysis. Reading, Mass.: Addison-Wesley (reprint: New York: Wiley 1992).

Bradley L 1971 Smallpox inoculation: an eighteenth century mathematical controversy. University of Nothingham: Adult Education Department

Cox DR, Hinkley DV 1974 Theoretical statistics. London: Chapman and Hall

Hassanizadeh SM, Carrera J 1992 Editorial: introduction to special issue on validation. Advances in Water Resources 15: 75-83

Gelman A, Carlin JB, Stern HS, Rubin DB 1995 Bayesian data analysis. London: Chapman and Hall

Kabat P, Marschall B, Van Den Broek BJ, Vos J, Van Keulen H 1995 Modelling and parameterization of the soil-plant-atmosphere system: a comparison of potato growth models. Wageningen: Wageningen Pers

Keesman K, Van Straten G 1990 Set membership approach to identification and prediction of lake eutrophication. Water Resources Research 26: 2643-2652

Kroese AH, Van Der Meulen EA, Poortema K, Schaafsma W 1995 Distributional inference, Statistica Neerlandica 49, 63-82

Metselaar K, Jansen MJW 1995a Evaluating parameter uncertainty in crop growth models, IMACS/IFAC First International Symposium on Mathematical Modelling and Simulation in Agriculture and Bio-industries, Brussels

Metselaar K, Jansen MJW 1995b The wicked if-then-else: Discontinuous model responses to parameter changes. Camase-News, july 1995.

Reilly PM 1976 The numerical computation of posterior distributions in Bayesian statistical inference. Appl.Statist 25: 201-209

Spitters CJT, Van Keulen H, Van Kraalingen DWG 1989, A simple and universal crop growth simulator: SUCROS87. Pages 145-181 in: Rabbinge R, Ward SA, Van Laar HH (eds) Simulation and systems management in crop protection, Wageningen: PUDOC

Wallach D, Goffinet B 1989 Mean squared error of prediction as a criterion for evaluating and comparing system models. Ecological modelling 44: 299-306

2.3 Application of the Bootstrap in Plant Genetics

Johan W. Schut

C.T. de Wit Graduate School for Production Ecology, Department of Plant Breeding, Wageningen Agricultural University, P.O.Box 386, 6700 AJ Wageningen, the Netherlands
Email: Johan.Schut@users.PV.WAU

1. Introduction

The bootstrap procedure was introduced by Efron (1979). With the advent of 'fast' computers it rapidly became popular for certain aspects of statistical inference. The basic idea is the use of observed data to construct the distribution function used for statistical inference. First an example is presented to show how the bootstrap technique works in practice.

1.1 Example 1: cultivar by location interaction

Cultivar by location interaction is characterised by a different reaction of cultivars to a change of environment. It is important for plant breeders to know how their cultivars respond to environments which are different from where those cultivars were selected. This simplified example is based on grain yield data (kg/ha; 0% moisture) of two Dutch spring barley cultivars, Apex and Prisma, grown at two locations, Flevoland (APM-hoeve, Swifterbant) and Wageningen, in 1995. The numbers of observations are given in Table 1. The average yields are presented in Fig.1.

Table 1. Numbers of observations of four cultivar by location combinations

	Apex	Prisma
Flevoland	10	11
Wageningen	12	12

In this simplified example the interaction I can be estimated as the difference between the differences of the two cultivars at both locations:

$$\hat{I} = (\bar{y}_{\text{Prisma}} - \bar{y}_{\text{Apex}})_{\text{Wag}} - (\bar{y}_{\text{Prisma}} - \bar{y}_{\text{Apex}})_{\text{Flevo}}$$

$$= (\bar{y}_{\text{Wag}} - \bar{y}_{\text{Flevo}})_{\text{Prisma}} - (\bar{y}_{\text{Wag}} - \bar{y}_{\text{Flevo}})_{\text{Apex}}$$

$$= 333.8 \text{ [kg dm / ha]}$$

with estimated variance:

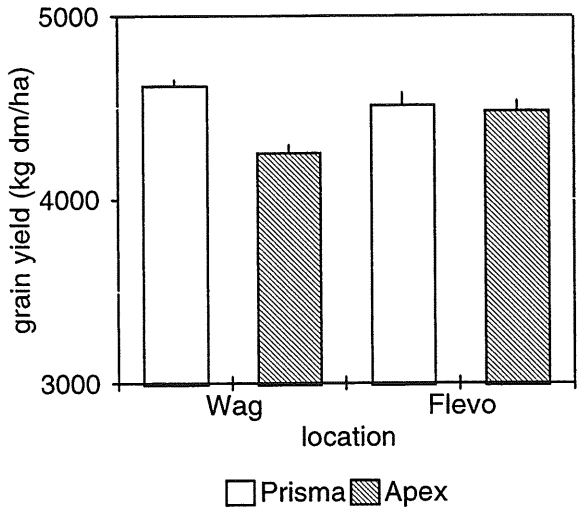


Fig.1. Grain yield (kg dm/ha) average and standard deviation for four cultivar by location combinations

$$s^2(\hat{I}) = \frac{(n_{PW}-1)s^2_{PW} + (n_{AW}-1)s^2_{AW} + (n_{PF}-1)s^2_{PF} + (n_{AF}-1)s^2_{AF}}{n_{PW}+n_{AW}+n_{PF}+n_{AF}-4} \cdot \left(\frac{1}{n_{PW}} + \frac{1}{n_{AW}} + \frac{1}{n_{PF}} + \frac{1}{n_{AF}}\right)$$

$$= (107.5)^2 \ [\text{kg}^2 \ \text{dm} / \text{ha}^2]$$

The question is whether the interaction is significant, i.e. whether $\hat{I}$ is significantly different from zero.

The usual approach to answer this question would be to approximate the distribution of $\hat{I}$ by Student's t-distribution with 41 degrees of freedom (45 minus 4). This assumes that yield observations y_{cl} from cultivar c at location l are normal distributed with mean $\bar{y}_{cl}$ and that variances are equal for all cultivar by location combinations. The resulting 95%-confidence interval is [116.7, 550.8], showing that cultivar by location interaction is significant in this case.

If it is not justified to make the assumptions mentioned above, for instance when yield data are not normally distributed, one could use a bootstrap approach (Algorithm 1.). Bootstrap samples are taken with replacement from the original data for each of the cultivar by location combinations. Bootstrap estimates for I are calculated and sorted resulting in a bootstrap distribution for $\hat{I}$. With this the bootstrap variance for $\hat{I}$ is estimated as $s^2(\hat{I}^*_b) = (106.5)^2$ [kg^2 dm / ha^2] and the 2.5-and 97.5-percentiles are used to determine the 95%-bootstrap confidence interval for $\hat{I}$: [125.0, 552.4], showing significant cultivar by location interaction.

Results of the analytical approach and the bootstrap appear to be quite similar. This is due to the fact that the normality and equal variance assumptions made for the analytical approach were appropriate in this case. When these assumptions do not hold the analytical approach would have given biased results.

Algorithm 1. Bootstrap algorithm for estimating the bootstrap confidence interval for $\hat{I}$

1. sample n_{cl} times *with replacement* from each set of original observations Y_{cl} to construct bth bootstrap sample $Y^*_{cl.b}$,
e.g. $Y^*_{AF.b} = (y_7^{AF}, y_3^{AF}, y_5^{AF}, y_9^{AF}, y_{10}^{AF}, y_2^{AF}, y_8^{AF}, y_9^{AF}, y_3^{AF}, y_2^{AF})$
2. calculate $\hat{I}^*_b = (\bar{y}^*_{PW.b} - \bar{y}^*_{AW.b}) - (\bar{y}^*_{PF.b} - \bar{y}^*_{AF.b})$
3. repeat 1. and 2. B times
4. order the B $\hat{I}^*_b$ -estimates
5. determine the confidence interval for $\hat{I}^*_b$ and/or estimate $s^2(\hat{I}^*_b)$

1.2. The bootstrap

The bootstrap procedure was introduced by Efron (1979). It is based on the so-called plug-in principle (Efron and Tibshirani, 1993) also called bootstrap principle (Hall, 1992). This states that parameter $\theta = t(F)$ with probability distribution function F can be estimated by: $\hat{\theta} = t(\hat{F})$ using the empirical distribution function $\hat{F}$. The empirical distribution function is based on the data. An example of such a distribution, based on six observations, is shown in Fig. 2. A simple and well-known example of the plug-in principle is the plug-in estimate for the expectation $\theta = E_F(x) = \int x\, dF(x)$ which is $\hat{\theta} = E_{\hat{F}}(x) = \frac{1}{n}\sum_{i=1}^{n} x_i$ (Efron and Tibshirani, 1993).

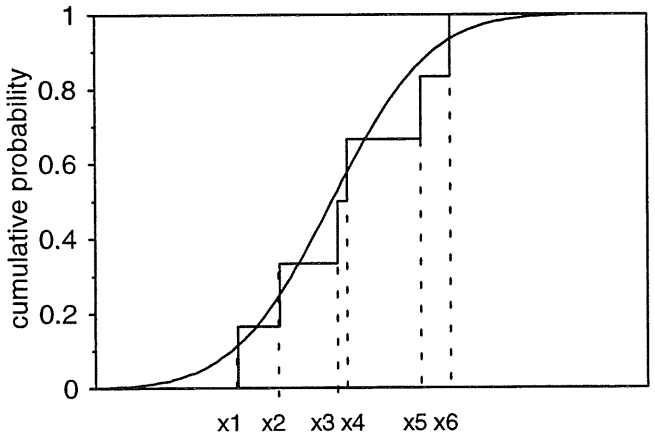


Fig.2. An example of an empirical distribution function based on six observations

Bootstrap methods are in fact resampling from approximated distributions and can be divided in two groups: nonparametric and parametric bootstrap. In the nonparametric bootstrap one resamples, according to the bootstrap principle, from the empirical or sample distribution function $\hat{F}$, i.e. from the observed data, resulting in the empirical distribution function $\hat{F}^*$ and one uses $\hat{\theta}^* = t(\hat{F}^*)$ as an estimate of $\hat{\theta} = t(\hat{F})$. In the parametric bootstrap one resamples from the estimated parametric distribution function $\hat{F}_{param}$, i.e. F as determined by parameters estimated from the observed data,

resulting in the empirical distribution function $\hat{F}^*_{\text{param}}$ and one uses $\hat{\theta}^*_{\text{param}} = t(\hat{F}^*_{\text{param}})$ as an estimate of $\hat{\theta} = t(\hat{F})$ (Hall, 1992).

There are several reasons why one would choose to use the bootstrap procedure to assess accuracy of $\hat{\theta}$. A very important advantage of the nonparametric bootstrap is the fact that no probability distribution functions are necessary for statistical inference. For instance in the case of simultaneous distributions this may be a great asset. Also a wide range of parameters can be handled; including parameters that can not be derived from the analytical probability distribution function, i.e. $\theta \neq t(F)$. Another merit is that bootstrapping usually is quite straightforward and comprehensible. A prerequisite is, of course, that a 'fast' computer is available.

2. Example 2: correlation between two similarity measures

2.1. Introduction

Plant breeders are interested in relatedness of genotypes. Therefore a genetic similarity measure, based on pedigree information, is compared with a genetic similarity measure, based on a new type of DNA-based information. From the pedigree data we calculated for a pair of genotypes i and j the so-called coefficient of coancestry (f_{ij} ; Malécot, 1948), i.e. the probability that a random allele at a random locus in genotype i is identical by descent to a random allele at the same locus in genotype j. The DNA-based information consisted of so-called AFLP-marker data (Vos et al., 1995). A small piece of an autoradiogram of an AFLP electrophoresis gel is shown in Fig.3. Horizontal bands represent DNA-fragments of specific sizes; a vertical lane contains bands derived from one genotype. Simultaneous presence of bands at equal heights in two lanes means that the two genotypes are identical at the position where the fragment is located in the DNA. The data-matrix derived from the autoradiogram reads as follows:

		genotypes					
		g_1	g_2	g_3	.	.	g_n
	m_1	1	0	1	.	.	1
	m_2	1	1	1	.	.	0
	m_3	0	0	1	.	.	0
markers	m_4	1	1	0	.	.	1
	m_5	0	1	0	.	.	1
	.	.	.	.	.	.	.
	.	.	.	.	.	.	.
	m_q	1	0	1	.	.	1

with 1 designating presence of a band and 0 designating absence of a band.

Genetic similarity between genotypes i and j (gs_{ij}) is calculated following Dice (1945):

$gs_{ij} = \dfrac{2N_{ij}}{N_i + N_j}$, where N_{ij} is the number of bands present in both genotypes, N_i is the number of bands present in genotype i and N_j is the number of bands present in genotype j.

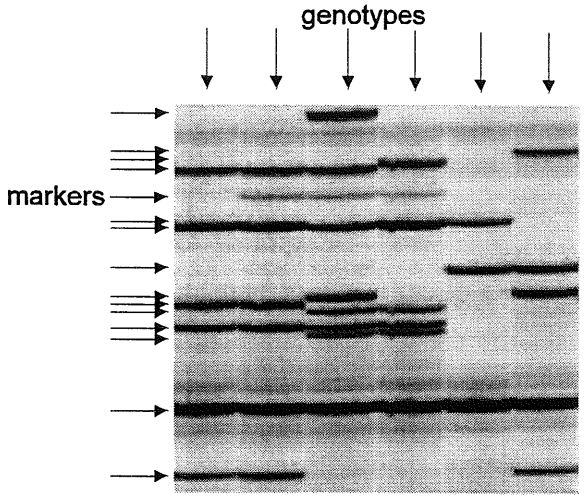


Fig. 3. Detail of an autoradiogram of an AFLP-gel. Horizontal bands represent DNA-fragments of specific sizes; a vertical lane contains bands derived from one genotype.

The aim was to test whether there was a significant correlation between f_{ij} and gs_{ij}, i.e. to check whether they are both relevant measures of relatedness. The null hypothesis in the test assumed absence of correlation. A second aim was to assess accuracy of $r(f_{ij}, gs_{ij})$. Twenty-five European spring barley genotypes (cultivars, lines) were used in this study. General answers for the European spring barley population should be provided.

2.2. Bootstrapping artificial data

To investigate whether a known distribution of $r(f_{ij}, gs_{ij})$ could be approximated by a non-parametric bootstrap procedure based on a given (f_{ij}, gs_{ij})-dataset, it was decided to create a number of artificial (f_{ij}, gs_{ij}) datasets using independent samples of artificial genotypes from an infinite population. In this way we could estimate the distribution for $r(f_{ij}, gs_{ij})$ based on the correlation coefficient estimate as calculated for each of these datasets.

The artificial genotypes are characterised by random positions in a c-dimensional sphere with radius 0.5. So each genotype is characterised by a c-dimensional vector and its maximum Euclidean distance to other genotypes is 1. The coefficient of coancestry f_{ij} is calculated as $1-d_{ij}$, where d_{ij} is the Euclidean distance between the position of genotype i and the position of genotype j. For each original set of genotypes the genotypes are randomly shifted around their original positions. This is done for each genotype by sampling a new position vector from a multivariate normal distribution that has the original position vector as expectation. Genotype positions beyond the above-mentioned sphere are not accepted. Based on the slightly shifted genotype positions the genetic similarities gs_{ij} are calculated similarly as f_{ij}, resulting in an (f_{ij}, gs_{ij}) dataset. Then $r(f_{ij}, gs_{ij})$ can be calculated. The variance of the multivariate normal distribution used to shift the genotype positions determines the value of $r(f_{ij}, gs_{ij})$. In this way we created 10,000 artificial (f_{ij}, gs_{ij}) datasets and calculated the 10,000 correlation coefficients $r(f_{ij}, gs_{ij})$. For this parametric bootstrap we used five dimensions (c=5) and twenty-nine genotypes (n=29). Afterwards one of these artificial (f_{ij}, gs_{ij}) datasets, with an $r(f_{ij}, gs_{ij})$ that was close to the average correlation coefficient of the 10,000 bootstraps, was used to perform a nonparametric bootstrap procedure with 2000 samples to investigate whether this procedure approximates the distribution of $r(f_{ij}, gs_{ij})$ well enough to make statistical inferences for

our real data. Genotypes were sampled with replacement, after which f_{ij} and gs_{ij} matrices were constructed. From these matrices bootstrap estimates for $r(f_{ij}, gs_{ij})$ were calculated (Algorithm 2.). Results presented in Fig. 4. show that the parametric and non-parametric distribution have a similar shape. This led us to the conclusion that the nonparametric bootstrap, as presented in Algorithm 2., can be used to assess accuracy of $r(f_{ij}, gs_{ij})$. In this case the genotypes were assumed to be a random and representative sample from a larger population. As a reference we show in Fig. 4. the bootstrap distribution of $r(f_{ij}, gs_{ij})$ as a result of sampling (f_{ij}, gs_{ij}) data directly (which is incorrect) instead of sampling genotypes. The incorrect distribution resembles the usual distribution (results not shown) used to make statistical inference on correlation coefficients assuming independence of (f_{ij}, gs_{ij}) data.

Algorithm 2. Non-parametric bootstrap procedure for estimating a confidence interval for a correlation coefficient between two similarity measures within a group of *n* genotypes

1. sample *with replacement* from *n* genotypes
2. construct an *nxn*-matrix of f_{ij} -values and an *nxn*-matrix of gs_{ij} -values
3. remove f_{ij} and gs_{ij}, whenever $i=j$
4. calculate $r_b^*(f_{ij}, gs_{ij})$
5. repeat 1.,2.,3. and 4. B=2000 times
6. order the $r_b^*(f_{ij}, gs_{ij})$ -estimates
7. derive the confidence interval for $r(f_{ij}, gs_{ij})$

2.3 Bootstrapping real data

Using the barley data mentioned in 2.1. we calculated the correlation coefficient $r(f_{ij}, gs_{ij})$ between the off-diagonal part of two symmetric 25 x 25 matrices - 25 being the number of genotypes

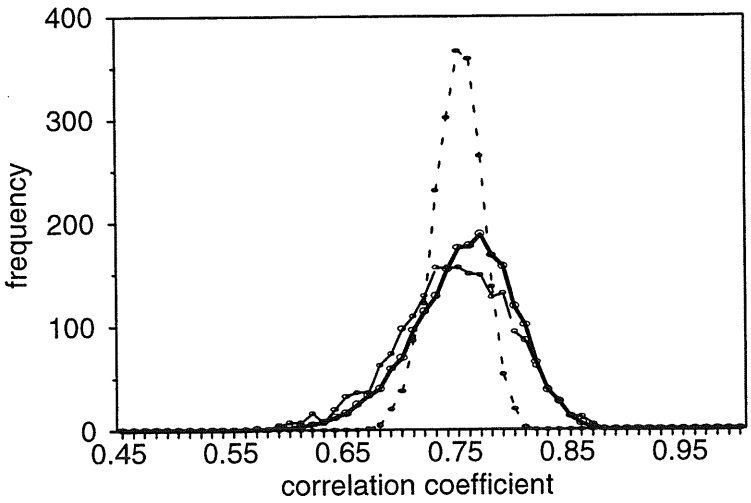


Fig.4. Bootstrap distribution of correlation coefficient estimates for artificial (f_{ij}, gs_{ij}) data (B=2,000 samples). Frequency classes are connected by lines for the sake of clearness and scaled by a factor 1/5 in the case of the parametric bootstrap (10,000 samples). solid line - parametric bootstrap; dashed line - nonparametric bootstrap sampling from genotypes; dotted line - nonparametric bootstrap sampling directly from (f_{ij}, gs_{ij}) data which is incorrect.

- containing f_{ij} and gs_{ij} respectively. The estimated correlation coefficient amounted to 0.389. Using the nonparametric bootstrap procedure described in algorithm 2. we estimated the 95%-confidence interval for $r(f_{ij}, gs_{ij})$, which is [0.097,0.600]. This showed that the correlation between marker-based similarity and pedigree based similarity is significant for α=0.05. However, the similarity measures f_{ij} and gs_{ij} gave partly different information and may not be very accurate, otherwise the correlation coefficient would have been much higher. The extent to which f_{ij} and gs_{ij} were different and/or inaccurate, as shown by the value of $r(f_{ij}, gs_{ij})$, could not be estimated very accurately on the basis of a sample of 25 genotypes, as can be seen from the large confidence interval.

3. Example 3: cross validation of groups of molecular markers

The AFLP-marker data as presented in section 2. are combined data sets based on several groups of markers. Each marker group is generated by a different combination of so-called PCR-primers, causing amplification of specific DNA-fragments based on the DNA-sequences at the ends of the fragment. The hypothesis tested is that the marker sets of each of the primer combinations yield the same gs_{ij}-estimates as the pooled marker sets of the other primer combinations. A bootstrap procedure approximating the simultaneous distribution of the gs_{ij}-estimates in a symmetric gs-matrix is used to answer this question.

The marker data matrix presented in section 2.1. has now been divided into p primer combination groups where the numbers of markers in primer combination h is q_h:

		genotypes				
		g_1	g_2	.	.	g_n
markers	m_{11}	1	0	.	.	1
	m_{12}	0	1	.	.	1
	.	.	.			.
	.	.	.			.
	m_{1q_1}	1	1	.	.	0
	.	.	.			.
	.	.	.			.
	m_{h1}	1	1	.	.	0
	m_{h2}	0	1	.	.	1
	.	.	.			.
	.	.	.			.
	m_{hq_h}	1	0	.	.	0
	.	.	.			.
	.	.	.			.
	.	.	.			.
	.	.	.			.
	m_{pq_p}	1	0	.	.	1.

For the analysis the genetic similarities between twenty-one European spring barley genotypes (n=21) were used to test whether each of the fourteen primer combinations (p=14) was equivalent to the aggregate of the 13 others. The average number of polymorphic bands per primer combination was 31.

A bootstrap procedure was used to determine a 95% simultaneous confidence interval for the gs_{ij}-estimates that were calculated using marker data of one primer combination. The confidence interval was based on the marker data provided by the other thirteen primer combinations. The number of markers sampled was equal to q_h , i.e. the number of markers in the primer combination h under investigation. Genetic similarities were calculated based on these samples. Using B bootstrap gs_{ij}^*-matrices a simultaneous confidence interval was constructed. The number of bootstrap samples B was not determined beforehand. Only the number of extreme bootstrap estimates at the lower and upper ends of each gs_{ij}^*-distribution, that had to be stored during the bootstrap procedure, was fixed at five. Besides the values of the extreme gs_{ij}^* estimates, also information about which bootstrap replications generated these extremes was stored. After every bootstrap replication the total number of bootstrap replications B_E that produced one or more of these extreme gs_{ij}^*-estimates, was counted. The bootstrap sampling was interrupted at the moment that $\frac{B_E}{B}$ was approximately equal to α=0.05. The simultaneous confidence interval was based on the values on the fifth and the (B-5)th estimate in each ordered gs_{ij}^*-distribution. If one or more of the gs_{ij}-estimates, based on the primer combination under investigation, exceeded the confidence limits, it was decided that that primer combination produced gs_{ij}-estimates significantly different from those of the other primer combinations (Algorithm 3.).

Algorithm 3. Cross validation of *gs*-estimates of primer combination *h* using simultaneous bootstrap confidence intervals

1. calculate gs_{ij}-matrix for primer combination h, which has q_h markers
2. sample q_h times from markers from all other primer combinations
3. calculate gs_{ij}^*-matrix
4. count the total number of bootstrap replications B_E that produced one or more extreme gs_{ij}^*-estimates, being the five smallest and the five largest gs_{ij}^*-values of each i,j-combination
5. repeat 2., 3. and 4. and stop at the moment that $\frac{B_E}{B}$ is approximately equal to α=0.05, where B is the total number of bootstrap replications.
6. construct bootstrap confidence intervals for every gs_{ij} using the fifth and the (B-5)th estimate of each ordered gs_{ij}^*-distribution
7. test whether all $\binom{q_h}{2}$ gs_{ij}-estimates of primer combination h lie in the 95%-confidence interval

Among the fourteen primer combinations one was found that produced significantly different gs_{ij}-estimates. After cross validation with another set of genotypes and markers it was decided that this primer combination may have been a false positive. Our final conclusion therefore is that there were no deviating AFLP primer combinations, when looking at gs_{ij}-estimation.

4. References

Dice LR 1945 Measures of the amount of ecologic association between species. Ecology 26: 297-302

Efron B 1979 Bootstrap methods: another look at the jackknife. Ann. Stat. 7: 1-26

Efron B, Tibshirani RJ 1993 An introduction to the bootstrap. New York: Chapman & Hall

Hall P 1992 The bootstrap and Edgeworth expansion. New York: Springer

Malécot G 1948 Les mathémathiques de l'hérédité. Paris: Masson & Cie

Vos P, Hogers R, Bleeker M, Reijans M, Van de Lee T, Hornes M, Frijters A, Pot J, Peleman J, Kuiper M, Zabeau M 1995 AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res. 23: 4407-4414

3. Data and information systems

Maximum utility from expensive data

What is the information content of a datum? Information systems must carry and process the data. Storing, retrieving and displaying spatial data, in particular interactively. In a spatial sense, data are representative for a particular scale. Methodology to transfer data from one scale to another is needed. This requires averaging, or diversification of data. Also new data are needed, maybe based upon data at another scale, but sometimes not. Data-based decision support apply to all scales of agricultural systems, where decisions have to be taken based on available information.

3.1 Multi-scale approaches for geodata

Martien Molenaar

International Institute for Aerospace Survey and Eartch Sciences (ITC)
PO Box 6, 7500 AA Enschede, The Netherlands
E-mail: molenaar@itc.nl

Topological object relationships in combination with object classification hierarchies appear to be fundamental in the definition of the aggregation rules for spatial objects. Such rules are essential building blocks for the construction of generalization procedures in spatial databases. A model for spatial database generalization can be formulated based on the syntax of the Formal Data Structure (FDS) as proposed in (Molenaar 1989). The syntax of the FDS will be formalised first, then database generalization procedures will be formulated with this syntax.

Four strategies will be explained for generalization, these are:
- *geometry driven generalization*, the change of geometric resolution cells determines the transition from entities at a large scale to new entities at a smaller scale,
- *class driven generalization*, spatial objects at a large scale forming a region under one thematic class are merged for representation at a smaller scale,
- *functional generalization* links objects that are considered as response units in processes defined at different scale levels,
- *structural generalization* gives a stepwise simplification of a spatial process description in an area.

These different strategies will be explained and compared. In the final discussion spatial data generalization will be presented as data transformation processes. For such a transformation we should specify which aspects of a terrain description should be invariant after generalization, this may have the effect that other aspects will not be invariant.

1. Introduction

1.1 Spatial Processes at Multi-Scale Levels

Multi-scale approaches are at present one of the focal points of the GIS research community. This is due to the rising awareness that many processes on the earth surface can only be monitored and managed if they are understood in their geographical context. Part of this context is defined by the scale range at which these processes work.

If we consider for instance the development of land use in a district, then we see that this is driven by actors at a lower aggregation level such as farmers, residents and companies. Their activities are constrained however by the socio-economic conditions and the infrastructure at regional level and by the macro economic planning at national and supra national level. An other example is the development of the natural vegetation cover of certain regions. The actual state of such a vegetation cover is defined by the co-occurrence of species that form vegetation types, which are part of eco-systems. Their development will be constrained by climatic conditions, the geologic and soil condition of the region and its hydrology. Here too we find hierarchical levels of organization.

The monitoring and management of such processes requires information at different scale levels. The research problem for the GIS community in this context is:

- to decide for each type of process which information should be handled at each scale level,
- to develop methods for transferring information between the different scale levels so that duplication of expensive data acquisition can be avoided as much as possible, and so that the consistency between data at the different scale levels can be maintained.

This second item is strongly related to the long standing research problem of map generalization, that is why it is often seen from that perspective. Researchers in this field become more and more aware of the fact, however, that multi-scale approaches in a GIS environment can be dealt with by data base generalization operations. These allow approaches that are quite different form the procedures applied in map generalization and they are more flexible.

This new research field derives it terminology and many of its concepts from both cartography and the object oriented approaches of computer science. This mixture of the idiom of different disciplines leads often to confusion so that the concepts that are covered by the terminology become fuzzy. This confusion may send researchers in the wrong direction when they want to solve multi-scale problems. In this paper a data base perspective for multi-scale approaches will be presented, emphasising the role of topologic and semantic (hierarchical) data models.
The different concepts that play a role in the generalization of spatial data will be discussed in relation to several strategies which can be used for solving multi-scale problems.

1.2 Spatial Databases and Multi-Scale Problems

A spatial database contains data that represent in principle elementary statements about some spatial situation. These elementary statements refer to the relationships between objects and geometric data and thematic data etc. Query operations are applied to derive other statements that contain more relevant information for the user, e.g. about the state of the objects and about their mutual relationships. The semantics of the derived statements is generally of a higher level of complexity than the stored data. They should help the user to understand the structure of the mapped area, therefore they often refer to spatial relationships between the mapped objects. If the area structure should be understood at a higher abstraction level though, these derived statements could also refer to relationships among aggregated objects. The understanding of the structure of an area at several abstraction levels is strongly related to the problem of spatial generalization and multi scale representations.

Aggregation hierarchies for spatial objects can serve as basic tools for multiple representations of geo-data within the context of conceptual generalization (information abstraction) processes. These aggregation hierarchies can be based on the formal data structure (FDS) for single valued vector maps (Molenaar 1989), which combines aspects of object-oriented and topologic datamodels. Point-, line- and area objects are represented with their geometric and thematic data. Their geometric representation supports the analysis of topologic object relationships, whereas their thematic description is structured in object classes that form generalization hierarchies. These class hierarchies together with the topologic object relationships support the definition of aggregation hierarchies of objects. The classification- and aggregation hierarchies play an important role in linking the definition of spatial objects at several scale levels. Accordingly, these structures are fundamental in the definition of rules for modelling generalization of spatial information at different resolution levels. The capacity of Geographical Information Systems (GIS) to register and handle topological information in combination with object hierarchies makes them very useful tools for the automation of conceptual generalization of spatial data.

In a cartographic context, generalization can be defined as the process of abstracting the representation of geographic information when the scale of a map is changed. It is a complex

process involving abstraction of thematic as well as geometric data of objects. The process usually involves two phases:

a) a conceptual generalization phase, which implies the determination of the content of a representation in the generalized situation (information abstraction), and the definition of rules how the generalized objects can be derived from the objects at lower generalization level

b) a graphical generalization phase (cartographic generalization), which implies the application of algorithms for geometric simplification of shapes and for symbolization to assure map legibility.

Information abstraction in these subprocesses is mainly determined by expert knowledge and can usually be expressed as logical rules. These rules are susceptible to be translated as database management procedures in a GIS environment (Martinez Casasnovas 1994, Richardson 1993). Regarding information abstraction, several processes are recognized: classification, association, (class) generalization and aggregation. Class generalization and aggregation are directly related to changes in the level of definition of objects when the mapping scale changes. Aggregation is the combination of elementary objects to build composite objects and will be based on two types of rules:

a rules specifying the classes of elementary objects building a composite object and

b rules specifying the geometric characteristics (such as minimum size) and topological relationships of these elementary objects (i.e. adjacency, connectivity, proximity, etc.).

The syntactic structure of a data model for handling topologic and hierarchical relationships between spatial objects will be explained in this article. Processes for database generalization will be formulated with this data model.

2. A Spatial Data Model For Multi-Scale Approaches

2.1 Topologic Structures for the Representation of Spatial Objects

Entity Types for Spatial Data

The spatial structure of an area can be expressed in terms of point-, line- and area objects. Their spatial extend and their topologic relationships will be expressed by means of a set of geometric elements. (Frank ea 1986) showed that the geometric structure of a vector map can be described by means of cell complexes. For a two dimensional map these consist of 0-cells, 1-cells and 2-cells. The 0-cells and 1-cells play similar roles as respectively the nodes and edges when the geometry of the map is interpreted as a planar graph. The 2-cells can then be compared to the faces related to the planar graph through Eulers formula (Gersting 1993). The terminology of the planar graph interpretation will be used here, but their relationships will be formulated as those for cells according to the concepts presented in (Molenaar 1994). This formulation is then based on

- three geometric types: *nodes, edges and faces* represented by the symbols *n, e and f*
- three geometric object types: *point objects, line objects, area objects.*

The further developments will only use line- and area objects which will be represented by the symbol O_l *and* O_a. Instances of these entity types will be indicated by suffixes.

The reader will recognise that the formalization explained in this paper is to a large extent isomorphic with topologic data structures defined for GIS such as ATKIS/DLM, DGF, TIGER and DIME etc., see (Hesse 1992, Walter 1994, Marx 1990, USBUREAU 1990). This formalization will be based on the FDS described in (Molenaar 1989).

Relationships Between Nodes, Edges and Faces

The following relationships can be defined between the geometric elements of a planar graph:

- Edge e_i has node n_j as the begin node
  → $Begin[e_i, n_j] = 1$ otherwise = 0
- Edge e_i has node n_k as the end node
  → $End[e_i, n_k] = 1$ otherwise = 0

We will consider edges as straight line segments. Each edge will always have one face at its left hand side and on at its right hand side. These relationships will be expressed by the following functions:

- Edge e_i has face f_a at its left-hand side
  → $Le[e_i, f_a] = 1$
  For any $f_b \neq f_a$ we get then $Le[e_i, f_b] = 0$
- Edge e_j has face f_a at its right-hand side
  → $Ri[e_j, f_a] = 1$
  and again for $f_b \neq f_a$ we get then $Ri[e_j, f_b] = 0$

If an edge e_i has face f_r at the right hand side and face f_l at the other side then these faces are adjacent at this edge, which will be expressed by the function

$ADJACENT[f_r, f_l | e_i] = 1$ (and = 0 otherwise)

the fact that there is some edge where the faces are adjacent can then be expressed by

$ADJACENT[f_r, f_l] = 1$ (and = 0 otherwise)

Line Objects

The geometry of a simple line object is represented by a chain of edges as in figure 1a. The fact that an edge e_p is part of the object can be established by the function $Part_{11}[e_p, O_l]$. The notation $Part_{uv}[\ ]$ means that an entity with spatial dimension u is a part of an entity with dimension v. If the edge is part of the object then $Part_{11}[\ ] = 1$, else it has a value = 0.

A line object will have a begin node
$n_b = BEG(O_l)$ and an end node
$n_e = END(O_l)$. These can be found through the edges of O_l, the direction of the object can then be specified by $Dir[O_l] = \{ n_b, n_e \}$

Area Objects

The geometry of an simple area object is represented by one or more adjacent faces as in figure 1b. If a face f_v is part of an area object O_a this will be represented by $Part_{22}[f_v, O_a] = 1$. The set of all objects from which a face f is a part is $OA(f) = \{ O \mid Part_{22}[f, O] = 1 \}$.

This function relates two two-dimensional entities. If there are overlapping area objects then each face might be part of several objects, but each object will also consist of one or more faces. Therefore this is a many-to-many relationship. Overlapping objects can be found through their common faces.

Now it is possible to check whether edge e_i is related through face f_v to an area object O_a. There is at most one face for which both $Le[e_i, f_v] = 1$ and $Part_{22}[f_v, O_a] = 1$. If such a face exists then the function relating the edge to the object will get the value 1, in all other cases it will be = 0. Hence if edge e_i has area object O_a at its left-hand side then $Le[e_i, O_a] = 1$ else = 0. Similarly if edge e_i has area object O_a at its right-hand side then $Ri[e_i, O_a] = 1$ else = 0.

The combination of these two functions gives for edge e_i:

$B[e_i, O_a] = Le[e_i, O_a] + Ri[e_i, O_a]$

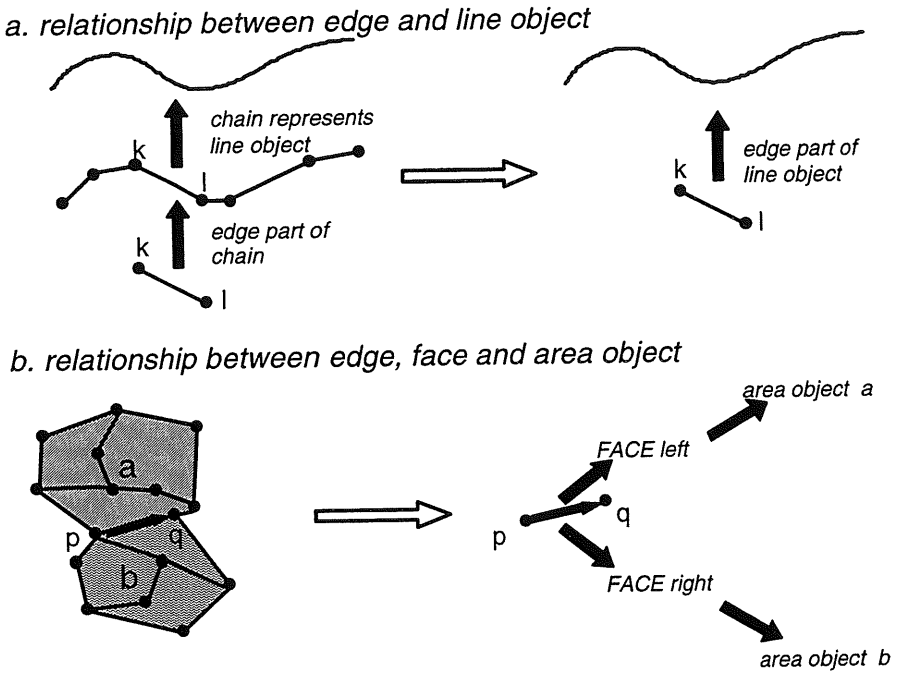


fig. 1: Relationships between edges and objects.

If an edge e_i is part of the boundary of O_a then only one of the functions Ri and Le is equal to 1 but not both, so for such an edge we find $B[e_i , O_a] = 1$. If e_i has O_a both at its left-hand side and at its right-hand side then
$B[e_i , O_a] = 2$, in that case it is running through O_a. If
$B[e_i , O_a] = 0$ there is no direct relationship between e_i and O_a.

Adjacent Area Objects

When an edge has an object O_a at its left hand side and not at its right hand side and object O_b at its right hand side and not at its left hand side then these objects are adjacent at this edge. If the objects overlap not at all, i.e. if they have no common faces and they are adjacent at least one edge, then they are adjacent which is expressed by the function $ADJACENT[O_a , O_b] = 1$ (and = 0 otherwise).

Line- and Area Objects

Several important relationships between a line object O_l and an area object O_a can be found by checking for each edge that is part of the line object how it is related to the area object. This will be expressed by the functions

$$Le[O_l , O_a \mid e_i] = MIN(Le[e_i , O_a], Part_{li}[e_i , O_l])$$
$$Ri[O_l , O_a \mid e_i] = MIN(Ri[e_i , O_a], Part_{li}[e_i , O_l])$$

For the relationship between a line object O_l and an area object O_a we can write

$$B[O_l , O_a \mid e_i] = Le[O_l , O_a \mid e_i] + Ri[O_l , O_a \mid e_i]$$

If this function has the value = 2 then the line object runs through the area object at edge e_i if the value = 1 then it is at the border and if it is = 0 then there is no relationship. The relationship between the two objects might be different at different edges.

A Hydrologic Example

For modelling hydrological systems three types of elementary objects will be defined according to (Martinez Casasnovas 1994), these are the water course lines, the drainage elements and their catchments, see figure 2. The drainage elements are gullies, each element has a catchment area from which it receives overland flow of water. Each element also receives water from upstream

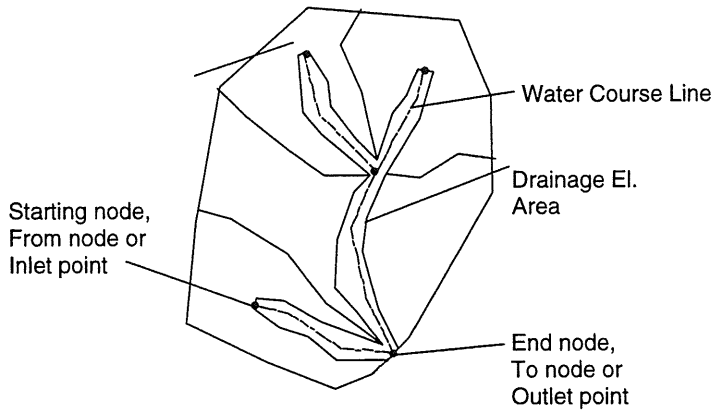


fig. 2: Elementary objects in a drainage system.

elements (if there are any) and it empties into a downstream element. The water flow through each element is represented by a water course line.

The relationship between these objects is one to one in the sense that each drainage element D_i contains exactly one water course line W_i and is embedded in exactly one subcatchment area C_i. A subcatchment area may be dissected by its drainage element, as can be seen in figure 2, but it is still considered as one subcatchment. The topologic relationships between these objects can be expressed by functions of section 2:

for water course line W_i and drainage element D_i is:
$(\forall e_k \mid Part_{11}[e_k, W_i] = 1) \Rightarrow B[W_i, D_i \mid e_k] = 2$
this will be written shortly as $B[W_i, D_i] = 2$
if $j \neq i$ then $B[W_i, D_j] = 0$

This means that W_i runs through D_i so that it has D_i at both sides and it is not related to any other drainage element. This is a topologic restriction due to a semantic constraint valid in the context of this hydrologic model. Another semantic constraint is

for drainage element D_i and catchment C_i is
$ADJACENT[D_i, C_i] = 1$
if $j \neq i$ then $ADJACENT[D_i, C_j] = 0$

so that D_i is only adjacent to C_i and to no other catchment.

Each drainage element is also connected to a downstream element and, depending on its position in the network, to one or more upstream elements. The relationship between the drainage elements can also be found through the watercourse elements. These should be directed according to the direction of the water flow, for each W_i we can find the upstream element W_h through the rule $END(W_h) = BEG(W_i)$. This relation between these water course lines will be expressed by $Upstr[W_i, W_h] = 1$, this function will have the value = 0 otherwise.

Due to the 1 to 1 relationships between W, D and C the upstream relationship can be transferred as follows

$$Upstr[W_j, W_i] = Upstr[D_j, D_i] = Upstr[C_j, C_i]$$

so that the order relationships between the water course lines can be translated into order relationships between the areas in which they are contained. We will assume here that the stream network structure is defined so that for each W_j with a Strahler number > 1 there are two or more upstream water lines W_i, but for each W_i there is only one downstream water line W_j.

2.2. *Object Classes and Class Hierarchies*

Terrain objects refer to features that appear on the surface of the earth and are interpreted in a systems environment with a thematic and geometric description. In most applications the terrain objects will be grouped in several distinct classes and a list of attributes will be connected to each class. Let Ci be a class, and let the list of its attributes be $LIST(Ci) = \{A_1, A_2, ..., A_n\}$ then

$$LIST(Ci) \neq LIST(Cj) \text{ for } i \neq j$$

i.e. these attribute lists will be different for different classes. Terrain objects inherit the attribute structure from their class, i.e. each object has a list containing a value for each class attribute, thus for member e of class C:

$$LIST(e) = \{a_1, a_2, ..., a_n\}$$
where: $a_k = A_k(e)$ is value of A_k for object e
$e \in C$
$A_k \in LIST(C)$

When two or more classes have attributes in common, then a superclass can be defined with a list containing these common attributes as "superclass-attributes" (Molenaar 1993). The original classes are subordinated to these super classes, for example, the class 'forest' is a superclass containing subclasses such as "deciduous", "evergreen", and "mixed forest". The terrain objects are then assigned to these classes.

With these observations we find the class hierarchical structure of figure 3. In literature on semantic modelling (Brodie 1984, Brodie e.a. 1984, Egenhofer e.a. 1989, Oxborow e.a. 1989) the upward links of the classification hierarchy are labelled respectively as "ISA" links. These links relate each particular object to a class and to super classes.

It is possible to add more hierarchical levels to the structure of figure 3. At each level the classes inherit the attribute structure of their superclass at the next higher level and propagate it normally with an extension to the next lower level. At the lowest level in the hierarchy are the terrain objects, at this level the attribute structure is not extended any more, but here the inherited attributes are evaluated. In this case we find for e:

$$LIST(e) = \{a_1, a_2, ..., a_n\}$$
where:
$a_1 = A_1(e)$ is value of A_1
$A_1 \in LIST(C) \cup LIST(SC) \cup .....$

thus A_1 is an attribute of the class *or superclass(es)* of e. If the classes at each level are disjoint so that the hierarchy has a tree structure then the terrain objects will get their attribute structure only through one inheritance line in the hierarchy, i.e. they have a unique thematic description. We will work under this assumption in this paper.

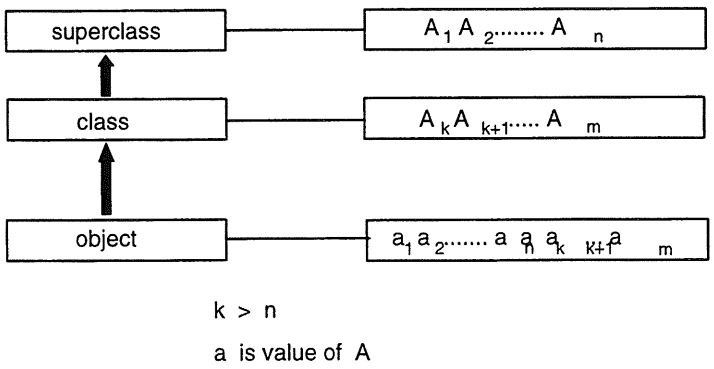


fig 3. The hierarchical relationships between objects and classes and their attributes.

The terrain objects occur at the lowest level in the classification hierarchy. They can be seen as the elementary objects within the thematic field represented by the classification system. This implies that the decision, whether certain terrain objects should be considered as elementary or not, should always be made within the frame work of a thematic field. Objects that are considered as elementary in one thematic field are, however, not necessarily elementary in another thematic field.

2.3 Object Aggregation

Objects can be aggregated to build composite objects at several levels of complexity. These may form aggregation hierarchies which are quite distinct from classification hierarchies. An aggregation hierarchy shows how composite objects can be built from elementary objects and how these composite objects can be put together to build more complex objects and so on. In literature on semantic modelling (Brodie 1984; Brodie e.a.1984; Egenhofer e.a.1989; Oxborrow e.a.1989) the upward relationships of an aggregation hierarchy are called PARTOF links. These links relate a particular set of objects to a specific composite object and on to a specific more complex object and so on. For example, 'James Park is PARTOF Westminster is PARTOF London.'

For composite spatial objects the PARTOF links might be based on two types of rules involving the thematic and the geometric aspects of the elementary objects. Consequently the generic definition of a type of an aggregation should consist of the following rules (Molenaar 1993):

- *rules specifying the classes of the elementary objects building an aggregated object of this type,*
- *rules specifying the geometric and topologic relationships among these elementary objects.*

Suppose that aggregated objects of a type T should be formed. To do that we should first identify the objects O_i that could be part of such aggregates. These objects should fulfil certain criteria, which according to the two sets of rules given earlier will often be based on the thematic data of the objects. Let these criteria be expressed by a decision function

$D(O_i, T) = 1$ if the object fulfils the criteria
$= 0$ otherwise

Regions can now be formed by applying two rules:

> all objects in the region satisfy the decision function for T
$(\forall O_i \mid O_i \in R_r) \Rightarrow D(O_i, T) = 1$

> All objects that satisfy the decision function for T and that are adjacent to objects of the region belong to the region
$(\forall O_i \mid D(O_i, T)=1)(\exists O_j \in R_r \mid ADJACENT[O_i, O_j]=1) \Rightarrow (O_i \in R_r)$

The second rule implies that a region can be formed when at least one object has been identified that fulfils the first rule. This object is then the seed around which the region can grow by identification of the other objects that fulfil both rules.

A region R_r can be expressed as a set of objects, i.e.:

$$R_r = \{...., O_i, ...\}$$

The objects of the region can be aggregated to form an aggregated or composite object O_{ar}, the suffixes express that the object is of aggregation type a and r is its identification number. The operation will be expressed by

$$O_{ar} = AGGR(R_r) = AGGR(\{...., O_i, ...\})$$

The fact that O_i is part of O_{ar} is expressed by

$$Part_{kl}[O_i, O_{ar}] = 1$$

The reverse relation expresses that the object O_{ar} consists of the region R_r, i.e. the function identifies the object that are the components of O_{ar}:

$$\begin{aligned} COMP(O_{ar}) &= R_r = \{...., O_i, ...\} \\ &= \{ O_i \mid Part_{kl}[O_i, O_{ar}] = 1\} \end{aligned}$$

The geometry of the aggregates can be found through the geometry of the original objects, for each geometric element we can check whether it will be part of an aggregated object of type T_a. This should be done in two steps, which will be explained for the faces of an area object O_i in relation to an aggregated area object O_{ar}. The first step evaluates the function:

$$Part_{22}[f_j, O_{ar} \mid O_i] = MIN(Part_{22}[f_j, O_i], Part_{22}[O_i, O_{ar}])$$

this function expresses whether the face is related to an aggregate through object O_i. If that is true then both functions in the expression at the right hand side of the equation will have the value = 1, and this value is assigned to the function at the left hand of the equation. If it is not the case then at least one of the functions at the left hand side will have the value = 0, so that also the function at the left hand side will get the value = 0. The second step is the evaluation of

$$Part_{22}[f_j, O_{ar}] = MAX_{oi}(Part_{22}[f_j, O_{ar} \mid O_i])$$

If there is any object through which the face will be part of an aggregate then this function will have the value = 1, otherwise it will be = 0. If this function has been evaluated for all faces of the map then the geometry of the object O_{ar} can be found through their adjacency graph. For the edges e_i of these faces the function $B[e_i, O_a]$ can be evaluated and with this function the boundary edges can be found (i.e. $B[e,O] = 1$) and through these the topologic relationships with the other objects.

The geometry of the aggregated area object O_a can sometimes be simplified by a reduction of the number of faces. Therefor the edges e_i should be identified for which $B[e_i, O_a] = 2$, that are the interior edges. If these edges are not part of some line object so that $LO(e_i) = \emptyset$ then they do not carry any semantic information at this aggregation level and could therefor be eliminated.

The example refers to the situation where a face is related through an area object to an aggregated are object, so that all involved elements are of dimension 2. Other combinations of dimensions might occur as well, this could be the case when for example an edge is related through a line object to an aggregated area object, e.g. it is related through a river to a country.

It is possible to define aggregation types by means of their construction rules. If elementary objects are combined to form a compound object, their attribute values are often aggregated as well (as in figure 6). We will see in section 3.3 that farm yield is the sum of the yields per field, and the yield per district is the sum of farm yields. The desaggregation of such values is usually quite difficult because it can only be done if information is added to the system. An aggregation hierarchy has therefore a bottom-up character, in the sense that the elementary objects from the lowest level are combined to compose increasingly complex objects as one ascends in the hierarchy. The compound objects inherit the attribute values from the objects by which they are composed.

The PARTOF relations connect groups of objects with a certain aggregate and possibly on a higher level with another even more complex aggregate, and so on. That means that an aggregation hierarchy expresses the relationship between a specific aggregated object and its constituent parts at different levels. This is different from class hierarchies where classes at several generalization levels can be defined with their attribute structured and their intentions, but where the objects can be assigned to these classes in a later stage of a mapping process.

3. Strategies For Object Generalization

The formalism of the previous chapter helps us to express the structure of spatial datasets. This can be done in an abstracted sense, i.e. without any reference to the logic model of any implemented spatial data base. Processes applied to such datasets could also be expressed through this formalism. The four basic operations that will be used in generalization processes are:

- the *selection* of objects to be represented at the reduced scale, this selection will be based on the attribute data of the objects,
- the *elimination* from the data base of objects that should not be represented,
- the *aggregation* of area objects that should not be represented individually,
- the *reclassification* of the generalized objects.

For these four operations information about the spatial structure of the mapped area will be required. Firstly to check which relationships the objects have with their environment and what the effect of their eventual elimination will be on the spatial structure of that environment. Secondly this information is required to formulate aggregation rules for the objects that are to be merged. Once the process has been formulated one can choose how to implement it in any suitable database environment. The hydrologic example presented in sections 2.1 and 3.4 of this paper has been implemented in an Arc\Info environment, but other students of the author have made implementations of similar applications in an Oracle database, and exercises with Prolog have been made as well.

Several strategies for database generalization can be formulated with this formalism and these basic database operations .These are:

- **geometry driven generalization**: in this strategy it is the geometric information that drives the aggregation process. A clear example of this case is when the geometry of the spatial data has a raster structure. If it is then decided that the resolution of the raster will be decreased, i.e. when the cell size increases, then the original, smaller cells are merged into new larger cells. The thematic information carried by the original cells should then be transferred to the new cell.
- **class driven generalization**: in this strategy regions are identified, consisting of mutually adjacent objects belonging to the same class. These objects will then be aggregated to form larger spatial units with uniform thematic characteristics. The generalization is then driven by the thematic information of the spatial data.

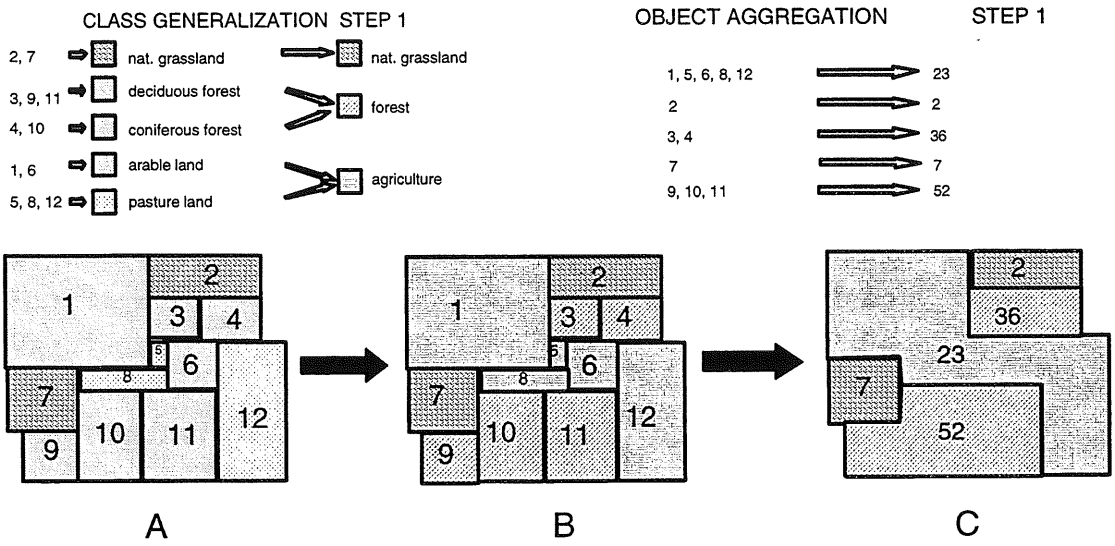


fig. 4: Class driven object aggregation.

- **functional generalization**: spatial objects at a low aggregation level are aggregated to form new objects at a higher level. The objects are functional units with respect to some process defined at their aggregation level, the processes at the different aggregation levels are related.
- **structural generalization**: the main aim of the process is to simplify the description of a spatial system, such as drainage networks, while keeping the overall structure intact. This to the fact that after generalization the total functioning of the same system can be understood at a less detailed level.

Each strategy has its own range of applications. Data base users should be well aware of why they are generalizing spatial data, so that they can chose which strategy is to be used. The first strategy is in most cases used when the geometric resolution of a spatial description is reduced without a clear semantic motivation. The latter three strategies, however, are semantically defined and motivated. They will be explained in some more detail now.

3.1. Class Driven Object Generalization

Suppose that a database contains the situation A of figure 4, this is a detailed description of a terrain situation with agricultural fields, forrest areas and natural grasslands. This description might be too detailed for a structural analysis which should give information about the areas covered by the different major types of land use and their spatial distribution. A less detailed spatial description can then be obtained, if the original objects area aggregated to form larger spatial regions per major land use class.

Figures 4 and 5 show that this less detailed description can be obtained in two steps:

- first the objects are assigned to more general classes representing the major land use types this results in situation B of figure 4,
- then mutually adjacent objects are combined per class to form regions, this results in situation C of figure 4.

These final regions can be considered as aggregated objects. The functions `D(O,T)` express then that objects should be aggregated per (super)class, i.e. if aggregated objects should be formed for agriculture then

```
if O ∈ Agriculture D(O,Agriculture) = 1
     else          D(O,Agriculture) = 0.
```

The output of the aggregation process are regions in the sense of section 2.3. Each region is an aggregate of objects that belong to one land use class, so if R_a is an agricultural region then:

- for all objects $O_i \in R_a$ is $D(O_i, \text{Agriculture}) = 1$
- if $O_i \in R_a$ and $\text{ADJACENT}[O_j, O_i] = 1$ and $D(O_j, \text{Agriculture}) = 1$ then $O_j \in R_a$

A consequence of this rule is that after the aggregation process there can be no two adjacent regions that are of the same type, i.e. that represent the same land use class.

3.2 Thematic and Geometric Resolution

The example represented in the figures 4 and 5 shows a situation where the thematic aspects of the newly aggregated objects can still be handled within the original class hierarchy. It might be that the same classes can be used as for the original objects, but the example shows a situation where it is quite clear that with each database generalization step the class hierarchy is adjusted; per step the occurring lowest level of classes is removed, only the more general classes remain, see also figures 5 and 12. That means that the thematic resolution is adjusted to the geometric resolution of the terrain description.

There might be situations where it is not necessary to jump to more general classes with each aggregation step. In those cases the new objects can be assigned to the original classes with consequence that they have the same attributes as the objects from which they have been composed. This is in fact the case if we consider the step from B to C in figure 4 in isolation. There the objects 1,5,6,8 and 12 all belong to the class "agriculture". Therefore they have the same attribute structure. They are distinct because they had different attribute values. Within this class they are aggregated to form the composite object 23, i.e.

$$O_{23} = \text{AGGR}(O_1, O_5, O_6, O_8, O_{12}).$$

This new object still belongs to the same class "agriculture" and has attributes in common with original objects. The attribute values of the original objects will then be transferred to the new object as in figure 6.

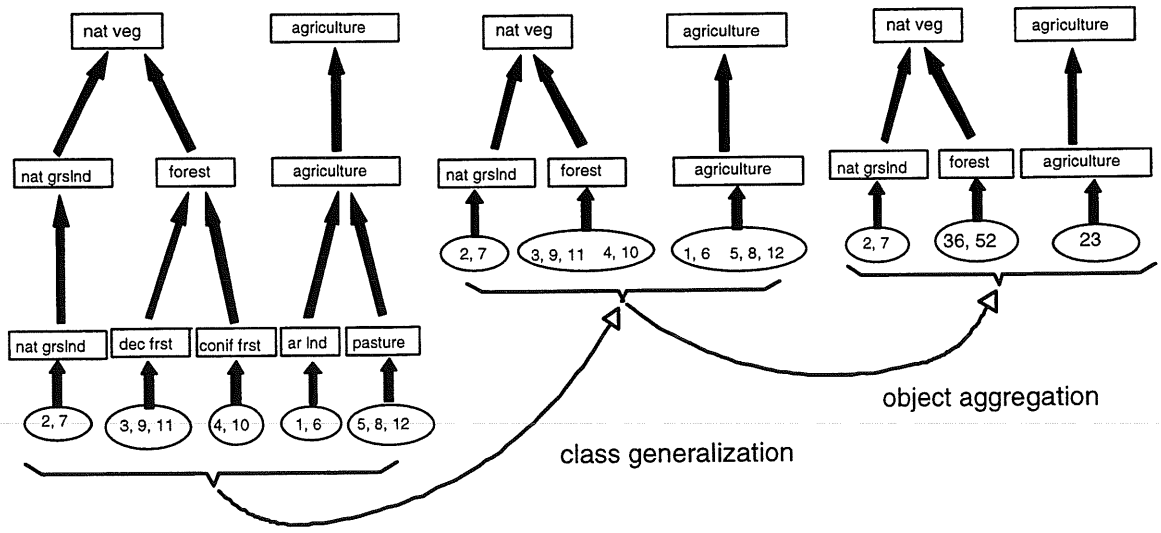


fig. 5: A diagram representing the generalization and aggregation steps of the object generalization process of figure 4.

When the attribute values are of the ratio scale type then the aggregated value can often be obtained by summation or by taken the average value over the objects that compose the new object, e.g.:

$$A[O_a] = \Sigma_{oi \in COMP(oi)} A[O_i]$$

Examples are attributes like wood volume and crop yield and population. For other attributes like vegetation cover or population density it might be that (weighted) averages should be computed.

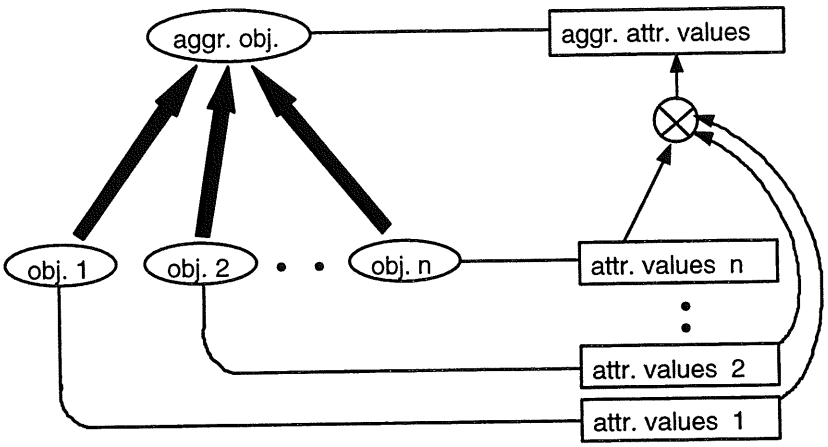


fig. 6: The aggregation of attribute values.

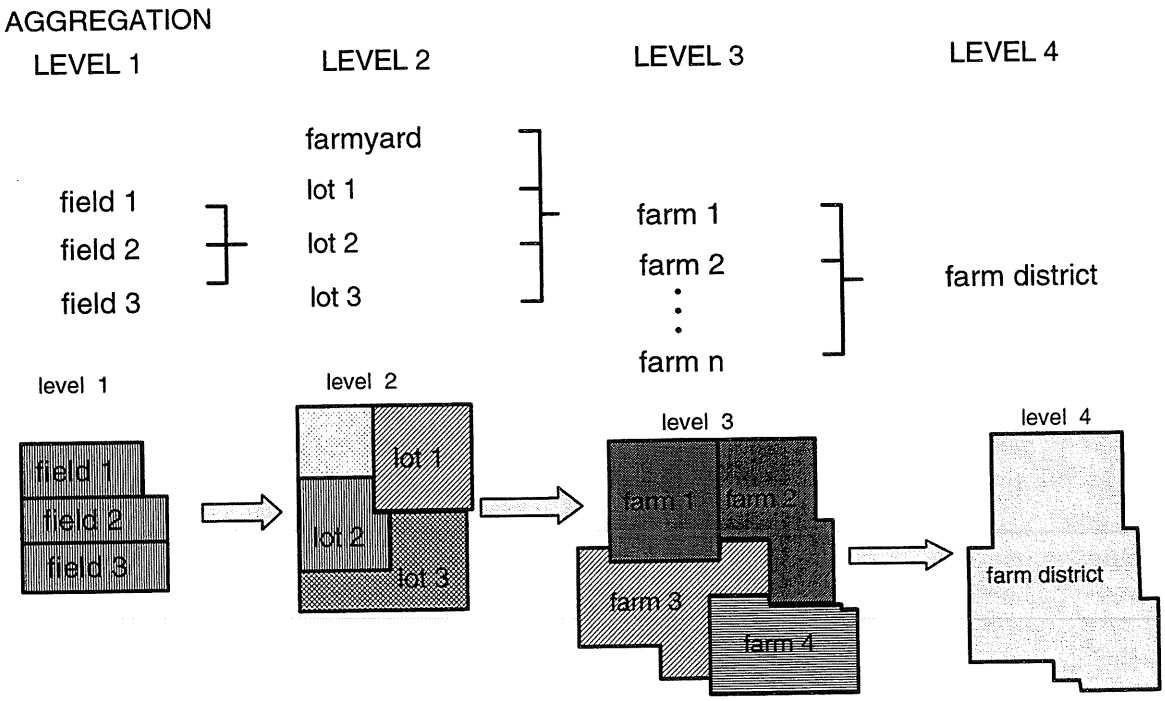


fig. 7: An example of a functional object generalization process.

3.3. Functional Object Generalization

That will always have the effect that the spatial variability of the attribute values will be reduced, because after each aggregation step the attribute values that were assigned per object will then be merged into one value for a larger object. That means that the relationship between spatial and thematic resolution is not only expressed through the link between class level and aggregation level, it also expressed by the spatial variability of the attribute values.

It is certainly not always so that object aggregation can be done within the framework on one class hierarchy. In many cases object aggregation will imply a completely different thematic description of the objects, so that new classes should be defined. This is illustrated in figure 7 where farm yards and fields have been aggregated into farms and these in their turn into farm districts. The aggregation hierarchy has a bottom up character in the sense that starting from the elementary objects composite objects of increasing complexity are constructed in an upward direction (in figure 7 from left to right). The farm districts should only consist of farms and the farms should be mutually adjacent so that the adjacency graph (see section 4) of the farms that belong to one district is connected.

The aggregation steps in figure 7 show how the fields are considered as elementary objects at level 1. They are defined per growing season as spatial units under one crop. For the farmer they are management units, because his management operations are planned and performed per field. They are aggregated to lots which are elementary objects at level 2, i.e. these objects belong to the extensions of classes such as "arable-lot" and "grass land". These are management units at a higher level; the farmer will maintain a drainage system per lot and he will decide per growing season how to partition each lot into fields. These lots might both belong to a superclass "farm lots" in a land use data base and these again might belong to an even higher superclass "lot" which also contains the classes "forest lot" and "residence lot". The aggregation step from level 1 to 2 and the next steps to the levels 3 and 4 where we have the farms and farm districts show that after each step new objects are created. At farm level the farmer will decide whether he will be a cattle farmer or whether he will grow arable crops, in the latter case he has to decide on a rotation scheme. At district level the infrastructure and irrigation schemes will be developed. The objects at each level have their own thematic description expressed in an attribute structure that should be defined in a class hierarchy according to section 2. In this example each aggregation level requires its own classification hierarchy. This should be structured so that the generated attribute structures provide the information to support the management operations defined at the aggregation levels of the objects. The diagram of figure 8 represents the fact that a classification hierarchy should be defined per aggregation level.

This is an example of a more general situation where objects at each aggregation level are functional units with respect to some process. In this case these were farm management processes, but we could also take examples like ecologic development, or demography and many more. Each aggregation level within such a hierarchy will have its own (sub) context within a thematic field, expressed through a class hierarchy with related attribute structures. The different (sub) contexts are related by the fact that sets of objects at one level can be aggregated to form the objects at the next higher level. There are often also relationships between various classes of the different class hierarchies related to the aggregation levels as was the case for the cover classes for the farm lots, the farm types and land use types of the farm districts at the levels 2,3 and 4 of figure 7.

There are bottom-up relationships between the objects at different levels in the sense that the state information of the lowest level objects, as contained in the attribute data, can be transferred through a process like figure 6, to give state information about the objects at higher levels. There are top-down relationships in the sense that the behaviour of lower level objects will be constrained by the information contained in the higher level objects.

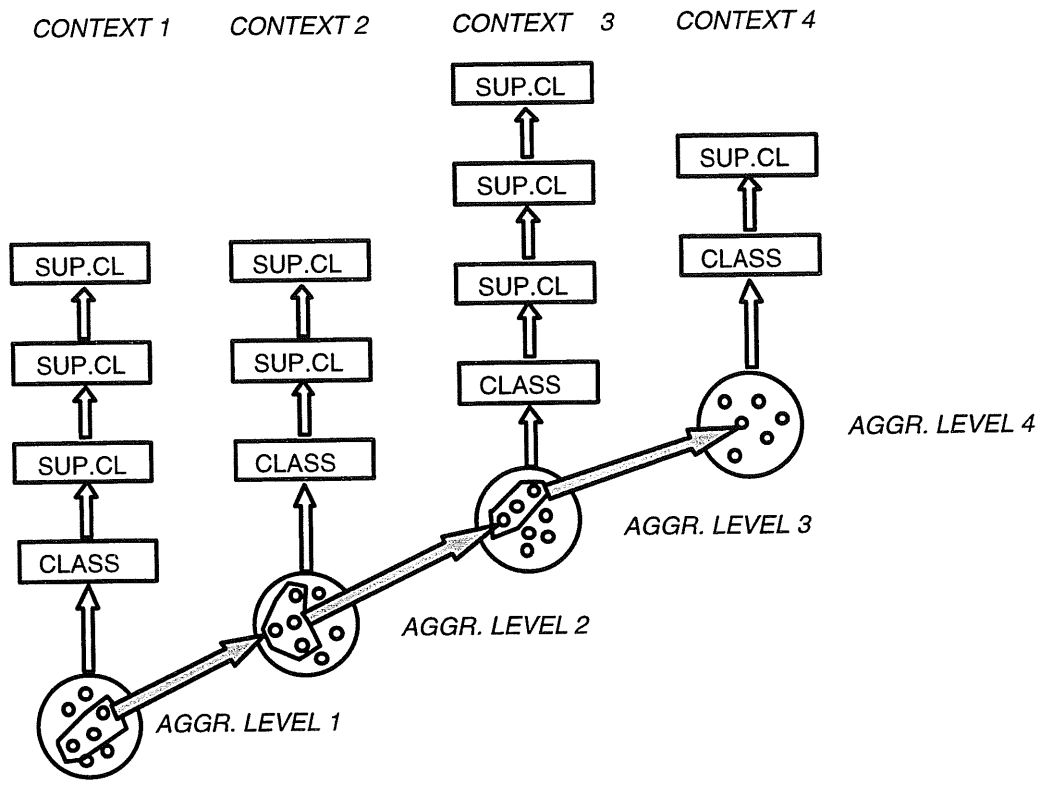


fig. 8: Classification hierarchies related to aggregation levels.

3.4. Structural Object Generalization

This strategy will be explained by means of an example based on a database where the spatial description at a 1:50.000 scale, of a drainage system. The database has been structured according the FDS as in section 3, see (Martinez Casasnovas 1994). A generalization process will be executed to derive the 1:100.000 representation, so that we reduce complexity to stress spatial structure. Here the spatial structure refers to the network structure of the drainage system in relation to the subcatchments. The generalization process will keep the area of the aggregated subcatchments and the network structure of the system invariant so that the computation of overland water flows per node in the network will not be effected significantly.

The database contains geometric data and thematic data of the elements of the system, as defined in the example of section 2.1. Let the attribute `ORDER` contain the Strahler order number of each drainage element. These numbers in combination with the function Upstr[ ] make it possible to analyze the stream network built by the drainage elements. Through this network aggregation steps can be defined for the catchment areas. The methodology for these aggregation steps will follow to a great extend procedures defined by (Richardson 1993 and 1994).

The process starts with the identification of the drainage elements that are not mappable at the target scale, those are the elements with Strahler number = 1 with an average width `aw < Thr(eshold)`. The minimum mapping width of the drainage elements will be put at 0.75mm, that gives a threshold `Thr = 0.75mm/scale` at terrain scale, hence `Thr = 75m` in this case. The average

width for an drainage element D_i can be computed from the $AREA_i$ of the element and the $LENGTH_i$ of its water line W_i hence

$$aw_i = AREA_i / LENGTH_i$$

The selection procedure applied to the drainage elements is then

> select the drainage elements D_h with $ORDER_h > 1$
> select from the class with $D_h = 1$ the elements D_i with $aw_i \geq Thr$

The set of elements that should be eliminated is then

$$S = \{D_i \mid ORDER_i = 1,\ aw_i < Thr\},$$

their catchments should be combined with adjacent catchments to form aggregates. The elimination of the drainage elements $D_i \in S$ should consist of the following steps

> eliminate W_i
> $AGGR(D_i, C_i) = C_{DHi}$
> find C_h for which $Upstr[C_h, C_i] = 1$
> $AGGR(C_{DHi}, C_h) = C_{h;i}$

where the notation C_{DHi} means that the area of Di has been merged into the area of its subcatchment, the notation $C_{h;i}$ means that the area of C_{DHi} has been merged into the area of C_i. When water line W_i joins the outlet point $END(W_i)$ with only one water line W_j of a drainage element that will not be eliminated then the catchment of W_j should be merged with $C_{h;i}$

> find C_j for which $Upstr[C_h, C_j] = 1$
if $D_j \notin S$ and there is no $D_{k \neq j} \notin S$ with $Upstr[C_h, C_k] = 1$
then
> $AGGR(D_j, D_h) = D_{jh}$
> $AGGR(C_j, C_{h;i}) = C_{jh;i}$

The notation D_{jh} means that D_j and D_h have been merged and $C_{jh;i}$ means that C_j and $C_{h;i}$ have been merged. When these steps have been done for each element $D_i \in S$, then new Strahler numbers can be assigned to the remaining elements according to their new position in the network. Then the selection procedure can be repeated and so on until no more elements are eliminated.

A Test Case

The drainage system represented in fig. 9a will be used as an example to demonstrate the generalization process. This figure is a schematic representation of the Romani Drainage system in the Anoia-Penedes Area in NE-Spain (Martinez Casasnovas 1994). The total area of this drainage system is 28.53 km^2, at a 1:50.000 scale it has 37 mappable drainage elements. The figure only shows the water lines W_i with their catchments, the drainage element D_i are not shown here.

The transition from the original scale to the scale 1:100.000 will be done through the generalization procedure explained before. So first the drainage elements with Strahler number 1 and width < 75m are eliminated, their catchments are aggregated with their downstream catchments. Then the remaining drainage elements are reclassified and the procedure is repeated until no more elements are eliminated.

The result of this procedure is shown in fig. 9b. The drainage system represented at 1:100.000 scale has only nine mappable drainage elements. Their catchments are aggregates of the catchments shown at the 1:50.000 scale. The fact that they are considered to be aggregated catchments implies that the information carried by the original catchments is now transferred to these aggregates.

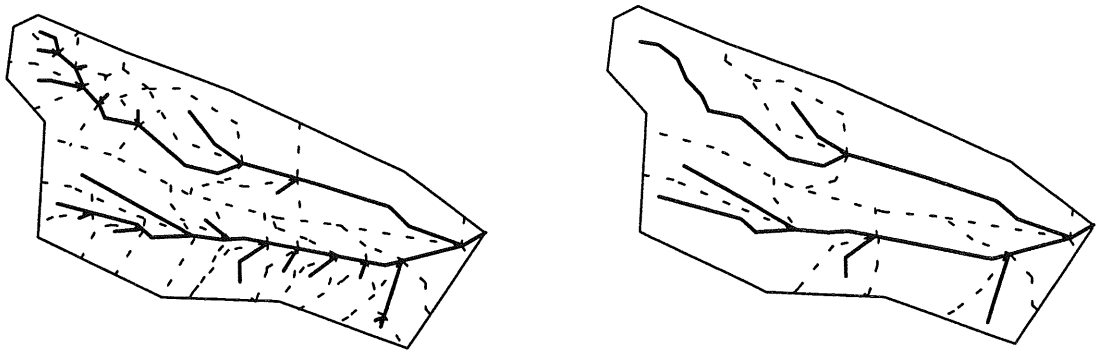


fig. 9: a) The drainage system at 1:50.000 scale representation
b) The drainage sustem at 1:100.000 scale representation

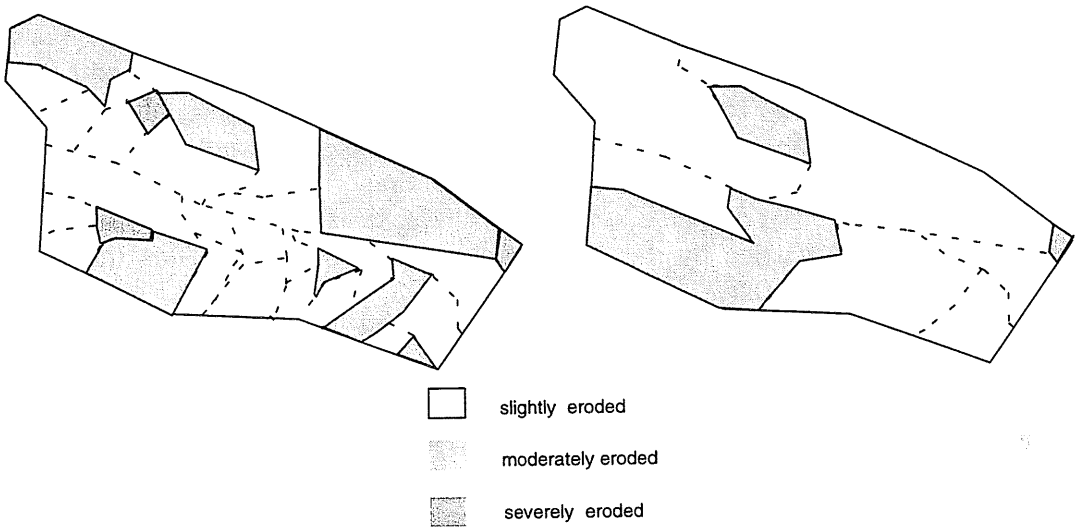


fig. 10: a) Erosion classes estimated in 1:50.000 scale representation
b) Erosion classes estimated in 1:100.000 scale representation

The Romani drainage system has been mapped for an erosion survey. Erosion classes are estimated per catchment from the information contained in the attributes of the drainage elements. These are used to compute per catchment the drainage density in km/km^2 and the crenellation ratio in km/km^2 These data combined with the depth and the activity class of the drainage elements determine the erosion class of each catchment. When the area of the catchments are summed per erosion class we find in the original situation at 1:50.000 scale that 68% of the area is slightly eroded, 30% is moderately eroded and about 2% shows severe erosion, see figure 10a.

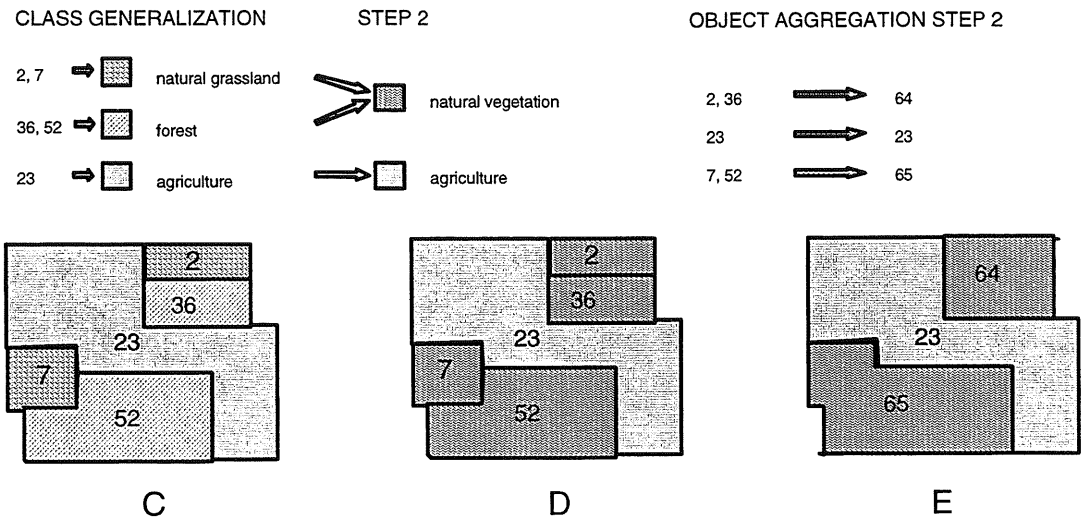


fig. 11: The second aggregation step for the objects of figure 4c.

After generalization aggregated catchments are formed for which the erosion classes have to be estimated. Although a large number of drainage elements have not been represented any more at the reduced scale, the information they carry has been transferred to the aggregated catchments. With these data we find now for the erosion classes that 79,5% of the area is slightly eroded, 20% is moderately eroded and 0.5% is severely eroded, see fig. 10b. These numbers deviate significantly from the original values, furthermore the spatial distribution of the occurrence of the erosion classes is quite different from the original distribution. The case is even worse if we had completely ignored the information carried by the eliminated drainage elements. Then the values would be respectively 99.5%, 0% and 0.5%.

The structural generalization of the drainage system kept considered its constituting entities as hydrologic units. This had the effect that the computation of hydrologic processes is invariant after the generalization. The generalized network could, however, not be used to formulate reliable statements about the erosion classes of the areas in the system. That would require another generalization process where we have to specify what statements about erosion should be invariant after transformation. A class driven or a geometry driven strategy might have been more useful in this case.

4. Object generalization and levels of spatial complexity

Chapter 3 discussed several strategies for the generalization of spatial databases. These strategies were based on the concept of spatial object aggregation in combination with class hierarchies. In the process of object aggregation the information of lower level objects is aggregated to higher level objects, but in principle the original detailed information is maintained so that it is possible to access the detailed information of the lower level objects through the aggregated objects. The result of such an aggregation process is a less detailed terrain description that may be compared to the result of a map generalization.

The output of this process could be used as the input for a following aggregation step. This has been illustrated in fig. 11, that shows a process starting from the situation of fig. 4c. The regions of situation c are assigned to more general classes in situation d and the aggregated to form the larger regions of situation e.

The class generalization and object aggregation steps of the approach of figs. 4 and 11 have been represented in a different way in fig. 13. This figure combines per database generalization step two steps like those of fig. 4. In the first step of fig. 13 the objects of the different classes are first assigned to the super classes at the next higher level in the hierarchy (compare the class generalization step of fig 4), then in the same step the objects that form a region per super class are aggregated to form a larger object (compare the object aggregation step of fig 4). This procedure is repeated in the second step of fig. 13.

This figure shows that the two steps of the example of section 3.1 reduce the number of objects, that is why we rather talk of database generalization because the process generated objects with a lower spatial and thematic resolution then the original objects. Due to the fact that the original objects formed a geometric partition of the mapped area and due to the fact that generalization process made use of the topologic and hierarchical structures in which the objects had been modelled, this process resulted in a new set of objects that also formed a geometric partition of the mapped space. But the result was a terrain description of a reduced spatial complexity as is shown in the stepwise reduction of the complexity of the adjacency graphs of fig. 12.

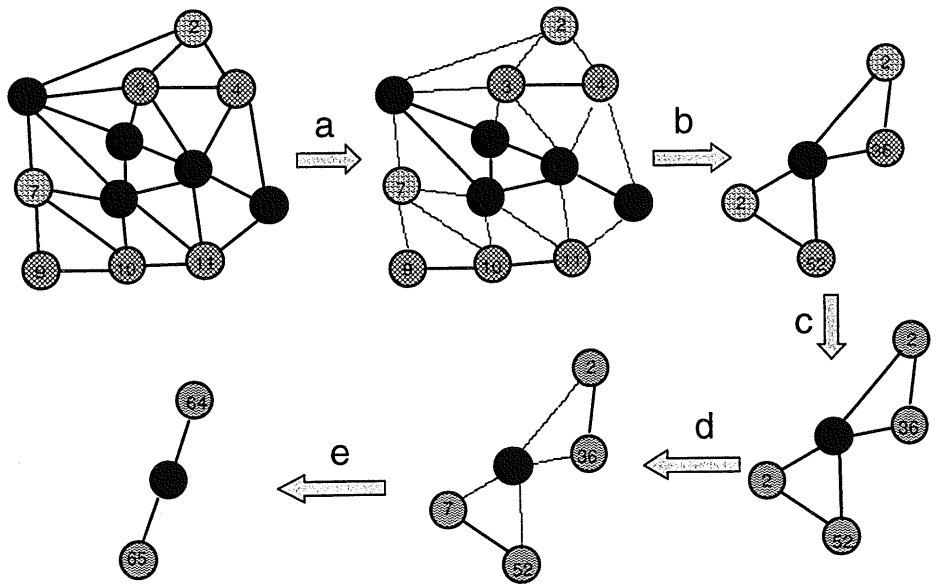


fig. 12: The adjacency graphs related to different stages of the generalization processes of figs. 4 and 11.

Each object is represented by a node in these graphs and the adjacency between two objects is represented by an arc. This figure gives the adjacency graphs related to each stage of a process that starts from the situation B of figure 3 where the original objects have been assigned to their super classes. If we follow the steps of fig. 12 then we see that:

- in step a the regions per class have been identified,
- in step b the objects in each region are aggregated to form a composite object that is represented by one node,

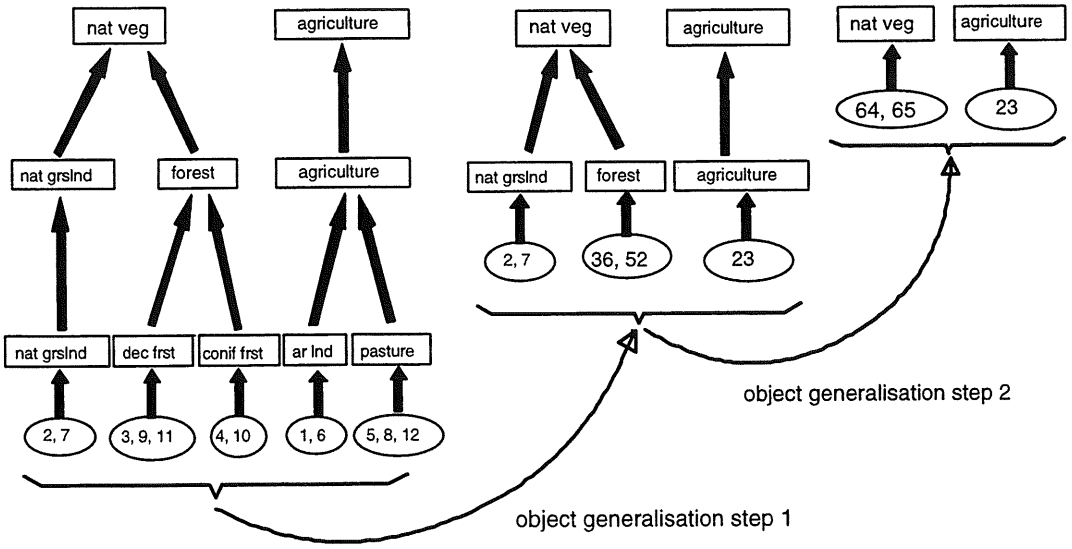


fig. 1: A diagram representing the object generalization steps of figs. 4 and 11.

- these regions are after step c assigned to more general classes,
- in step d regions at this higher class level have been identified, these are composed of the objects obtained after step b,
- then finally after step e each of these regions have been aggregated again to form the objects at the higher aggregation level which is then represented by one node, this is the adjacency graph of situation e of figure 11.

The reduction of spatial complexity is one of the important aspects of generalization processes as they are known in mapping disciplines. This process has traditionally been applied in the form of map generalization to reduce the information content of a map so that a mapped area could be represented at a smaller map-scale. This process has two steps, the conceptual generalization and the graphic generalization. The conceptual generalization results in a redefinition of the mapped spatial features or objects to reduce their number for the terrain description at the smaller scale. The graphical generalization is in fact a simplification of the graphical representation of these features or objects, including such aspects as geometric simplification, object displacement, resymbolization etc.

5. Conclusion

When we deal with spatial database generalization in a GIS environment then this might include the graphical representation as well, but that is not necessarily so. The main aim will be a simplified terrain description, i.e. a lower spatial complexity to emphasize spatial patterns and relationships that might be difficult to find in a more detailed terrain description. That means that this process is very much related to the conceptual generalization step mentioned before, the main aim of this step is to obtain a data reduction. We have seen that it can to a large extend be defined in the form of database operations for databases that are implementations of the formal data structure (FDS) as explained in chapter 2. Such databases will be called shortly FDS-databases.

A spatial database may contain information about different aspects of a particular area, as we saw in the example of section 3.4. A generalization process may keep one aspect invariant, let us call that the primary interpretation of the database. The other aspects may be affected so that the information is not reliable after the process, we will call these the secondary interpretations of the

database. If we consider generalization operations as a type of transformation of a spatial data base, then we should make explicit decisions about which aspects of the original data bases are to remain invariant, so we should decide what is to be considered as the primary interpretation of the data base. This choice will be made within some users context of the data base, i.e. the user will be interested in the correct representation of some spatial characteristics, while others may be deformed by the transformation.

A good understanding of database generalization may be useful for the design of procedures for spatial data acquisition. Information extraction from images is partly a reverse process to generalization. Generalization is a process with a stepwise data reduction, going from high resolution to low resolution. The information of the high resolution objects is merged into low resolution objects. Image interpretation can often be formulated as a process where data are produced stepwise. We can learn from generalization processes what information low resolution objects carry about their constituting high resolution objects. This knowledge may help us in image interpretation, where large image segments can be seen as low resolution objects. These should then contain thematic information in addition to the radiometric and spectral information of the image itself, to identify smaller segments that may represent high resolution objects.

6. References

Brodie ML 1984 On the Development of Data Models. In Brodie ML, Mylopoulus, Schmidt (Eds) On Conceptual Modelling.New York: Springer Verlag

Brodie ML, Ridjanovis D 1984 On the design and specification of database transactions. In: Brodie, Mylopoulos, Schmidt (Eds) On Conceptual Modelling. New York: Springer-Verlag

Egenhofer MJ, Frank AU 1989 Object-Oriented Modelling. In GIS: Inheritance and propagation. Auto-Carto 9, p588

Frank AU, Kuhn, W 1986 Cell Graphs: A Provable Correct Method for the Storage of Geometry. Proceedings of the 2nd International Symposium on Spatial Data Handling, Seattle.

Gersting JL 1992 Mathematical structures for computer science - third edition. New York: Computer Science Press.

Hesse W, Leahy FJ 1992 Authorative Topographic-Cartographic Information System ATKIS. Bonn: Landes Vermessungamt Nordrhein-Wetsfalen.

Marx RW 1990 The TIGER system: automating the geographic structure of the United States census. In: Peuquet DJ, Marble DF (Eds) Introductory readings in GIS. London: Taylor and Francis, pp120-141.

Martinez Casanovas JA 1994 Hydrographic information abstraction for erosion modelling at regional level. MSc. thesis, Dept of Landsurveying and Remote Sensing, Wageningen Agricultural University, Wageningen.

Molenaar M 1989 Single valued vector maps - a concept in GIS. Geo-Informationssysteme 2:18-26

Molenaar M 1993 Object Hierarchies and Uncertainty in GIS or Why is Standardization so Difficult. Geo-Informations-Systeme 6

Molenaar M 1994 A syntax for the representation of fuzzy spatial objects. In: Molenaar M, De Hoop S (Eds) Advanced geographic data modelling. Netherlands Geodetic Commission, New Series, Nr.40, Delft, pp155-169

Oxborrow E, Kemp Z 1989 An object-oriented approach to the management of geographical data. Conference on Managing Geographical Data and Databases, Lancaster

Richardson DE 1993 Automatic Spatial and Thematic Generalization Using a Context Transformation Model. (Doctoral Dissertation, Wageningen Agricultural University) R&B Publications, Ottawa, Canada

Richardson DE 1994 Contextual Transformations and Generalizations of Remotely Sensed Imagery for Map Generation. In: Molenaar M, De Hoop S (Eds) Advanced geographic data modelling, Netherlands Geodetic Commission, New Series, Nr. 40, Delft, pp170-178

U.S.BUREAU OF THE CENSUS 1990 Technical description of the DIME System. In: Peuquet DJ, Marble DF (Eds) Introductory readings in GIS, London: Taylor and Francis, pp100-111

3.2 Modern GIS and model linking

Michael F. Goodchild

Department of Geography, and National Center for Geographic Information and Analysis
University of California Santa Barbara, CA 93106-4060, USA
Email: good@ncgia.ucsb.edu

The traditional approach to analysis sees it as part of a larger scheme that begins with problem formulation and ends with interpretation of results. Techniques of spatial analysis form a well-defined subset of the larger set of analytic methods, defined by an invariance property. For many reasons this view of spatial analysis, and the larger field of analysis in general, is undergoing profound change, brought on in part by the advent of integrated computing environments such as geographic information systems (GIS). The paper reviews the trends contributing to this change, and its possible effects on the role of spatial analysis, and the broader context of GIS, in the future.

1. Introduction

GIS and spatial analysis have enjoyed a long and productive relationship over the past decades (for reviews see Fotheringham and Rogerson 1994; Goodchild 1988; Goodchild et al. 1992). GIS has been seen as the key to implementing methods of spatial analysis, making them more accessible to a broader range of users, and hopefully more widely used in making effective decisions and in supporting scientific research. It has been argued (e.g. Goodchild 1988) that in this sense the relationship between spatial analysis and GIS is analogous to that between statistics and the statistical packages. Much has been written about the need to extend the range of spatial analytic functions available in GIS, and about the competition for the attention of GIS developers between spatial analysis and other GIS uses, many of which are more powerful and better able to command funding. Specialized GIS packages directed specifically at spatial analysis have emerged (e.g. IDRISI, and see Bailey and Gatrell 1995). Finally, implementation of spatial analysis methods in GIS is leading to a new, exploratory emphasis.

The purpose of this paper is to explore new directions that have emerged recently, or are currently emerging, in the general area of GIS and spatial analysis, with particular emphasis on the practical issues that arise in making use of today's capabilities for spatial analysis in GIS. In the next section, it is argued that in the past GIS and spatial analysis have followed a very clearly and narrowly defined path, one that has more to do with the world of spatial analysis prior to the advent of GIS than with making the most of both fields—the path is, in other words, a legacy of prior conditions and an earlier era. The following section identifies a number of trends, some related to GIS but some much more broadly-based, that have changed the context of GIS and spatial analysis over the past few years, and continue to do so at an increasing rate. The third section identifies some of the consequences of these trends, and the problems that are arising in the development of a new approach to spatial analysis. The paper concludes with some comments about the complexity of the interactions between analysis, data and tools, and speculation on what the future may hold, and what forms of spatial analysis it is likely to favor. A further elaboration of these ideas will appear in a forthcoming chapter written jointly with Paul Longley (Goodchild and Longley 1997).

2. Traditions in Spatial Analysis

2.1 The Linear Project Design

In the best of all possible worlds, a research project (the term 'research' will be interpreted very broadly to include both scientific and decision-making activities) begins with a clearly stated problem. Some decision must be made, some question of scientific theory resolved by resorting to experiment or real-world evidence. An experimental design is developed to resolve the problem, data are collected, analyses are performed, and the results are interpreted and reported. This simple structure has underlain generations of student dissertations, government reports, and research papers. The sequence is strictly linear, implying that the availability of data has no influence on problem definition; availability or awareness of methods of analysis no influence on collection of data; etc. Indeed, the terms 'data-driven' and 'technique-driven' are highly perjorative in research generally, as are such phrases as 'a technique in search of a problem'—in this ideal world, the statement of the problem strictly precedes the collection of data and the performance of analysis.

In this simple, sequential world the selection of methods of analysis can be reduced to a few simple rules (in the context of statistical analysis, see for example Levine 1981: Ch 17; Marascuilo and Levin 1983: inside cover; Siegel 1956: inside cover). Choice of analytic method depends on the type of decision to be made (e.g. whether two samples are drawn from the same, unknown population, or whether two variables are correlated), and on the characteristics of the available data (e.g. scale of measurement—nominal, ordinal, interval or ratio; but see Chrisman 1997 for a discussion of this simple four-way classification in the context of GIS).

2.2 Spatial analysis

Spatial analysis, or spatial data analysis, is a well-defined subset of the methods of analysis available to a project. One might define spatial analysis as a set of methods useful when the data are spatial, in other words when the data are referenced to a two-dimensional frame. More narrowly, the Earth's surface provides a particular instance of such a frame, the geographic frame, with its peculiar properties of curvature. This definition of spatial analysis is arguably too broad, because in basing the definition on the properties of data it does not address the question of whether the two-dimensional frame actually matters—could the same results have been obtained if the frame were distorted in some way, or if objects were repositioned in the frame? More precisely, then, spatial analysis can be defined as that subset of analytic techniques whose results depend on the frame, or will change if the frame changes, or if objects are repositioned within it. To distinguish analytic methods from more mundane operations they might be defined as methods for processing data with the objective of solving some scientific or decision-making problem.

Methods of spatial analysis have accumulated in a literature that spans many decades, indeed centuries. They have been invented in many disciplines, including mathematics, and particularly geometry; statistics, and particularly spatial statistics and statistical geometry; and in geography and other Earth sciences. Compendia have been published (among others, see Bailey and Gatrell 1995; Berry and Marble 1968; Haining 1990; Taylor 1977; Unwin 1981), and various approaches proposed for structuring this body of technique. Many of the earlier methods could be described as confirmatory, mirroring the hypothesis-testing tradition of statistics in seeking to confirm or deny some formally stated hypothesis through the analysis of empirical data. Others are better described as exploratory, subjecting data to manipulations selected for their ability to expose patterns and anomalies that might not otherwise be evident to the analyst, or manipulating the data in ways designed to enhance the investigator's intuition.

2.3 The well-informed analyst

Traditionally, the responsibilities of the inventor of a technique ended when the technique had been tested and described. Even the testing of a technique can be suspect in an academic world that often values theory over empiricism, and is suspicious of empirical results that cannot be demonstrated to be generally true. The advent of the digital computer changed this world fundamentally, because it became possible for a scientist to perform a method of analysis automatically, without taking personal responsibility for every aspect of the performance. It was now possible using the 'black box' of the computer to perform an analysis that one did not know everything about—that one could not perform by hand. Methods emerged, beginning in the 1970s and particularly in the area of multivariate statistics, that would be impossibly impractical to perform by hand. Pedagogically, a fundamental shift became possible in how analysis was taught—that one might learn about a technique by studying the nature of its response to particular inputs, rather than by studying how the response was generated. But there is a fundamental difference between these two positions: between whether one understands the results of a principal components analysis, for example, as the extraction of eigenvalues from a specific matrix, or the generation of statistics that broadly indicate some concept of 'relative importance'.

Exactly where this change occurred is open to debate, of course. It may have occurred when students were no longer required to perform statistical analyses by hand before being let loose on computer packages; or when Fortran appeared, making it necessary to understand less about how instructions were actually carried out; or when the growth of the scientific enterprise had reached such a level that potential replication of every result was a practical impossibility.

In the early days of statistical analysis all calculations had to be carried out by hand. Although the intensity of the necessary calculations must clearly have had some influence on the choice of method, in principle the paradigm had no way of including this factor as a criterion that could affect the choice of method. In this somewhat monastic world the cost of the scientist's labor was simply not a factor in his or her science. Highly routine tasks could be assigned to an apparently inexhaustible supply of unpaid or very cheap student labour. Thus the intense numerical calculations needed in the early days of factor analysis seem to have had surprisingly little negative impact on the development or adoption of the technique (Harman 1976).

Of course the digital computers that were introduced to the scientific community beginning in the late 1950s produced rapid change in the labour demands of many statistical methods. The intricate calculations of factor analysis could be performed by a fully automatic machine, provided the researcher could command sufficient computer time, and provided labour was available to punch the necessary cards. Computers and the brains of young children are similar in many ways; both begin essentially empty; both must acquire the primitive elements of reasoning; but having done so, both can build enormously complex structures out of simpler ones, apparently ad infinitum. What began in the 1960s as a set of uncoordinated efforts by individual scientists writing their own programs had developed by the 1990s into a complex of enormously sophisticated tools, each integrating a large number of methods into an easy-to-use whole.

2.4 Extending the functions of analytic software

Although they show clear evidence of their roots, the packages used by the scientists of the 1990s are different in fundamental respects from the programs of the 1960s. Besides implementing large numbers of statistical methods, today's packages also provide support for the creation and maintenance of data. There will be tools for documenting data sets, and describing their properties, such as accuracy and history. Other tools will support the sharing of data, in the form of format converters or interfaces to the Internet. In short, the functions of today's digital computers in

supporting research go far beyond those of a simple calculating machine, carrying out well-defined methods of analysis. The same digital computer may now be involved in the selection and formulation of a problem, by providing access to automated library catalogs and on-line literature; in the collection of data through support for real-time data acquisition; in management of data, performance of analysis, visualization of results, writing of conclusions; and even in publication through access to the Internet and the World Wide Web. The computer is no longer part of the research environment—we are rapidly approaching a world in which the computer *is* the research environment.

These trends are echoed strongly in geographical information systems. Although a particular scientist might use a GIS in ways that are more analogous to the early days of statistical computing, by performing a single buffering operation, for example, scientific applications are much more likely to include integration of many GIS functions. Today's scientist or decision-maker is likely to see a GIS as an environment for research, rather than as a means of automating analysis. The GIS is likely to be involved in the project from beginning to end, and to be integrated with other tools and environments when these are needed. GIS will be used for collecting, assembling, verifying and editing the data; performing some of the analyses required by the project; and presenting and interpreting the results. Moreover, much GIS use may not be tied to a specific project—GIS finds extensive use in the collection of data for purposes that may be generic, or not well-defined, or may be justified in anticipation of future demand. Even though these may not be projects in the sense of the earlier discussion, analysis may still be necessary as part of the data production process—for example, when a soil scientist must analyze data to produce a soil map.

2.5 When to choose GIS

If GIS has multiple roles in support of science and problem-solving, then one might not be surprised to find that the choice between GIS alternatives is complex and often daunting. The many GIS packages offer a wide range of combinations of analysis functions, housekeeping support, alternative ways of representing the same phenomena, different levels of sophistication in visual display, and performance. In addition, choice is often driven by the available hardware, since not all GIS run on all platforms; on the format in which the necessary data has been supplied, the personal preferences and background of the user, and so forth. Even the extensive and frequently updated comparative surveys published by groups such as GIS World Inc can be of little help to the uninitiated user.

The existence of other classes of analytic software complicates the scene even more. Under what circumstances is a problem better solved using a package that identifies itself as a GIS, or using a statistical package, or a mathematical package, or a scientific visualization package? Under what circumstances is it better to fit the square peg of a real problem into the round GIS hole? GIS are distinguished by their ability to handle data referenced to a two-dimensional frame, but such capabilities also exist to a more limited extent in many other types of software environment. For example, it is possible to store a map in a spreadsheet array, and with a little ingenuity to produce a passable 'map' output; and many statistical packages support data in the form of images.

Under what circumstances, then, is an analyst likely to choose a GIS? The following conditions are suggested, although the list is certainly not complete, and the items are not intended to be mutually exclusive:

- when the data are geographically referenced, and when geographical referencing is essential to the analysis (see earlier discussion of the definition of spatial analysis);
- when the data include a range of vector data types (support for vector analysis among non-GIS packages appears to be much less common than support for raster analysis);
- when topology—representation of the connections between objects—is important to the analysis;

- when the curvature of the Earth's surface is important to the analysis, requiring support for projections and for methods of spatial analysis on curved surfaces;
- when the volume of data is large, since alternatives like spreadsheets tend to work only for small data sets;
- when data must be integrated from a variety of sources, requiring extensive support for reformatting, resampling, and other forms of format change;
- when geographical objects under analysis have large numbers of attributes, requiring support from integrated database management systems, since many alternatives lack such integration;
- when the background of the investigator is in geography, or a discipline with strong interest in geographical data;
- when the project involves several disciplines, and must therefore transcend the software traditions and preferences of each;
- when visual display is important, and when the results must be presented to varied audiences;
- when the results of the analysis are likely to be used as input by other projects, or when the data are being extensively shared.

3. Elements of a New Perspective

This section reviews some of the changes that are altering the context and face of spatial analysis using GIS. Some are driven by technological change, and others by larger trends affecting society as we approach the millennium.

3.1 The costs of data creation

The collection of geographical data can be extremely labor-intensive. Early topographic mapping required the map-maker to walk large parts of the ground being mapped; soil mapping requires the exhausting work of digging soil pits, followed often by laborious chemical analysis; census data collection requires repeated visits to a substantial proportion of all households; and forest mapping requires 'operational cruise', the intensive observation of conditions along transects. Although many new methods of geographical data creation have replaced the human observer on the ground with various forms of automated sensing, there is no alternative in those areas that require the presence of expert interpreters in the field.

Many of the remaining stages of geographical data creation are also highly labor-intensive. There is still no alternative to manual digitizing in cases where the source document is complex, compromised, or difficult to interpret. The processes of error detection and correction are difficult if not impossible to automate, and the methods of cartographic generalization used by expert cartographers have proven very difficult to formalize and replace. In short, despite much technical progress over the past few decades, geographical data creation remains an expensive process that is far from fully automated.

Labor costs continue to rise at a time when the resources available to government, the traditional source of geographical data, continue to shrink. Many geographical data sets are collected for purposes which may be far from immediate, and it is difficult therefore to convince taxpayers that they represent an essential investment of public funds, especially in peacetime. Governments in financial straits call for evidence of need, and many have moved their mapping operations onto a semi-commercial basis in order to allow demand to be expressed through willingness to pay. To date, the U.S. Federal mapping agencies have resisted the trend, but internationally there is more and more evidence of the emergence of a market in geographical information.

Within the domain of geographical data the pressures of increased labor costs favor data that can be collected and processed automatically. Given a choice between the labor-intensive production of vector topographic data, and the semi-automated generation of such raster products as digital elevation models and digital orthophotos, economic pressures can lead only in one direction. It is easy to imagine a user trading off the ability to identify features by name against the order of magnitude lower cost, and thus greater potential update frequency, of raster data.

Of course, the principle of information commerce is alien to the scientific community, which is likely to resist strongly any attempt to charge for data that is of interest to science, even peripherally. But here too there are pressures to make better use of the resources invested in scientific data collection. Research funding agencies now increasingly require evidence that data collected for a project have been disseminated, or made accessible to others, while recognizing the need to protect the interests of the collector.

But trends such as these, while they may be eminently rational to dispensers of public funds, nevertheless fly directly in the face of the traditional model of science presented earlier. How can projects fail to be driven by data, if data are forced to obey the economic laws of supply and demand? Where in traditional science are the rules and standards that allow scientists to trade off economic cost against scientific truth? It seems that economic necessity has forced the practice of science to move well beyond the traditions that are reflected in accepted scientific methodologies and philosophies of science.

3.2 The life of a data set

In the traditional model presented earlier data were collected or created to solve a particular problem, and had no use afterwards except perhaps to historians of science. But many types of geographical data are collected and maintained for generic purposes, and may be used many times by completely unrelated projects. For other types, the creation of data is itself a form of science, involving the field skills of a soil scientist, for example, or a biologist. Thus a data set can be simultaneously the output of one person's science, and the input to another's. These relationships have become further complicated by the rise of multidisciplinary science, which combines the strengths and expertise of many different sciences, and partitions the work among them. Once again, the linear model of science is in trouble, unable to reflect the complex relationships between projects, data sets, and analytic techniques that exist in modern science. The notion that data are somehow subsidiary to problems, methods and results is challenged, and traditional dicta about not including technical detail in scientific reports may be counterproductive.

In this new world a given set of data is likely to fall into many different hands during its life. It may be assembled from a mixture of field and remote sensing sources, interpreted by a specialist, cataloged by an archivist or librarian, used by scientists and problem-solvers, and passed between its custodians using a range of technologies. It is quite possible in today's world that the various creators and users share little in the way of common disciplinary background, leaving the data set open to misunderstanding and misinterpretation. Recent interest in metadata, or ways of describing the contents of data sets, is directed at reducing some of these problems, but the easy access to data provided by the Internet and various geographical data archives has tended to make the problem worse.

These issues are particularly prominent in the case of data quality, and the ability of the user of a data set to understand its limitations, and the uncertainty that exists about the real phenomena the data are intended to represent. To take a simple example, suppose information on the geodetic datum underlying a particular data set—potentially a very significant component of its metadata—was lost in transmission between source and user; or alternatively suppose that the user simply assumed the wrong datum, or was unaware of its significance. This loss of metadata, or specification

of the data content, is equivalent in every respect to an actual loss of accuracy equal to the difference between the true datum and the datum assumed by the user, which can be several hundreds of meters. In short, the quality of a data set to a user is a function of the difference between its contents and the *user's* understanding of its meaning, not the creator's.

3.3 Data sharing

In this new world of shared data the term *metadata* has come to function as the equivalent of documentation, catalog, handling instructions, and production control. The U.S. Federal Geographic Data Committee's Content Standards for Digital Geospatial Metadata (FGDC 1994) have been very influential in providing a standard, and have been emulated frequently. If the custodian of a large collection of geographical data sets provides metadata in this form, it is possible for others to search its records for those that match their needs. The FGDC's National Geospatial Data Clearinghouse (http://www.fgdc.gov) is one such directory (and see also the Alexandria Digital Library project to provide distributed library services for geographically referenced data sets, Smith et al. 1996; and see http://alexandria.ucsb.edu).

The user of a traditional library will rarely know the exact subject of a search—instead, library search has an essential fuzziness, which is supported by the traditional library in several essential ways. By assigning similar call numbers to books on similar subjects, and shelving by call number, the traditional library is able to provide an environment that allows the user to browse the collection in a chosen area. But this support is missing when the records of a metadata file are searched using simple Boolean methods. It would make better sense to model the search process as one of finding the best fit between a metadata record representing the user's ideal, and metadata records representing the data sets available. It is very unlikely, after all, that data exist that perfectly match the needs of a given problem, especially in the ideal world of problem-solving represented earlier.

3.4 New techniques for analysis

Many new methods of spatial analysis have emerged in the rich computational environment now available to scientists. These include neural nets, new methods of optimization such as simulated annealing and genetic techniques, and computationally intensive simulation. The term *geocomputation* has been suggested. Methods of exploratory spatial data analysis have extended the principles of exploratory data analysis (Tukey 1977) to spatial data.

In science generally, the combination of vast new sources of data and high-speed computation have led to an interest in methods of *data mining*, which implies the ability to *dredge* data at very high speed in a search for patterns of scientific interest. In a geographical context, the very vague notion of 'scientific interest' might suggest the need for methods to detect features or measurements that are inconsistent with their surroundings, in apparent violation of Tobler's 'first law of geography' (Tobler 1970). Linearities in images are of potential interest in geological prospecting; and one can imagine circumstances in which atmospheric scientists might want to search large numbers of images for patterns consistent with weather events. Such techniques of pattern recognition were pioneered many years ago in particle physics, to search vast numbers of bubble-chamber photographs for the tracks characteristic of rare new particles.

One might argue that such techniques represent a renewal of interest in inductive science—the search for regularities or patterns in the world that would then stimulate new explanatory theories. Inductivism has fallen out of fashion in recent decades, at least in disciplines that focus on geographical data, leading one to ask whether a renewal of interest represents a fundamental shift in science, or merely a response to the opportunities offered by more powerful technology. On this

issue the jury is clearly still 'out'—geocomputation has not yet provided the kinds of new insights that might support a broad shift to inductivism.

3.5 New computer architectures

The communication technologies that have emerged in the past decade have allowed a fundamental change in the architecture of computing systems. Instead of the early mainframes and later stand-alone desktop systems, today's computers are linked with high-speed networks that allow data, software, and storage capacity located in widely scattered systems to be integrated into functioning wholes. Data can now be 'served' from central sites on demand, avoiding the need to disseminate many copies, with subsequent confusion when updates are needed.

The new approaches to computing that are possible in this interconnected environment are having a profound effect on spatial analysis. Because it is no longer possible to assume a lifetime association between a user and a particular system design, there are mounting pressures for standards and interoperability between systems to counter the high costs of retraining of staff and reformatting of data.

The proprietary GIS that once dominated the industry attempted to provide a full range of GIS services in one homogeneous environment. Data were stored in proprietary formats, often kept secret by vendors to maintain market position, but making it difficult for others to expand the capabilities of the system by programming extra modules. The 'open GIS' movement (Buehler and McKee 1996; and see http://www.ogis.org) mirrors efforts in other areas of the electronic data processing world to promote interoperability, open standards and formats, and easy exchange from one system to another. While such ideas were often regarded as counter to the commercial interests of vendors, there is now widespread acceptance in the industry that they represent the way of the future.

The implications of open systems for spatial analysis are likely to be profound. First, they offer the potential of a uniform working environment, in which knowledge of one system is readily transferrable to another. To make this work, however, it will be necessary to achieve a uniform view, and its acceptance across a very heterogeneous user community. There is no prospect of interoperability and open systems without agreement on the fundamental data models, terminology and objectives of GIS-based analysis. Thus much effort will be needed on the part of the inventors and implementors of spatial analysis to develop this uniform view.

Second, the possibility of easy sharing of data across systems gives even greater momentum to efforts to make geographical information more shareable, and even greater demands on the existence and effectiveness of metadata.

Third, interoperability is likely to create an environment in which it is much easier to implement methods of spatial analysis in GIS. Traditionally, vendors of monolithic systems have added functions when market demand appears to justify the development costs. It has been impossible, in a world of proprietary systems, for third parties to add significant functionality. Thus expansion of spatial analytic capabilities has been slow, and has tended to reflect the needs of the commercial market, rather than those of science and problem-solving, when these diverge. In a world of open systems it will be much easier to add functions, and the new environment will encourage the emergence of small companies offering specialized functionality in niche markets.

Finally, new interoperable approaches to software will encourage the modularization of code. It is already possible in some mainstream software environments to launch one specialized application within another—for example, to apply spreadsheet functions to information in a word processing package. This 'plug and play' environment offers enormous scope to GIS, since it will lead ultimately to a greater integration of GIS functions, and map and imagery data in general, into mainstream electronic data processing applications.

The scientific world has grown used to a more or less complete separation between data, and the functions that operate on and manipulate data. Functions are part of 'analysis', which plays a role in the traditional approach to problem-solving outlined earlier that is clearly distinct from that of data. But it has already been argued that in a world of extensive data sharing and interaction between disciplines it is impossible to think of data in isolation from its description, or metadata, which allows the meaning of information to be shared.

In the abstract world of object-oriented methods it is argued that the meaning of data lies ultimately in the operations that can be performed. If data sets exist in two systems, and pairs of functions exist in both systems that produce the same answers, then the two data sets are the same in information content, irrespective of their specific formats and arrangements of bits. It makes sense, then, to *encapsulate* methods with data. When more than one method is available to perform a given function, it makes sense for the choice to be made by the person best able to do so, and for the method thereafter to travel with the data. For example, a climatologist might encapsulate an appropriate method for spatial interpolation with a set of point weather records, because the climatologist is arguably better able to select the best method of spatial interpolation, given his or her knowledge of atmospheric processes.

In future, and especially given the current trend in computing to object-oriented methods, it is likely that the distinction between data and methods will become increasingly blurred. Commonly used techniques of spatial analysis, such as spatial interpolation, may become encapsulated with data in an extension of the concept of metadata to include methods. Of course this assumes that methods are capable of running in a wide variety of host systems, which takes the discussion back to the issue of interoperability introduced earlier.

4. Spatial Analysis in Practice

At this stage, it seems useful to introduce a discussion of the practical problems which face the users of today's GIS. While it is now possible to undertake a wide range of forms of spatial analysis, and to integrate data from a range of sources that would have seemed inconceivable as little as five years ago, there continue to be abundant limitations that impede the complete fulfilment of the technology's promise. The following subsections discuss several of these current impediments.

4.1 Absolute and relative position

First, and perhaps foremost, are problems of varying data quality. In science generally it is common to express quality in terms such as 'accurate to plus or minus one degree'. But while such methods are useful for many types of data, they are much less so when the data are geographical. The individual items of information in a geographical data set are typically the result of a long and complex series of processing and interpretation steps, and bear little relationship to the independent measurements of traditional error analysis. The following discussion is limited to the particular problems encountered when merging data sets.

While projections and geodetic datums are commonly well-documented for the data sets produced by government agencies, the individual scientist digitizing a map may well not be in a position identify either. The idea that lack of specification could contribute to uncertainty was discussed earlier, and its effects will be immediately apparent if a data set is merged with one based on another projection or datum. In practice, therefore, users of GIS frequently encounter the need for methods of *conflation*, a topic discussed in detail below.

The individual items of information in a geographical data set often share lineage, in the sense that more than one item is affected by the same error. This happens, for example, when a map or photograph is registered poorly—all of the data derived from it will have the same error. One indicator of shared lineage, then, is the persistence of error—all points derived from or dependent on the same misregistration will be displaced by the same or a similar amount. Because neighboring points are more likely to share lineage than distant points, errors tend to show strong positive spatial autocorrelation (Goodchild and Gopal 1989).

Rubber-sheeting is the term used to describe methods for removing such errors on the assumption that strong spatial autocorrelations exist. If errors tend to be spatially autocorrelated up to a distance of x, say, then rubber-sheeting will be successful at removing them, at least partially, provided control points can be found that are spaced less than x apart. For the same reason, the shapes of features that are less than x across will tend to have little distortion, while very large shapes may be badly distorted. The results of calculating areas, or other geometric operations that rely only on relative position, will be accurate as long as the areas are small, but will grow rapidly with feature size. Thus it is important for the user of a GIS to know which operations depend on *relative* position, and over what distance; and where *absolute* position is important (of course the term *absolute* simply means relative to the Earth frame, defined by the Equator and the Greenwich meridian, or relative over a very long distance).

When two data sets are merged that share no common lineage (for example, they have not been subject to the same misregistration), then the relative positions of objects inherit the absolute positional errors of both, even over the shortest distances. While the shapes of objects in each data set may be accurate, the relative locations of pairs of neighboring objects may be wildly inaccurate when drawn from different data sets. The anecdotal history of GIS is full of such examples—data sets which were perfectly adequate for one application, but failed completely when an application required that they be merged with some new data set that had no common lineage. For example, merging GPS measurements of point positions with streets derived from the U.S. Bureau of the Census TIGER files may lead to surprises where points appear on the wrong sides of streets. If the absolute positional accuracy of a data set is 50m, as it is with parts of TIGER, then such surprises will be common for points located less than 50m from the nearest street.

4.2 Semantic integration

Some of the most challenging problems in GIS practice occur in the area of semantic integration, where integration relies on an understanding of meaning. Such problems can occur between geographic jurisdictions, if definitions of feature types, or classifications, or methods of measurement vary between them. It is common, for example, for schemes of vegetation classification to vary from one country to another, making it difficult to produce horizontally merged data (Mounsey 1991). 'Vertical' integration can also be problematic, as in the problems of merging maps produced of the same area by different agencies.

While some of these problems may disappear with more enlightened standards, others are eminently reasonable. The problems of management of ecosystems in Florida are clearly different from those of Montana, and it is reasonable that standards adopted by the two states should be different. Even if it were possible to standardize for the entire U.S., one would be no further ahead in standardizing between the U.S. and other countries. Instead, it seems a more reasonable approach is to achieve interoperability without standardization, by more intelligent approaches to system design.

4.3 Conflation

Conflation appears to be the term of choice in the GIS community for functions that attempt to overcome differences between data sets, or to merge their contents. Conflation attempts to replace two or more versions of the same information with a single version that reflects the pooling of the sources; it may help to think of it as a process of weighted averaging. The complementary term *concatenation* refers to the integration of the sources, so that the contents of both are accessible in the product. The polygon overlay operation familiar to many GIS users is thus a form of concatenation.

Two distinct forms of conflation can be identified, depending on the context: (1) conflation of feature geometry and topology, and concatenation of feature attributes; and (2) conflation of geometry, topology and attributes. As an example of the first case, suppose information is available on the railroad network at two scales, 1:100,000 and 1:2 million. The set of attributes available is richer at the 1:2 million scale, but the geometry and topology are more accurate at 1:100,000. Thus it would be desirable to combine the two, discarding the coarser geometry and topology. As an example of the second case, consider a situation in which soils have been mapped for two adjacent counties, by two different teams of scientists. At the common border there is an obvious problem, because although the county boundary was defined by a process that was in no way dependent on soils, the border nevertheless appears in the combined map. Thus it would be desirable to 'average' the data at and near the boundary by combining the information from both maps in compatible fashion. As these two examples illustrate, the need for conflation occurs both horizontally, in the form of edgematching, and 'vertically'.

4.4 Perfect positioning

It is easy to imagine that the need for conflation and for discussions of relative and absolute positional accuracy will eventually go away, as positioning becomes more and more accurate, leading eventually to 'perfect' positioning. Unfortunately there are good reasons why that happy state will never be reached. Although the positions of the Greenwich meridian and various geodetic control points have been established by fixing monuments, seismic motions, continental drift, and the wobbling of the Earth's axis all lead to fundamental uncertainty in position. Any mathematical representation of the Earth's shape must be an approximation, and different approximations have been adopted for different purposes. Moreover, there will always be a legacy of earlier, less accurate measurements to deal with. Thus it seems GIS will always have to deal with uncertainty of position, and with the distinctions between relative and absolute accuracy, and their complex implications for analysis.

Instead, strategies must be found for overcoming the inevitable differences between databases, either prior to analysis or in some cases 'on the fly'. Consider, for example, the problems caused by use of different map databases for vehicle routing. Systems are already available on an experimental basis that broadcast information on street congestion and road maintenance to vehicles equipped with map databases and systems to display such information for the driver. In a world of many competing vendors such systems will have to overcome problems of mismatch between different databases, in terms both of position and of attributes. For example, two databases may disagree over the exact location of 100 Main St, or whether there *is* a 100 Main St, with potentially disastrous consequences for emergency vehicles, and expensive consequences for deliveries. Recent trends suggest that the prospects for central standardization of street naming by a single authority are diminishing, rather than growing.

5. Conclusion

The prospects for spatial analysis have never been better. Data are available in unprecedented volume, and easily accessed over today's communication networks. More methods of spatial analysis are implemented in today's GIS than ever before, and GIS has made methods of analysis that were previously locked in obscure journals easy and straightforward to use. Nevertheless, today's environment for spatial analysis raises many issues, not the least of which is the ability of users to understand and to interpret correctly. Questions are being raised about the deeper implications of spatial analysis, and the development of databases that verge on invasion of individual. And our expectations may be unreasonable given the inevitable problems of spatial data quality.

Technological developments have further muddied the methodological waters, by confusing what was once a simple linear sequence of problem formulation, data collection, analysis, and conclusion. It seems clear that tomorrow's science will be increasingly driven by complex interactions, as data become increasingly commodified, technology increasingly indispensible to science, and conclusions increasingly consensual. New philosophies of science that reflect today's realities are already overdue.

If science and problem-solving are to be constrained by these new realities, then what kinds of spatial analysis are most likely to dominate in the coming years? The points raised in this chapter's discussion suggest that the future environment will favor the following:

- data whose meanings are widely understood, making it easier for multidisciplinary teams to collaborate;
- data with widespread use, generating demands that can justify the costs of creation;
- data with commercial as well as scientific and problem-solving value, allowing costs to be shared across many sectors;
- methods of analysis with commercial application, making it more likely that such methods will be implemented in widely available form;
- methods implemented using general standards, allowing them to be linked to other methods using common standards and protocols.

6. References

Bailey TC, Gatrell AC 1995 Interactive Spatial Data Analysis. New York: Wiley.

Berry BJL, Marble DF (eds) 1968 Spatial Analysis: A Reader in Statistical Geography. Englewood Cliffs, NJ: Prentice-Hall.

Buehler K, McKee L (eds) 1996 The OpenGIS Guide. Wayland, MA: The Open GIS Consortium Inc

Chrisman NR 1997 Exploring Geographic Information Systems. New York: Wiley.

Federal Geographic Data Committee 1994 Content Standards for Digital Geospatial Metadata. Washington, DC: Federal Geographic Data Committee, Department of the Interior. http://www.fgdc.gov.

Fotheringham AS, Rogerson PA (eds) 1994 Spatial Analysis and GIS. London: Taylor and Francis.

Goodchild MF 1988 A spatial analytic perspective on geographical information systems. International Journal of Geographical Information Systems 1: 327–334.

Goodchild MF, Gopal S 1989 Accuracy of Spatial Databases. London: Taylor and Francis.

Goodchild MF, Haining RP, Wise S and 12 others 1992 Integrating GIS and spatial analysis: problems and possibilities. International Journal of Geographical Information Systems 6(5): 407–423.

Goodchild MF, Longley P 1997 The future of GIS and spatial analysis. In Longley P, Goodchild MF, Maguire DJ, Rhind DW (eds) Geographical Information Systems: Principles, Techniques, Management, and Applications. Cambridge: Geoinformation International.

Haining RP 1990 Spatial Data Analysis in the Social and Environmental Sciences. New York: Cambridge University Press.

Harman HH 1976 Modern Factor Analysis. Chicago: University of Chicago Press.

Levine G 1981 Introductory Statistics for Psychology: The Logic and the Methods. New York: Academic Press.

Marascuilo LA, Levin JR 1983 Multivariate Statistics in the Social Sciences: A Researcher's Guide. Monterey, CA: Brooks/Cole.

Mounsey H 1991 Multisource, multinational environmental GIS: lessons learned from CORINE. In Maguire DJ, Goodchild MF, Rhind DW (eds) Geographical information systems: principles and applications. Harlow: Longman Scientific and Technical, 2: 185–200.

Siegel S 1956 Nonparametric Statistics for the Behavioral Sciences. New York: McGraw-Hill.

Smith TR, Andresen D, Carver L, Dolin R, and others 1996 A digital library for geographically referenced materials. Computer 29(5): 54, 29(7):14.

Taylor PJ 1977 Quantitative Methods in Geography: An Introduction to Spatial Analysis. Boston: Houghton Mifflin.

Tobler WR 1970 A computer movie simulating urban growth in the Detroit region. Economic Geography supplement 46: 234–240.

Tukey JW 1970 Exploratory Data Analysis. Reading, MA: Addison Wesley.

Unwin DJ 1981 Introductory Spatial Analysis. London: Methuen.

3.3 Using GIS and models for decision support in Costa Rican farming

Jetse J. Stoorvogel[1]

Department of Soil Science and Geology, Wageningen Agricultural University,
P.O. Box 37, 6700 AA Wageningen, The Netherlands
Email: jetse.stoorvogel@bodlan.beng.wau.nl

GIS and agricultural models comprise important tools to support agricultural decision making. In Costa Rica two decision support systems are developed for two distinct scale levels. The USTED methodology supports policies interventions through an *ex ante* evaluation of the agricultural policies. BanMan, a decision support system for banana management, works on the farm level for the identification and analysis of local differences in yields and yield forecasting. Main constraints for the development of decision support systems are related to data availability. A black approach seems to be unavoidable due to the complexity of the systems.

1. Introduction

Typically, decision makers are faced with data limitations. The required data are in many cases not available and additional data collection is costly. Nevertheless, the decisions need to be taken and decision support systems (DSS) have to work with the limited data sets that are available. In particular cases, data may be identified that need to be collected before decisions can be taken. A cost/benefit analysis should reveal whether the investment is worthwhile.

Fortunately, science increasingly develops models that describe simple and/or complex processes in the real world in simplified representations. On the basis of these models alternative scenarios may be evaluated without carrying out expensive experiments. On the basis of model runs we are now able to direct data collection very specifically. Additionally, new techniques enable us to collect geographic information with lower costs, *e.g.* global positioning systems, remote sensing and yield mapping. The rapid development and adoption of these new techniques is accelerated through the development of geographical information systems (GIS) that enable the processing and analysis of spatial data. Literature provides us with several studies showing the use of GIS in decision making (e.g. Hassan and Hutchinson 1992; Maguire *et al.* 1991; Fotheringham and Rogerson 1994). Similarly, model development in combination with GIS has been applied frequently (see Goodchild, Section 3.2).

This paper identifies the use of GIS and models for decision support on two distinct scale levels: the farm level and the regional level. The theoretical framework is illustrated with examples from two Costa Rican case studies. For Guácimo county in the northeast of Costa Rica, a methodology has been developed for an *ex ante* evaluation of policy interventions. The example at farm level originates from a banana farm, where a DSS was developed to support farm management.

[1] The research of Dr J. Stoorvogel has been made possible by a fellowship of the Royal Netherlands Academy of Arts and Sciences.

2. GIS and models for decision support

2.1 Decision support systems

DSS are developed to support decisions of individuals. Note that DSS do not take decisions, they might analyse the effects of certain decisions or look for an optimum given certain objectives and constraints as listed by the decision maker.

Decisions need to be taken at different spatio-temporal scales each with specific requisites in terms of input data and problem definition. At the regional level, for example, policy makers will have to deal with general agricultural statistics to select between a number of policy interventions. At the farm level on the other hand, farmers base their decisions on past experiences to take their tactical and operational decisions.

The development of DSS traditionally follows a pathway as illustrated in Figure 1. After a first phase of problem identification, a certain methodology will be selected and elaborated. The methodology will have certain data requirements, if these data are not available they need to be collected. It is realised that not all data can be collected and that data requirements easily exceed the available resources. In such cases, one should return to adapt the methodology, until the methodology and the collected data (+ available data) are well balanced. On the basis of the collected data, certain models which may be included in the methodology may require calibration and/or validation. If problems arise during calibration/validation one may need to return to collect additional data or adapt the methodology. In a final stage the methodology may be tested and applied as a decision support system.

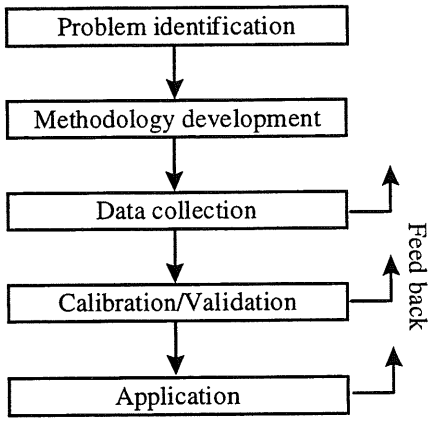


Figure 1 The development of DSS

GIS and models are, amongst others, important tools in the development of DSS. Although GIS are powerful tools for the analysis of geo-information, commercial GIS packages do not always meet the specific needs of users (O'Kelly, 1994). User requirements often comprise very specific disciplinary operations and user oriented shells. To a limited extent, GIS software enables the development of applications that may include (simple) models. Implementation of additional procedures into GIS packages or modifications of GIS packages usually coincides with high costs due to the complexity of GIS software and additional modelling systems (Abel *et al.*, 1994). Therefore, GIS can be considered to be a closed system, i.e. no changes in the internal schemes of the software can be made. Specific disciplinary analysis like crop growth simulation need, therefore, external models, which work independently from the GIS and perform the analysis which the GIS package is unable to handle. For operationalisation, the GIS needs to be linked to these external models. Although the necessity to link GIS with models is generally recognised, many practical problems are known to occur (Burrough, 1989; Abel *et al.*, 1994). Part of the problems originate in the incompatibility of data formats, data organisation or semantics which respectively requires reformatting, restructuring and data analysis before the GIS database can be used in combination with external models. No GIS architecture has yet been developed that conceptualises the link between GIS and external models. At present, therefore, the link between models and GIS is often established in an ad-hoc manner. Specially designed structures may facilitate this link and can be included in the GIS for operationalisation (e.g Stoorvogel 1995).

3. Decision support at the regional level

3.1 Problem identification

Agricultural policies and economic incentives can be important tools to achieve a more sustainable use of natural resources. In developing countries the number of policy interventions is relatively limited compared to developing countries as a result of a limited control system to check on limits of, for example, the use of agro-chemicals. Nevertheless, even in those countries we find a considerable array of incentives and policies that may influence the agricultural sector (Lutz and Daily 1991). *A priori*, it is extremely difficult to make prognoses on the reaction of a complex sector on the changes in the socio-economic environment as a result of the interventions. In developed countries, special government agencies are occupied with the analysis of agricultural statistics and data to make prognoses of trends and analysis of trade-offs. Developing countries typically lack data and expertise for this kind of analysis.

Policies are defined by national or regional governments but the ultimate decision on how to use agricultural land is made on individual farms. Different farms may respond differently to certain policies or incentives. These differences in response are related to differences between farms in physical properties (size, soil types, climate), in socio-economic conditions (market access and prices of inputs and outputs, possibilities to work off farm), and in objectives and preferences of the persons managing the farm. In many situations, interaction exist between different farms, e.g. through labour exchange. Various types of models exist that enable analysis of farmers' reactions to policy incentives (e.g. Schipper 1996; Kruseman et al. 1995; Antle 1996). Although land use and many of its consequences clearly have spatial dimensions, until recently the available literature did not offer a methodology that allows flexible georeferencing of inputs to, and outputs from, land use evaluation models.

3.2 Methodology development

The USTED methodology (*Uso Sostenible de Tierras En el Desarrollo*: Sustainable Land Use in Development) has been developed by REPOSA (Research Programme on Sustainability in Agriculture), a collaborative project of Wageningen Agricultural University (The Netherlands), the Tropical Agronomic Research and Higher Education Center (Costa Rica) and the Costa Rican Ministry of Agricultural and Livestock (Stoorvogel et al. 1995.

USTED is developed for an *ex ante* evaluation of agricultural policies and incentives. Through the analysis it may support policy makers in their decisions. It evaluates the effects of external factors (e.g. labour availability and market prices) on agricultural land use at the sub-regional level. Sustainability aspects of land use can be expressed by a number of quantitative criteria, thus presenting a pragmatic approach to the concept of sustainability. The output of USTED is the selection and distribution of land use according to a goal and a set of constraints.

The general structure of the USTED methodology is presented in Figure 2. Various models and database management systems are integrated in the USTED methodology (Fig. 2). In USTED, options for land use are defined as combinations of a land unit and a land use type with a specified technology, called LUSTs (Land Use Systems with a defined Technology). Each LUST describes a specific quantitative combinations of physical inputs and outputs, thus representing fixed input-output technologies (Jansen and Schipper, 1995). Technical coefficients for each LUST are calculated by combining the physical quantities of relevant inputs and outputs with their specific attributes (*e.g.,* data on prices, chemical composition, toxicity). Different crop and livestock related agricultural activities can be performed by different technologies which have different income generating capabilities as different sustainability implications.

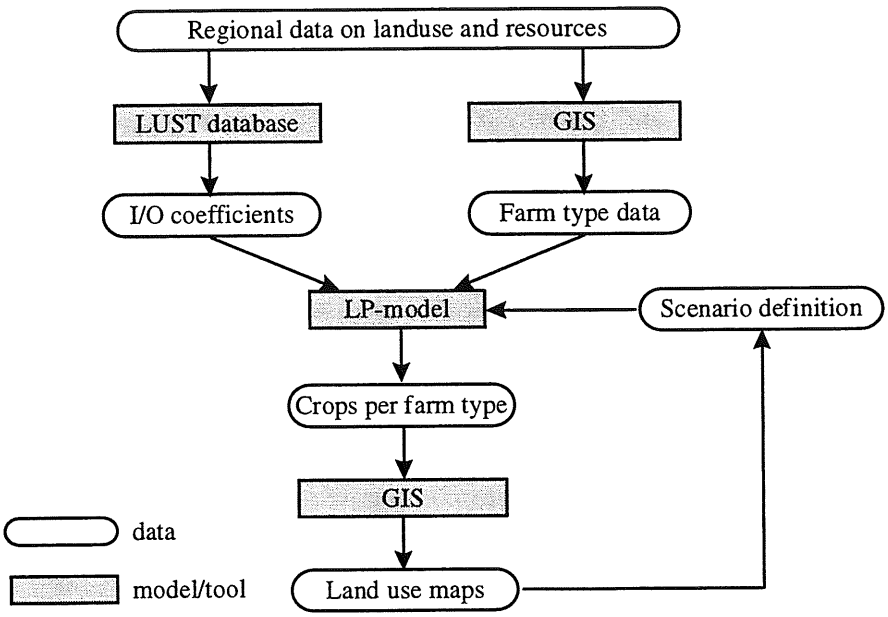


Figure 2 **The set up of the USTED methodology**

Actual systems are described by farm surveys whereas alternatives may be developed using crop growth simulation models.

The technical coefficients of the LUSTs are offered to a sub-regional linear programming (LP) model (Schipper et al, Stoorvogel et al). Because of the relatively small size of the region, all prices are assumed exogenous, thus avoiding the issue of price decreases caused by supply increases. Even though policies influencing land use are defined by national or regional governments, individual farms are included as the ultimate decision makers by grouping the farms of the canton into several farm types, defined on the basis of farm size and the dominating soil group. A constant average labor availability is assumed, making each farm type homogeneous in its land and labor ratio. In this way, aggregation bias is less than if the canton would have been considered as one large farm.

The LP model contains sub-matrices for each farm type which encompass the constraints for a specific farm-type and the specific coefficients for the LUSTs. The model maximizes the sum of net farm income over the different farm types subject to resource constraints.

Trade-offs between different policy goals are studied through scenarios in which the effects on land use are analyzed for different (hypothesized) changes in the socioeconomic environment or policy instruments. Specific attention is given to the trade-off between economic growth (or maximization of aggregate income) and sustainability. The latter is operationalised in terms of soil nutrient balances and biocide use. Using an adapted version of the NUTBAL model (Stoorvogel 1993), nitrogen depletion from the soil nutrient stock was estimated for each of the systems. For each biocide, an environmental impact index (BEII) was calculated on the basis of active ingredients, toxicity and persistency (Jansen *et al* 1995).

3.3 Data collection

Like any analysis of the agricultural sector, DSS at the regional level require a considerable amount of data at the farm and regional level. For most analysis a minimum data set can be given:

- A general purpose soil survey describing the variation of one of the main natural resources.
- Numbers, sizes and soil types of individual farms or generalized farm types.

- Quantitative descriptions of the principal land use systems in the area and the corresponding inputs and outputs.
- Data on the region for detailed descriptions of the constraints and alternatives for different land use systems.
- Insight in a number of relevant sustainability indicators.

USTED has been applied to Guácimo county in the Atlantic Zone of Costa Rica, with a total of 58,000 ha. Altitude of cultivated areas is between 10 and 50m above sea level in a region which climate is classified as humid tropical, without dry months (Herrera and Gómez 1993). A 1:150,000 soil survey (Wielemaker and Vogel 1993) was used which was generalized resulting in three main soil groups for Guácimo county. A total of 122 actual and alternative land use systems were described (Jansen and Schipper 1995). The agricultural census for Costa Rica (DGEC 1987) in combination with a 1992 land use map and the soil map yielded 9 farm types defined on the basis of the physical resources size and soil type (Belder 1994).

3.4 Calibration and validation

The calibration and validation of regional land use models is extremely difficult. Data sets representing the actual trade off between economic growth and sustainability are in most cases not available. If they are available they will probably form the basis for the model runs and as a result they can not be used for validation. The results of the analysis are therefore only indications of changes and should not be used as actual forecasts.

3.5 The application of USTED

Analysis with USTED comprises of base run for a reference year and additional runs for changing conditions. The model run for the reference year (here 1990) is called the base scenario Policies interventions are translated into possible changes in the socio-economic or bio-physical environment and subsequently into new technical coefficients for the LP model. After the optimization the results were evaluated in contrast to the base scenario. In Table 1 the results for 4 alternative scenarios are presented:

- Palm heart price: The area planted to palm heart continues to increase at a fast pace. Palm heart is largely exported, though international markets are quite small. Consequently, alternative scenario runs were performed with a reduction in the palm heart price. At a 25% price decrease, palm heart is replaced by cassava, pasture and tree plantations (Fig. 3), resulting in a 12% decrease in net income per farm. The selected cassava, pasture and tree plantation technologies use significantly less biocides than the palm heart systems, resulting in a 64% decrease in the BEII.
- Environmental tax: Prices increases for the various types of biocides were differentiated through an environmental tax which level increases progressively with the value of an individual biocide's BEII. The impact on crop choice is negligible with little change in income indicators. However, technology choice is pushed towards more environmentally-friendly cultivation methods, partially substituting herbicides by manual weeding. As a result, the aggregated BEII decreases significantly.
- Biocide regulation: Legislation aimed at a quantitative reduction in biocide use was modeled by imposing a ceiling (50% of the aggregated BEII value per ha in the base scenario) for each individual farm type. In addition to farm types with relatively high biocide use in the base scenario, also farm types with average BEII below 21 in the base scenario are affected because of labor interactions. The overall results are similar to an environmental tax in terms of income, but

sustainability is improved much less. Also, implementation of direct regulatory measures may prove more difficult than introduction of a tax.

- Capital availability: The effects of increased capital scarcity are twofold. First, relatively input-intensive crop-technologies are replaced by alternatives with lower initial costs (including labor). Palm heart has relatively high initial costs due to its initial labor requirements, and its area decreases steadily as capital becomes more limiting. The area under cassava, which uses significantly more biocides and fertilizers than palm heart, initially increases, but when capital becomes insufficient to finance the necessary inputs, cassava is partially replaced by relatively less input-intensive tree plantations. Second, decreases in income from crop cultivation are increasingly compensated by working more off-farm. As a result, net income decreases less than gross margins from crop cultivation. In addition, sustainability improves significantly, both in terms of lower biocide use and less depletion of the soil nutrient stock.

Table 1 Results of alternative scenarios (in % changes from base scenario)

Scenario		Gross margin	Net income	N-loss	BEII
Palm heart price	-25%	-37	-12	+47	-64
Environmental tax		-7	-2	-10	-99
Limit on biocide use		-6	-1	-8	-50
Capital availability	-20%	-11	-2	-18	-33
	-40%	-27	-8	-34	-51
	-60%	-46	-16	-55	-67

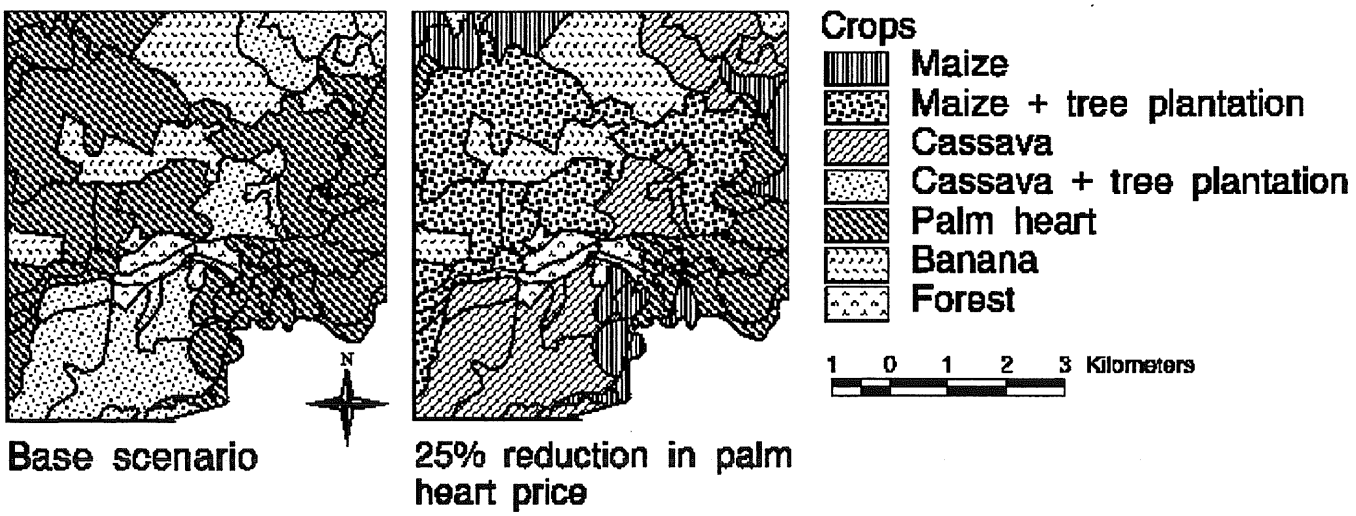


Figure 3 **Spatial distribution of LUSTs in base scenario and 25% decrease in palm heart price scenario in a small part of Guácimo county.**

3.6 Discussion

The USTED methodology for land use analysis integrates an LP programming model and a module for the storage of data on combinations of land use types and soils with specified technologies, linked to a GIS. The LP model accounts for differences in resource availability at the farm level. However, different farm types are not optimized in isolation, since each type has to take into account the sub-regional labor supply and demand. Since the methodology explicitly considers both income and sustainability related factors, it can be used to provide the policy maker with information regarding trade-offs between different goals. Due to data limitations, studies at the regional level generally have to take assumptions concerning certain variations at field and farm level. Soil types have to be generalized to general soil groups, farms have to be classified as farm types.

4. The farm level

4.1 Problem definition

Farmers depend on experiences from previous years, general literature and the information provided by extension services. The introduction of site-specific yield monitoring may bring important changes in the near future. Yield monitoring enables the farmer to register his harvest on a site specific basis, resulting in yield maps with differences on small distances (Robert *et al* 1995).

Bananas are the dominant crop in the perhumid coastal plains in the northeast of Costa Rica. The perennial crop is cultivated on large farms of over 100 ha and covers approximately 52,000 has (1994 figure, Corbana) or 10% of the area. Over 50% of the plantations are owned by large multinational companies with international research programmes to support farm management. National producers, however, rely on research carried out by Corbana (The National Cooperation of Banana Producers) as well as the international research community (often not very specific for the Costa Rican situation).

A research programme was started in cooperation with Corbana and Reposa, to deal with a number of very specific decisions being taken at the Rebusca farm. The Rebusca farm (84°01' E, 10°28'N) covers an area of 107 ha and started the cultivation of banana in 1991 during the large expansion of banana in the Atlantic zone of Costa Rica.

The decision support system, denominated BanMan (decision support system for banana management), was developed after the identification of two clear questions:

- How can we reduce the production costs of bananas? The production of bananas is generally considered to be extremely efficient. Nevertheless, it can be observed that the large fields (>100ha) are extremely heterogeneous in terms of soils and at the same time are managed as a single unit. Site specific management in other crops has proven to deal efficiently with this variation.
- How can we make accurate yield predictions? Bananas produced in Costa Rica are exported to the Europe and the USA. Farmers need to reserve space on the container ships three months in advance. This means that they need to predict their harvest. Underestimating or overestimating production has serious financial consequences. In case of an underestimate the farmer may not be able to export his product whereas in case of overestimation farmers still have to pay for the reserved containers.

4.2 *Methodology development*

BanMan is based on site specific yield monitoring. Yield maps enable farm management a better analysis of the production (backward looking) and by doing so improving future farm management (forward looking). To explain differences in crop performance, a detailed 1:5,000 soil survey was carried out. By overlaying the soil survey with the production maps, one may filter for differences in soil type (typically static properties, which can only be changed on the long term). Other differences, *i.e.* differences within the soil units are likely to be the result of planting material and/or management. Through the identification of site specific problems, farm management may improve these local limitations, improving the performance of the farm and at the same time reducing the costs.

Yield prognoses in BanMan are based on the hypothesis that the expected harvest $Y_{t=3}$ in three months time equals a maximum obtainable yield Y_{max} minus a constant C times the number of stress days in the past 6 months. Stress days are defined as the total number of days d where the soil is extremely wet ($\theta > \theta_{max}$) or extremely dry ($\theta < \theta_{min}$). In short:

$$\overline{Y}_{(t=0)} = Y_{max} - C*\left(\sum_{t=-6}^{t=0} (d|\theta > \theta_{max}) + \sum_{t=-6}^{t=0} (d|\theta < \theta_{min}) \right)$$

The daily estimates of the soil moisture content are based on daily rainfall and the LEACHM model (Wagenet and Hutson 1992). The regression parameters Y_{max}, C, θ_{min} and θ_{max} are estimated for each management unit in the plantation on the basis of specific model runs and production figures. After each harvest, an optimization model re-estimates the different parameters for the different management units. The estimates will improve by using the system, whereas the database is increasingly expanded.

4.3 *Data collection*

Besides a soil map and information on the farm infrastructure the decision support system does not require specific data. Yield data are generated through site specific yield monitoring which is an integral part of the system. Yield monitoring is increasingly being applied in grain crops where continuous grain flow monitors in conjunction with global positioning systems are installed on combines resulting in detailed yield maps (see e.g. Robert *et al.* 1995). In other crops, however, yield monitoring is not being applied through the lack of efficient register systems. Experiments revealed, however, that yields in beets, potatoes, and millet are similarly variable.

The banana crop is harvested almost continuously with 2-3 harvests per week. The banana bunches are harvested and transported to the packing plant by trains using a intensive cable system throughout the plantation. Each train comprises 25-30 bunches and originates from a specific location in the plantation (for a detailed description of the banana management in Costa Rica see Soto 1994). By registering the origin of the trains and weighing the bunches of each train (by using a balance in the cable) yields are monitored.

The soil map is the result of a detailed soil survey which has been carried out for almost all banana plantations in Costa Rica, whereas it is, in combination with a soil suitability classification for banana, obligatory for credit and subsidies.

4.4 *Calibration and validation*

BanMan mainly focuses on the identification of problem areas, *i.e.* areas where the production is low compared to the average for that soil type. Validation is carried out through detailed observations in the problem areas. Yield forecasting is a result of a regression procedure, which is automatically being checked by future production. The LEACHM model has not been calibrated specifically for the

local conditions in the plantations. However, whereas the results of the modelling are not directly used, but input in a regression equation this seems no problem. After 7 months of operation an r^2 of 0.78 has been reached. The r^2 is still going up as new yield data are included. Whereas climatic conditions in the last three months before harvest will remain uncertain, the system will always contain an error.

4.4 Application

BanMan has been operational for six months. Farm management is satisfied whereas it meets the original requirements of the project. Yield mapping functions properly (illustrated in Figure 4). BanMan indicates problem areas, and in most cases farm management is subsequently able to identify the problems. In most cases, the problems are linked to local water stagnation and poor planting material. Currently a program of replanting has been started.

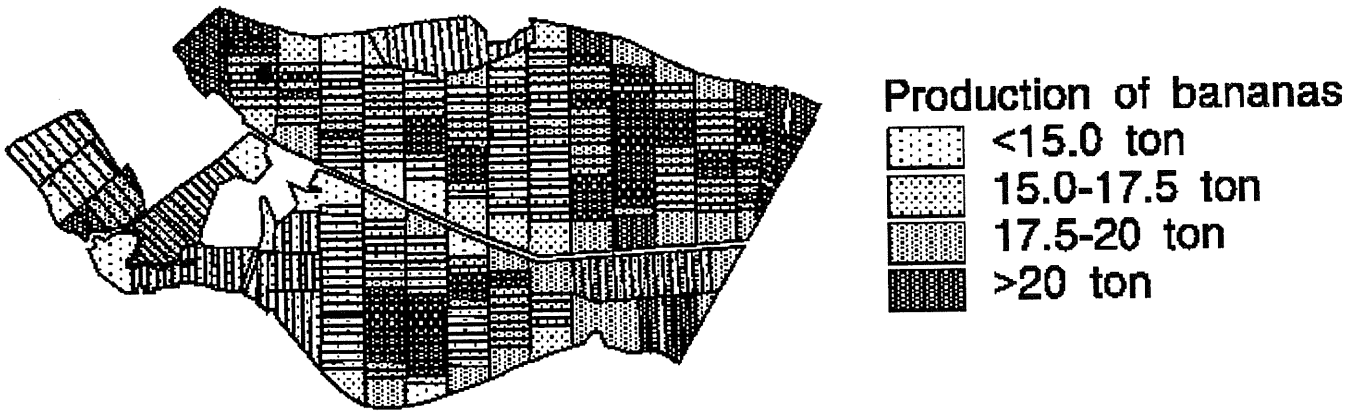


Figure 4 Yield map for the Rebusca plantation

4.5 Discussion

Although yield mapping seems to work well in bananas, a number of problems are encountered. Temporal problems originate from the almost continuous production. Yields will have to be aggregated in time, but the criteria still have to be defined. Spatial problems relate to the way the bananas are harvested and registered in the packing plant. Each train, the basic unit, includes 25-30 bunches originating from a relatively large area (±0.4 ha). Additionally the blocks do not have fixed boundaries and differ per harvest.

Changing farm management in terms of fertilization and pest and disease management may lead to misleading forecasts whereas the regression equations are based on past management and coinciding productions.

Banman is relatively easy to implement and in most cases will only require the installation of a balance in the cable near the packing plant.

5. Conclusions

- Notwithstanding the increase in new techniques that enable us to collect data relatively cheap and fast, we still face a data crisis. The extremely complex systems that we are trying to describe and understand require new approaches and techniques. GIS and modelling may contribute in the management and analysis of the data.
- The decision support systems that are being developed on different levels of detail encompass complex models and sub-systems. These systems can not be managed by the decision maker, resulting in a prevalence of black box approaches. Clearly, this goes along for the risk of mismanagement of these systems.
- Although scientist prefer a thorough analysis of the problem, followed by an intensive phase of experimentation and finally recommendations, decision makers need to take their decisions now. Science needs therefore to take a pro-active approach. Demand-driven research is OK, but we have to look ahead, what are the problems of the future and how can we tackle them when decision support is required.
- Typically, there is no single answer. For most problems there is a large array of possible alternative approaches each with its specific advantages and disadvantages. The decision support systems evaluate the different alternatives, the decision maker will take the final decision on the basis of his criteria.

6. References

Abel DJ, Kilby PJ and Davis RJ 1994 The Systems Integration Problem. International Journal on Geographical Information Systems 8:1-12.

Belder M 1994 Land Use and Land Use Dynamics in the Atlantic Zone of Costa Rica. Guápiles (Costa Rica): Atlantic Zone Programme (CATIE-WAU-MAG).

Burrough PA 1989 Modelling Land Qualities in Space and Time: the Role of Geographical Information Systems. In: Bouma J and Bregt AK (eds) Land Qualities in Space and Time. Wageningen: PUDOC, 317-320.

DGEC, 1987 Censo Agropecuario 1984. San José: Ministerio de Economia, Industria y Comercio.

Fotheringham AS, Rogerson PA (eds) 1994 Spatial Analysis and GIS. London: Taylor and Francis.

Hassan HM, Hutchinson C (eds) 1992 Natural resourceand environmental information for decisionmaking. Washington: The World Bank.

Herrera W and Gómez LD 1993 Mapa de unidades bioóticas de Costa Rica. San José: Instituto Geográfico de Costa Rica.

Jansen DM and Schipper RA 1995 A Static, Descriptive Approach to Quantify Land Use Systems. Netherlands Journal of Agricultural Science 43: 31-46.

Jansen DM, Stoorvogel JJ and Schipper RA 1995 Using Sustainability Indicators in Agricultural Land Use Analysis: an example from Costa Rica. Netherlands Journal of Agricultural Science 43: 61-82.

Lutz E and Daily H 1991 Incentives, Regulations and Sustainable Land Use in Costa Rica. Environmental and Resource Economics 1: 179-194.

Maguire DJ, Goodchild MF and Rhind DW (eds) 1991 Geographical Information Systems. Harlow (UK): Longman Scientific & Technical.

O'Kelly ME 1994 Spatial Analysis and GIS. In: Fotheringham AS, Rogerson PA (eds) 1994 Spatial Analysis and GIS. London: Taylor and Francis, 65-79.

Robert PC, Rust RH and Larson WE (eds) 1995 Site-Specific Management for Agricultural Systems Madison: ASA-CSSA-SSSA.

Schipper RA, Jansen DM and Stoorvogel JJ 1995 Sub-Regional Linear Programming Models in Land Use Analysis: a Case Study of the Neguev Settlement, Costa Rica. Netherlands Journal of Agricultural Science 43: 83-109.

Stoorvogel JJ 1995 Linking GIS and Models: Structure and Operationalisation for a Costa Rican Case Study. Geoderma 43: 19-29.

Stoorvogel JJ 1996 Geographical Information Systems as a Tool to Explore Land Characteristics and Land Use with Reference to Costa Rica. PhD Thesis, Wageningen: Wageningen Agricultural University.

Stoorvogel JJ, Schipper RA and Jansen DM 1995 USTED: a Methodology for a Quantitative Analysis of Land Use Scenarios. Netherlands Journal of Agricultural Science 43: 5-18.

Wagenet RJ and Hutson JL1992 LEACHM, Leaching Estimation and Chemistry Model, Version 3. Ithaca: Cornell University.

Wielemaker WG and Vogel AW (eds) 1993 Un Sistema de Información de Suelos y Tierras para la Zona Atlántica de Costa Rica. Report No. 22 (Phase 2), Turrialba: Atlantic Zone Programme (CATIE-UAW-MAG).

4. Data and scale

Pitfalls in large scale, real time data usage

Reduction of models becomes relevant for real-time data usage. A combination of deterministic (mechanistic) and stochastic (statistical) models is used to achieve improved predictions. Neural networks and geographic information systems are used in relation to agricultural models. The search for relationships between variates in large datasets is laborious and one usually overlooks certain aspects. 'Data mining' tries to find all kinds of relationships in large databases.

4.1 Finding and using data for small scale applications of agrometeorological models such as yield forecasting at a European scale

Paul Vossen

Joint Research Centre of the European Community, 21020 - Ispra (VA), Italy

The application at the scale of large regions and countries of agrometeorological models for yield forecasting, goes along with a number of problems and constraints resulting from the non adequacy in input data requirements between the scale of application and the local scale at which the models were initially developed. Using the example of the European Commission's system for crop yield forecasting as a reference, this paper presents and discusses the specific aspects related to the collection and processing of the input data such as meteorological and soils data, the estimation of derived parameters such as potential evapotranspiration and global solar radiation and the statistical validation of model outputs. Special attention is paid to the constraints related to the introduction of region-specific crop knowledge in a simulation model, the interpolation of input data, the validation of the outputs and the limitations in the use of the results. These limitations are such that almost no downscaling of the system outputs is permitted and that, what initially starts as a deterministic simulation of crop biomass and grain development, eventually results in indices whose inter-annual variability varies along with the inter-annual variability of grain yield at a regional or national scale, but not necessarily in an absolute simulation of grain and biomass.

1. Introduction

Since 1993, the European Commission is operating a system for yield forecasting at the level of the E.U. countries and of large sub-national regions. The crops of interest are the major cereals, the oil and protein crops and olive and grapevine. The results are valid at the level of the EU as a whole and of the EU Member States. The system is operated by the Joint Research Centre as part of the Project for the Application of Remote Sensing to Agricultural Statistics created in 1988 by the EC Council of Ministers. The main clients of the Project are the EC's Directorate General for Agriculture and the European Statistical Office (EUROSTAT). Since 1996, a similar system is being developed for the central and eastern European countries and for the Maghreb area.

The following implementation strategy was opted for:

1. Proven methods that relate satellite imagery to quantitative crop yield forecasts at the national or regional scale not yet being available, the crop yield forecasting would - at least in the initial stages - mainly be based on more traditional methods such as agrometeorological crop growth simulation models, the aeropalynological method and technological time trend extrapolation. The latter techniques have indeed shown their applicability at a regional or national scale (Brochet et al, 1975; Dagneaud et al, 1981; FAO, 1986; Haun, 1982; Motha et al, 1986; Place and Brown, 1987; Sakamoto, 1978; Thompson et al, 1981; U.S.D.A., 1987; Vossen, 1990a;).
2. Simultaneously, the possible use of satellite remote sensing techniques for crop monitoring and yield forecasting and for the improvement the precision and spatial resolution of the agrometeorological model outputs would be investigated.

This report presents how the previous strategy was followed for finding and using data for quantitative yield forecasting at national scale.

2. The models for crop yield forecasting at the scale of the EU

Three type of models have been developed tested and implemented:
- Deterministic crop growth simulation models for annual crops and grassland;
- The aeropalynologic method for grapevine;
- A simple water balance model for the perennial grapevine and olive tree crops;

Each of the models is briefly summarised hereafter.

2.1 Deterministic models for annual crops and grassland

For annual crops, deterministic agrometeorological crop growth simulation models each derived from the WOFOST model (van Keulen and Wolf, 1986; van Diepen et al, 1989; Supit et al, 1994) have been validated for cereals, grain maize, rice, soybean, field bean, potato, sugar beet, oilseed rape and sunflower (Vossen, 1990, 1992). The models are driven by a combined energy balance / water balance module which compares real transpiration with calculated potential transpiration through a light interception / CO2-assimilation / water requirements / water availability sub-module. The models are calibrated to make them region-specific (e.g., initial dry matter at emergence, mean planting date, calibration of the length of phenological stages as a function of sums of temperatures, etc.) on the basis of site specific field data available from various research institutes in the EC and on the basis of the crop knowledge bases which were specially established for this purpose (see the annex in the Literature list). Both adaptation and calibration have been reported on in Boons-Prins et al (1993).

For grassland, an approach for the areawise monitoring at the level of the EU of the state of grassland was derived from 2 existing models. For the northern part of the EU, the LINTUL (Spitters, 1987; Kooman, 1995) was adapted and validated. This model applies to grasslands with winter dormancy and restart of the vegetative growth after winter. In the simulation of biomass production, the exploitation of the grass by cattle or farmer is taken into account, as well as the fact that the grassland almost never flowers. For the southern parts of the EU, with mainly annual grasses, the ARID CROP model (van Keulen, 1975) was validated. The grass cover, composed of a mixture of annual grasses, generally passes through a global cycle emergence, vegetative growth, reproduction, ripening and death. The software module which integrates those two approaches is referred to as the LINGRA Model (Bouman et al, 1996.).

2.2 Simple water balance models for the perennial grapevine and olive tree crops

For most perennials in the EC, deterministic growth simulation models are not available and their environmental growing conditions, phenological calendar, and farming practices (including the main cultivated varieties) have, in most cases, only partly be described in the literature. The same applies for the soil types on which the various perennials are grown and for the rooting depths which are highly variable according to the general climate, soil type and farming type (irrigation, for example). It is thus difficult to propose for perennial crops of major economic importance such as olive tree and grapevine, a simulation model that needs -as an input or for model validation- information such as sprouting period, available water capacity, period of flowering, etc.

A simple water balance model for the monitoring and forecasting of the national yields of olive oil and grapevine had therefore also to be developed. It is referred to as the OLIWIN model

(Bories, 1996) and is largely based on Vossen (in: Riou, 1993; Vossen and Rijks, 1995), Doorenbos and Kassam (1979) and Doorenbos and Pruit (1984). In its present state, the model estimates the soil moisture content and its variations on a monthly basis from April to September, which are the average sprouting and ripening months for the EC To estimate the water requirements and water consumption, a crop factor K_c expressing the ratio between water requirements during a development stage and potential evapotranspiration, and an estimate for the profile available water capacity are used (King and Le Bas, 1994). The soil moisture at the end of March is estimated from the October-March rainfall. However, additional data collection is ongoing, to make the model more region- and soils specific, and to take into account variable dates of sprouting and flowering.

2.3 The European Commission's Crop Growth Monitoring System

The 3 models WOFOST, LINGRA and OLIWIN are integrated in the EC's Crop Growth Monitoring System (Vossen, 1990, 1992; Hooijer and van der Wal, 1994; Vossen and Rijks, 1995). This system has been designed to use daily weather data (rainfall, temperature, vapour pressure, 24-hour wind run, sunshine duration or cloud cover) that are interpolated to a regular grid covering the whole of Europe. It permits to run the various models for the different soil types on which a crop is cultivated, taking into account region-specific average planting or sprouting dates and crop parameters. For yield forecasting at the national scale, model outputs are weighted according to the relative importance of the soil types and of the crop acreages in the various regions of a country. The validation procedure for forecasting purposes is done once every month, between the month of expected planting and the harvest month. In this procedure, time trend and model outputs at the end of a given month, are simultaneously regressed against historical yield series. This approach assumes that the weather conditions during the remaining part of the season will be "normal". The approach also implies that the forecasts become more and more reliable according as the season progresses.

The following outputs are available as tables or cartographic products a few days after data acquisition: biomass and grain production, under the actual rainfall conditions and as if all required moisture were available; estimated actual soil moisture reserve and state of advancement of the cycle; derived meteorological parameters such as sums of temperature, climatic water balance, etc.; percentage departure from the long term mean given decade or period within the growing season, for all of the previous indicators.

2.4 The aeropalynologic method for olive tree and grapevine

Simultaneously, but only for grapevine, the aeropalynologic method for the forecasting at flowering of the potential grapevine production was further validated and implemented at the scale of the wine producing regions of the EU. The method is referred to as the POLLEN method (Besselat and Cour, 1990) and is based on the assumption that the number of pollen liberated into the atmosphere during the flowering period of certain crops, is a good indicator of the yield *potential* of the crop. In fact, the intensity of the pollination may in some cases integrate the growing conditions previous to flowering. Further details on this method and its results are provided in Besselat and Cour (1996).

3. Inventory of problems related to the small scale application within the EU context of models for yield forecasting.

The implementation of a system for the production of timely yield forecasts at a national or regional scale, goes along with the following problems:

1. A change of scale: from site-specific input information, to assessments or forecasts that are valid for large regions or countries;

2. A limited precision of the input information. For example, the information contained in the EC's 1:1.000.000 soil map (CEC, 1985) is already a generalised synthesis of the original field observations; the meteorological conditions at the level of the weather stations are not always comparable with the weather conditions in the farming areas, nor with the weather conditions that were recorded at the research sites were the relations between crop growth and weather conditions were established; planting dates, crop cycle lengths, etc., can only be estimated with a precision of approx. 5 to 10 days; part of the available input information is only valid within national boundaries (e.g., the soil classification criteria and the type and installation characteristics of the meteorological instruments).

3. Non-availability of part of the input information. For example, the exact profile available water capacity of the soil (PAWC); the depth of the soil and rooting depth; the soil occupation by the various crops in a region; the relative importance of the various crop varieties that are planted in a given region.

4. A limited spatial resolution of the input information. For example, the number of reliable, daily reporting weather stations in the EC is limited to approximately 650; complete and reliable time series of crop yield, needed for the validation of any model or method, are only available for the E.U.; for what concerns crop growth conditions, farming practices, etc., information is in most cases only available for very large regions of several 1000nds of square kilometres.

5. Knowledge on the relations between crop yield and the agro-pedo-meteorological growth conditions, is often only available at the level of individual research sites and it was never integrated into information that is valid at the regional scale.

6. Non timely availability of the information on farming practices and farmers decision making mechanisms, for example the planting dates during a given season, the fertiliser doses, etc.

From what precedes, two conclusions must be made: first of all, quantitative yield forecasts based on the above listed limited information, can never be valid for a specific locality: they can only be valid or reliable for (very) large areas such as countries or regions, provided the information and model outputs were first carefully weighted for the relative importance, within the large region, of soil types, groups of varieties, common farming practices, etc. Secondly, the yield forecasting methods will have to be validated per country and/or per large region; these validations have to be done with the help of statistical techniques (regression, analysis of variance), to quantify effects such as the technological trend and to take into account the variable impact of environmental conditions according to the farming level and the farming conditions.

4. Activities undertaken to solve part of the listed problems

The above listed problems and constraints are mainly related to the non adequacy in input data requirements between the scale of application and the local scale for which the models were initially developed. To remedy at least partly to them, a number of applied research activities was undertaken.

4.1 Establishment of agro-pedo-meteorological crop inventories.

For most of the crops that are commonly cultivated in the E.U., literature and expert-knowledge based information was collected on (1) statistics on cultivated surfaces and yields, (2) regionalized phenological crop calendars and (3) the environmental requirements of the crops. The list of available crop inventories is given in annex. Each inventory contains:

a. general crop data such as rooting depth, base temperature, crop coefficients expressing the ratio between crop waiter requirements and potential evapotranspiration during a givens development stage, etc. This information was mainly obtained through literature searches.
b. agronomic data, such as the earliest and latest dates of harvest in a region, the most common farming practices, the major varieties grown in a region, maximum altitude at which a crop is grown, technological evolution over the last years, etc. This information was mainly obtained from postal questionnaires.
c. detailed physiological information from individual trials, such as the duration of phenological trials. This information was also mainly obtained through literature searches.

Limitations of the knowledge bases (Russell, 1990) One of the important problems is that information is often available at the wrong scale. For example, the sowing dates for a region are often only known for a small sample of fields which may or may not have been biased. Information about the representativity of the subsamples, needed for accurate scaling up from the side-specific information to a regional or national level is usually unavailable. Similarly, crop modelling parameter values obtained from individual trials will differ from those that would have been obtained if a complete enumeration had been achieved. This is of particular importance wherever the relations are not linear (de Wit and van Keulen, 1987). Also, certain information is frequently unavailable for certain parameters and in certain regions or countries. In such cases a 'best guess' needs to be made.

It is thus clear from the preceding comments that both the time and space resolution of the outputs will be limited: first of all, it is not indicated to exploit daily model outputs. For example, estimates of the date of occurrence of phenological stages, most likely will be not precise and probably only suitable to be expressed in relative terms such as departures from normal or "very late", "late", "normal", etc. Secondly, using mapped crop state monitoring products for spot evaluations is not permitted. The output products are best not presented as isolines, but assigned to grid cells (for the qualitative monitoring) or as tables with summaries at the national scale (for the quantitative yield forecasts). In practice a resolution of 50km x 50km is used which corresponds also to the resolution to which the daily meteorological data, available in real time from the European network, can be interpolated with confidence (see further).

4.2 Estimation of solar radiation and potential evapotranspiration

Both PET and solar radiation are key input variables in many crop yield forecasting models. However, the regional validity of most estimation methods is in most cases limited. Methods for

the calculation of both parameters, that at the same time result in improved estimates as compared to the use of an already existing, but only locally validated formula and are valid for all of the EC-regions have therefore been developed by Jones et al (1991), Choisnel et al (1992) Hough and Parker (1992) Supit (1994) and Supit and van Kappel (1997). Their approaches could only be validated on the basis of limited sets of observations. For solar radiation: 100 stations in the EU, Central and Eastern Europe, the Near East and North Africa countries for which observed solar radiation data were available. For PET: 6 sites for which observed PET data were available (Wallinford in the UK Gembloux in Belgium, Guyancourt in France, Avignon-Montfavet in France, Roma and Bari in Italy. It has thus to be noted that large areas are devoid of data points, so that in some parts of the maps the isolines are rather subjective. However, given the scarcity of literature in this fields, the here presented results are the presently the most indicated ones one can use:

1. ***For the estimation of global radiation*** from visual observed cloud cover at meteorological station level, Ångström (1924), Hargreaves (1985) or Supit & Van Kappel (1997) are being used, depending upon which surface meteorological data are available in real time:

   a) If daily sunshine duration is available (Ångström, 1924; preference choice):

   $$R_{g,d} = R_{o,d}(c_a + c_b \frac{n}{D}) \quad (1)$$

   b) If both cloud cover and temperature data are available (Supit & Van Kappel, 1997):

   $$R_{g,d} = R_{o,d}\left\{c_a\sqrt{T_{\max} - T_{\min}} + c_b\sqrt{1 - cloud/8}\right\} + c_c \quad (2)$$

   c) If only temperature data are available (Hargreaves et al.,1985):

   $$R_{g,d} = c_a R_{o,d}\sqrt{T_{\max} - T_{\min}} + c_b \quad (3)$$

   where $\underline{R}_{g,d}$ is the observed daily global radiation ($MJ.m^{-2}.d^{-1}$), $\underline{R}_{o,d}$ are daily values of the extra terrestrial radiation ($MJ.m^{-2}.d^{-1}$), $\underline{n}$ is the daily sunshine duration (h), $\underline{D}$ is the astronomical day length (h) and $\underline{c}_a$, $\underline{c}_b$ are empirical constants.

   The above 3 approaches were tested for more than 100 stations ranging from Northern Finland to Southern Egypt. Generally the Ångström method provides the best estimates. However Hargreaves et al. (1985) and Supit and van Kappel (1997) are most acceptable alternatives when no daily global radiation and sunshine duration observations are available, which is mostly the case for the daily reporting meteorological stations.

2 ***For potential evapotranspiration*** the original Penman formula (1948), is being used, but by using the above method for the estimation of solar radiation (Choisnel et al, 1992). These modifications to the original Penman formula give acceptable estimates of PET totalled over 10-day periods, between March and October. During this period, the mean deviation between daily lysimeter observation and estimated PET, the mean of the absolute deviations and the mean quadratic deviation are respectively 0.1, 0.55 and 0.71 mm/day (average of the mean values of the 6 lysimeter stations). Although the Penman-Monteith approach give comparable results in terms of difference between observed and estimated PET, it has not been implemented in the system because part of the necessary input information (crop canopy resistance) can not (yet) be made available for use on a regional scale as the system requires.

4.3 Soils data

For use at the European scale as a whole, only the existing 1:1.000.000 EC soils data base (C.E.C., 1985, FAO classification) is available. This data base was initially a compilation of national maps, which goes along with problems such as lack of cross-boundary harmonisation, differences in classifications, errors in digitalisation and map attributions, different levels of precision in the description and aggregation of soil types into soil mapping units, etc. The soils data base was therefore corrected, harmonised, updated and completed with information available in the original archives that were used for its elaboration. This eventually resulted in Version 3.1 of the EC soil map (King et al, 1993). The soils data base as such is mainly used in conjunction with the crop knowledge bases, to identify the areas where a given crop can possibly grow. From this version, a Profile Available Water Capacity Map has been derived (King et al, 1993). The latter data base contains for each soil typological unit (STU) in a given soil mapping unit (SMU): a PAWC estimate for the first 3 main horizons (or less if an non penetrable horizon is present) obtained from a number of pedotransfer rules, an estimated maximum possible rooting depth, an indication whether moisture could become available through capillary rise and the importance of the STU expressed as a percentage of the total area of the SMU.

4.4 Meteorological data and their processing.

The number of stations for which actual daily data of rainfall, temperature, vapour pressure, wind speed and cloud cover or sunshine duration are available for the EU as a whole is approx. 650. But for only 350 stations long term historical series are available. These networks are relatively sparse considering the large range of reliefs and elevations. As at present no proven techniques for the interpolation of daily meteorological data from the synoptic network and applicable to the EC as a whole exist (such techniques are only available for a limited number of variables and only for a few countries), a new technique was developed for the interpolation of data from the existing network of meteorological stations on a regular grid. (Van der Voet et al., 1993).

The technique can be summarised as follows: the values of the variables are estimated by means of simple inverse distance weighting of the data of an optimum set of stations, surrounding the centre of a the cell of a regular grid mesh over Europe. A set of surrounding stations is selected on the basis of criteria of similarity between the station characteristics and the grid cell to which is being interpolated. The criteria of similarity are: proximity, similarity in terms of altitude, distance to the coast, the position relative to climatic barriers such as the Alps and the Pyrenees, the degree to which the selected stations are surrounding the location of a grid cell and the number of stations in a set. Initially, the 7 nearest stations surrounding a grid cell are selected. For each combination of 4 of the 7 chosen nearest stations, a 'suitability score' is then calculated on the basis of a number of empirically validated weights attributed to the values taken by the similarity criteria. And a simple interpolation to the centre of each grid cell is then carried out from those stations. It may be mentioned that increasing the number of chosen stations beyond 4 does significantly increase the calculation time but does not increase anymore the quality of the results.

In the case of rainfall, one single station, with the lowest difference score is selected. The use of more than one station may improve the mean prediction error, but it also overestimates the number of wet days considerably, hence modifies the temporal distribution of rainfall which is of major importance for crop modelling.

According to the results of the analysis of the errors of estimate, the meteorological conditions that are relevant for crop growth, are fairly accurately estimated at a spatial resolution of 50 × 50 km.

Figures 1 and 2 give the network of the meteorological stations providing daily data in real time and illustrates the 50 × 50 km grid over the EU to which both the historic and actual data are

interpolated. Only those stations are represented from which on a daily basis, the input data required for running the models can be obtained. In practice it represents the SYNOP network with at least 4 available observations per day. The station list was submitted to an expert group, for screening the stations that should not be used for agrometeorological crop state monitoring nor alarm warning, mainly because of their location. (For example stations located on a small island or at the edge of the coast line, surrounded by the sea, in the centre of a very large city or in too close proximity of an airport runway were excluded.) To identify the stations that can be used, the data from all available SYNOP stations are once a year submitted to quality tests (Meteo Consult, 1991) and the stations that reported too irregularly or too unreliably during the previous year are excluded for use during an actual season.

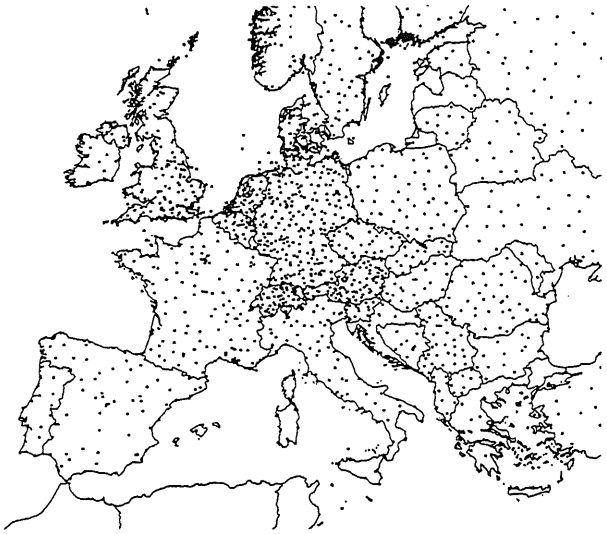


Figure 1: the network of actually available meteorological stations (providing daily data in real time)

The consequences of the fact that number of stations for which actual data are available is far higher then the number of stations for which long term historical series are available are partly compensated for by the fact that crop model runs (and the calculation of other derived parameters such as sum of temperatures, climatic water balance, etc.) are, both for the historical as for the actual data, performed at the level of each grid cell. To run the models at the level of each cell is necessary, because of the non-linearity of a majority of the relations between plant growth and agrometeorological growing conditions and because of the non-linearity of spatial evolution of many of the meteorological and pedological conditions (de Wit and van Keulen, 1987).

From what precedes follows again, that the yield forecasting methods that use meteorological information, can not give reliable outputs at the local scale and that the agrometeorological data or model outputs obtained for a given station or grid cell, can only serve as areawise or regional indicators of the quality of a cropping season. To use the information for quantitative yield forecasts, it is necessary to calculate regional averages of local outputs, that are preferably weighted for the area occupied by the crop in a region and the occupation of the different soil types by a crop.

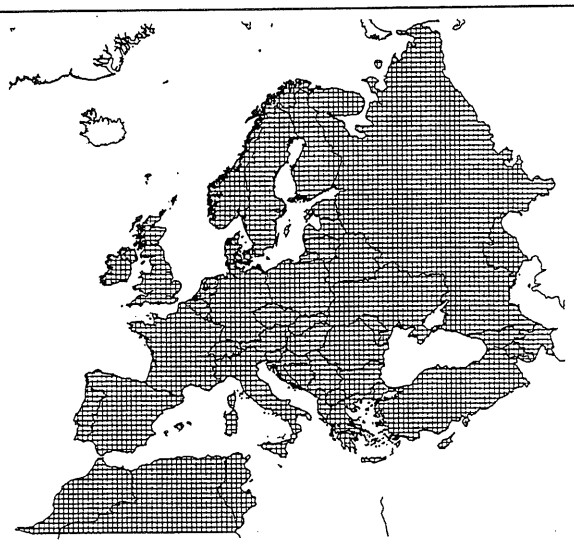


Figure 2: the 50 × 50km grid over the E.U. to which meteorological data are interpolated.

4.5 Time trend analysis.

Crop yields largely depend upon the farming practices and they may vary more or less rapidly, according to the techniques that farmers adopt and their flexibility to introduce innovations. However, this information is not timely available for yield forecasting purposes at a national scale. Crop growth modelling is thus done on the assumption that environmental growing conditions other than weather are stable and do not change from year to year. This assumption is obviously not true for farming practices such as fertiliser application and variety choice. On the other hand, a correctly quantified possibly existing time trend of regional and national crop yields, may to some extent be used as a replacement for part of the farming conditions that affect crop yield but that are not easily available at the time the crop yield forecasts are made. The use of a time trend includes also that, for a given year, changing farming practices from one site to another within the area, compensate each other and result in a stable general situation for the region as a whole. A consequence will be, again, that the yield predictions can only be valid for the areas for which the trend (existing or not) was quantified.

The form itself of the trend is different from crop to crop, from region to region and, often, from period to period within a given time series. The correct description of time trends is thus extremely important as part of the development of methods for crop yield forecasting. Palm and Dagnelie (1993) carried out the study on this subject for the different member states and for most of the crops. Their main conclusion was that, although more complicated models resulted in a (very) limited gain of precision, the technological trend was best approached by the quadratic function:

$Y = a + b*T + c*T^2$, adjusted by stepwise regression. (4)

4.6 Validation of the model outputs.

The general procedure that is followed for the validation of the agromet models as been reported on in detail in (Vossen, 1990, 1992). It basically is a regression analysis of the series of regional and national yields against the model outputs and (possible) time trend:

$$YIELD = constant + \Psi \{ CROP\ MODEL\ OUTPUTS \} + \Theta\{TREND\} + error \quad (5)$$

The possible time trend and model outputs are simultaneously taken into account and validated on

the basis of the hypothesis that weather conditions can suppress or modify the expression of a time trend (Nyankori, 1979; Vossen, 1990a). The tested trend function is of the form $+/-b*T +/- c*T^2$, where b and/or c are zero. Usually, analyses including time trends are carried out for long time series k of 20-30 years. However, because of the recent agricultural policy of the EC, aiming at a reduction of the total production of and/or the subsidies for certain crops (e.g., wheat), the length of the time series used for the statistical model validation has been set to $k = 9$. This is length long enough to give a sufficient number of degrees of freedom in a regression analysis with at least 2 and possibly up to 5 independent variables.

These variables are: (a) the trend; (b) one of the following outputs, which are interdependent: biomass, grain, leaf area index, development stage; (c) one (possibly more?) of the following variables: the ratio between theoretical water requirements and real water consumption, the rainfall in excess to the water requirements (e.g., during a given development stage), the climatic water balance during winter (rainfall minus P.E.T). The possible number of variables here is almost unlimited and the selection must be careful and based upon objective considerations reflecting real constraints to crop production.

At present, only the above two variables (a) and (b) have been validated. For the forecasting, one, the statistically most significant, model output is selected. Model precision and stability and the trustworthiness of the forecasts are evaluated with the technique of independent estimates, based on the model validation for the previous 9 years.

What precedes has as major consequence that the simulation model outputs, in reality, are not anymore considered as reflecting a reality (for example a truly simulated biomass) but have become indicators whose inter-annual variability possibly accounts for a more or less important part of the inter-annual variability of national or regional crop yields. These indicators, because of all the constraints linked to data availability and the resulting imposed simplifications and work approaches, still have a biological basis, but are not anymore an exact representation of a field reality.

5. Results

From Table 1 appears that the May-September yield predictions are in most cases more reliable and earlier available then the Eurostat figures which are mainly derived from time trend analyses completed by inputs provided by national statistical offices.

Further in-depth analysis (Vossen and Rijks, 1995; Genovese, 1997) shows however that the time trend component is often more significant then the agrometeorological model outputs. On the national scale, equation [5], including the time trend component, significantly (1% level) account for inter-annual yield variability for 77% of the crop.member state combinations. In only 53% of the combinations, crop simulation model outputs significantly improve the result. This does however not necessarily mean that in all those cases the prediction errors also significantly decrease. It does however mean that in a large majority of cases, the system contributes to an improved monitoring and yield forecasting of crops, especially on the national scale, as compared to the use of time trend extrapolations alone or assessments based on a subjective interpretation of general weather conditions.

Table 1 (from: Genovese, 1997) compares the May-September yield forecasts at the level of the EU and for the years 1993-1996, with the estimates available from the European Statistical office.

crop	year	May	June	July	August	Sept.	Final
wheat	1993 (E)	5.60	5.46	5.47	5.46	5.21	**5.32**
	1993 (M)	5.30	5.21	5.18	5.21	5.20	
	1994 (E)	5.45	5.38	5.55	5.57	5.47	**5.42**
	1994 (M)	5.23	5.26	5.22	5.18	5.18	
	1995 (E)	5.68	5.73	5.33	5.35	n.a.	**5.31**
	1995 (M)	5.39	5.38	5.34	5.35	5.35	
	1996 (E)	n.a.	n.a	n.a.	n.a.	5.84	**5.88**
	1996 (M)	5.50	5.41	5.37	5.47	5.63	
Barley	1993 (E)	4.10	3.96	3.97	4.14	4.17	**4.21**
	1993 (M)	4.10	4.17	4.14	4.06	4.07	
	1994 (E)	3.88	3.93	3.95	4.04	4.05	**4.01**
	1994 (M)	4.00	4.10	4.00	4.00	4.00	
	1995 (E)	4.47	4.09	3.80	3.83	n.a.	**3.98**
	1995 (M)	4.00	3.90	3.90	3.90	3.90	
	1996 (E)	n.a.	n.a.	n.a.	n.a.	4.62	**4.63**
	1996 (M)	4.16	4.16	4.18	4.29	4.47	
Maize	1993 (E)	7.70	7.68	7.68	7.53	7.84	**7.96**
	1993 (M)	7.20	7.17	7.16	7.55	7.58	
	1994 (E)	7.87	7.67	7.59	7.42	7.86	**7.69**
	1994 (M)	7.80	7.80	7.60	7.60	7.50	
	1995 (E)	8.05	8.09	7.88	7.53	n.a.	**7.90**
	1995 (M)	n.a.	7.60	7.70	7.50	7.50	
	1996 (E)	n.a.	n.a.	n.a.	n.a.	8.30	**8.32**
	1996 (M)	7.70	7.76	7.80	7.85	8.22	

Table 1: Summary of the 1993-1996 monthly yield predictions established by the MARS Project (M) and compared with the predictions available from Eurostat (E). Yield figures are given in tons per hectare. (Genovese, 1997).

6. Discussion

The relative weak contribution of the agrometeorological model outputs to the quality of the yield forecasts is somewhat disappointing. The reasons are probably double:

a) The importance of the time trend function, which in reality is not one single smooth line of increasing (decreasing) yields corresponding to gradual but steady improving (deteriorating) farming practices. In fact, the mathematical expression of the trend represents as a line, the average affect of improvements, but which are introduced as irregular jumps. This fact may in certain cases mask the effects in the regression analysis of agrometeorological growing conditions; for example, if for a given country a strong positive technological trend is present, then favourable agrometeorological growing conditions during a given season, combined with and unchanged level of farming practices, may result in an "artefact" that the time trend extrapolation appears as a better yield predictor. Moreover, as a result of advanced farming practices and the increasing capacity of farmers to adjust their crop husbandry to actual weather conditions, the inter-annual variability of yield data is already relatively limited. The following percentage coefficients of variation can for example be given for the 1975-1989 wheat yields: Germany (14.7), France (16.1), Italy (7.24), The Netherlands (14.2), Belgium (18.6), Luxembourg (21.1), United Kingdom (17.3), Ireland (21.0), Denmark (14.8), Greece (12.7), Spain (24.9), Portugal (26.3).

b) Shortcomings of the crop simulation model and the system as a whole, for example:

- at present, the water balance module does not take in to account possible capillary rise of

soil moisture which can become available to a crop; the present water balance model is also a one-layer model.
- in the present version of the model, the planting date is fixed and has been put equal to the inter-annual mean value;
- the interpolation approach has without any doubt shortcomings, especially for what concerns the estimation of rainfall and radiation (from sunshine duration or visual cloud cover observations at station level);
- the fact that, at present, maximum two variables are taken into account (one trend variable and one model output);
- errors in the model parameters (e.g., the initial dry matter at emergence; sum of temperatures required to reach a phenological stage, repartition of biomass towards roots leaves, stems, grain, etc.).

c) The quality of the series of yield data that were used for validation. The precision as such of the data is, according to specialists of the European Statistical Office, of the order of +/- 5% in the northern EC countries and less in the countries with a Mediterranean climate. Moreover, the precision and reliability of the national and regional statistics varies according to the fact whether those statistics are aimed to provide a good estimate at the regional level (e.g., Germany) or at the national level (e.g., Italy).

7. Perspective: the use of remotely sensed information.

At present, the entire system is based on information generated from surface observations. However, this has as two major disadvantages: first of all, the input data have to be obtained from the relatively sparse European meteorological network, which does not allow to retrieve reliable weather information for areas smaller than approx. 50 km × 50 km. Secondly, because ground observed information (farming practices, planting date, stand density, variety, soil type, crop development stage, etc.) can not be made timely available for the whole of the E.U., the model outputs themselves for each 50kmx50km grid cell have a limited precision. As a result of both facts, the quantitative yield forecasts can only become reliable when they are applied to large regions or countries. This approach was initially (in 1988) opted for because of the non availability - at least within the E.U. with its typical agricultural pattern- of proven techniques that operationally relate remotely sensed information to national yields.

However, it is clear that the introduction of remotely sensed information into the models will significantly improve the spatial representativity and the trustworthiness of both the model inputs and outputs. Such improvements are expected from an appropriate use of VEGETATION, METEOSAT and/or AVHRR information, especially for the following applications (see also: Loudjani, 1995):

a. the spatialisation of inputs such as meteorological data and of certain outputs such as, for example, drought severity indicators;
b. the introduction of land utilisation information into the system;
c. an improved fitting of crop cycle lengths and major phenological stages, for example by an improved assessment of the start of the development of the vegetation after winter.
d. an improved, more reliable and spatially correct depiction of alarm situations such as droughts, extreme colds and abnormal high temperatures;
e. the direct input of estimated global solar radiation from METEOSAT imagery and of remotely sensed surface temperatures and evaporation.

8. References

Ångström A 1924 Solar and terrestrial radiation. Quarterly Journal of the Royal Meteorological Society 50:121-125

Besselat B, Cour P 1990 La prévision de la Production Viticole à l'Aide de la Technique 'Capture de Pollen'. In: Proceedings of the Conference on The Application of Remote Sensing to Agricultural Statistics, Varese, Italy, 10-11 October 1989, pp 261-268. Publ. EUR 12581 of the Office for Official Publications of the EC, Luxembourg

Besselat B, Cour P 1996 Guide pratique: Elaboration d'une prévision de récolte à partir du dosage pollinque de l'atmosphère. Publ. EUR 16422 de l'Office des Publication Officielles de la Commission Européenne, Ispra

Boons-Prins ER, De Koning GHJ, Van Diepen CA, Penning de Vries FWT 1993 Crop specific simulation parameters for yield forecasting accross the European Community. Simulation Reports N° 32. Centre for Agrobiological Research (CABO) and Joint Research Centre of the EC Wageningen

Bories L, 1996 The OLIWIN Project: Agro-meteorological models for the estimation of olive and grapevine yield. Internal report. Space Applications Institute, Joint Research Centre, Ispra

Bouman BAM, Schapendonk AHCM, Stol W, van Kraailingen DWG Description of the growth model LINGRA as implemented in CGMS. DLO Research Institute for Agrobiology and Soil fertility. Series Quantitative Approaches in Systems Analysis N° 7, Wageningen

Brochet P, Gerbier N, Bedel J 1975 Contribution à l'étude agrométéorologique du mais: applications à la prévision des phases phénologiques et des rendements. Monographie No.95. Météorologie Nationale, Paris

CEC (Commission of the European Communities) 1985 Soil Map of the European Communities. Office for Official Publications and Directorate General VI-Agriculture of the European Communities, Luxembourg

Choisnel E, de Villele O, Lacroze F 1992 Une approche uniformisée du calcul de l'évapotranspiration potentielle pour l'ensemble des pays de la Communauté Européenne. Publ. EUR 14223 of the Office for Official Publications of the EC, Luxembourg

Dagneaud J-P, Tranchefort J, Couvreur F 1981 Blé tendre d'hiver: une méthode opérationnelle de prévisions de rendements. Perspectives Agricoles (Document ITCF) 48

De Wit CT, van Keulen H 1987 Modelling production of field crops and its requirements. Geoderma 40:253-265

Doorenbos J, Kassam AH 1979 Yield response to water. FAO Irrigation and Drainage Paper N°33 FAO, Rome

Doorenbos J, Pruit WO 1984 Guidelines for predicting crop water requirements. Irrigation and Drainage Paper N° 24. FAO, Rome

FAO (Food and Agriculture Organisation of the United Nations) 1986 Early agrometeorological crop yield assessment. FAO Plant Production and Protection Paper N°73, Rome

Genovese G 1997 The evaluation of the MARS crop yield forecasting model. Report submitted to the Committee for Agricultural Statistics of the EC. Session of October 1997. European Statistical Office Eurostat, Luxembourg (In print)

Haun JR 1982 Early prediction of corn yields from daily weather data and single predetermined seasonal constants, Agric Meteorol 27:191-207

Hargreaves GL, Hargreaves GH, Riley P 1985 Irrigation water requirement for the Senegal River Basin. Journal of Irrigation and Drainage Engineering, ASCE 111:265-275

Hooijer AA, Van der Wal T 1994 CGMS Version 3.1 user manual. Winand Staring Centre, Wageningen and Joint Research Centre, Ispra

Hough MN, Parker J 1992 The estimation, for the EC, of global solar radiation and sunshine hours from cloud cover for 1-day and 10-day periods. Final report to the Institute for Remote Sensing

Applications, MARS Project, London
King D, Daroussin D, Jamagne M 1994 Development of a soil geographic database from the soil map of the European Communities, Catena 21:37-56
King D, Le Bas C 1994 Programme d'estimation des réserves en eau des sols à partir des paramètres de base de données géographiques des sols de l'Union Européenne. Final report of the Contract N° 5538-93-11 EP ISP F for the EC Joint Research Centre, Ispra
Kooman PL 1995 Yielding ability of potato crops as influenced by temperature and daylenght. PhD Thesis, Wageningen Agricultural University
Loudjani P 1995 Remote Sensing and Crop Yield estimation: an Overview. COST 77 Document. Publ. EUR 16275 of the Office for Official Publications of the EC Luxembourg
Motha RP, Heddinghaus TR 1986 The Joint Agricultural Weather Facility's Operational Assessment Program. Bulletin of the American Meteorological Society: 67:1114-1122
Meteo Consult BV 1991 AMDaC User Manual. Software package for Actual Meteorological Database Construction, Wageningen
Palm R, Dagnelie P 1993 Tendance et effets du climat dans la prévision des rendements agricoles des différents pays de la CE. Joint Research Centre of the EC. Publ. EUR 15106 of the Office for Official Publications of the EC, Luxembourg
Place RE, Brown DM 1987 Modelling corn yields from soil moisture estimates: description, sensitivity analysis and validation. Agric.For.Meteorol. 41:31-56
Sakamoto CM 1978 The Z-index as a variable for crop yield estimation. Agric.For.Meteorol. 19:305-313
Spitters CJT 1987 An analysis of variation in yield among potato cultivars in terms of light absorption, light utilisation and dry matter partitioning. Acta Horticulturae 214:71-84
Supit Y 1994 Global radiation. Publ. EUR 14767 of the Office for Official Publications of the EC, Luxembourg
Supit Y, Hooijer AA, Van Diepen CA (Ed) 1994 System description of the WOFOST 6.0 crop simulation model implemented in CGMS. Publ. EUR 15956 of the Office for Official Publications of the EC, Luxembourg
Supit I, Van Kappel RR 1997 A simple method to estimate global radiation. Submitted for publication
Swanson ER, Nyankori JC 1979 The influence of weather and technology on corn and soybean yield trends. Agric. Meteorol. 20:327-342
Thompson N, Barrie IA, Ayles M 1981 The Meteorological Office Rainfall and Evaporation Calculation System MORECS (July, 1981). Hydrological Memorandum No.45. Meteorological Office, Bracknell
USDA (US-Departments of Commerce and of Agriculture) 1987 Weekly weather and crop bulletin: National Weather Summary. Vol.74, No.40 (Oct.6, 1987). Washington, 25 pp
Van der Voet P, Van Diepen C, Oude Voshaar J 1993 Spatial interpolation of daily meteorological data: a knowledge based procedure for the regions of the European Communities. Report 53/3. The Winand Staring Centre, Wageningen
Van Keulen H, Wolf J 1986 Modelling of agricultural production: weather, soils and crops. Centre for Agricultural Publishing and Documentation, Pudoc, Wageningen
Van Keulen H 1975 Simulation of water use and herbage growth in arid regions. Simulation Monographs, Pudoc, Wageningen
Van Diepen CA, Wolf J, Van Keulen H, Rappoldt C 1989 WOFOST: a simulation model of crop production. Soils use and management 5:16-24
Vossen P 1990a Comparative statistical validation of two ten day water use models and three yield reduction hypotheses for yield assessment in Botswana. Agric. For. Meteorol. 51:177-195
Vossen P 1990b Modèles Agrométéorologiques pour le Suivi des Cultures et la Prévision de Rendements des grandes Régions des Communautés Européennes. In: Proceedings of the

Conference on The Application of Remote Sensing to Agricultural Statistics, Varese, Italy, pp. 75-84. Publ. EUR 12581 of the Office for Official Publications of the EC, Luxembourg

Vossen P 1992 Forecasting national crop yields of EC countries: the approach developed by the Agriculture Project. In: Proceedings of the Second Conference on The Application of Remote Sensing to Agricultural Statistics, Belgirate, Italy, pp 159-176. Publ. EUR 14262 of the Office for Official Publications of the EC, Luxembourg

Vossen P, Rijks D 1995 Early crop yield assessment of the EC countries: the system implemented by the Joint Research Centre. Publ. EUR 16318 of the Office for Official Publications of the EC Luxembourg, 182 pp

Annex : Crop knowledge bases

Bignon J 1990 Agrométéorologie et physiologie du mais grain dans la Communauté Européenne. Publ. N° EUR 13041 of the Office for Official Publications of the EC, Luxembourg

Carbonneau A, Riou C, Guyon D, Riom J 1992 Agrométéorologie de la vigne en France. Publ. EUR 13911 of the Office for Official Publications of the EC, Luxembourg

Falisse A 1992 Aspects Agrométéorologiques du Développement des Cultures dans le BENELUX et les Régions Voisines. Publ. EUR 13910 of the Office for Official Publications of the EC, Luxembourg

Hough MN 1990 Agrometeorological Aspects of Crops in the United Kingdom and Ireland. Publ. EUR 13039 of the Office for Official Publications of the EC, Luxembourg

MacKerron DKL 1992 Agrometeorological aspects of forecasting yields of potato within the E.C. Publ. EUR 13909 of the Office for Official Publications of the EC, Luxembourg

Narciso G, Ragni P, Venturi A 1992 Agrometeorological aspects of Crops in Italy, Spain and Greece. Publ. EUR 14124 of the Office for Official Publications of the EC, Luxembourg

Riou C (Ed) 1994 The effect of climate on grape ripening: application to the zoning of sugar content in the European Community. Publ. EUR 15863 of the Office for Official Publications of the EC, Luxembourg

Russell G 1990 Barley knowledge base. Publ. EUR 13040 of the Office for Official Publications of the EC, Luxembourg

Russell G, Wilson G 1994 Wheat knowledge base. Publ. EUR 15789 of the Office for Official Publications of the EC, Luxembourg

4.2 Pathways in crop modelling for cultivation control

Gerrit van Straten

Wageningen Agricultural University, Systems and Control Group, Department of Agricultural Engineering and Physics, Bomenweg 4, 6703 HD Wageningen, The Netherlands
E-mail:Gerrit.vanStraten@user.AenF.WAU.NL

Developing crop growth models suitable for cultivation steering is the issue of this paper. Research oriented models tend to be too large or too complex to be directly used. Various options to remedy this situation are discussed. Mechanistic models can be reduced, either by reasoning or by formal identification methods from the original prototype model. Alternatively, mechanistic black-box models (e.g. artificial neural nets) can be derived directly from the data. If prediction errors of mechanistic models have non-random components, an enhancement of the system equations with a (data-based) stochastic model may be an alternative over expanding the models with more equations and more state variables. In cases where further detailing of the models cannot be avoided, e.g. in describing resource allocation over components of the plant, internal-optimization (cybernetic) models which incorporate the optimizing behaviours of plants can be an interesting alternative.

1. Introduction

Crop growth models have been developed to gain insight in the dynamic behaviour of crops, and to make predictions about potential crop yield under various circumstances. In this paper we are interested in the use of crop growth models for cultivation steering. Cultivation steering is of interest in protected environments, i.e. in greenhouses, for balancing crop yield versus costs of inputs - like heating costs - but also for quality control and to control the timeliness of crops on the market. Control of cultivation is also of interest in open field crops, e.g. in precision farming, in order to balance crop yield versus inputs of fertilizers, crop protection measures and environmental damage.

Cultivation steering requires that predictions can be made about the behaviour of the crop in the future. Models are the most versatile tool to this end. However, the intended use for control sets a number of requirements. An important condition is that a clear distinction is made between input variables and state or output variables. Also, the model should not contain or require inputs that cannot be measured with reasonable effort (e.g. sky temperature). The model must be self-consistent, i.e. it should not require inputs that depend upon the state itself (e.g. LAI as input is unacceptable, except perhaps in a feed-back situation, as it depends on the state). The model outputs must be expressed in terms that are important for the final goal. Also, it must be valid over the intended range of control. And for optimization purposes and on-line control purposes, the model must not be unduly complex.

Efforts in crop modelling in the past decades have lead to useful mechanistic, deterministic models, but they usually do not fulfil the conditions desirable for control. The aim of this paper is to discuss possible pathways to remedy this. Two major issues are addressed: (i) how to get reduced models: mechanistic and non-mechanistic, and (ii) how to handle model mismatch without further reductionistic detailing of the model: the introduction of stochasticity, and mechanisms to cope with time variability and adaptation.

2. Off-the-shelf crop growth models

The vast majority of crop-growth models are mechanistic and deterministic. They have been developed for research purposes and are being used for prediction of potential crop yield. The pathway followed to arrive at these models can be characterized as the 'reductionistic' approach. Separate experiments are set up to isolate parts of the crop growth and development process, and to build sub-models for these. Next, the various submodels are combined. If overall performance is not satisfactory, further details are added, either by detailing sub-pro cesses, or by adding other sub-processes.

Typical examples of such models are SUCROS (Dayan et al., 1993) and LINTUL (Spitters 1990) for field crops, and TOMGRO (Dayan et al., 1993), HORTISIM and others (De Koning 1994) for fruit crops like cucumber, sweet pepper and tomato in greenhouses. SUCROS and LINTUL are models with a limited number of state variables (10 and 5, resp.) and are not complex for this reason. However, since they attempt to describe different development stages of the plant, they are still fairly complex in the sense that they are, in fact, hybrid models, where the structure changes discontinuously as a function of the development stage state variable (see Figure 1). The greenhouse models are complex models because they contain a large number of state variables, in an attempt to describe the growth and development of trusses or even individual fruits. All these models are set up as discrete time models, with a time step of one day, so that variat ions over a day are averaged out.

Although important as tools in guiding scientific re search, the direct applicability of such models for cultivation steering is limited. In particular, the model complexity makes them hard to use in on line control or in optimization calculations. We will investigate pos sible remedies in the next section. In addition, application of generalist models to a particular situation is problematic, because actual conditions usually differ from the assumptions underlying the general param eterization. This will be substantiated in the sections 4 and further.

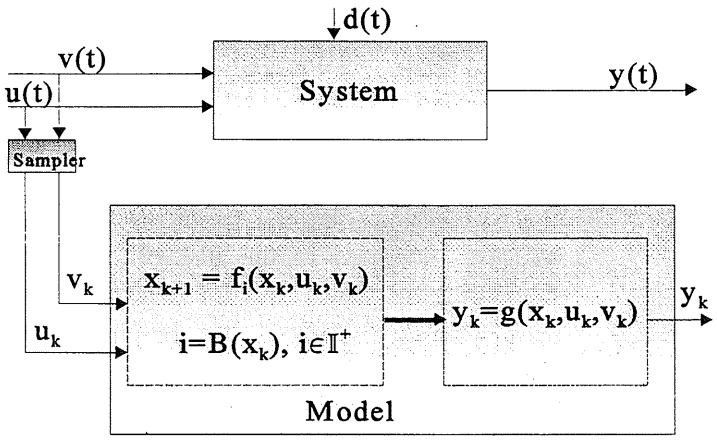


Figure 1. Crop (hybrid) system model. Inputs u,v,d: controlled, measured environment and disturbance inputs, resp.; States x; Outputs y. Subscript k indicates discrete time t_k, i is a positive integer, selected by the switching function B.

3. Model reduction and meta-modelling

Take a complex model and assume that its predicted behaviour is actually an accurate image of reality in a given validity range. We can then try to find reduced models, with a similar behaviour, but which are easier to handle. There are two basically different procedure to achieve such reduced models.

The first is to reduce the equations by lumping, aggregating, simplificating and pruning of the original models by reasoning. As an example, Tap (Tap and Van Straten 1995) developed a big-fruit-big-leaf approximation of the tomato model with more than 300 states developed by De Koning (De Koning 1994). This was mainly achieved by lumping the individual fruits and leaves, while maintaining the basic structure. The final model structure is depicted in Figure 2. Also, the model was made continuous, rather than on a daily basis, for use in an optimal controller setting.

This was achieved by introducing a dynamic assimilate buffer, and formulating a smooth switching function to prevent fruits and leafs from drawing upon assimilates when the buffer is empty. Therefore, the model still accomodates sink limited and source limited growth. Sink limited growth occurs when the assimilate demand is less than the photosynthetic production. Source limited growth occurs when the assimilate production is less than the demand from leafs and fruits. However, due to the continuization, the system can be source limited during the night, and sink limited during the day. This is important from the point of view of control, because adjusting environmental conditions in response to the actual state of the plant in stead of according to the averaged state over a day can improve the efficiency of the process. A comparison of model and data in an optimal control experiment are shown in Figure 3. Despite considerable model reduction a description of fruit weight is obtained, which is satisfactory in view of the rather large variance of the measurements.No attempt has been made to compute the behaviour by using the original, large model, because this would have required extensive re-calibration, which is difficult to perform.

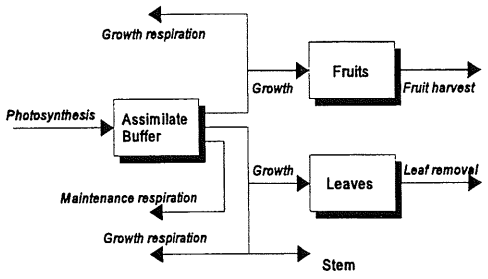


Figure 2. Reduced model: Big-leaf-big-fruit tomato model structure (Tap and Van Straten 1995).

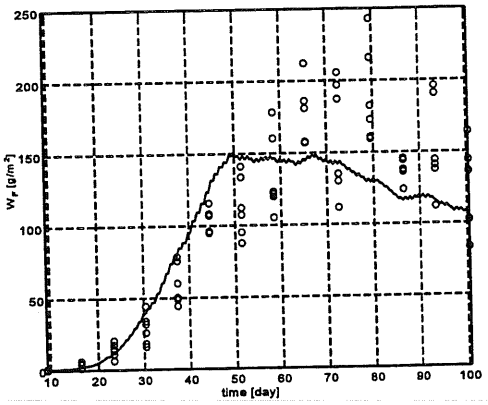


Figure 3. Calibration of big-leaf-big-fruit model. Fruit dry weight. Notice the large variation in the data (circles).

A completely different approach is to approximate the behaviour of the complex model by black box, i.e. data oriented models. The procedure is to first generate an extensive set of pseudo-data in the range of interest from the complete model. Next, a black box model is developed from the data by ordinary system identification methods. In this way Young and Lees approximate the greenhouse physics by a data-based mechanistic model of low order (Young and Lees, 1996). In order to cope with non-linearities, a non-linear input-output model may be chosen. Neural nets are very convenient non-linear mappings. Seginer and Ioslovich (1996) showed that TOMGRO having 69 state variables, could as well be described by a neural net with an equivalent state vector with dimensions one order of magnitude less.

4. Validity domain and need for calibration

One may think that in the ideal case parameter estimation is unnecessary. This is incorrect for several reasons. Each model has associated to it a validity domain. Validity has several attributes:

- the amplitude and frequency range of input signals;
- the assumptions about the non-modelled environment of the system
- the properties and internal structure of the system itself.

More often than not the validity domain of a model is not well documented. Greenhouse crop models usually have been developed for environmental inputs prevailing under 'normal' operation. However, in a search for optimal control, it may easiliy happen that the model needs to be used in an input range not yet observed, e.g. for temperatures lower than usual. The model is than stretched outside its original validity domain, and re-calibration (on the basis of newly designed experiments) is needed to cover this. Most crop models assume that the root environment is non-limiting. When in an actual situation this condition is not fulfilled the model predictions will be erroneous. Some differences in environment, for instance the soil type, that have not been modelled, may be accomodated by re-calibrating the model. Finally, a model may be designed for a specific crop or cultivar. When used for another one, re-calibration of some of its parameters is unavoidable.

The need for re-calibration restricts the general applicability of a model. Efforts can be made to collect as many re-calibrations as possible, in order to create a table from which the likely parameters under for various crop/cultivars, under various non-modelled environmental conditions can be read. Such a table is, in fact, a static data-based black box submodel, mapping application conditions to parameters of the model.

General applicability need not be required in a control environment. It is quite conceivable to calibrate a model for the particular situation at hand, and use it for subsequent instances of control and operation with the same crop in the same environmental setting.

5. The exploitation of data

In the above, mechanistic models have been taken as the starting point. What is the role of data? First, data can be used for model calibration. An obvious requirement is that calibration of the model is, indeed, possible. In complex models, this can be quite problematic, because of structural identifiability problems, and problems of identifiability from the available data. This is another reason to look for simplified model structures. However, there has been little systematic research to see how parameter identifiability affects the predictive power of models. The fact that a single parameter may not be uniquely identifiable from the data indicates that the output is not sensitive to this parameter. Consequently, its precise knowledge is not required for making predictions *in the same output domain.* Similarly, correlations between parameter estimates hamper the identification of each parameter separately, but cross-correlation reduces the uncertainty in prediction as compared to the uncertainty obtained if each parameter was sampled from its own confidence interval.

The above assumed that data were already present. The situation is better if we have control over the system. In fact, experiments can be designed to obtain the largest possible information with respect to parameter identifiability. In bioprocess operation identification-optimal steering inputs have been computed by minimizing the condition number of the Fisher information matrix (Munack, 1989).

$$u^*(t) = arg\min \frac{\lambda_{max}(F)}{\lambda_{min}(F)} \quad with \quad F = \int_0^{t_f} (\frac{\partial y}{\partial p})^T Q \frac{\partial y}{\partial p} dt$$

Here, y is the model output, u the controllable input, p the model parameters to be identified, Q the weighting matrix, λ the eigenvalues of the Fisher information matrix F. Such dynamic designs are not very common in crop growth modelling, but can be well rewarding. The method could also be applied in the situation of section 1, where pseudo-data generated by a large simulation model are used for calibrating a reduced model.

In well defined situations, e.g. a specific greenhouse, with a specific crop/cultivar, an attractive alternative to mechanistic modelling is to use the input-output data to generate a black-box dynamical model. A particularly suitable form is given by

$$y_{k+1} = f(y_k, y_{k-1}, \ldots, y_1, u_k, u_{k-1}, \ldots, u_{k-m}$$

This is known as a NARMAX form (Non-linear Auto-Regressive, Moving-Average with eXogenous variables). The function f represents a non-linear mapping, and can be conveniently realized by an artificial neural or neural-fuzzy net. Particularly convenient are networks that are linear in the parameters. This can be obtained by taking the non-linear input part randomly (e.g. the Random Activation Network RAWN, (Te Braake and Van Straten 1995), or by choosing it on the basis of qualitative knowledge (e.g. the selection of centres and widths of radial basis functions, or of fuzzy membership functions in a fuzzy-neural

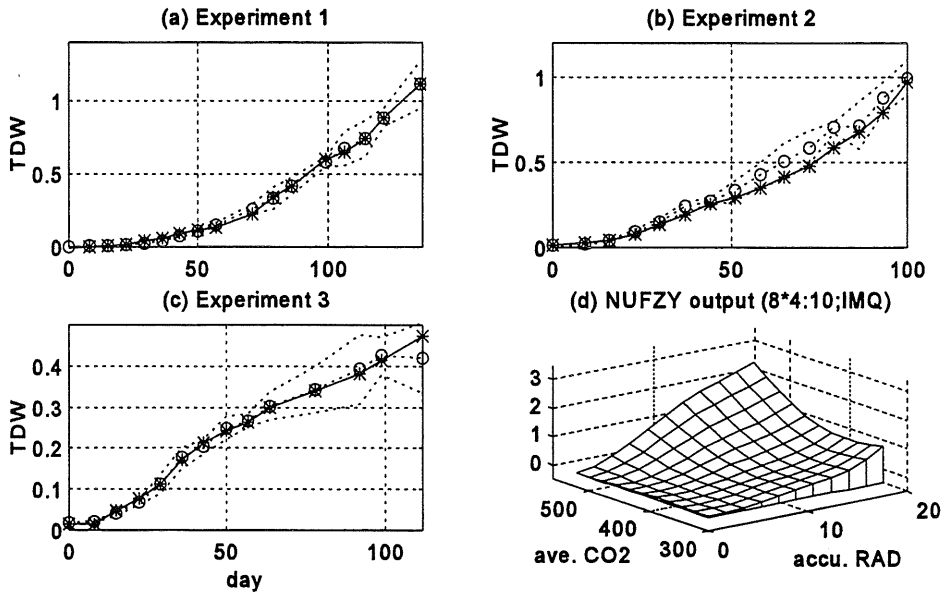


Figure 4. The measured TDW of tomato (circle-dotted line) with 95% confidence interval (dotted line). Experiment 1 (a) and 3(c) are used for calibration of NUFZY (star-dotted line), experiment 2 (b) for validation. The predicted TDW by NUFZY is related with accumulated radiation (kW m^{-2}) and the CO_2 concentration (ppm) averaged over the period drom starting day until the measurement day.

approach). Figure 4 shows the performance of the fuzzy-neural model NUFZY with 4 inputs with 8 inverse multiquadratic membership functions each, and having 10 rules, to describe the growth of tomato (Tien and Van Straten 1995). The number of free parameters is equal to the number of rules in this type of model.

6. Handling mismatch

Even after calibration there is usually mismatch between model and data. The first source of mismatch is in the variance of the data, as shown in the example of Figure 3. This is not a problem of the model, but indicates that real systems are subject to variability. Some of this may be traced back to observable causes. For instance, spatial variability is known to be a major source of variability among plants in a greenhouse. One remedy would be to design and operate a greenhouse such that spatial variability is reduced to a minimum. CFD (Computational Fluid Dynamics) techniques can be used to study the options to accomplish this. The other approach is to model - and perhaps consciously exploit - the spatial variability. It is obvious that this would tremendously increase the modelling effort.

The other source of mismatch is an inherent attribute of modelling. Modelling, by its very nature, means simplification. Some discrepancy, therefore, must be accepted if the model is to serve a useful purpose.

The first task when dealing model mismatch is to make a judgement about the acceptability of the discrepancies. In the ideal case, this should be done with respect to the usefulness for the final goal. In the case of cultivation control, we need to check how model errors affect the final net income of the farmer. Research regarding the sensitivity of the final operation and objective function to modelling errors is almost absent, but is urgently needed.

6.1 Modelling the error: Stochastic models.

In order to judge the effect of modelling error on prediction performance it is necesssary to have a description of the error. The most common way to do this is by adding noise terms to the model, as follows:

$$x_{k+1} = f(x_k, u_k, v_k) + G\eta_k$$

$$0 \quad y_{k+1} = g(x_{k+1}, u_{k+1}, v_{k+1}) + \nu_k$$

Here G is a weighting matrix, η_k and ν_k are random processes, called systems noise and output noise, respectively. The model is now called a stochastic model, rather then deterministic, although deterministic with stochastic terms would be a more appropriate name. The systems noise can be thought off as a signal generated by a noise filter driven by a uniform independent stochastic signal. If it is assumed that the driving signal has a normal distribution, then the approach is probabilistic. As an alternative it can be assumed that the driving signal is just unknown but bounded. In that case we have a set-theoretic setting (cf. e.g. Schweppe, 1973). The art of modelling is to deduce the distribution properties of the systems noise from the data, in the presence of measurement noise. This can be done by systems identification techniques (cf. Ljung and Glad, 1994).

Predictions made with the model can only be made by generat ing a realisation. A number of realisations yield an output with a mean and a certain distribution. Figure 5 shows the behav iour of a stochastic mechanistic lettuce model for ten different realizations of the systems noise in a hypothetical example. Assuming systems noise in stead of just output noise often provides a better descrip tion of the expected variability. An indicator for the presence of systems noise is the prediction error sequence, i.e. the sequence that results when predictions are made from the previous data. If this sequence still contains signal information, there are most likely modelling errors. The parameters estimates obtained from erron eously assuming an output error structure are biased. Enhancing the systems equations by a stochastic term often dramatically improves the expected prediction error, and reduces the parameter bias.

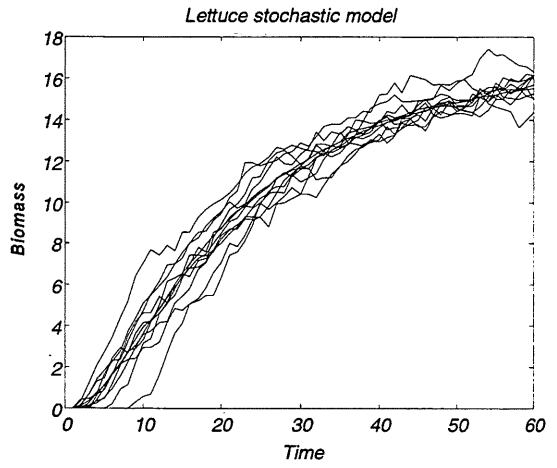


Figure 5. Ten different realizations of a simple stochastic lettuce model, illustrating the effect of system error.

6.2 Reducing the error: Model expansion.

Suppose that by analysis or qualitative judgement the model mismatch is deemed to be unacceptable. In that case it is conceivable to expand the model, or revise the formulation of its sub-processes. This pathway is the choice if it comes to increasing our understanding. Modelling is a powerful tool to organize our thoughts and to guide experimentation. A disadvantage of model expansion is that it increases the complexity, and entails the need for further experimentation, associated with a parameter estimation excercise. If operational control is the purpose it would be desirable to develop methods that will assess the control consequences of an expanded model in order to balance the degree of detail against practical usefulness.

6.3 Reducing the error: Internal optimization models.

A common source of considerable modelling difficulties in crop models is the need to describe the allocation of resources to different parts of the plant. Similarly, in bioprocesses, modelling the preference of micro-organisms to various substrates is an issue of considerable difficulty. It appears that biosystems have the tendency to adapt to changed environment, by changing their internal structure, both on the species level, as well as on the community level, so as to maximize their biomass or chances for survival. In a full mechanistic model it would be necessary to model these internal regulation mechanisms, so that optimizing behaviour on the macroscopic level is an emergent property. This, however, is an enormous task, and would involve an approach perhaps on the molecular or gen level.

An interesting option to avoid explicit detailed modelling of adjustment mechanisms and self-organization, is to incorporate the optimizing behaviour into the models. In this way internal-optimization models, sometimes called cybernetic models (Kompala et al., 1986), are obtained. The procedure is to introduce time-variant parameters that act as control variables into the equations of change, and then compute the time pattern of these control parameters so that a specific goal is optimized over time. As an example, consider the system

$$\dot{x} = p(t)[f\{x\} - \alpha x - \beta y]$$

$$\dot{y} = (1 - p(t))[f\{x\} - \alpha x - \beta y]$$

where x is the vegetative plant part (leafs), y the generative part (seeds), $f\{x\}$ the photosynthesis, and αx and βy the loss due to maintenance of leafs and seeds, resp. Adopt the optimization criterion that at the end T of the growing season the seed yield should be optimum. Then, the task is to find the optimal pattern $p^*(t)$ such that $y(T)$ is maximized, or

$$p^*(t) = arg\min\, y(T; p(t))$$

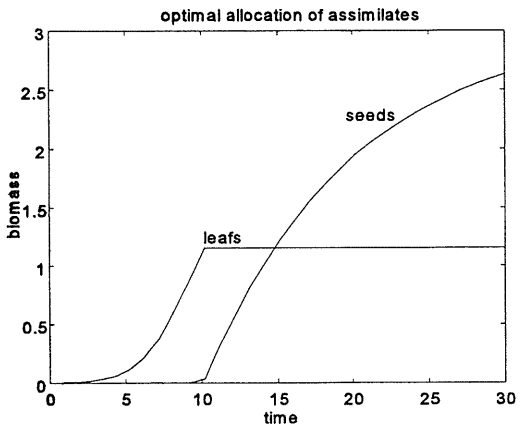


Figure 6. Optimal allocation. The allocation function (not shown) switches from 1 to 0 at $t = 10$.

This problem can be solved with optimal control theory. A particular useful method is the optimization with the Hamiltonian function, introducing co-states (Lewis, 1986), which represent the marginal value of a unit of leafs or seeds at any time. Figure 6 shows that first all assimilates are assigned to the leafs ($p = 1$). Then, after a time determined by the optimal solution, all efforts are put to produce seeds ($p = 0$). This result is achieved without explicitely modelling how p depends upon the system states or the environment.

Iwasa and Roughgarden discuss several more elaborate cases in their seminal paper (Iwasa and Roughgarden, 1984). Velten and Richter (1995) have shown that maximizing biomass as a goal function yields similar results.

7. The introduction of feed-back

Having a suitable model, cultivation control can be achieved by computing optimal steering patterns over a season. This is an open loop solution. If the actual weather conditions deviate from the expected, this may become sub-optimal. Also, modelling errors accumulate. The classical solution to abate uncertainty is the introduction of feed-back. In crop cultivation, this would mean the on-line observation of the behaviour of the plants. Several approaches have been proposed, e.g. the speaking plant concept (Hashimoto, 1989), and the use of image processing techniques (Van Henten and Bontsema, 1995). Feed-back together with self-learning models will be a challenging issue for further development of crop cultivation control in the future.

8. Summary and Conclusions

Crop cultivation steering requires concise but sufficiently accurate models. Several options to arrive at these control oriented models have been discussed. Reduced mechanistic models can be developed by reasoning, or by formal identification techniques applied to large comprehensive crop growth models. Identification methods can also be applied directly to the data to derive non-linear black-box models, such as neural nets or fuzzy-neural models. Enhancing deterministic models with a stochastic part does not reduce uncertainty, but it reduces the bias in parameter estimates and increases the reliability of the forecasts. Finally, if detailing models is unavoidable, cybernetic modelling may be an interesting option to solve the problem of resource allocation.

9. References

Dayan E, Van Keulen H, Jones JW, Zipori I, Shmuel D, Challa H 1993 Development, calibration and validation of a greenhouse tomato growth model: I. Description of the model. Agricultural Systems 43: 145-163

Dayan E, Van Keulen H, Jones JW, Zipori I, Shmuel D, Challa H 1993 Development, calibration and validation of a greenhouse tomato growth model: II. Field calibration and validation. Agricultural Systems 43: 165-183

De Koning ANM 1994 Development and dry matter distribution in tomato: a quantitative approach. PhD Dissertation. Wageningen Agricultural University

Hashimoto Y 1989 Recent strategies of optimal growth regulation by the speaking plant concept. Acta Horticulturae 260: 115-121

Iwasa Y, Roughgarden J 1984 Shoot/root balance of plants: optimal growth of a system with many vegetative organs. Theoretical Population Biology 25: 78-105

Kompala DS, Ramkrishna D, Jansen NB, Tsao GT 1986 Investigation of bacterial growth on mixed substrates: experimental evaluation of cybernetic models. Biotechnology and Bioengineering 28: 1044-1055

Lewis FL 1986 Optimal Control. Wiley

Ljung L, Glad T 1994 Modeling of Dynamic Systems. Prentice Hall

Munack A 1989 Optimal feeding strategy for identification of Monod-type models by fed-batch experiments. In: NM Fish, RI Fox, NF Thornhill (eds) Computer Applications in Fermentation Technology: Modelling and Control of Biotechnological Processes (Elsevier) p 195-204

Schweppe FC 1983 Uncertain Dynamic Systems. Prentice-Hall

Seginer I, Ioslovich I 1996 Crop model reduction and simulation in reduced space.Acta Horticulturae 406: 63-71

Spitters CJT, Van Keulen H, Van Kraalingen DWG 1989 A simple and universal crop growth simulator: SUCROS87. In: R Rabbinge, SA Ward, HH van Laar (eds) Simulation and Systems Management in Crop Protection. Simulation Monograph 32, PUDOC, Wageningen p 147-181

Spitters CJT 1990 Crop growth models: their usefulness and limitations. Acta Horticulturae 267: 349-368

Tap RF, Van Straten G 1995 Development of a reduced order tomato model. 1st IMACS/IFAC Symposium on Mathematical modelling and simulation in agriculture & bio-industries. Brussels Vol. III VI.A.2-1:6.

Te Braake HB, Van Straten G 1995 Random activation weight neural net (RAWN) for fast non-iterative training. Engineering Applications of Artificial Intelligence 8: 71-80

Tien BT, Van Straten G 1995 Neural-fuzzy systems for non-linear system identification - orthogonal least squares training algorithms and fuzzy rule reduction. Preprints 2nd IFAC/IFIP/EurAgEng Workshop on Artificial Intelligence in Agriculture, Wageningen, The Netherlands, May 29-31, (IFAC-Elsevier) p 249-254

Van Henten EJ, Bontsema J 1995 Non-destructive crop measurements by image processing for crop growth control. J Agric Engng Res 61: 97-105

Velten K, Richter O 1995 Optimal root/shoot-partitioning of carbohydrates in plants. Bulletin of Mathematical Biology 57: 99-107

Young PC and Lees MJ 1996 Simplicity out of complexity in glasshouse climate modelling. Acta Horticulturae 406: 15-28

4.3 From fields to models: the data story - Feeding models that should not be used.

Johan Bouma

Dept. Soil Science and Geology, C.T.de Wit Graduate School of Production Ecology
Agriculture University Wageningen
Email: johan.bouma@bodlan.beng.wau.nl

1. Introduction

The subtitle of these series of seminars refers to the pittfalls that are associated with using point data in models to represent the behaviour of large areas of land. In this paper, we will focus on farmer's fields, which represent relatively small areas as compared with a watershed or a large region or country. But even so, point data obtained by augerings characterize only a very small volume of the area to be studied. The line for this paper appears, therefore, to be logical: how to select a minimal quantity of proper data to feed the proper model. Also, what is the accuracy obtained when using different quality data?

Before this question can be analysed we should define our problem. Unfortunately, this aspect is often either overlooked or ignored. Of course, the objectives of any study should govern its execution and planning. Here, we will focus on Precision Agriculture (PA) which is a rapidly developing field focusing on the adjustment of land management within fields to accomodate the needs of the growing plant during the growing season while also considering environmental threshold values for land and water. This adjustment is possible now because of information technology and availability of technical equipment, such as Global Positioning Systems, Yield Monitoring, proximal and remote sensing techniques etc. A recent review is provided in Bock and Lake (1997). We will consider two case studies, a high-tech one in the Netherlands and a low-tech one in the Sahel. These studies have been and are being documented thoroughly in accessible publications. We will therefore provide a review here and refer the reader to the more detailed source publications for specific information.

2. Characterizing fields

In their studies on precision agriculture, using a field in the Dutch polder the Wieringermeer, Finke and Bosma (1993), Finke and Stein (1994) and Verhagen (1995) have developed a standard procedure for soil characterization, focused on using the mechanistic, quantitative models LEACHM and WAVE. This includes the following steps:

1. Exploratory geostatistical survey to determine optimal observation densities. This uses geomorphological and soil expertise to distinguish different subareas withn the field which are to be characterized seperately.
2. Construction of variograms and determination of optimal boring densities.
3. Soil survey of the area, focused on obtaining an operational database.
4. Functional characterization of soil data in terms of their physical or chemical properties. Basically, this implies that horizons that are pedologicaly different are put together when they function identically.

5. Running of the model for points considering actual and potential weather and management conditions with a focus on plant growth and leaching of nitrates.
6.Interpolation of point data to obtain expressions for areas of land.

2.1 The high-tech case

Discussions during recent workshops indicate that two broad approaches to PA are being considered by different researchers (e.g. Bock and Lake, 1997). There is the reactive approach which has a strong technical focus and implies the following steps:

1. Every year yields are monitored site-specifically within the fields of the farm. This will provide over the years maps indicating local variability.
2. Remote sensing will be available to characterize crop conditions within the field during the growing season, providing indications as to when some management action is needed, in terms of crop protection, fertilization, irrigation etc.
3. When indications are given, machines are sent into the field. They are programmed according to patterns observed during yield monitoring or - and this is likely to occur increasingly in future - they use in-situ proximal sensing of nutrients and water. Management is adjusted on-the-go.

The question is raised by some engineers as to why we would need soil maps and models to calculate crop growth? The above procedure would work fine and would result in the particular advantages of precision agriculture: doing the right type of management at the right time and vary this within fields to avoid local over- and underkill!

In terms of the topic of our lecture series, this offers an interesting proposition. If we don't need models, we don't need data for models. So it turns out that the question being raised here about data needs for models is irrelevant. We do need data, of course, but they are all measured by a host of new techniques. How to handle such data in databases is a very interesting question that will not be discussed here.

Of course, we feel that modelling can and should play a pro-active role in PA but this discussion has been presented elsewhere (Bock and Lake, 1997).

2.2 The low-tech case

Spatial and temporal variability is much greater in the Sahelian region than it is in the Netherlands. Monitoring data on field level at ICRISATSC has been summarized by Brouwer and others in various recent publications (e.g. Bouma et al, 1995; 1997; Brouwer and Bouma, 1997).

The variability has been studied in farmer's fields by observing patterns that are associated with differences in microrelief and biological activity. Farmers have no money to buy fertilizers, so manipulation of animal manure and crop residues are the only means to try to maximize production of millet. Powel has reported results of some experiments in Bouma et al. (1997). No concerns here about pollution of groundwater with agrochemicals!

There is no clear role for modelling because the differences in soil conditions within the field are in terms of pH values and different nutrient contents, while the mineralization processes of manure are highly complex at the prevailing high temperatures. There are no models available to express the effects of such subtle differences.The kind of empirical experiments, reported by Powell are therefore most relevant at this time to determine which application rates of manure and crop residues are appropriate within subareas of the fields.

3. Data in relation to monitoring and empirical experimentation, rather than modelling.

So, in the context of this seminar series, we see again no clear relation between soil data and simulation models in the low-tech case, and the conclusion can be the same as for the high-tech case: consider the problem to be studied first and plan accordingly. We may find, as we do here, that monitoring and empirical experimentation may be needed to solve the problems most efficiently. Then, of course, data are very much needed as well but from a different perspective.

4. References

Bock G, Lake J (Eds) 1997 Precision Agriculture. International CIBA-Foundation/EERO Conference, Wageningen, Netherlands (in press).

Bouma J, Brouwer J, Verhagen A, Booltink HWG 1995 Site specific management on field level: high and low tech approaches. In: Bouma J, Kuyvenhoven A, Bouman B, Luyten J and Zandstra H (Eds) Eco regional approaches for sustainable land use and food production. Kluwer Publ., Dordrecht, Netherlands pp 453-475.

Bouma J, Verhagen J, Brouwer J, Powell JM 1997 Using systems approaches for targetting site specific management on field level. In: Kropff MJ, Teng PS, Aggerwal PK, Bouma J, Bouman BAM, Jones JW, Van Laar HH (Eds) Applications of systems approaches at the field level. Kluwer Acad. Publ., Dordrecht, Netherlands pp 25-37.

Brouwer J, Bouma J 1997 Soil and crop growth variability in the Sahel. Info. Bull. 49 ICRISAT-Sahelian Center and Agric. Univ. Wageningen, Netherlands. Patenchern 502324. Andhra Pradesh, India.

Finke PA, Bosma WJP 1993 Obtaining basic simulation data for a heterogeneous field with stratified marine soils. Hydrol. Process. 7:63-75

Finke PA, Stein A 1994 Application of disjunctive cokriging to compare fertilizer scenarios on a field scale. Geoderma 62:247-263

Verhagen A, Booltink HWG, Bouma J 1995 Site specific management: balancing production and environmental requirements at farm level. Agricult. Systems 49:369-384