MATHEMATICAL MODELS FOR GEOMETRICAL INTEGRATION

Stefan Kampshoff

Geodetic Institute, RWTH Aachen University, Germany –
E-mail: stefan.kampshoff@gia.rwth-aachen.de

KEY WORDS: geometrical integration, homogenization, fragmentation, positional accuracy improvement, topological transform, spatial autocorrelation, collocation, hybrid approach

ABSTRACT

We define a simple model of a map and formulate geometrical integration as a transformation problem between one idealized true map only known partially and multiple realizations of the map perturbed by random noise and unknown systematic effects.
In the statistical model the connection between the true map and its realizations is interpreted as a multivariate spatial random process and further investigated by geostatistical and classical estimation methods. As a main result a model of the distance-dependent relative accuracy (neighborhood accuracy) in spatial datasets is obtained. On the basis of this statistical hypothesis two models for the estimation of the true map, namely intrinsic kriging and collocation, are suggested.
As an alternative to the stochastical model we suggest the deterministic model of the topology of the euclidean plane where the connection between the maps is treated as a homeomorphism. The consideration of constraints in form of a bijection for homologous points and nonlinear functions for geometrical constraints leads to the formal definition of the homogenization of maps.
An extension of the collocation model allows us to estimate the unknown parameters (coordinates) in the system of the true map and to simultaneously consider geometrical constraints, the linear trend, the nonlinear signal and the random noise. Due to the high density of its design matrix the suggested model is unsuitable for practical applications with mass data.
The hybrid approach of Benning appears to be an optimal compromise between the extended collocation model and other alternatives. It has the advantage that statistical least squares methods can be combined with (efficient) deterministic interpolation methods and furthermore leads to a sparse design matrix.

1 INTRODUCTION

The increasing degree of syntactical interoperability in GIS and spatial data-infrastructures enables the combination of data from different sources and providers. These datasets, though syntactically integrated, may comprise heterogeneity at the geometrical, schematical and semantical level (Laurini, 1998).

2 GEOMETRIC HETEROGENEITY IN FRAGMENTED DATASETS

Geometric heterogeneity is observed in fragmented datasets, where the two cases of zonal or horizontal fragmentation and layer or vertical fragmentation can be distinguished (Laurini, 1998, Gröger and Kolbe, 2003, Kampshoff and Benning, 2005).

Zonal fragmentation typically occurs when thematically similar data are captured independently by different organizations (local authorities, states etc.) and no organizational framework like a commitment to a common fixed border-geometry is arranged beforehand.

An example for zonal fragmentation is given in fig. 1. In this case, the homologous geometries (points) from different data sources do have varying coordinates. Furthermore, homologous lines may intersect, neighboring areas can overlap each other or gaps in tessellations can emerge when fragmented data are superimposed. The simple (purely syntactical) integration of zonally fragmented datasets can therefore lead to geometrical and topological inconsistencies.

Layer fragmentation occurs when a base-map (base 3-d model), as *external* part of an integrated multi-layer dataset, is replaced by a novel or updated version (positional accuracy improvement, PAI). In this context the term *external* indicates, that the base-map has been captured and updated by an external organization, e.g. surveying agencies or other data providers. On the other hand the remaining user-layers are usually captured and updated by an operating company from e.g. the utility or telecoms sector. The data capture of the user-layers is commonly done relative to the base-map layer, which results in a high relative inter-layer spatial accuracy.

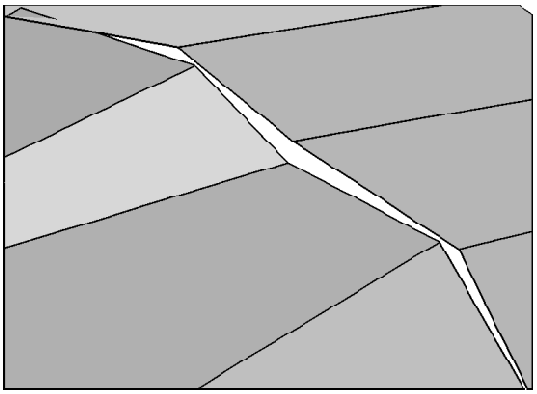

Figure 1: *Varying coordinates of homologous points from fragmented datasets*

Thus, the associations between spatial objects in multi-layer datasets are of a less explicit nature, than those in the case of zonal fragmentation. In particular, they can not be detected entirely by the standard transaction rules for geometric consistency in GIS (Gröger and Plümer, 1997, Gröger and Plümer, 2005). In general there is no direct relation (incidence, congruence) between heterogeneous geometries from different layers. Instead, the inter-layer geometrical and topological relations are only given implic-

itly through the coordinates of the geometries. In special cases the relative position of multi-layer geometries is given by dimensioning objects (cf. fig. 2).

The spatial dependency in multi-layer spatial datasets is also called associativity. The maintenance of the multi-layer consistency during base-map updates is hence called associativity problem (Wan and Williamson, 1994a, Wan and Williamson, 1994b, Scheu et al., 2000).

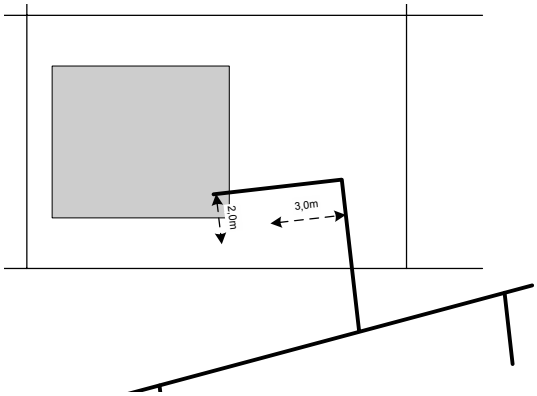

Figure 2: *Geometric heterogeneity of a network model and an updated base-model*

3 THE PROCESS OF GEOMETRICAL INTEGRATION

The example given in figure 2 is from the 2-D GIS domain, but the described problems of geometrical interoperability can obviously be transferred to the domain of 3-D city models in a straightforward manner. In both domains it is necessary to overcome geometrical heterogeneity by an adaptation of the coordinate-geometry that we call geometrical integration or homogenization.

Geometrical integration should be seen in the broader context of spatial database integration as described in (Laurini, 1998, Parent and Spaccapietra, 1998). Before the core steps of geometrical integration can be performed, a general schema integration and correspondence analysis has to be done and semantic conflicts of the datasets should be solved on the model level.

As three core steps of geometrical integration we identify the

- correspondence problem for homologous spatial objects from multiple maps (instance level),
- the identification of single and multi layer implicit spatial relations (associativity) and
- the computation of a transform that warps heterogeneous source datasets into one consistent target dataset.

The procedure can be followed by a non-geometric feature integration between corresponding classes and attributes. A detailed use case analysis as well as corresponding process models and algorithms for geometrical integration can be found in (Kampshoff, 2005).

The remainder of this paper deals with mathematical models for the transformation step of geometrical integration.

4 FORMAL MODEL OF A 3-D MAP

We define a map K, as a description of a fixed state of the real world by points and connected sets of points of a mathematical space. As embedding space for 3-D city models we use the euclidean $\Re^3$. The map may consist of four sets $K = (P, L, S, V)$ of points, lines, surfaces and volumes, that are given in form of a continuous parameter function which ensures the topological connectivity of the elements of L, S and V.

We distinguish between a unique true map K^w of K, whose points represent a fixed state of the real world exactly, and one or multiple realizations k of K, which emerge from measurements of limited accuracy (soft data).

The points of the true map, which are also called hard data, are not available in practice. Instead the best available values are used. The cost of the determination of such (quasi-)true values is high, they are known for a small subset of points only.

The process of realization assigns exactly one point of the realization $\boldsymbol{x}^k \in k$ to each point of the true map $\boldsymbol{x} \in K^w \subset \Re^3$ and thereby fulfills the characteristics of a mapping (Fischer, 1989). For the mapping T of the realization we have

$$T : K^w \rightarrow k, \quad \boldsymbol{x}^w \mapsto T(\boldsymbol{x}^w). \tag{1}$$

4.1 Geometrical integration in the map model

The variable source datasets are modeled as realizations k of the map K. In the case of zonal fragmentation we have a set of neighboring realizations $k^1, \ldots, k^l$ containing some homologous points that link the borders of the realizations whereas in the case of layer fragmentation the realizations do strongly overlap each other.

The problems of geometrical integration are caused by the fact that the coordinates of the realizations of different datasets are based on different measurements and adjustment techniques. The coordinate values are perturbed by unknown irregular errors, which results in the described contradictions between spatial objects from fragmented datasets.

What is needed for the solution of geometrical integration is a model and an estimation procedure for the mapping $T : K^w \rightarrow k$ from incomplete data. In practice we will have multiple realizations $k^1, \ldots, k^l$ and one incomplete true map K^w. The formal aim of geometrical integration can thus be described as the inversion of the realization-process (T^{-1}) and finally the estimation of the unknown coordinates of the complete true map of K.

In the following we discuss various stochastic and deterministic approaches as possible models for T. For a complete derivation of the approaches we refer to (Kampshoff, 2005).

5 STOCHASTICAL MODELS FOR GEOMETRICAL INTEGRATION

5.1 Maps as spatial random processes

In the stochastical model the coordinates $\{\boldsymbol{x}_i^k\}$ of a realization k of a map K are treated as realizations of a 3-D random process.

The points of the true map K^w correspond to the domain of the random process

$$\{\boldsymbol{x} : \boldsymbol{x} \in D \subset \Re^3\}. \tag{2}$$

For each position $\boldsymbol{x}_0 \in D$ we define a m-dimensional vector of observations called *regionalized variable*, representing the $m = 3$ observations at the position $\boldsymbol{x}_0$ with

$$\{|z_1(\boldsymbol{x}), \ldots, z_m(\boldsymbol{x})|' = \boldsymbol{z}(\boldsymbol{x}), \boldsymbol{x} \in D\}. \quad (3)$$

The vectors of regionalized variables are treated as outcome of a vector field of random variables, where one random vector of $m = 3$ random variables $(Z_1(\boldsymbol{x}_0), \ldots, Z_m(\boldsymbol{x}_0))$ is attached to each position $\boldsymbol{x}_0$ of the domain. Hence the regionalized variable is a realization of a vector random function (also: random process)

$$\{|Z_1(\boldsymbol{x}), \ldots, Z_m(\boldsymbol{x})|' = \boldsymbol{Z}(\boldsymbol{x}), \boldsymbol{x} \in D\}. \quad (4)$$

The transition from the random process to our map model is achieved by interpreting the points of the realization of a map k as realizations of a random process. Each position $\boldsymbol{x} \in D$ is at the same time a point $\boldsymbol{x} := \boldsymbol{x}^w$ of the true map K^w. The process of the realization of the random field does therefore correspond to the definition of a mapping T between the true map and its realizations, because with

$$\boldsymbol{Z}(\boldsymbol{x}) \overset{Realization}{\longrightarrow} \boldsymbol{z}_k(\boldsymbol{x}) := \boldsymbol{x}^k \quad (5)$$

we have one vector of observations $\boldsymbol{x}^k \in k$ for each position $\boldsymbol{x} \in D$. In the stochastical model of the map we thus have direct observations for the coordinates of the domain D.

5.2 Stationarity in the map model

For a description of the theory of (geo-)statistical stationary random processes we refer to (Wackernagel, 1998). We assume, that the multivariate distribution of the random vector is symmetric and has got a single unique maximum. Then for the expectation value of (the random process) of a map we get

$$E[\boldsymbol{Z}(\boldsymbol{x})] = \boldsymbol{\mu}(\boldsymbol{x}) = \boldsymbol{x}^w = \boldsymbol{x}. \quad (6)$$

It is further possible that a random process is given, whose expectation values are given through bijective, differentiable functions of the positions of the true map by

$$E[Z_i^*(\boldsymbol{x})] = h_i(\boldsymbol{x}^w, \boldsymbol{\beta}_h). \quad (7)$$

In both cases the expectation values of the random functions are neither stationary nor intrinsically stationary. To retrieve an intrinsically stationary random process we define the residual function of the random process, which is defined as the difference of the coordinates to their expectation values (in the case of (6)).

$$\boldsymbol{R}(\boldsymbol{x}) := E[\boldsymbol{Z}(\boldsymbol{x})] - \boldsymbol{Z}(\boldsymbol{x}) = \boldsymbol{x}^w - \boldsymbol{Z}(\boldsymbol{x}). \quad (8)$$

The expectation of the residual function $\boldsymbol{R}(x)$ is stationary, as we have

$$E[\boldsymbol{R}(\boldsymbol{x})] = E[\boldsymbol{x}^w - \boldsymbol{Z}(\boldsymbol{x})] = \boldsymbol{x}^w - \boldsymbol{x}^w = 0. \quad (9)$$

The central moments of the residual function are identical to those of $\boldsymbol{Z}(\boldsymbol{x})$, as they are related to the expectation value.

5.3 Absolute spatial accuracy

The absolute positional accuracy of a point $\boldsymbol{x}_j^k$ of a realization (map) k is defined as its distance to its true value (Croitoru and Doytsher, 2003)

$$\boldsymbol{\delta}(\boldsymbol{x}_j^k) = \boldsymbol{x}_j^k - \boldsymbol{x}_j^w. \quad (10)$$

As an empirical measure of the mean absolute accuracy of a set of n points the mean quadratic distance can be used with

$$\Delta_i^{abs} = \frac{\sum_{j=1}^n \delta_i^2(\boldsymbol{x}_j^k)}{n}. \quad (11)$$

Per definition the absolute positional accuracy does not take into account the relative spatial dependencies of points. Due to the identity of true values and expectation values in our map model, we can interpret the mean absolute accuracy (11) as an estimate of the variance of the random function $V[Z_i(\boldsymbol{x})] = E[(Z_i(\boldsymbol{x}) - x_i^w)^2]$ of a map with known expectation value $E[Z_i(\boldsymbol{x})] = x_i^w$ and location independent variance from a set of uncorrelated observations.

5.4 Relative spatial accuracy

For geometrical integration it is in particular important to have a model of the associativity and local dependency of the points of a map. This dependency emerges in an increased relative accuracy (neighborhood accuracy) of neighbored points. In a set of points we speak of a locally increased relative accuracy if the standard deviation of the distance of two points grows with their distance.

5.4.1 Empirical measures of relative spatial accuracy An empirical measure of the relative accuracy of two points $\boldsymbol{x}_i^k$ and $\boldsymbol{x}_j^k$ of the realization k can be derived from the deviation of the distance vector

$$\boldsymbol{d}(\boldsymbol{x}_i^k, \boldsymbol{x}_j^k) = \boldsymbol{x}_j^k - \boldsymbol{x}_i^k \quad (12)$$

from its true value

$$\boldsymbol{\delta}(\boldsymbol{x}_i^k, \boldsymbol{x}_j^k) = \boldsymbol{d}(\boldsymbol{x}_i^k, \boldsymbol{x}_j^k) - \boldsymbol{d}(\boldsymbol{x}_i^w, \boldsymbol{x}_j^w). \quad (13)$$

A mean value of the relative accuracy can be given by the mean square deviation over all distances of a point set by

$$\Delta_i^{rel} = 2\frac{\sum_{j=1}^n \sum_{l=i}^n \delta_i^2(\boldsymbol{x}_j^k, \boldsymbol{x}_l^k)}{n(n-1)}. \quad (14)$$

For the analysis of the distance dependency of the relative spatial accuracy the empirical variogram can be utilized. In the empirical variogram γ_i^* the deviation $\gamma_i^*(d_{jl}) = \delta_i^2(\boldsymbol{x}_j^k, \boldsymbol{x}_l^k)/2$ is drawn against the distance d. The resulting graph is called variogram cloud, an example is given in fig. 3.

The empirical variogram is derived from the variogram cloud by the definition of a piecewise constant function on intervals of equal length D_k. A locally increased relative accuracy leads to a continuous enlargement of the differences and an increasing variogram function. For the opposite case of an almost constant variogram over all distance intervals there is no distance dependency in the data.

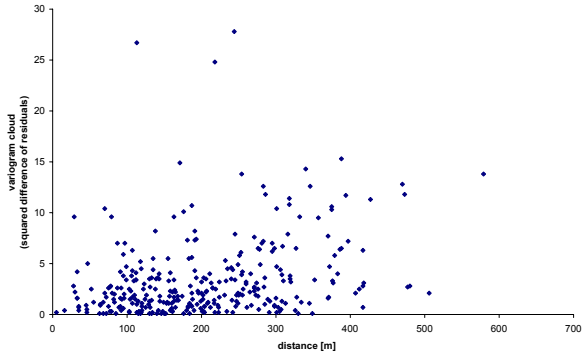

Figure 3: *Variogram cloud of the residuals of layer-fragmented datasets*

5.4.2 Theoretical analysis of relative spatial accuracy The statistical analysis is performed on base of the increment of the residual random function

$$\boldsymbol{R}(\boldsymbol{x}_j) - \boldsymbol{R}(\boldsymbol{x}_i) = (\boldsymbol{Z}(\boldsymbol{x}_j) - \boldsymbol{Z}(\boldsymbol{x}_i)) - \boldsymbol{h}_{ij} \qquad (15)$$

with the distance vector $\boldsymbol{h}_{ij} = \boldsymbol{x}_j - \boldsymbol{x}_i$. The increment can be interpreted as the difference of the observed distance of two points from their true distance. For the expectation value of the increment of the residual function we obtain

$$E[\boldsymbol{R}(\boldsymbol{x}_j) - \boldsymbol{R}(\boldsymbol{x}_i)] = E[(\boldsymbol{Z}(\boldsymbol{x}_j) - \boldsymbol{Z}(\boldsymbol{x}_i)) - \boldsymbol{h}_{ij}] = 0. \quad (16)$$

We assume intrinsic second order stationarity for the random function of the residuals, so that the (theoretical) variogram of the residual function is given by

$$\gamma_i(\boldsymbol{h}) = \frac{1}{2} E[(Z_i(\boldsymbol{x}+\boldsymbol{h}) - Z_i(\boldsymbol{x}) - h_i)^2]. \qquad (17)$$

The theoretical variogram $\gamma_i(\boldsymbol{h})$ of the residual function can be interpreted as the variance of the i-th coordinate of the random distance vector of two points. Thus we obtain a measure of the relative accuracy of two points with

$$\begin{aligned} \gamma_i(\boldsymbol{h}) &= \frac{1}{2} E[(Z_i(\boldsymbol{x}+\boldsymbol{h}) - Z_i(\boldsymbol{x}) - h_i)^2] && (18)\\ &= \frac{1}{2} E[(H_i(\boldsymbol{x}+\boldsymbol{h},\boldsymbol{x}) - E[H_i(\boldsymbol{x}+\boldsymbol{h},\boldsymbol{x})])^2] && (19)\\ &= \frac{1}{2}\sigma^2_{H_i}. && (20) \end{aligned}$$

With an estimate of the theoretical variogram $\hat{\gamma}_i(\boldsymbol{h})$, which can be computed from the experimental variogram γ^*_i, we are able to predict values of $\hat{\sigma}^2_i(\boldsymbol{h}) = 2\hat{\gamma}_i(\boldsymbol{h})$ as measures of the relative accuracy in fragmented datasets.

If we do further assume full stationarity of order two for the random process, we obtain the covariance function $C_i(\boldsymbol{h}) := C[R_i(\boldsymbol{x}+\boldsymbol{h}), R_i(\boldsymbol{x})]$ which is only dependent on the distance vector $\boldsymbol{h}$. The covariance function has a supremum at $\boldsymbol{h} = \boldsymbol{0}$ with

$$|C_i(\boldsymbol{h})| \leq C_i(\boldsymbol{0}) = V[R_i(\boldsymbol{x})]. \qquad (21)$$

Given the covariance function we obtain the variogram from

$$\gamma_i(\boldsymbol{h}) = C_i(\boldsymbol{0}) - C_i(\boldsymbol{h}) = V[R_i(\boldsymbol{x})] - C_i(\boldsymbol{h}). \qquad (22)$$

Vice versa we can construct the corresponding covariance function from a given variogram if and only if there exists a finite upper bound for the variogram $\gamma_i(\boldsymbol{\infty})$ with

$$\gamma_i(\boldsymbol{\infty}) - \gamma_i(\boldsymbol{h}) = C_i(\boldsymbol{h}). \qquad (23)$$

The dependence of the covariance function and the variogram can be transfered to our problem of relative accuracy of the random process of a map.

- If the spatial random variables of the points of a map are spatially uncorrelated, then the relative spatial accuracy is independent of the distance. A distance-dependent relative accuracy can, in the stochastical model, only be explained by a spatial correlation of the random variables.
- If the points are spatially independent (uncorrelated) and directly observed without systematical deviations (7), then we have $E[R_i(\boldsymbol{x})] = E[Z_i(\boldsymbol{x}) - \boldsymbol{x}^w] = E[Z_i(\boldsymbol{x})] - \boldsymbol{x}^w = 0$, and the relative spatial accuracy is constantly equal to the mean absolute accuracy of the point set.
- If the spatial correlation of the points decreases with increasing point distance, then the relative accuracy $\gamma_i(\boldsymbol{h})$ does asymptotically converge to the variance $C_i(\boldsymbol{0}) = V[Z_i(\boldsymbol{x})]$ of the points.

The proves of these propositions can be found in (Kampshoff, 2005).

5.5 Estimation of T in the stochastical model

Equipped with a model of relative spatial accuracy for our map we now discuss various possibilities for the determination of the realization-mapping T as defined in (1).

The aim of the determination of T is the prediction of the residual function $\boldsymbol{r}^*(\boldsymbol{x}_0)$ at positions $\boldsymbol{x}_0 \in D$ for which a realization $\boldsymbol{x}^k_0 = \boldsymbol{z}_k(\boldsymbol{x}_0) \in k$ but no corresponding true coordinate $\boldsymbol{x}^w_0 = E[\boldsymbol{Z}(\boldsymbol{x}_0)] \in K^w$ is known (cf. fig. 4).

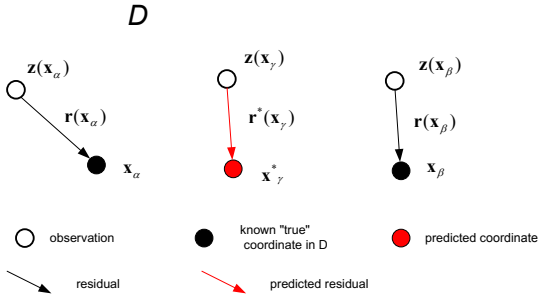

Figure 4: Prediction of the residual function in the stochastical model

5.6 Intrinsic kriging with random error

Intrinsic kriging is used as an estimator in random processes that do not fulfill the assumption of second order intrinsic stationarity. Then the random function is composed of the stationary random function of the residuals and an unknown drift function $m(\boldsymbol{x})$ (cf. 7)

$$Z(\boldsymbol{x}) = m(\boldsymbol{x}) + R(\boldsymbol{x}). \qquad (24)$$

It is assumed, that the drift can be represented by a linear combination of deterministic functions f_l with

$$m(\boldsymbol{x}) = \sum_{l=0}^{L} a_l f_l(\boldsymbol{x}). \qquad (25)$$

It is further assumed that the functions f_l build a translation invariant vector space and that the variance of the linear combination of f_l is given by a linear combination of a symmetric generalized covariance function[1]

$$V[\sum_{\alpha=0}^{n} w_\alpha Z(\boldsymbol{x}_\alpha)] = \sum_{\alpha=1}^{n}\sum_{\beta=1}^{n} w_\alpha w_\beta K(\boldsymbol{x}_\alpha - \boldsymbol{x}_\beta). \quad (26)$$

Based on the generalized covariance function the intrinsic kriging system can be build with

$$\begin{pmatrix} \boldsymbol{K} & \boldsymbol{F} \\ \boldsymbol{F}' & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{w} \\ -\boldsymbol{\mu} \end{pmatrix} = \begin{pmatrix} \boldsymbol{k} \\ -\boldsymbol{f} \end{pmatrix} \quad (27)$$

where $\boldsymbol{K} = (K_{ij}) = (K(\boldsymbol{x}_i - \boldsymbol{x}_j))$, $\boldsymbol{F} = (F_{ij}) = (f_j(\boldsymbol{x}_i))$, $\boldsymbol{k} = (k_i) = (K(\boldsymbol{x}_i - \boldsymbol{x}_0))$ and $f = (f_i) = (f_i(\boldsymbol{x}_0))$.

The kriging equations are dependent on the position of interpolation $\boldsymbol{x}_0$ and thus have to be solved for every position. Alternatively the dual kriging system with the vector of the data $\boldsymbol{z}$ (coordinates of realizations) can be derived, which is independent from the interpolation position and thus has to be solved only once.

$$\begin{pmatrix} \boldsymbol{K} & \boldsymbol{F} \\ \boldsymbol{F}' & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{b} \\ \boldsymbol{d} \end{pmatrix} = \begin{pmatrix} \boldsymbol{z} \\ \boldsymbol{0} \end{pmatrix}. \quad (28)$$

As a solution to the system we obtain

$$\boldsymbol{d} = (\boldsymbol{F}'\boldsymbol{K}^{-1}\boldsymbol{F})^{-1}\boldsymbol{F}'\boldsymbol{K}^{-1}\boldsymbol{z} \quad (29)$$

$$\boldsymbol{b} = \boldsymbol{K}^{-1}(\boldsymbol{z} - \boldsymbol{F}\boldsymbol{d}) \quad (30)$$

and the interpolation (prediction) of a value of the true map is finally done with

$$z^*(\boldsymbol{x}) = \boldsymbol{b}'\boldsymbol{k}_{\boldsymbol{x}} + \boldsymbol{d}'\boldsymbol{f}_{\boldsymbol{x}}. \quad (31)$$

Intrinsic kriging in the form of (28) is an exact interpolator. It can be extended by a spatially independent random noise (observational error) to the model

$$Z(\boldsymbol{x}) = m(\boldsymbol{x}) + R(\boldsymbol{x}) = m(\boldsymbol{x}) + S(\boldsymbol{x}) + \epsilon, \quad (32)$$

$$\text{with} \quad C[R(\boldsymbol{x}+\boldsymbol{h}), R(\boldsymbol{x})] = \boldsymbol{K}(\boldsymbol{h}), \quad (33)$$

$$\boldsymbol{\Sigma}_{\epsilon\epsilon} = \text{diag}(\sigma_1^2, \ldots, \sigma_n^2). \quad (34)$$

The dual kriging system of intrinsic kriging with random error can be derived from (28) by adding the variance matrix $\boldsymbol{\Sigma}_{\epsilon\epsilon}$ of the random error to the spatial covariance matrix $\boldsymbol{K}$.

5.7 Least squares collocation

The method of least squares collocation has already been suggested for residual spreading in transformation problems (Moritz, 1973). The collocation model consists of a random vector $\boldsymbol{y}$ that is explained by a linear mapping of unknown parameters $\boldsymbol{X}\boldsymbol{\beta}$, a random vector $\boldsymbol{s}$ (signal), a random vector of observational errors $\boldsymbol{n}$ with

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{s} + \boldsymbol{n} \quad (35)$$

and the expectation value of the signal and the random error $E[\boldsymbol{s}] = \boldsymbol{0}$ and $E[\boldsymbol{n}] = \boldsymbol{0}$. Signal and random error are uncorrelated $\boldsymbol{\Sigma}_{sn} = \boldsymbol{0}$ and the covariance matrices are given with $C[\boldsymbol{s}, \boldsymbol{s}] = \boldsymbol{\Sigma}_{ss}$ and $C[\boldsymbol{n}, \boldsymbol{n}] = \boldsymbol{\Sigma}_{nn}$.

[1] For a detailed definition of the intrinsic random function of order k we refer to (Wackernagel, 1998).

The signal $\boldsymbol{s}$ is a deterministic part of the realization of a map. For repeated measurements of coordinates it will be constant. It is treated as a stochastical random variable with known expectation value, as we have no precise knowledge about its analytical nature.

The solution of the collocation model for the parameter vector $\boldsymbol{\beta}$ and the signal $\boldsymbol{s}$ is given by (Moritz, 1973, Koch, 1997)

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'(\boldsymbol{\Sigma}_{ss} + \boldsymbol{\Sigma}_{nn})^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{\Sigma}_{ss} + \boldsymbol{\Sigma}_{nn})^{-1}\boldsymbol{y} \quad (36)$$

$$\hat{\boldsymbol{s}} = \boldsymbol{\Sigma}_{ss}(\boldsymbol{\Sigma}_{ss} + \boldsymbol{\Sigma}_{nn})^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}). \quad (37)$$

The comparison of this solution with the solution of the kriging system (29) and (30) shows the formal correspondence of both approaches.

The prediction of observations $\boldsymbol{y}^*$ at a position $\boldsymbol{x}_0$ can be done with the design matrix $\boldsymbol{X}^*_{\boldsymbol{x}_0}$ and the covariance matrix of the predicted signal $\boldsymbol{s}^*$ at the position $\boldsymbol{x}_0$ and the signal at the observed positions $\boldsymbol{s}$ with $C[\boldsymbol{s}^*, \boldsymbol{s}] = \boldsymbol{\Sigma}_{s^*s}$ and

$$\hat{\boldsymbol{y}}^* = \hat{\boldsymbol{s}}^* + \boldsymbol{X}^*\hat{\boldsymbol{\beta}} = \boldsymbol{\Sigma}_{s^*s}\boldsymbol{b} + \boldsymbol{X}^*\hat{\boldsymbol{\beta}} \quad \text{with} \quad (38)$$

$$\boldsymbol{b} = (\boldsymbol{\Sigma}_{ss} + \boldsymbol{\Sigma}_{nn})^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}). \quad (39)$$

6 DETERMINISTIC MODELS FOR GEOMETRICAL INTEGRATION

6.1 Formal definition of the homogenization problem

In the deterministic model we define the problem of the determination of the mapping T from the true map to its realization as follows:

Definition 6.1 (Homogenisation) *Given are a realization of a map k, the corresponding true map K^w and two sets of points $P^k \subseteq k$ and $P^w \subseteq K^w$ out of k and K^w. A bijection $b : P^w \rightarrow P^k$ is given so that every point of P^w corresponds to exactly one point in P^k. Further a set of geometrical constraints $C = \{c_1, \ldots, c_n\}$ with $c_i : \Re_3 \rightarrow \Re$ is defined on the points. The construction of a homeomorphism $h : k \rightarrow K^w$, that does fulfill the bijection b and the geometrical constraints C we call homogenization of the realization k and its true map K^w.*

In the above definition geometrical constraints are defined as nonlinear scalar functions on $\Re^3$. All kinds of constraints that relate to a metric or angular quantity can be defined on the base of these equations. Examples are

- distance of (point-point), (point-line), (point-plane), (line-line), (line-plane), (plane-plane) etc.
- angle between (line-line), (line-plane), (plane-plane) etc.

Some types of topological constraints can not be translated into an equation based on a metric quantity. An example for this is a containment relation between a point and a cube. For this class of problems the definition 6.1 has to be extended to constraints given as inequality equations.

For the case of multiple realizations the definition can easily be extended. Additional constraints like links for homologous points of neighboring realizations and geometrical constraints with references to points from multiple realizations can occur in those cases.

6.2 Construction of T in the deterministic model

The problem of constructing a homeomorphism for a given bijection of points is also known as rubber sheeting. There exist a number of different exact interpolation techniques that have been proposed for rubber sheeting in the past. Among them are

- the inverse distance weighted interpolation (Shepard, 1964, Hettwer, 2003),
- the piecewise linear transformation in triangles (Merkel, 1932, White and Griffen, 1985, Fagan and Soehngen, 1987, Gillman, 1985, Saalfeld, 1985),
- the natural neighbor interpolation (Sibson, 1981, Owen, 1992, Roschlaub, 1999, Hettwer and Benning, 2003),
- the multiquadratic method (Hardy, 1972, Göpfert, 1977, Wolf, 1981) and
- the thin plate spline interpolation (Bookstein, 1989).

A comparison of these techniques in the context of general interpolation problems can be found in (Mitas and Mitasova, 1999).

It can be shown that the interpolation techniques based on radial basis functions (multiquadratic and thin plate spline) are equivalent to intrinsic kriging without random error (Wackernagel, 1998, Kampshoff, 2005). If a random error is added to the deterministic models, full correspondence is achieved, whereby the covariance function in the stochastical model corresponds to the kernel matrix of the radial basis functions in the deterministic case.

From fig. (5) it can be seen, that the interpolations based on radial basis functions are able to reproduce non-linear smooth surfaces from scattered data points. The drawback of these techniques is their high computational complexity, as a dense system of linear equations has to be solved.

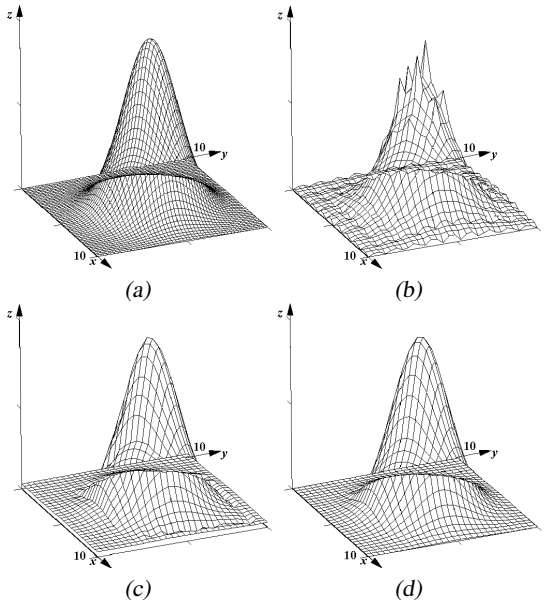

Figure 5: *(a) original (b) inverse distance (c) thin plate spline (d) multiquadratic (*AHMED ET.AL.*)*

The only interpolation technique, that can be proved to be a homeomorphism is the linear interpolation in triangles. A solution to this *homeomorphism extension problem* in 2-D is given in (Saalfeld, 1993). Furthermore the linear interpolation in triangles can be performed in $o(nlog(n))$ operations in 2-D[2]. The resulting transform is continuous, but not differentiable as it has blips at the edges of the triangles.

The natural neighbor interpolation appears to be a good compromise between the interpolations with radial basis functions and the linear interpolation in triangles. It is a smooth interpolator and differentiable everywhere besides in homologous data points. The delaunay triangulation of the points can be computed in linear time for equally distributed point sets (Tsai, 1993), and the interpolation of coordinate values can be performed in constant time, if a spatial index is used.

7 COMPARISON, EXTENSION AND INTEGRATION OF BOTH APPROACHES

7.1 Exact or approximate interpolation?

The main difference between the stochastical an deterministic approaches lies in the consideration of a spatially uncorrelated random error in the data points. The stochastical models do not reproduce the values of the true map, as the random error is filtered in these locations. At first sight this seems to be a drawback of the stochastical models. Anyway, if we keep in mind the problem of the maintenance of associativity in the map, only those parts of the residual function should be transfered (interpolated) to neighboring points, that show a spatial correlation.

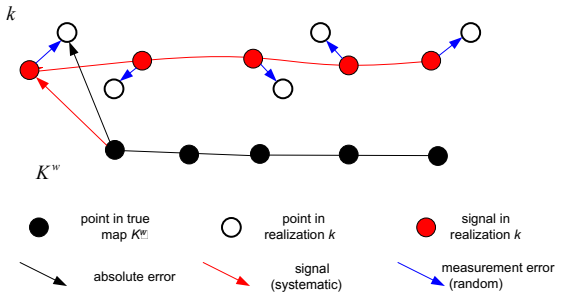

Figure 6: *Residuals consisting of a random uncorrelated error and a correlated signal*

Thus, as can be seen from figure (6), only the (filtered) signal should be used for interpolation. The random error can not be assumed be of similar size and direction for neighboring points.

In general the stochastical model should therefore be used to solve the integration problem, as simple rubber sheeting leads to a spreading of random errors in the map. Nonetheless, exact deterministic interpolators can be applied when the random error is of neglectable magnitude compared to signal.

7.2 Extensions for geometrical constraints

There are basically two ways to integrate geometrical constraints into our map models: the sequential and the simultaneous approach.

The sequential approach consists of two separate steps. In the first step the interpolation problem is solved by one of the described

[2]This will be sufficient for 2.5D graph-surfaces

interpolation techniques (rubber sheeting). In a second independent step the realization of geometrical constraints is performed by a least squares adjustment, as there can be contradictions in the definition of the constraints. Due to the independence of both steps, the realization of the geometrical constraints may lead to inconsistencies in the map. In particular it is possible that a subset of points is strongly moved in step two, which leads to a decrease of the relative spatial accuracy for those points not included in the constraints.

A consistent geometrical integration can only be achieved if the maintenance of the relative spatial accuracy (interpolation) and the realization of geometrical constraints are done simultaneously. This can be reached by an extension of the collocation model. First we reformulate the collocation model as a Gauß-Markoff-Model (Koch, 1997)

$$E[\begin{pmatrix} \boldsymbol{y}_h \\ \boldsymbol{0} \end{pmatrix}] = \begin{pmatrix} \boldsymbol{X}_t & \boldsymbol{I} \\ \boldsymbol{0} & \boldsymbol{I} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_t \\ \boldsymbol{s}_h \end{pmatrix} \text{ with} \quad (40)$$

$$D[\begin{pmatrix} \boldsymbol{y}_h \\ \boldsymbol{0} \end{pmatrix}] = \sigma^2 \begin{pmatrix} \boldsymbol{\Sigma}_{y_h y_h} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}_{s_h s_h} \end{pmatrix}. \quad (41)$$

This reformulation enables us to add the unknown coordinates of the true map as unknown parameters in the vector $\boldsymbol{\beta}_n$ and the additional observation equations for those points

$$\boldsymbol{f}(\boldsymbol{x}_n^w, \boldsymbol{\beta}_t) = \boldsymbol{X}_t^n \boldsymbol{\beta}_t + \boldsymbol{X}_n \boldsymbol{\beta}_n + \boldsymbol{s}_n = \boldsymbol{y}_n + \boldsymbol{e}_n. \quad (42)$$

These additional observations do not lead to a change in the estimates of the collocation model, as for every observation a new unknown parameter is defined (no increased redundancy).

The model can easily be extended with additional observations for geometrical constraints

$$\boldsymbol{b}(\boldsymbol{\beta}_n) = \boldsymbol{X}_b \boldsymbol{\beta}_n = \boldsymbol{y}_b + \boldsymbol{e}_b \quad (43)$$

and we finally obtain the extended collocation model with a spatially correlated signal, a random observational error and observations for geometrical constraints with

$$E[\begin{pmatrix} \boldsymbol{y}_h \\ \boldsymbol{0} \\ \boldsymbol{y}_n \\ \boldsymbol{y}_b \\ \boldsymbol{0} \end{pmatrix}] = \begin{pmatrix} \boldsymbol{X}_t & \boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{X}_t^n & \boldsymbol{0} & \boldsymbol{X}_n & \boldsymbol{I} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{X}_b & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_t \\ \boldsymbol{s}_h \\ \boldsymbol{\beta}_n \\ \boldsymbol{s}_n \end{pmatrix} \quad (44)$$

$$D[\begin{pmatrix} \boldsymbol{y}_h \\ \boldsymbol{0} \\ \boldsymbol{y}_n \\ \boldsymbol{y}_b \\ \boldsymbol{0} \end{pmatrix}] = \quad (45)$$

$$\begin{pmatrix} \boldsymbol{\Sigma}_{y_h y_h} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}_{s_h s_h} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{\Sigma}_{s_h s_n} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{\Sigma}_{y_n y_n} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{\Sigma}_{y_b y_b} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}_{s_n s_h} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{\Sigma}_{s_n s_n} \end{pmatrix}.$$

It is possible to extend this model for the integration of multiple realizations of maps. For each map a new transformation block with a separate signal and covariance matrix is introduced.

7.3 Scalability and the hybrid approach

For the integration of mass data it is important to choose a scalable approach. The scalability of the solution of (44) is determined by the covariance matrix of the signal, which typically is a dense matrix. An obvious improvement is thus to determine the signal externally by an efficient approach like natural neighbor interpolation. The results of the interpolated signal r_i^* can then be integrated into a simplified stochastical model by uncorrelated observations, that replace the correlated signal in (44).

The difficulty lies in the choice of the observational model for the introduction of the interpolated signal. Direct observation, like

$$r_i^* - r_i = r_i^* - (\boldsymbol{X}_t^i \boldsymbol{\beta}_t + \boldsymbol{X}_n^i \boldsymbol{\beta}_n - y_i) = 0 + e_i, \quad (46)$$

imply the negative effects of the sequential approach. Therefore BENNING suggests the introduction of differential observations between neighboring points with (Benning, 1995)

$$\Delta \boldsymbol{x}_j - \Delta \boldsymbol{x}_i - (\boldsymbol{r}_j^* - \boldsymbol{r}_i^*) = \boldsymbol{0} + \boldsymbol{e}. \quad (47)$$

The definition of neighborhood is taken from the edges of a delaunay triangulation of the points. Equation (47) can be interpreted as the difference of the distance vector $\boldsymbol{h}_{ij} = \boldsymbol{x}_j^w - \boldsymbol{x}_i^w$ of the estimated coordinates from the vector $\boldsymbol{h}_{ij}^*$ which results from the deterministic interpolation. Thus, if no geometrical constraints are defined, the result of the adjustment will be congruent with the result of the deterministic interpolation of the signal. If, on the other hand, geometrical constraints lead to a displacement of points, the difference equations (47) ensure the maintenance of relative spatial accuracy in the map.

The complete model of the hybrid approach is given by

$$\begin{pmatrix} \boldsymbol{X}_t & \boldsymbol{0} \\ \boldsymbol{X}_t^n & \boldsymbol{X}_n \\ \boldsymbol{0} & \boldsymbol{X}_b \\ \boldsymbol{0} & \boldsymbol{X}_\Delta \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_t \\ \boldsymbol{\beta}_n \end{pmatrix} = E[\begin{pmatrix} \boldsymbol{y}_h \\ \boldsymbol{y}_n \\ \boldsymbol{y}_b \\ \boldsymbol{0} \end{pmatrix}] \quad (48)$$

$$D[\begin{pmatrix} \boldsymbol{y}_h \\ \boldsymbol{y}_n \\ \boldsymbol{y}_b \\ \boldsymbol{y}_\Delta \end{pmatrix}] = \begin{pmatrix} \boldsymbol{\Sigma}_{y_h y_h} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}_{y_n y_n} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{\Sigma}_{y_b y_b} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{\Sigma}_{\Delta\Delta} \end{pmatrix} \quad (49)$$

The normal equations of the hybrid approach result in a sparse matrix, that can be solved efficiently for large datasets with more than 10^6 unknown parameters (Kampshoff and Benning, 2005). A further improvement of the scalability can be achieved by a fragmentation of the dataset into zones or layers that are treated individually (Kampshoff, 2005).

8 CONCLUSIONS

On the basis of a theoretical analysis of the absolute and relative accuracy in spatial datasets we have derived a mathematical model for geometrical integration, that allows us to simultaneously integrate heterogeneous spatial objects from fragmented sources into one consistent dataset. The model can be applied to large integration problems, and is therefore suitable for problems of positional accuracy improvement of spatial datasets like 3-D city models. The hybrid approach is superior to a sequential combination of rubber sheeting techniques and adjustment of geometrical constraints, as the associativity of the data is consistently considered in the whole process. In the hybrid approach the estimation of the unknown coordinates is done in a stochastical model that allows us to asses the quality of the data and to apply statistical outlier tests.

REFERENCES

Ahmed, M.N., S. M. F. A., n.d. Function Approximation for Free-Form Surfaces.
http://citeseer.ist.psu.edu/77808.html.

Benning, W., 1995. Nachbarschaftstreue Restklaffenverteilung für Koordinatentransformationen. Zeitschrift für Vermessungswesen 120, pp. 16–25.

Bookstein, F., 1989. Principal Warps: Thin Plate Spline and the Decomposition of Deformations. IEEE Transactions on Pattern Analysis and Machine Intelligence 11(6), pp. 567–585.

Croitoru, A. and Doytsher, Y., 2003. Accounting for Discontinuities in Cadastral Data Accuracy: Toward a Patch-Based Approach. In: FIG Working Week 2003, Paris, International Federation of Surveyors (FIG).

Fagan, G. and Soehngen, H., 1987. Improvement of GBF/DIME File Coordinates in a Geobased Information System by Various Transformation Methods and Rubbersheeting Based on Triangulation. In: AutoCarto 8, Baltimore, Md, ACSM/ASP, Falls Church, Va, pp. 481–491.

Fischer, G., 1989. Lineare Algebra. Vieweg, Braunschweig.

Gillman, D., 1985. Triangulations for Rubber Sheeting. In: AutoCarto 7, Washington, DC, ACSM/ASP, Falls Church, Va, pp. 191–199.

Göpfert, W., 1977. Interpolationsergebnisse mit der Multiqudratischen Methode. Zeitschrift für Vermessungswesen 102, pp. 457–460.

Gröger, G. and Kolbe, T., 2003. Interoperabilität in einer 3D-Geodateninfrastruktur. In: L. Bernard, A. Sliwinski and K. Senkler (eds), IfGI Prints: Beiträge zu den Münsteraner GI-Tagen 2003, Institut für Geoinformatik, Uni Münster, pp. 325–343.

Gröger, G. and Plümer, L., 1997. Provably Correct and Complete Transaction Rules for GIS. In: R. Laurini, P. Bergougnoux and N. Pissinou (eds), Proceedings of the 5th International Workshop on Advances in Geographic Information Systems (ACM-GIS'97), Las Vegas, ACM Press, pp. 40–43.

Gröger, G. and Plümer, L., 2005. How to Get 3-D for the Price of 2-D — Topology and Consistency of 3-D Urban GIS. Geoinformatica 9(2), pp. 139–158.

Hardy, R., 1972. Geodetic Applications of Multiquadratic Analysis. Allgemeine Vermessungs-Nachrichten 79, pp. 398–406.

Hettwer, J., 2003. Numerische Methoden zur Homogenisierung großer Geodatenbestände. In: Veröffentlichungen des Geodätischen Instituts der RWTH Aachen, Geodätisches Institut der RWTH Aachen.

Hettwer, J. and Benning, W., 2003. Restklaffenverteilung mit der Natural-Neighbour-Interpolation. Allgemeine Vermessungs-Nachrichten 110, pp. 122–129.

Kampshoff, S., 2005. Integration heterogener raumbezogener Objekte aus fragmentierten Geodatenbeständen. In: Mitteilungen des Geodätischen Instituts der RWTH Aachen, Aachen.

Kampshoff, S. and Benning, W., 2005. Homogenisierung von Massendaten im Kontext von Geodaten-Infrastrukturen. zfv–Zeitschrift für Geodäsie, Geoinformation und Landmanagement 130, pp. ?

Koch, K.-R., 1997. Parameterschätzung und Hypothesentests in linearen Modellen. Dümmler, Bonn.

Laurini, R., 1998. Spatial Multi-database Topological Continuity and Indexing: a Step Towards Seamless GIS Data Interoperability. International Journal of Geographical Information Science 12(4), pp. 373–402.

Merkel, H., 1932. Koordinatenumformung durch maschenweise Abbildung. Allgemeine Vermessungs-Nachrichten 44, pp. 114–124, 131–143.

Mitas, L. and Mitasova, H., 1999. Spatial Interpolation. In: P. Longley, M. Goodchild, D. Maguire and D. Rhind (eds), Geographical Information Systems, 2 edn, Vol. 1: Principles and Technical Issues, Wiley, New York, pp. 481–492.

Moritz, H., 1973. Neuere Ausgleichungs- und Interpolationsverfahren. Zeitschrift für Vermessungswesen 98, pp. 137–146.

Owen, S.J., 1992. An Implementation of Natural Neighbor Interpolation in Three Dimensions. Master's Thesis, Brigham Young University.

Parent, C. and Spaccapietra, S., 1998. Issues and Approaches of Database Integration. Communications of the ACM (CACM) 41(5), pp. 166–178.

Roschlaub, R., 1999. Klassifikation und Interpolation mittels affin invarianter Voronoidiagramme auf der Basis eines Wahrscheinlichkeitsmaßes in großmaßstäbigen Geoinformationssystemen. In: Reihe C, Deutsche Geodätische Kommission, München.

Saalfeld, A., 1985. A Fast Rubber Sheeting Transformation Using Simplicial Coordinates. The American Cartographer 12(2), pp. 169–173.

Saalfeld, A., 1993. Conflation: Automated Map Compilation. PhD thesis, University of Maryland College Park, Computer Vision Laboratory. CS-TR-3066.

Scheu, M., Effenberg, W. and Williamson, I., 2000. Incremental Update and Upgrade of Spatial Data. Zeitschrift für Vermessungswesen 4, pp. 115–120.

Shepard, D., 1964. A Two-Dimensional Interpolation Function for Irregularly Spaced Data. In: ACM National Conference, pp. 517–524.

Sibson, R., 1981. A Brief Description of Natural Neighbor Interpolation. Interpreting Multivariate Data, pp. 21–36.

Tsai, V., 1993. Delaunay Triangulations in TIN Creation: an Overview and a Linear-Time Algorithm. International Journal of Geographical Information Systems 7, pp. 501–524.

Wackernagel, H., 1998. Multivariate Geostatistics: an Introduction with Applications. 2 edn, Springer, Berlin.

Wan, W. and Williamson, I., 1994a. Problems in Maintaining Associativity in LIS with Particular Reference to the Needs of the Utility Industry. The Austrailian Surveyor 39(3), pp. 187–193.

Wan, W. and Williamson, I., 1994b. Solutions to Maintaining Associativity in LIS with Particular Reference to the Needs of the Utility Industry. The Austrailian Surveyor 39(4), pp. 290–296.

White, M. and Griffen, P., 1985. Piecewise Linear Rubber-Sheet Map Transformation. The American Cartographer 12(2), pp. 123–131.

Wolf, H., 1981. Multiquadratische Methode und Kollokation. Allgemeine Vermessungs-Nachrichten 88, pp. 89–95.