OPTIMAL DESIGN FOR MOVING LOCAL REGRESSIONS

WERNER G. MÜLLER

Forschungsbericht/
Research Memorandum No. 281

May 1991

Abstract

After the introductory sections, where nonparametric regression methods and basics of design theory are overviewed, this paper describes the so-called *moving local regression*, a special nonparametric statistical tool. Its incorporation into the design framework is given, including the derivation of the necessary formulae. An example from practice is given, illuminating the interrelations of the basic ingredients of the method.

Den Einführungskapiteln, wo nichtparametrische Methoden der Regression und Grundkonzepte der Versuchsplanung kurz umrissen werden, folgt die Beschreibung der sogenannten *gleitenden lokalen Regression*, einem speziellen nichtparameterischen Verfahren. Seine Einbindung in die Versuchsplanungproblematik wird diskutiert und die notwendigen Formeln werden abgeleitet. Danach folgt ein praktisches Beispiel zur Illustration des Zusammenwirkens der Komponenten dieser Methode.

1 Introduction

Among the countless approaches to nonparametric regression, some of them like spline methods or kernel estimators shortly described in Subsection 2.3, stands one, which is intuitively simple as well as it shows up most desirable properties. It was independently developed for either smoothing or interpolation purposes. Though Cleveland, (1979) used it for extracting information from fuzzy scatterplots and somehow claimed the invention of the referred method, it was probably Pelto et al., (1968), who firstly used a variant implemented in an automatic contouring algorithm (the moving average, which is a particular case, is already known for a much longer time). Their aim was to interpolate a given surface η with a sparse amount of irregulary spaced data points by local regression surfaces. They called the method 'moving weighted least squares' estimation and derived some important properties of which, like the interpolation of data points under certain conditions. Obviously norm L_q generalizations could equivalently be used.

It was Cleveland, (1979) who first used the method in a univariate context: the smoothing of strongly scattering time-series. He considered computational algorithms and statistical properties of the method, which he called 'locally weighted regression' (or *loess*-regression).

To avoid confusion the discussed approach will be referred to as 'moving local regression' throughout this paper, though the differences of the Pelto versus the Cleveland approach will always be remarked.

Its statistical as well as its design properties will be shown and derived and a practical example will be given.

Mathematical development in this area is not yet completed and a lot of questions still remained unanswered. The problems of optimal design or optimal weighting function are only examples. Both of them are addressed here as well as in Müller, (1987), who provides a result for the univariate case, that shows asymptotic equivalence between certain moving local regression and kernel smoothers.

The main principle behind a well designed experimental plan is obvious: Asserting hypothesis-testing or parameter estimation as the aim of the experiment, the task remains to take a (possibly given) number of measurements in a way, that either the power of the test or the precision of the estimation is maximized. Additionally restrictions on the experimental conditions (i.e. to a certain experimental region) have to be considered.

In what follows the core of experimental design theory will mainly be referred. Namely the methods for response surface design, initiated by Kiefer, (1959) and the resulting developments in the area of experimental design for regression experiments, which are naturally linked to the discussed problem.

2 Statistical structures

2.1 Definitions

One calls the set of variables:

$$\begin{pmatrix} y_{11}, y_{12}, ..., y_{1N_1}; & y_{21}, y_{22}, ..., y_{2N_2}; & ; ...; & y_{n1}, y_{n2}, ..., y_{nN_n}; \\ \sigma_1^2; & \sigma_2^2; & ; ...; & \sigma_n^2; \\ x_1; & x_2; & ; ...; & x_n; \end{pmatrix}$$

an experiment $\mathcal{E}(n, N)$, where y_{ij} denotes the observed values, σ_i^2 its variances and x_i the design-points (the so-called spectrum) ($N = \sum_{i=1}^{n} N_i$).

The set:

$$\begin{pmatrix} p_1, & p_2, & ..., & p_n \\ x_1, & x_2, & ..., & x_n \end{pmatrix}$$

where $p_i = N_i/N$ and $\sum_{i=1}^{n} p_i = 1$ one calls the normalized (experimental) design $\xi(n)$. The weights p_i can be regarded as precision or duration of the measurements.

If the number of observations is large enough, one can use a continuous design ξ instead of the discrete $\xi(n)$ and proceed with the more comfortable tools of continuous mathematics. This continuous (normalized) design ξ is then characterized through a probability measure $\xi(x)$:

$$\int_{\mathcal{X}} \xi(x)dx = 1, \qquad \xi(x) > 0, \quad \forall x \in \mathcal{X} \tag{2.1}$$

where $\mathcal{X}$ stands for the experimental region.

Such designs are called approximative designs, in contrary to exact discrete designs applied in practise, since only after rounding one gets a suitable experimental plan.

2.2 A general model

If one wants to apply methods from optimal design of experiments theory, he or she has to pose the following question initially:

> *Which mathematical model serves as as sufficiently good description of the observed process ?*

Such a (stochastic) model can be formally defined as:

$$y_i = \eta(x_i, \theta) + \epsilon_i \qquad \text{with} \quad E[\epsilon_i] = 0 \tag{2.2}$$

or

$$E[y|x] = \eta(x, \theta) \tag{2.3}$$

where $y = (y_1, ..., y_i, ..., y_n)$ is a vector of observations $x_1, ..., x_i, ..., x_n$, at the n design points, and $\eta(x, \theta)$ is the prior given structure called response-function.

If we impose the distributional assumption $\epsilon \sim N(0, \sigma^2)$, then for an observation y we can construct the so-called Fisher information matrix (continuous version) in the following way:

$$I(\theta, x) = \sigma^{-2} \partial\eta/\partial\theta (\partial\eta/\partial\theta)^T.$$

The average information matrix $M(\xi)$ associated with a certain design ξ (probability measure on $\mathcal{X}$) is given by:

$$M(\xi) = \int_{\mathcal{X}} I(\theta, x) \xi(dx).$$

which is (for independent observations) inverse proportional to the asymptotic covariance matrix of the ML-estimator for θ.

For a deeper analytical treatment one has either to simplify the model (2.2) through restricting assumptions about the structure of η or the underlying covariance structure, or to turn to even more general models as discussed below.

The question of optimal design now refers to finding a certain ξ that maximizes the information matrix (e.g. a scalar functional of it, the so-called design criterion).

One of the main consequences of approximate theory is the possibility of identifying relations between different optimality criteria. In this context it is usually referred to solving the so-called 'dual' problem. Fundamental results stem from the well-known paper of Kiefer & Wolfowitz, (1960) with the publication of their famous general equivalence theorem. A generalized result for all convex optimality criteria gives Fedorov, (1980), which will be applied for calculations in this paper.

In spite of the attractivity of this approach one has to keep in mind that its application is restricted to approximative designs (or asymptotically exact designs), whilst exact designs have to be solved combinatorically.

Concise discussion of this subject can be found in Fedorov, (1972) or Silvey, (1980), a short review is given in Müller, (1990).

2.3 Nonparametric approaches

If little is known about the structure of the process generating model, one has to confine oneself to so-called nonparametric approaches, which afford much less assumptions (just about smoothness and differentiablity) than the discussed parametric methods.

Nonparametric methods are frequently applied in the first stage of a study, as exploratory analysis to find the general shape of the underlying function. Nevertheless if it is impossible to justify strong assumptions, those methods remain the only to be reasonably used.

Most of these approaches (like kernel estimators, spline functions, moving local regression estimators, running means, bin smoothers, some of them discussed below) and also least squares regression belong to the class of linear estimators, which Buja et al., (1989) give a detailed comparison of. They have the common linear form:

$$\hat{\eta}(x) = \sum_{i=1}^{n} l_{i,n}(x) y_i \tag{2.4}$$

i.e. they are locally weighted averages of the data, with the weights $l_{i,n}(x)$ independent from y_i.

In the regression context the probably most common approach is the so-called kernel method, firstly introduced by Priestley & Chao, (1972) in a univariate context. Their weights have general form, with:

$$l_{i,n}(x) = \frac{1}{b}\int_{s_{i-1}}^{s_i} K(\frac{x-u}{b})du \tag{2.5}$$

which is for instance described in Müller, (1988). Here s_i are interpolating the x_i's, b is the so-called band-width and K the kernel function. Details and examples of kernel estimators can e.g. also be found in Müller, (1988).

A different approach are smoothing splines. They also belong to the class defined by (2.4) and can be thought of as kernel estimators with variable bandwith (for a recent survey see Silverman, (1985)). Principally they can be regarded as minimizing a somehow modified sum of squares (incorporating a penalty function for roughness defined via the second derivative η''):

$$\hat{\eta}(x) = \arg\min_{\eta} \sum_{i=1}^{n}[y_i - \eta(x_i)]^2 + \alpha \int_{\mathcal{X}} \eta''(x)^2 dx, \tag{2.6}$$

where α is the so-called smoothness parameter. Hence the function $\hat{\eta}$ is a cubic polynomial in each interval (x_i, x_{i+1}) with continuous first and second derivatives at the design points. The design problem for spline estimation is addressed, but also just in a univariate context, by Micchelli & Wahba, (1981).

Smoothing splines loose their applicability in higher dimensions, since this affords tesselation of the region and handling of a growing number of (mixed) second (or higher) derivatives. For high-dimensional generalizations of non-parametric methods some notes can be found in Ripley, (1981). For a simple polynomial problem see Spruill, (1988).

The third branch of non-parametric regression techniques, moving local least squares estimation, is discussed thoroughly in Section 3, from analysis as well as from design point of view. It avoids some drawbacks of its most serious alternatives: spline and kernel methods. The former lack computational simplicity and calculation speed, the latter introduce high bias due to local constancy.

A related method for fitting segmented polynomials regression models that in principle is a non-moving version, or a spline-estimator without differentiability conditions, is fully described in Park, (1978).

3 Moving local regression

3.1 About the model

The main idea behind the approach is that a considerably smooth function can always be approximated by a 'simpler', say polynomial function over a small region of the regressors space. At any point ${}_jx$ where an interpolated or smoothed value is desired, a weighted least squares regression is performed, with a weight function decreasing when the distances from ${}_jx$ to the data points increase (leading indices indicate fixed points, contrary to design points). A good choice of this function, fulfilling certain requirements, introduces locality, the construction of which is a major point to the problem.

It is easy to imagine, that the method operates in a moving fashion similar to kernel estimators or moving averages, introducing much computerwork, which explains the delay of developing the method fully in the past years.

The proposed nonparametric estimation method, moving local regression can effectively be used for various purposes, either to build up ideas about a possible parametric model during the explorative phase of a study or due to its rich properties itself could be used for inference about a (presumably nonlinear) process. McLain, (1971) shows its usefulness, as already mentioned via experimental results.

However it should not be concealed from the method's only serious drawback referred to as *the curse of dimensionality.* If the number of independent variables p becomes large, while n is fixed, the interpoint distances grow, which tends to increase the bias of the estimator. Although this effect can be reduced by altering the locality of the method via the weight function higher dimensional problems should be handeled with caution.

3.2 Some useful refinements - local linear approximation

Suppose that the data y_i, $i = 1, \ldots, n$, that are observed at the p-dimensional points x_i, $i = 1, \ldots, n$, are generated by the following mechanism:

$$y_i = \eta(x_i) + u_i, \tag{3.1}$$

where $\eta(x)$ is any considerably smooth response from x and the u_i are assumed to be random variables with $E[u_i] = 0, E[u_i u_{i'}] = \sigma_u^2 \delta_{ii'}$.

Then at any point x_i the response $\eta(x_i)$ can for instance be approximated by some Taylor-expansion from some arbitrary point ${}_jx$ in the vicinity of x_i:

$$\eta(x_i) = \theta_{0j} + \theta_j^T d_{ij} + r_{ij}, \qquad i = 1, \ldots, n, \tag{3.2}$$

where $d_{ij} = x_i - {}_jx$ ($\bar{d}_{ij} = |d_{ij}|^{1/2}$ is the distance between x_i and ${}_jx$) and r is the remainder term of the approximation, vanishing as $o(d_{ij})$ when $x_i \to {}_jx$.

Equivalently r could be thought of being a local stochastic disturbance term, with variance σ_r^2 following the natural properties: $\sigma_r^2 = 0$ if $x_i = {}_jx$ (or $\bar{d}_{ij} = 0$) and $\lim_{\bar{d}_{ij}\to\infty} \sigma_r^2 = \infty$. It simulates the fact that the approximation gets worse and r gets large, when the distance increases.

Obviously the choice of this approximation being linear is arbitrary, polynomial terms of higher order or even nonlinear functions, which are supported by physical considerations, could be used instead. Though Cleveland, (1979) spends some efforts in investigating the form of this function, any complication at that stage does not add to the richness of the method enough to justify the loss in simplicity. Anyhow this point will be discussed in a later section.

The regression function can now be rewritten as

$$y_i = \theta_{0j} + \theta_j^T d_{ij} + \epsilon_{ij}, \qquad i = 1, \ldots, n, \tag{3.3}$$

where $\sigma_\epsilon^2 = \sigma_u^2 + \sigma_r^2$, because of independence of u_i and r_{ij}.

A sensible estimator for θ_0 and θ is defined as

$$\{\hat{\theta}_{0j}, \hat{\theta}_j\} = \arg\min_{\theta_0,\theta} \sum_{i=1}^{n} \lambda(\bar{d}_{ij})[y_i - \theta_{0j} - \theta_j^T d_{ij}]^2, \tag{3.4}$$

where the so-called weight function λ reflects the reliability of the given Taylor-expansion (or the influence of local stochastics).

From Equation (3.2) it clearly follows that $\hat{\eta}({}_jx) = \hat{\theta}_{0j}$. Since $\hat{\theta}_{0j}$ is a linear estimator, the given estimator can be considered a moving locally weighted average of the data (compare Cleveland, (1979) and (2.4)):

$$\hat{\eta}({}_jx) = \sum_{i=1}^{n} l(\bar{d}_{ij}) y_i \tag{3.5}$$

or

$$\hat{\eta}({}_jx) = L_j y, \tag{3.6}$$

where $y^T = [y_1, \ldots, y_n]$ and $L_j = [l(\bar{d}_{1j}), \ldots, l(\bar{d}_{nj})]$. L_j depends upon the x_i's and λ but not on the y_i's. This fact allows to derive distributional results given in the sections below. Sometimes estimators with that property are referred to as linear smoothers, see Buja et al., (1989) (compare (2.4)).

3.3 How to compute

As already mentioned moving local regression, applied rigorously, affords a lot of tedious computations, since at every point ${}_jx$ to be evaluated a standard weighted least squares regression has to be performed. Fedorov, (1989) gives the necessary formulae derived from the well-known technique:

$$\begin{Bmatrix} \hat{\theta}_{0j} \\ \hat{\theta}_j \end{Bmatrix} = M_j^{-1} Y_j, \tag{3.7}$$

where

$$Y_j = \frac{1}{\dot{\lambda}_j}\begin{pmatrix} \lambda_j^T \\ d_j^T \Lambda_j \end{pmatrix} y, \qquad M_j = \frac{1}{\dot{\lambda}_j}\begin{bmatrix} \dot{\lambda}_j & \lambda_j^T d_j \\ d_j^T \lambda_j & d_j^T \Lambda_j d_j \end{bmatrix},$$

and $\dot{\lambda}_j = \sum_{i=1}^{n} \lambda(\bar{d}_{ij})$, $d_j^T = [d_{1j}, \ldots, d_{nj}]$, as well as $\lambda_j^T = [\lambda(\bar{d}_{1j}), \ldots, \lambda(\bar{d}_{nj})]$ and $\Lambda_j = \mathrm{diag}(\lambda_j)$.

Following (2.4) or (3.6) it is possible to express L_j for this particular model (local linear approximation) explicitely:

$$L_j^T = \left[\left(\dot{\lambda}_j - \lambda_j^T d_j \left(d_j^T \Lambda_j d_j\right)^{-1} d_j^T \lambda_j\right)^{-1} \lambda_j^T - \lambda_j^T d_j \left(\dot{\lambda}_j d_j^T \Lambda_j d_j - d_j^T \lambda_j \lambda_j^T d_j\right)^{-1} d_j^T \Lambda_j\right].$$

Proof:

Inversion of M_j, following Proposition (31) of Dhrymes, (1984) yields a symmetric matrix $A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$, where

$$\begin{aligned} A_{11} &= \dot{\lambda}_j \left(\dot{\lambda}_j - \lambda_j^T d_j \left(d_j^T \Lambda_j d_j\right)^{-1} d_j^T \lambda_j\right)^{-1} \\ A_{12} &= -\lambda_j^T d_j \left(d_j^T \Lambda_j d_j - \frac{1}{\dot{\lambda}_j} d_j^T \lambda_j \lambda_j^T d_j\right)^{-1} d_j^T \Lambda_j\Big]. \end{aligned}$$

Then by multiplication of Y_j it easily follows that:

$$L_j^T = \left[\left(\dot{\lambda}_j - \lambda_j^T d_j \left(d_j^T \Lambda_j d_j\right)^{-1} d_j^T \lambda_j\right)^{-1} \lambda_j^T - \lambda_j^T d_j \left(\dot{\lambda}_j d_j^T \Lambda_j d_j - d_j^T \lambda_j \lambda_j^T d_j\right)^{-1} d_j^T \Lambda_j\right].$$

qed

Buja et al., (1989) compare different linear estimators by plotting L_j against x_i. These plots, referred to as equivalent kernels, reflect the form of the neighbourhood and the influence of the weight function.

3.4 About the weight function

The usefulness of the given method enormously depends upon the used weight function λ. Two properties are essential to allow a sensible interpretation:

(*a*) $\lambda(\bar{d})$ is a nonincreasing function for $\bar{d} \geq 0$ ($\lambda(0) = \max \lambda$).

(*b*) $\lim_{\bar{d} \to \infty} \lambda(\bar{d}) = 0$.

Altering the weight function within this framework gives us a wide range of possibilities to adjust the method to specific problems.

If the given approach is used for interpolation of a surface, an additional condition on the weight function is needed to ensure that the surface fits exactly to the data. Pelto et al., (1968) provide the proof, that

(c_1) $\lim_{\bar{d}\to 0} \lambda(\bar{d}) = \infty$, s.t. $\int_{-\infty}^{\infty} \lambda(\bar{d}) = 1$

ensures interpolation. It is clear that through this condition the smoothing of the surface can be governed, as Cleveland et al., (1988) and Cleveland & Devlin, (1988) do in their applications. Buja et al., (1989) mention the problem of choosing smoothness parameters according to the data, which destroys the linear character of $\hat{\eta}$. This is generally avoided by the assumption, that the parameters are fixed a priori.

Another important notion is the concept of locality of the method. As Ripley, (1981) indicates it is not guaranteed that $\hat{\eta}({}_j x)$ is a weighted average of mainly local values if the region of interest is expanded, unless

(d_1) $\lim_{\bar{d}\to\infty} \bar{d}^p \lambda(\bar{d}) = 0$,

where p has for instance to be defined by the order of the Taylor's expansion.

Proof:

For a homogenous arrangement of data-points it is assumed that there are about $\text{const}\bar{d}^{p-1}\Delta\bar{d}$ data points from a distance $\bar{d}$ to $\bar{d} + \Delta\bar{d}$ from a point $x \in D$, since the volume of a p-dimensional sphere of radius $\bar{d}$ is $\text{const}\bar{d}^p$.

Hence the volume of the given area is $\text{const}[(\bar{d} + \Delta\bar{d})^p - \bar{d}^p]$. Since the first term of the left hand polynomial $\bar{d}^p$ vanishes, its dominating part is the second term $\bar{d}^{p-1}\Delta\bar{d}$.

These points contribute about $\text{const}\bar{d}^{p-1}\Delta\bar{d}\lambda(\bar{d})$ to the total of weights. Thus unless the integral $\int_1^{\infty} \bar{d}^{p-1}\lambda(\bar{d})d\bar{d}$ is not finite, $\hat{\eta}$ entirely depends upon how large D is chosen.

But if now $\lambda(\bar{d}) = o(\bar{d}^{-p})$ is guaranteed as $\bar{d} \to \infty$, then if for instance $\lambda(\bar{d}) = \bar{d}^{-p-\alpha}$ it follows:

$$\int_1^{\infty} \bar{d}^{p-1}\bar{d}^{-p-\alpha}d\bar{d} = \int_1^{\infty} \bar{d}^{-1-\alpha}d\bar{d} = -\frac{\bar{d}^{-\alpha}}{\alpha}\Big|_1^{\infty} = -\frac{1}{\alpha\bar{d}^{\alpha}}\Big|_1^{\infty} = \frac{1}{\alpha} < \infty$$

qed

Following this condition independence of the results from the choice of the region of interest is assured. Cleveland et al., (1988) and Cleveland & Devlin, (1988) avoid such difficulties by assigning weight zero to a certain percentage f of data points due to their remoteness

(c_2) $\lambda(\bar{d}) = 0$ if $\bar{d})\bar{d}_f$,
where $\bar{d}_f$ is the distance of the $f.n$ nearest point to ${}_j x$.

Thus only points in the neighbourhood of ${}_j x$ contribute to the estimation, which of course provides another tool for handling locality problems, although it could cause a lot of computational problems in multidimensional cases. This also helps in resolving the problem of clusters in the data-points, which could otherwise lead to undesirable bias effects. On the other hand discretizing the weight function as done above may lead to a loss of smoothness of the estimated surface, which will be shown below.

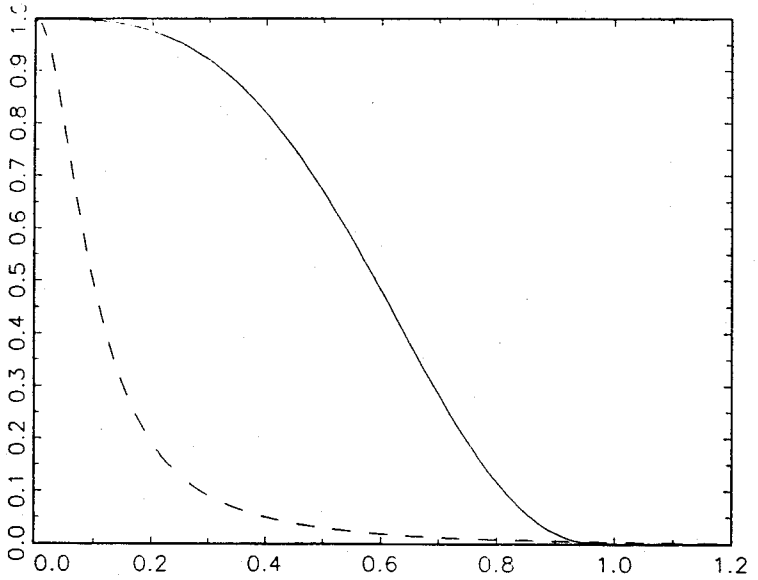

Figure 1: dotted - normalized McLain's (4.8), solid - Cleveland's (4.9) weight function

One of the most important questions related with the weight function is, which λ to choose to guarantee a considerably smooth, as many times as possible differentiable estimated surface $\hat{\eta}$? Silvey, (1980) states, that in general surfaces generated by considering conditions (a),(b),(c_1),(d_1) are as many times differentiable as $1/\lambda(|d|)$ is $(d \in \mathbf{R})$. Pelto et al., (1968) prove, that if

(e_1) $\lambda(\bar{d}) \propto \bar{d}^{-p}$, $\quad p$ even,

the surface will be infinitely differentiable, which could be elementary checked through using (3.7) and finding the derivatives $\partial\hat{\theta}_{0j}({}_jx)/\partial_j x$.

Out of practical considerations McLain, (1971) suggested the weight function

$$\lambda(d) = \exp(-\bar{d}^2/\bar{d}_n^2)/(\bar{d}^2 + \delta), \tag{3.8}$$

where $\bar{d}_n$ is the average distance between neighbouring data points and constant δ is used to avoid arithmetic overflow.

A computationally simpler function, the so-called tricube, fulfilling (a), (b), (c_2) is used by Cleveland, (1979):

$$\lambda = \begin{cases} (1 - (\bar{d}/\bar{d}_f)^3)^3 & 0 \le \bar{d}/\bar{d}_f \le 1 \\ 0 & \text{else} \end{cases} . \tag{3.9}$$

This function smoothly decreases from 1 to 0 in the interval [0,1]. Therefore it is believed to almost always provide a smooth enough surface.

The cited result by Müller, (1987) for the univariate case, that shows asymptotic equivalence between certain moving local regression and kernel smoothers implies the following additional restriction to the weight function: (c_3) $\lambda(\bar{d}) = 0$ if $\bar{d} > \delta$, with constant δ, which is equivalent to (c_2) with equally spaced data points.

3.5 Choice of the local fitting - the bias problem

As the remark to (3.2) indicates there exists some freedom in choosing alternative local fitting schemes versus the discussed linear approach. The point is that insufficient local

	f=0.1	f=0.3	f=0.5	f=0.7	δ=0.001	δ=0.01
m.s.e.	4.979	2.653	2.788	6.083	3.860	2.979
bias^2	0.013	0.179	1.339	5.042	0.076	0.244
variance	4.966	2.473	1.449	1.041	3.783	2.735
bias^2%	0.255	6.761	48.028	82.887	1.976	8.200
variance%	99.745	93.239	51.972	17.113	98.024	91.800

Table 1: Performance characteristics for local approximation schemes.

fitting may introduce a bias to the estimator, especially if the curvature of the original surface is high. It is well known that the variance and bias of $\hat{y}$ is related to its mean square error as:

$$E[(\hat{y}-y)^2] = (E[\hat{y}]-y)^2 + \mathrm{Var}[\hat{y}].$$

Cleveland, (1979) argues that increasing the neighbourhood (see section 5) tends to decrease the variance term of the given relation. On the other hand the bias keeps growing along with the neighbourhood, which of course is a fundamental problem to deriving statistical properties, as will be shown below.

A proof by C.L. Mallows, cited by Cleveland, (1979) for a constant local fitting, that is

$$y_i = \theta_0({}_j x) + \epsilon_{ij}, \qquad i = 1,\ldots,n, \tag{3.10}$$

confirms the above given statement. Note that (3.10), which McLain, (1971) calls *Shepard*'s method, with a uniform weight function is equivalent to the traditional moving average smoothing.

Proof:

Taking the usual steps model (3.10) leads us to the following estimation scheme:

$$\hat{\theta}_{0j} = M_j^{-1} Y_j,$$

where of course, see Equation (3.7) $Y_j = \frac{1}{\lambda_j}\lambda_j^T y$ and elementary $M_j = 1$, so that

$$\hat{\theta}_{0j} = \frac{1}{\lambda_j}\sum_{i=1}^{n}\lambda(\bar{d}_{ij})y_i.$$

That is for a $\lambda(\bar{d}_{ij}) = \begin{cases} 1 & \bar{d}_{ij} < c \\ 0 & \text{else} \end{cases}$, it is similar or in equidistant case equal to moving averaging.

qed

A simulation experiment with 5000 trials for different weight functions, the results of which are reported in Table 1, provides additional evidence. The investigated relationship was a polynomial of second order ($\eta = x^2$) at point 0 with $\sigma_\epsilon^2 = 0.1$.

It is clear, that through introducing more sophisticated local fitting schemes it will be possible to absorb the bias effect, though of course the computational burden may increase tremendously (as for local quadratic fitting $2p + p(p-1)/2$ parameters are to be estimated).

Cleveland, (1979),Cleveland et al., (1988) and Cleveland & Devlin, (1988) use the so-called M-plot as a tool for deciding between local polynomial approximation methods, while Fedorov, (1989) demands support for a specific scheme by prior knowledge.

Experimental results for deterministic surfaces, that is (3.1) with $\sigma_u^2 = 0$, given by McLain, (1971) favorize moving local regression with local second degree polynomial fit not only versus moving local regression with lower degree, but also versus other parametric and nonparametric estimation methods.

3.6 Statistical properties

To investigate distributional properties of moving local regression two additional assumptions to the model are needed:

[a] the errors are normally distributed, that is $\epsilon \sim N(0, \sigma_\epsilon^2)$, and

[b] $\hat{\eta}$ estimates η without bias.

The latter assumption is guaranteed only if the fitting scheme corresponds to the structure of the true relationship. On the other hand it was shown in Subsection 3.5 that through altering the neighbourhood via the weight function high bias may be prevented and assumption [b] will at least approximately hold.

The key point to deriving distributional results for the given methodology is the linear structure of $\hat{\eta}$. It clearly follows from the assumptions and formula (3.6) that

$$\hat{\eta}({}_j x) \sim N\Big(\eta({}_j x), \sigma_\epsilon^2 L_j L_j^T\Big) \tag{3.11}$$

Thus if $\hat{y}_i = \hat{\eta}(x_i)$ are the fitted values, it is possible to write in matrix presentation:

$$\hat{y} = Ly \qquad \text{and} \qquad \hat{\epsilon} = (I - L)y, \tag{3.12}$$

where $L^T = [L_1^T, \ldots, L_i^T, \ldots, L_n^T]$. Both $\hat{y}$ and $\hat{\epsilon}$ are then multivariate normal with covariance matrices $\sigma_\epsilon^2 LL^T$ and $\sigma_\epsilon^2 (I-L)(I-L)^T$ respectively.

Cleveland, (1979) remarks the analogy to parametric least squares and uses well known techniques from that field to derive further results:

Let $\pi_k = \mathrm{tr}[(I-L)(I-L)^T]^k$, then because of unbiasedness the residual sum of squares $E[\hat{\epsilon}^T\hat{\epsilon}] = \sigma_\epsilon^2 \pi_1$ and thus the variance of $\eta({}_j x)$ can be estimated by $\hat{\sigma}_\epsilon^2 L_j L_j^T$. Now since $(\pi_1^2 \hat{\sigma}_\epsilon^2 / \pi_2 \sigma_\epsilon^2)$ can be approximated by a χ^2 distribution with π_1^2/π_2 degrees of freedom, it becomes possible to derive approximate confidence intervals for $\hat{\eta}({}_j x)$ following the traditional methodology.

Buja et al., (1989) also introduced the notion degrees of freedom into this context. The most straightforward approach is to borrow the definition from the traditional linear model ($d.f. = \mathrm{tr}(LL^T)$).

3.7 How to design

In practise the following (design) problem could frequently occur: At some points of interest ${}_jx$ of a region $\mathcal{X}$ the unknown response η has to be interpolated (smoothed). The question is where to take observations (at which x_i's) within the prescribed region $\mathcal{X}$ (which in principle may be infinitly expanded) in order to estimate $\eta({}_jx)$ in a most efficient way using the scheme given above.

Such a set ξ_n^* of x_i's minimizing a so called optimality criterion $\phi(\xi_n)$ is referred to as an optimal design, following traditional experimental design theory (see for instance Silvey, (1980)).

If one looks at presentation (3.12) of the estimator, one could guess that it is reasonable to use a scalar function of $\mathrm{Cov}(\hat{\eta}) = L\,\mathrm{Cov}(y)L^T$ as the criterion of the design problem. Since the D-criterion in this context would be cumbersome to handle Fedorov, (1989) indicates that we can choose a weighted sum of the variance of the estimates $\hat{\eta}({}_jx)$ as a sensible objective function ϕ (e.g. the A-criterion).

The optimal design is then given by:

$$\xi_n^* = \arg\min_{\xi_n} \sum_{j=1}^{m} a_j \ \mathrm{var}(\hat{\eta}({}_jx)), \tag{3.13}$$

where a_j reflects the importance of ${}_jx$, the arguments coming from the diagonal of $\mathrm{Cov}(\hat{\eta})$.

If we now impose a new assumption about the error structure of model (3.3):

$$E[\epsilon_{ij}] = 0, \qquad E[\epsilon_{ij}\epsilon_{i'j}] = V_j, \tag{3.14}$$

then the variance of an estimator can be easily expressed as:

$$\mathrm{var}(\hat{\eta}({}_jx)) = \mathrm{tr}\, AM_j^{-1}, \tag{3.15}$$

where $A = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{pmatrix}$.

Proof:

Using (3.7) the variance-covariance matrix of $\begin{Bmatrix} \hat{\theta}_0({}_jx) \\ \hat{\theta}({}_jx) \end{Bmatrix}$ is given by

$$(F_j^T V_j^{-1} F_j)^{-1} F_j^T V_j^{-1}\, \mathrm{Cov}(y) V_j^{-1} F_j^T (F_j^T V_j^{-1} F_j)^{-1}.$$

Following (3.14), hence:

$$\mathrm{Cov}(\begin{Bmatrix} \hat{\theta}_0({}_jx) \\ \hat{\theta}({}_jx) \end{Bmatrix}) = (F_j^T V_j^{-1} F_j)^{-1}.$$

Extracting the upper left element yields:

$$\mathrm{Var}(\hat{\theta}_0) = \mathrm{tr}\, A(F_j^T V_j^{-1} F_j)^{-1},$$

which due to (3.4) is equivalent to (3.15).
qed

Using (3.15) the optimization criterion can directly be formulated:

$$\phi(\xi_n) = \operatorname{tr}\sum_{j=1}^{m} A_j (F_j^T V_j^{-1} F_j)^{-1} \tag{3.16}$$

$$= \operatorname{tr}\sum_{j=1}^{m} A_j M_j(\xi_n)^{-1}, \tag{3.17}$$

where $A_j = a_j A$, with a_j being a scalar.

Theorem 1 *Following the general theory, the minimizing optimality criterion (3.17) is equivalent to obeying the condition:*

$$\min_i \varphi(x_i, \xi) \geq 0, \tag{3.18}$$

with

$$\varphi(x_i, \xi) = tr\sum_{j=1}^{m} A_j M_j(\xi)^{-1} - \sum_{j=1}^{m} \frac{\lambda(\bar{d}_{ij})}{\lambda_j} F_{ij}^T M_j(\xi)^{-1} A_j M_j(\xi)^{-1} F_{ij}, \tag{3.19}$$

where F_{ij} is a particular row from F_j.

Proof:

For the sake of simplicity we now turn from exact designs ξ_n to approximate designs ξ, which impose a measure on the space spanned by the regressors.

It is clear that if ξ^* is the optimal design and $\bar{\xi}$ is any other, that the following inequality holds at point $\xi = \xi^*$:

$$\frac{\partial \phi(\xi)}{\partial \alpha} \geq 0, \tag{3.20}$$

if $\xi = (1-\alpha)\xi^* + \alpha\bar{\xi}$, with $M_j(\xi) = (1-\alpha)M_j(\xi^*) + \alpha M_j(\bar{\xi})$ (see for instance Fedorov, (1980) or Silvey, (1980)).

Evaluating (3.17) yields:

$$\frac{\partial \phi(\xi)}{\partial \alpha} = \frac{\partial \operatorname{tr}\sum_{j=1}^{m} A_j M_j(\xi)^{-1}}{\partial \alpha}$$

$$= \operatorname{tr}\sum_{j=1}^{m} A_j \frac{\partial M_j(\xi)^{-1}}{\partial \alpha} \geq 0,$$

and following Corollary (41) of Dhrymes, (1984)

$$= -\operatorname{tr}\sum_{j=1}^{m} A_j M_j(\xi)^{-1} \frac{\partial M_j(\xi)}{\partial \alpha} M_j(\xi)^{-1} \geq 0,$$

after inserting $M_j(\xi) = (1-\alpha)M_j(\xi^*) + \alpha M_j(\bar{\xi})$:

$$\operatorname{tr}\sum_{j=1}^{m} A_j M_j(\xi)^{-1} \frac{\partial(1-\alpha)M_j(\xi^*) + \alpha M_j(\bar{\xi})}{\partial\alpha} M_j(\xi)^{-1} \leq 0$$

$$= \operatorname{tr}\sum_{j=1}^{m} A_j M_j(\xi)^{-1}[M_j(\bar{\xi}) - M_j(\xi^*)]M_j(\xi)^{-1}.$$

If one now lets $\alpha \to 0$ then:

$$\operatorname{tr}\sum_{j=1}^{m} A_j M_j(\xi)^{-1} M_j(\bar{\xi}) M_j(\xi)^{-1} \leq \operatorname{tr}\sum_{j=1}^{m} A_j M_j(\xi^*)^{-1}$$

Now it is assumed that $\bar{\xi}$ only consists of one point x_i then $M_j(\bar{\xi}) = X_{ij}^T \frac{\lambda(\bar{d}_{ij})}{\Lambda_j} X_{ij}$ which yields:

$$\varphi(x_i, \xi) = \operatorname{tr}\sum_{j=1}^{m} A_j M_j(\xi)^{-1} - \sum_{j=1}^{m} \frac{\lambda(\bar{d}_{ij})}{\Lambda_j} X_{ij}^T M_j(\xi)^{-1} A_j M_j(\xi)^{-1} X_{ij}.$$

qed

It is obvious that for the usefulness of the criterion (3.13) the assumptions, that $\hat{\eta}$ is an unbiased estimator, which of course holds only in special cases, is fundamental, since otherwise the summed mean square error instead of the variance has to be used. Nevertheless it seems justified to derive further properties still keeping this assumption, since the amount of the bias can to some extent be governed by the weight function.

For his specially restricted moving local regression estimate Müller, (1984) and Müller, (1988) also addresses the problem of optimal design. He uses the integrated mean squared error (no special points of interest) as criterion function and finds the optimal design:

$$\xi^*(x) = \frac{\sigma(x)}{\int_0^1 \sigma(u)du} \tag{3.21}$$

where σ denotes a smooth variance function.

3.8 Optimizing the weight function

Very frequently when applying the discussed approach, the design at the first step might be given, and the aim of the statistician might not be the optimizing of the design, but in a given framework to utilize the weight function, that serves his aims best for a given design.

If we rewrite the basic model (3.3) as:

$$y_i = \theta_{0j} + \theta_j^T d_{ij} + \delta_j^T \varphi(d_{ij}) + \epsilon_{ij}, \qquad i = 1, \ldots, n, \tag{3.22}$$

where $\delta_j^T \varphi(d_{ij})$ is now the remainder term of the expansion and ϵ_{ij} real error, then by means of minimizing the dispersion matrix of the estimators one might deduce a somehow optimal weight function.

Fedorov derived in a proof, transmitted by personal communication, this weight function for the simplest case (moving average), when $\theta_j = 0$. Here the dispersion matrix of the intercept turns out to be the scalar $(\delta_j^2(\lambda^T\varphi(d_{ij}))^2 + \sigma_\epsilon^2\lambda^T\lambda)/\sum_{i=1}^n \lambda_i^2$. From straightforward minimization follows:

Theorem 2

$$\lambda_{ij}^* = n^{-1}[1 - \frac{\gamma^2(\varphi(d_{ij}) - \bar{\varphi}_j)\bar{\varphi}_j}{n^{-1} + \gamma^2 d(\varphi_j)}] \tag{3.23}$$

where $\gamma^2 = \delta^2/\sigma_\epsilon^2$, $\bar{\varphi}_j = n^{-1}\sum_{i=1}^n \varphi(d_{ij})$ *and* $d(\varphi_j) = n^{-1}\sum_{i=1}^n(\varphi(d_{ij}) - \bar{\varphi}_j)^2$.

The analysis could become complicated, if the λ_{ij}^* get negative, which can happen if for instance point n:

$$\varphi(d_{nj}) \leq \frac{\sum_{i=1}^{n-1}\varphi(d_{ij})^2}{\sum_{i=1}^{n-1}\varphi(d_{ij})}$$

Proof:

It is obvious, that λ_{nj}^* in Theorem 2 can only become negative, if:

$$\begin{aligned}
\gamma^2(\varphi(d_{nj}) - \bar{\varphi}_j)\bar{\varphi}_j &> n^{-1} + \gamma^2 d(\varphi_j), \\
\gamma^2(\varphi(d_{nj})\bar{\varphi}_j - \bar{\varphi}_j^2) &> n^{-1} + \gamma^2(n^{-1}\sum_{i=1}^n \varphi(d_{ij})^2 - \bar{\varphi}_j^2), \\
\varphi(d_{nj})\bar{\varphi}_j - \bar{\varphi}_j^2 &> n^{-1}\sum_{i=1}^n \varphi(d_{ij})^2 - \bar{\varphi}_j^2, \\
\varphi(d_{nj})\bar{\varphi}_j &> n^{-1}\sum_{i=1}^n \varphi(d_{ij})^2, \\
\varphi(d_{nj})\sum_{i=1}^n \varphi(d_{ij}) &> \sum_{i=1}^n \varphi(d_{ij})^2, \\
\varphi(d_{nj})^2 + \sum_{i=1}^{n-1}\varphi(d_{nj})\varphi(d_{ij}) &> \varphi(d_{nj})^2 + \sum_{i=1}^{n-1}\varphi(d_{ij})^2,
\end{aligned}$$

Hence it easily follows, that positivity of the weight function is guaranteed by:

$$\varphi(d_{nj}) \leq \frac{\sum_{i=1}^{n-1}\varphi(d_{ij})^2}{\sum_{i=1}^{n-1}\varphi(d_{ij})}$$

qed

For the more general case of models of local fit of arbitrary order one can suggest the minimization of the dispersion matrix utilizing algorithms similar to those described for design optimization. For the case of local linear approximation and symmetrical three-point design and true quadratic relationship for instance one can find the weight function:

$$\lambda_{ij}^* = \frac{1}{1 + 2\gamma^2 d_{ij}^4}, \tag{3.24}$$

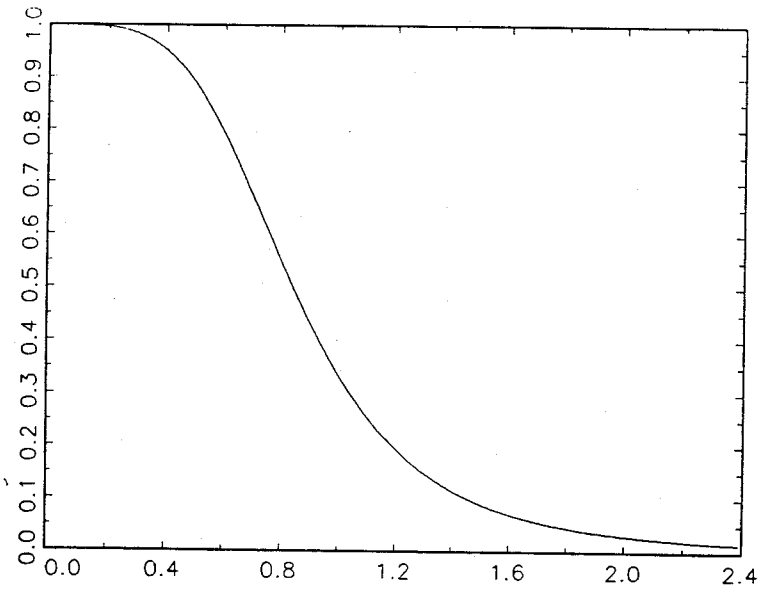

Figure 2: normalized Fedorov's (4.24) weight function

which will be referred to as Fedorov's weight function in the section, where respective examples will be calculated.

The question of optimal weight function was also in the univariate case addressed by Müller, (1987). He suggests symmetric polynomial weight functions of the form:

$$\lambda_i^* = \sum_{k=0,\text{even}}^{2\mu} \frac{(-1)^{k/2}(k+2)(2\mu+2)!}{(1+k/2)!(\mu-k/2)!(\mu+1)!2^{2\mu+3}} x_i^k 1_{[-1,1]}$$

since they not only minimze the asymptotic variance of the estimator, but also the asymptotic variances of the μ-th derivative.

4 Example: Optimal location of air-quality measurement stations

The applied example is an extension of the results by Fedorov & Müller, (1989). Here the design of an air pollution monitoring network is considered. Note that the use of moving local regression is more common in environmental statistics than in other application fields (see for instance Cleveland & McRae, (1988)).

It has to be emphasized that all designs have been calculated thoroughly numerically and the corresponding algorithm was stopped after about 100 iterations, which may lead to small differences to the 'true' optimal designs. For optimization algorithms refer to Silvey, (1980).

In Upper-Austria exists an SO_2-monitoring network (see Figure 3), that was obviously designed following political and cost considerations, and it was already shown in Fedorov & Müller, (1989), that this network does not allow optimal inference in various ways.

The probably most important question for observers, if they decide to alter the existing network is how to relocate the existing equipment in an optimal way. Some solutions for this question, assuming a rather rigid model (polynomial of second order) for the distribution of SO_2 concentrations can be found in Fedorov & Müller, (1989).

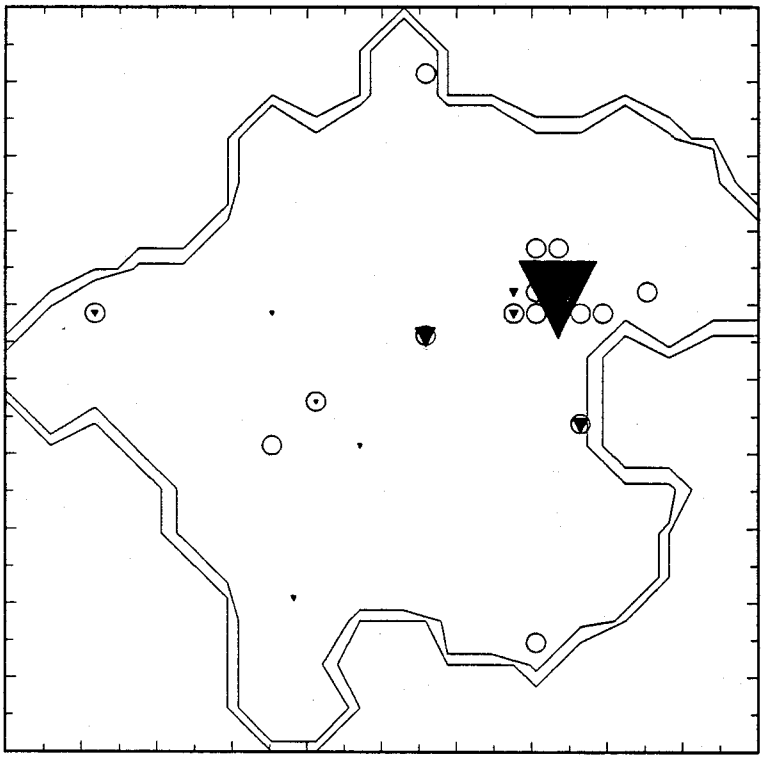

Figure 3: Upper-Austrian SO_2 monitoring network

Now, with the help of the methods developed in the previous sections, it is possible to solve this problem in a more elegant way.

If we define sensible points of interest and weight function, we can immediately use the moving local regression approach for estimation of the SO_2 concentration. This will allow the concentrations to form a nonlinear surface over the given region. The methods for optimal design, then help us to find the optimal location of stations, if for instance we use the so-called exchange type algorithm, which allows only equally weighted points at different locations in space.

Again only local linear approximation was used, since this is flexible enough and quadratic approximation induces rather slow convergence of the design optimization algorithm.

The points of interest were chosen at the cities with more than 10000 inhabitants and their influence was weighted with the city size. In the graphs this can be seen as the different sizes of the black triangles (the largest at the capital Linz).

McLain's weight function (see Figure 4) gives an optimal design closest to the existing network, due to its high locality and the fact that originally most of the design points were chosen near large cities. It can be seen, that with reducing the locality (Cleveland (Figure 5) and even more Fedorov (Figure 6)) points tend to the boundaries of the region as in the standard regression case.

In general one has to say, that despite no mathematics were presumably involved in the creation of the original network, it is fairly good choice, seen from its resemblance to the calculated examples.

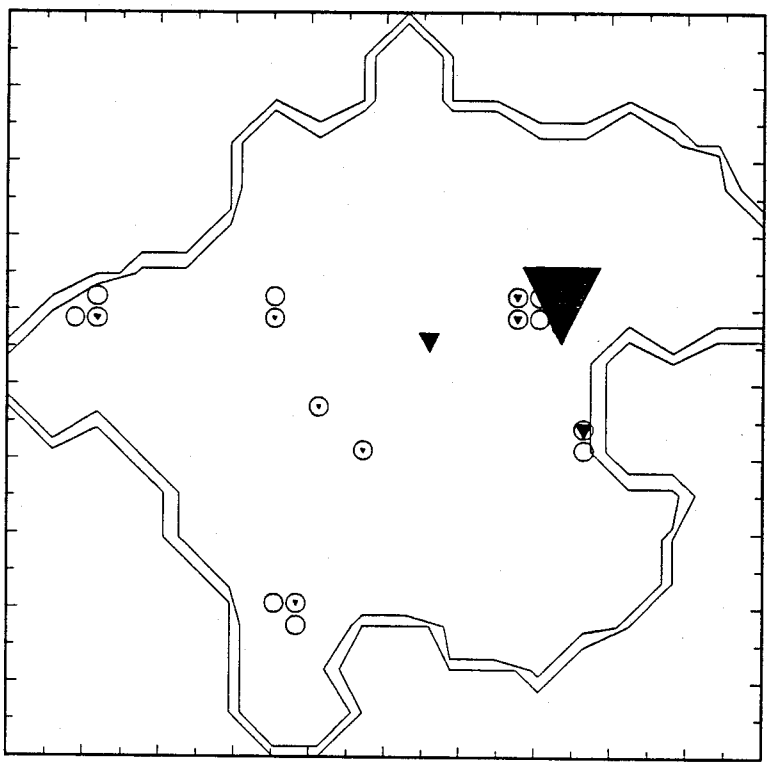

Figure 4: Optimal network with McLain's weight function

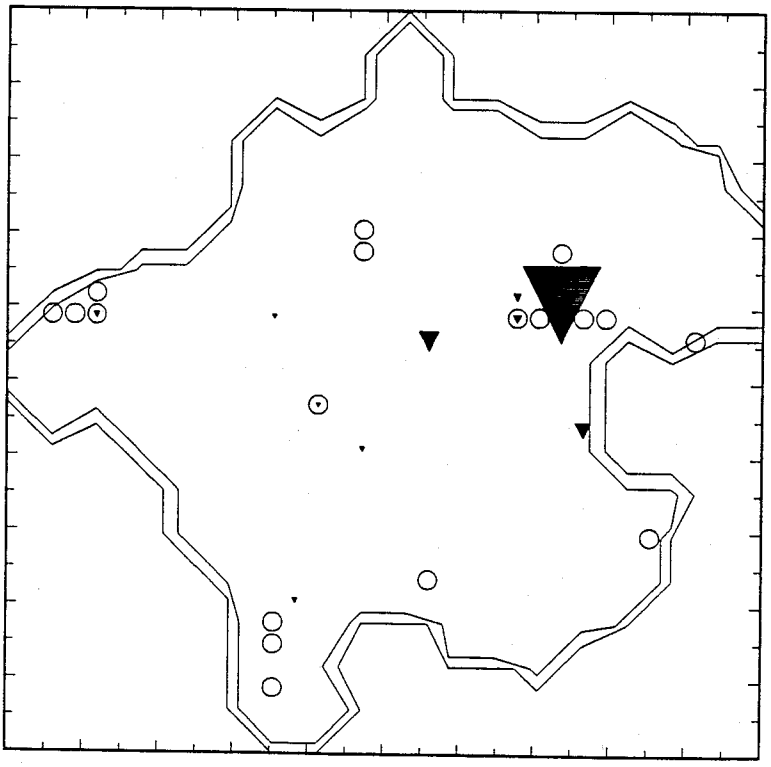

Figure 5: Optimal network with Cleveland's weight function

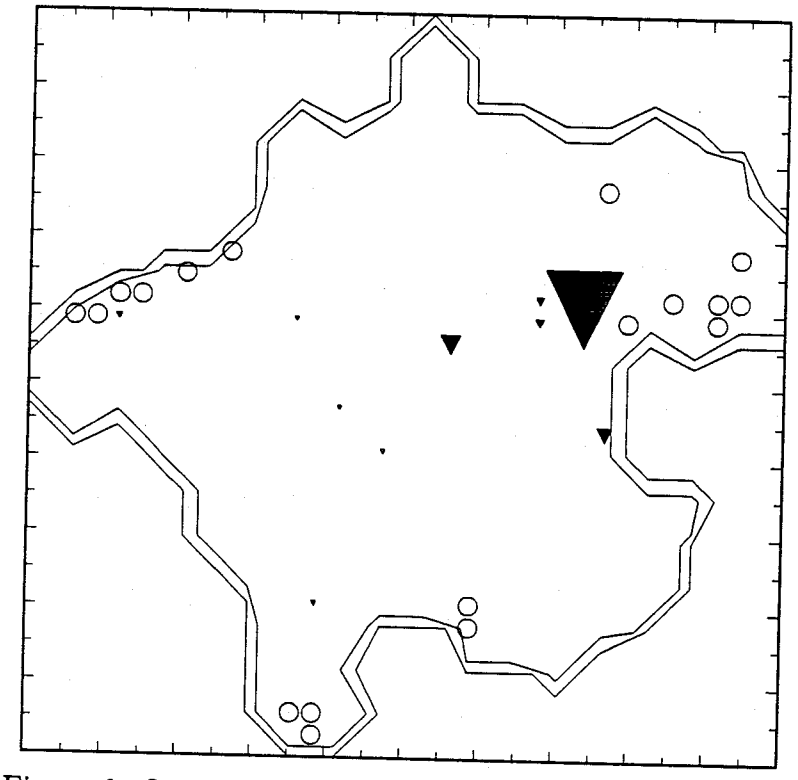

Figure 6: Optimal network with Fedorov's weight function

References

A. Buja, T. Hastie, and R. Tibshirani. Linear smoothers and additive models with discussion. *The Annals of Statistics*, 17(2):453–555, 1989.

W.S. Cleveland and S. Devlin. Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83(403):596–610, 1988.

W.S. Cleveland and J.E. McRae. *The Use of Loess and STL in the Analysis of Atmospheric CO_2 and related data.* Technical Report 67, AT&T Bell Laboratories, Statistical Research Reports, Murray Hill, N.J. 07974, 1988.

W.S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836, 1979.

W.S. Cleveland, S.J. Devlin, and E. Grosse. Regression by local fitting. *Journal of Econometrics*, 37:87–114, 1988.

P.J. Dhrymes. *Mathematics for Econometrics.* Springer Verlag, New York, 1984.

V.V. Fedorov and W. Müller. Design of an air-pollution monitoring network. an application of experimental design theory. *Österreichische Zeitschrift für Statistik und Informatik*, 1:5–18, 1989.

V.V. Fedorov. *Theory of Optimal Experiments.* Academic Press, New York, 1972.

V.V. Fedorov. Convex design theory. *Mathematische Operationsforschung und Statistik, Series Statistics*, 11(3):403–413, 1980.

V.V. Fedorov. Kriging and other estimators of spatial field characteristics. *Atmospheric Environment*, 23(1):175–184, 1989.

J. Kiefer and J. Wolfowitz. The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 14:363–366, 1960.

J. Kiefer. Optimal experimental designs (with discussion). *Journal of the Royal Statistical Society*, B:272–319, 1959.

D.H. McLain. Drawing contours from arbitrary data points. *The Computer Journal*, 17(4):318–324, 1971.

C.A. Micchelli and G. Wahba. Design problems for optimal surface interpolation. In *Approximation Theory and Applications*, pages 329–349, Academic Press, 1981.

H.G. Müller. Optimal designs for nonparametric kernel regression. *Statistics & Probability Letters*, 2:285–290, 1984.

H.G. Müller. Weighted local regression and kernel methods for nonparametric curve fitting. *Journal of the American Statistical Association*, 82(397):231–238, 1987.

H.G. Müller. *Nonparametric Regression Analysis of Longitudinal Data.* Volume 46 of *Lecture Notes in Statistics*, Springer-Verlag, 1988.

W.G. Müller. *Zur Versuchsplanung in Ökonomie und Ökonometrie.* Technical Report 270, Institute for Advanced Studies, Research Memorandum, Vienna, 1990.

S.H. Park. Experimental designs for fitting segmented polynomial regression models. *Technometrics*, 20(2):151–154, 1978.

C.R. Pelto, T.A. Elkins, and H.A. Boyd. Automatic contouring of irregulary spaced data. *Geophysics*, 33(3):424–430, 1968.

M.B. Priestley and M.T. Chao. Nonparametric function fitting. *Journal of the Royal Statistical Society*, B, 34:385–392, 1972.

B.D. Ripley. *Spatial Statistics.* Wiley, New York, 1981.

B.W. Silverman. Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal of the Royal Statistical Society*, B, 47(1):1–52, 1985.

S.D. Silvey. *Optimal Design.* Chapman and Hall, London, 1980.

M.C. Spruill. *Optimal Designs for Multivariate Interpolation.* Technical Report, School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332, U.S.A., 1988.