SPATIAL EXTRAPOLATION OF ANISOTROPIC ROAD TRAFFIC DATA

Hans Braxmeier[1,2], Volker Schmidt[2] and Evgueni Spodarev[2]
[1]Department of Applied Information Processing, [2]Department of Stochastics, University of Ulm, D–89069 Ulm, Germany
e–mail: hans.braxmeier@mathematik.uni–ulm.de,
schmidt@mathematik.uni–ulm.de, spodarev@mathematik.uni–ulm.de

ABSTRACT

A method of spatial extrapolation of traffic data is proposed. The traffic data is given by GPS signals over downtown Berlin sent by approximately 300 taxis. To reconstruct the traffic situation at a given time spatially, i.e. in the form of traffic maps, kriging with moving neighborhood based on residuals is used. Due to significant anisotropy in directed traffic data, the classical kriging has to be modified in order to include additional information. To verify the extrapolation results, test examples on the basis of a well–known model of stochastic geometry, the Boolean random function are considered.

Keywords: anisotropy, asymptotic Gaussian test, Boolean model, kriging, moving neighborhood, random field.

INTRODUCTION

A common difficult problem of large cities with heavy traffic is the forecasting of traffic jams. In this paper, a first step towards mathematical traffic forecasting, namely the spatial reconstruction of the present traffic situation from point measurements is done. To describe the traffic states, models of stochastic geometry and spatial statistics (or geostatistics) are used. A corresponding Java software that implements efficient algorithms of spatial extrapolation is developed.

This research is based on real traffic data originating from downtown Berlin. They were provided by the Institute of Transport Research of the German Aerospace Center (DLR). Approximately 300 test vehicles (taxis) were equipped with GPS sensors transmitting their geographic coordinates, velocity and status line (e.g. “free”, “hired”, “at the taxi rank”, etc.) to a central station within regular time intervals from 30 sec. up to 3 min. The regularity of these signals depends on the taxi’s status. Thus, a large data base of more than 13 million positions was

formed since April 2001 (see Fig. 1).

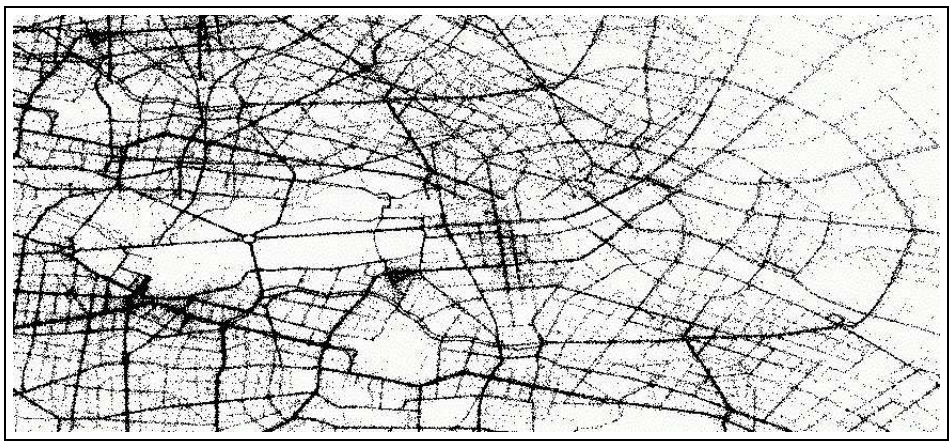

Fig. 1: Observed positions of test vehicles in downtown Berlin

In the present paper, a smaller data set (taxi positions on all working days from 30.09.2001 till 19.02.2002, 5.00–5.30 p.m., moving taxis only) is considered. The observation window was reduced to downtown Berlin in order to avoid inhomogeneities in the taxi positions. To study traffic jams, the rush hour (5.00–5.30 p.m.) was chosen.

To produce road traffic maps, the velocities of all vehicles at time t are assumed to be induced by a realization of a spatial random field $V(t) = \{V(t,u)\}$ where $V(t,u)$ is a traffic velocity vector at position $u \in \mathbb{R}^2$ and time $t \geqslant 0$. The spatial structure of such random velocity fields makes the analysis of traffic–jam mechanisms possible. Thus, the spatial localization of traffic jams can be obtained by a threshold operation on the grey–scale image of the map of velocities $V(t,u)$: a point u lies within the traffic jam region at time t if $|V(t,u)|$ is smaller than a given threshold value, e.g. 15 kph.

Since $V(t,u)$ can be measured just pointwise at observation points u_i, a spatial extrapolation of the observed data is necessary. Notice that the velocities strongly depend on the movement directions, e.g. the speed limits and consequently the mean velocities are higher on motorways than in downtown streets. Furthermore, the formation of traffic jams is also directional since a vehicle can influence only those vehicles moving behind it along the same road in the same direction. Moreover, the traffic speed at position u clearly depends on the traffic direction on the road, e.g. in directions of the city center or sub-

urbs.

The classical extrapolation methods of geostatistics such as the *ordinary kriging* (see e.g. Stoyan *et al.*, 1997, Wackernagel, 1998) either make no use of additional information or provide measurements $V(t, u + u_i)$ and $V(t, u - u_i)$ with equal weights. Both these features are not relevant to the above problem setting. An extrapolation method designed for directional data, the so–called *complex cokriging* of velocities and their directions (see e.g. Wackernagel, 1998) can not be used here as well since there is no one–to–one correspondence between measurement positions u and traffic directions. An obvious counterexample is a crossroads. Thus, the standard extrapolation methods had to be adapted to our specific problem. Therefore, a modified ordinary kriging with moving neighborhood is described that allows to extrapolate directed velocity fields. First, the original data set should be split into N directionally homogeneous subsets. A data unit $(u, V(t, u))$ belongs to the data set i $(i = 1, \ldots, N)$ if the polar angle of the vector $V(t, u)$ lies within the directional sector

$$S_i = [2\pi(i-1)/N, 2\pi i/N)\,.$$

By convention, the zero polar angle corresponds to the eastward direction on the city map. Throughout this paper, we put $N = 4$. From the practical point of view, this is sufficient for the separation of opposite traffic directions and, simultaneously, keeps the amount of resulting data sets small. Nevertheless, in principle, any other $N \geqslant 4$ could be used instead.

The above data sets should be extrapolated separately from each other. This yields N velocity maps corresponding to N directional sectors.

In what follows, the data from a given time interval $[t_1, t_2]$ will be taken for extrapolation. To be precise, we put $t_1 = 5.00$ p.m. and $t_2 = 5.30$ p.m. Keeping this in mind, we shall omit the time parameter t in further notation. The observed velocities are not spatially homogeneous. Hence, the mean velocity field $\{m(u)\}$ obtained by averaging the traffic velocities over all working days from 5.00 p.m. till 5.30 p.m. should be considered. As far as this mean field is subtracted from the original data, the deviations of actual velocity values are extrapolated in order to create the spatial field of velocity residuals.

This extrapolation method has been implemented in Java. Thus, a software library was developed comprising the estimation and fitting of variograms as well as the ordinary kriging with moving neighborhood. As far as it is known to the authors, it is the first complete implementation of such kriging methods in Java. An advantage of the Java programming language lies in its platform independence. Great attention was paid to the efficient implementation of fast algorithms. In contrast to

classical geostatistics operating with relatively small data sets, this efficiency is of great importance for larger data sets with more than 10000 entries. For instance, the Java package for variogram fitting described in Faulkner, 2002 can not be used for data sets with more than 1000 entries due to unacceptable runtimes. Efficient image processing and computational algorithms (see e.g. Mayer *et al.*, 2004) enabled us to drastically reduce the runtimes of the Java library.

The extrapolation method itself as well as the software quality are verified on the test example of a *Boolean random function*; see Serra (1988). Remarkable features of this model are its simplicity of simulation and nice analytical description. For test purposes, 90 independent realizations of a Boolean model with a deterministic drift have been simulated. The quality of extrapolation is proved by means of statistical significance tests of the area fraction. It is shown that extrapolated images perfectly retain the essential structure of original test images.

This justifies the application of the above method to traffic data. First, the mean velocity fields are estimated for all directional sectors. Then, the deviations from the mean of actual speed values are extrapolated for particular days and time intervals. On their basis, traffic–jam maps are created; see Figs. 19–20.

There are several interesting perspectives for further research. In particular, using methods recently developed in Heinrich *et al.*, 2004, Klenk *et al.*, 2004, and Schmidt and Spodarev, 2004, models of stochastic geometry can be statistically fitted to extrapolated traffic maps. In a next step, the fitted models can be used in order to predict future traffic states on the basis of currently incoming traffic data. Such space-time prediction models as well as their applications to forecasting of traffic states will be discussed in a forthcoming paper.

SOME PRELIMINARIES

Random fields

To model traffic maps, non–stationary random fields composed of a deterministic drift and an intrinsically stationary random deviation field, the so-called residual, are used. See e.g. monographs Cressie, 1993 and Wackernagel, 1998 for details.

Drift and deviation field

Let $X = \{X(u),\, u \in \mathbb{R}^2\}$ be a non–stationary random field with finite second moments

$$E[X(u)^2] < \infty, \quad u \in \mathbb{R}^2.$$

Then X can be decomposed into a sum

$$X(u) = m(u) + Y(u)$$

where $m(u) = E[X(u)]$ is the mean field (*drift*) and $Y(u) = X(u) - m(u)$ is the deviation field from the mean or

residual. Clearly, it holds $E[Y(u)] = 0$ for all u. Assume that Y is intrinsically stationary of order two. Denote by

$$\gamma(h) = \frac{1}{2}E[(Y(u) - Y(u+h))^2] \quad (1)$$

its variogram function. In practice, the field X can be observed in a compact (say, rectangular) window $W \subset \mathbb{R}^2$. Let $x(u_1), \ldots, x(u_n)$ be a sample of observed values of X, $u_i \in W$ for all i. The extrapolation method described in the next section yields an "optimal" estimator $\widehat{X}(u)$ of the value of $X(u)$ for any $u \in W$ based on the sample random variables $X(u_1), \ldots, X(u_n)$. Among the variety of extrapolation techniques for non–stationary random fields (see e.g. the *universal kriging* in Cressie, 1993, Kitanidis, 1997, Wackernagel, 1998), our approach is similar to the so–called *kriging based on the residuals*; see Cressie, 1993, p. 190. The main idea of the method is straightforward. First of all, an estimator $\widehat{m}(u)$ for the drift $m(u)$ has to be constructed. Then, the deviation field $Y^* = \{Y^*(u),\, u \in \mathbb{R}^2\}$ defined by

$$Y^*(u) = X(u) - \widehat{m}(u) \quad (2)$$

is formed and its kriging estimator $\widehat{Y}^*(u)$ is computed. Finally, the estimator $\widehat{X}(u)$ is given by

$$\widehat{X}(u) = \widehat{m}(u) + \widehat{Y}^*(u). \quad (3)$$

If we suppose that the drift is known, i.e. $\widehat{m}(u) = m(u)$ for all u then we know the exact values $Y(u_1), \ldots, Y(u_n)$ of the deviation field at $u_1, \ldots, u_n$ since

$$Y^*(u) = Y(u) = X(u) - m(u).$$

Let

$$y(u_i) = x(u_i) - m(u_i), \quad i = 1, \ldots, n$$

be a realization of the sample values of Y. The extrapolation of $Y(u)$ can be performed either by *simple kriging* based on the covariance function

$$C(h) = E[Y(u)Y(u+h)]$$

or by *ordinary kriging* making use of the variogram $\gamma(h)$; see Cressie, 1993, Kitanidis, 1997, Wackernagel, 1998. In what follows, the second method is used.

ORDINARY KRIGING WITH MOVING NEIGHBORHOOD

The kriging estimator

A simpler version of the following *ordinary kriging with moving neighborhood* can be found in Chilès and Delfiner, 1999, pp. 201–210, Kitanidis, 1997, pp. 71 and Wackernagel, 1998, pp. 101–102. Denote by $\mathbf{1}$ the usual indicator function

$$\mathbf{1}\{x \in B\} = \begin{cases} 1 & \text{if } x \in B, \\ 0 & \text{otherwise.} \end{cases}$$

Introduce the estimator $\widehat{Y}(u)$ of $Y(u)$ at $u \in W$ as a linear combination of the sample random variables $Y(u_i)$ with unknown weights $\lambda_i = \lambda_i(u)$ by

$$\widehat{Y}(u) = \sum_{i=1}^{n} \lambda_i Y(u_i) \mathbf{1}\{u_i \in A(u)\}. \quad (4)$$

The estimation involves only the sample random variables $Y(u_i)$ such that u_i is positioned in the "neighborhood" $A(u)$ of u, i.e. $u_i \in A(u)$. Being an arbitrary set, this moving neighborhood $A(u)$ contains *a priori* information about the geometric dependence structure of the random field Y. For instance, it could be designed to model the formation of traffic jams. In the case of a Boolean model, this set $A(u)$ is influenced by the shape of the primary grain. In general, $A(u)$ can be a random closed set, i.e. $A(u) = A\,(Z(u,\omega), u)$ where $Z = \{Z(u),\, u \in \mathbb{R}^2\}$ is a random field containing extra information about Y. Under such general assumptions on $A(u)$, the system of linear equations on the weights λ_i looks much more complicated than (7) considered below. In order to solve it, additional parameters such as crosscovariances of Y and Z should be estimated. Even in the case of uncorrelated fields Y and Z, it makes the extrapolation unnecessary complex. To avoid this, the present paper uses only deterministic sets $A(u)$.

The normalizing condition on the weights λ_i

$$\sum_{i=1}^{n} \lambda_i = 1 \qquad (5)$$

ensures the unbiasedness of the estimator given in (4). In other words, it holds

$$E[\widehat{Y}(u)] = E[Y(u)]$$

even if the mean of Y is not zero. Moreover, this condition makes it possible to use variograms in (7) since variograms are negative conditionally semidefinite (see e.g. Wackernagel, 1998, pp. 52–53). The "optimality" of the estimator $\widehat{Y}(u)$ means that its variance should be minimal, i.e.

$$E[(\widehat{Y}(u) - Y(u))^2] \longrightarrow \min\,. \qquad (6)$$

This classical minimization problem yields further conditions on λ_i which can be written together with (5) in the following system of linear equations. For all $i = 1, \ldots, n$ with $u_i \in A(u)$ it holds

$$\begin{aligned} \sum_{j=1}^{n} \lambda_j \gamma(u_j - u_i)\mathbf{1}\{u_j \in A(u)\} \\ +\mu = \gamma(u - u_i), \\ \sum_{j=1}^{n} \lambda_j \mathbf{1}\{u_j \in A(u)\} = 1. \end{aligned} \qquad (7)$$

In order to solve this system of equations, the knowledge of the variogram function $\gamma(h)$ is required. However, in most practical cases $\gamma(h)$ is unknown and has to be estimated from the data $y(u_1), \ldots, y(u_n)$.

Estimation of the variogram

In applications, a variogram estimation method to be used should always be chosen in accordance with the data framework. The most simple and popular one is undoubtedly the estimator of Matheron (see e.g. Chilès and Delfiner, 1999, Wackernagel, 1998). Its drawback is sensitivity to outliers. Among robust

estimation methods, the trimmed mean estimator (see e.g. Lehmann and Casella, 1998) as well as the estimators of Cressie–Hawkins (see Cressie, 1993) and Genton (see Genton, 1998a, Genton, 2001) should be mentioned. These methods are designed for noisy data but they are biased.

Since the traffic data seem to be not contaminated with outliers, the estimator of Matheron is used here. It is defined by

$$\widehat{\gamma}(h) = \frac{1}{2N(h)} \times \\ \times \sum_{i,j:u_i-u_j\approx h} (Y(u_i) - Y(u_j))^2 \quad (8)$$

where $u_i - u_j \approx h$ means that $u_i - u_j$ belongs to a certain neighborhood $U(h)$ of vector h and $N(h)$ denotes the number of such pairs (u_i, u_j) for $i, j = 1, \ldots, n$. The choice of $U(h)$ depends on the problem. In the present paper, the following segment of a circle is used:

$$U(h) = \{x \in \mathbb{R}^2 : x = (|x|, \varphi), \\ | \, |h| - |x| \, | < \delta, \, |\varphi - \varphi_0| < \varepsilon\}, \quad (9)$$

where $(|x|, \varphi)$ and $(|h|, \varphi_0)$ are the polar coordinates of x and h; δ, $\varepsilon > 0$. If γ is continuous then the estimator in (8) is asymptotically unbiased, i.e.

$$\lim_{\delta,\varepsilon\to 0} E[\widehat{\gamma}(h)] = \gamma(h).$$

Under further assumptions on Y such as ergodicity, it is also strongly consistent, i.e. it holds

$$\lim_{N(h)\to\infty} \widehat{\gamma}(h) = \gamma(h)$$

almost surely.

Variogram models

In practice, the estimated variogram $\widehat{\gamma}$ can not be substituted directly for γ in the system of linear equations (7). Trying this would make the numerical computation in (7) unstable because of the singularity of its coefficient matrix. Even in the case when this computation is possible its result is not correct. The reason for that is simple: $\widehat{\gamma}(h)$ is not a valid variogram function since it is not conditionally negative semidefinite. Hence, a valid parametric variogram model γ (the so–called *theoretical* variogram) should be fitted to the *empirical* estimator $\widehat{\gamma}$. In the following, some valid variogram models are considered. The corresponding fitting procedures are discussed later on. A popular isotropic variogram model is the *exponential* one (see e.g. Cressie, 1993, pp. 61–63, Wackernagel, 1998, pp. 244–246):

$$\gamma(h) = \begin{cases} 0, & h = 0, \\ a + b(1 - e^{-|h|/c}), & h \neq 0, \end{cases}$$

where $a \geqslant 0$, $b \geqslant 0$ and $c > 0$ are parameters with the following geometric meaning. The value of the *nugget effect* a measures the discontinuity of the realizations of Y at the microscopic scale. If $a > 0$ then the realizations of Y are not continuous. The *sill* b describes the variability of the data for greater distances $|h|$. The

third parameter c is the *range* of correlation of Y which implies that the random variables $Y(x)$ and $Y(x+h)$ are almost uncorrelated for $|h| > 3c$.

A parametric variogram model γ is called *geometrically anisotropic* if the range value c (and none of the other parameters) depends on the direction of h. If, in addition, the sill value b depends on the direction of h, the variogram is called *zonally anisotropic*.

As shown in Fig. 17, the traffic data lead to empirical variograms that are clearly zonally anisotropic. Below, we consider zonally anisotropic variogram models constructed from isotropic ones (see Cressie, 1993, Wackernagel, 1998). Introduce

$$\gamma(h) = \gamma_1(h) + \gamma_2(h), \qquad (10)$$

where $\gamma_1(h)$ is an exponential isotropic variogram model with nugget effect $a_1 > 0$, sill b_1 and range c_1. The second term

$$\gamma_2(h) = b_2(1 - e^{-\sqrt{h^\top C h}/c_2}) \qquad (11)$$

is a geometrically anisotropic exponential variogram model with sill $b_2 > 0$ and further parameter $c_2 > 0$. Here C is the quadratic matrix of a linear transformation of the observation window, i.e.

$$C = Q^\top \Lambda Q,$$

where

$$Q = \begin{pmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{pmatrix} \qquad (12)$$

is a rotation by the angle α around the origin and

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \qquad (13)$$

is a scaling transformation with scaling factors λ_1, λ_2 along the coordinate axes. For a vector $h = (h_1, h_2)$, we have

$$h^\top C h = \lambda_2 h_1^2 + \lambda_1 h_2^2 + (\lambda_2 - \lambda_1) \times \\ \times \left(\cos^2\alpha (h_2^2 - h_1^2) - h_1 h_2 \sin(2\alpha)\right).$$

Level curves of $\gamma_2(h)$ are ellipses with main axes of polar angles α and $\alpha + \pi/2$. The range values in these directions are equal to

$$\frac{3c_2}{\sqrt{\lambda_1}}, \quad \frac{3c_2}{\sqrt{\lambda_2}}. \qquad (14)$$

Figure 2 shows the level curves of the variogram model (10) with parameter values $a_1 = 130$, $b_1 = 20$, $c_1 = 0.03$, $b_2 = 70$, $c_2^2/\lambda_1 = 10^9$, $c_2^2/\lambda_2 = 5 \cdot 10^{-5}$, $\alpha = 5°$. Higher values of γ are marked red.

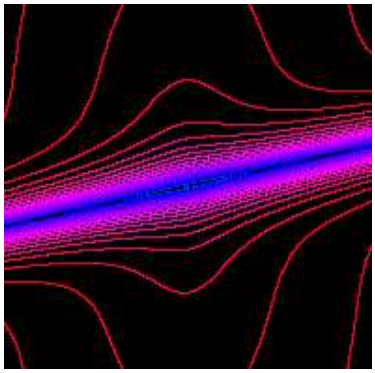

Fig. 2: Zonally anisotropic variogram

Variogram fitting

Let $\widehat{\gamma}(h)$ be an empirical variogram estimated from the experimental data $\{y(u_i)\}$ for the field Y and let $\gamma_\beta(h)$ be a theoretical parametric variogram model with parameter vector

$$\beta = (\beta_1, \ldots, \beta_k).$$

In the example mentioned above, we have

$$\beta = (a_1, b_1, c_1, b_2, \lambda_1/c_2^2, \lambda_2/c_2^2, \alpha).$$

In practice, only a finite number m of values

$$\widehat{\gamma}(h_1), \ldots, \widehat{\gamma}(h_m)$$

can be computed. For two reasons, it is enough to confine computations to vectors h_i of length $|h_i| < \mathrm{diam}(W)/2$. First, in most cases the behavior of the variogram in a small neighborhood of the origin is decisive for the adequate choice of the model. Second, for large distances $|h| > \mathrm{diam}(W)/2$ the estimated values $\widehat{\gamma}(h)$ are contaminated by noise due to edge effects.

In order to estimate the parameter vector β, the least-squares method is used. The generalized least-squares method (see Genton, 1998b) minimizes the following function of β

$$F(\beta) = \sum_{i,j=1}^{m} w_{ij}(\gamma_\beta(h_i) - \widehat{\gamma}(h_i)) \times \\ \times (\gamma_\beta(h_j) - \widehat{\gamma}(h_j))$$

where the weights can be chosen in accordance with the a priori assumptions on Y (see Cressie, 1993 for Gaussian random fields). If the distribution of Y is unknown, the classical weighting scheme can be applied:

$$w_{ij} = \begin{cases} 1, & i = j, \\ 0, & i \neq j \end{cases}$$

with the function

$$F(\beta) = \sum_{i=1}^{m} (\gamma_\beta(h_i) - \widehat{\gamma}(h_i))^2 \quad (15)$$

to be minimized. In the case of traffic data, no a priori assumptions on the structure of Y have been made. Thus, the classical least-squares methods is used.

For isotropic random fields, one fits the one–dimensional curve of a parametric variogram model $\gamma_\beta(|h|)$ to an empirical one. In the aniso–tropic case, $\widehat{\gamma}(h)$ is computed for vectors h on a square grid with m points and is fitted by a two–dimensional parametric surface $\gamma_\beta(h)$, $h \in \mathbb{R}^2$. This can be done either by summing in (15) over all grid points h_i or only over vectors h_i in a certain direction of interest φ, i.e.

$$h_i = |h_i|(\cos\varphi, \sin\varphi).$$

Since traffic data is substantially anisotropic, the variogram model (10) has to be fitted to the data on the whole grid as well as in two directions with polar angles α and $\alpha + \pi/2$.

DRIFT ESTIMATION

The mean field $\{m(u)\}$ can be estimated from the data by various methods ranging from radial extrapolation (see e.g. zu Castell *et al.*, 2002 and references therein) to smoothing techniques such as moving average and edge preserving smoothing (see e.g. Tomasi and Manduchi, 1998). In what follows, the moving average is used because of its ease and computational efficiency for large data sets.

By moving average, the value $m(u)$ is estimated as

$$\widehat{m}(u) = \frac{1}{N_u} \sum_{u_i \in W(u)} X(u_i) \qquad (16)$$

where $W(u)$ is the "moving" neighborhood of the point u and N_u denotes the number of measurement points $u_i \in W(u)$. The choice of the neighborhood $W(u)$ is arbitrary. For fast computation, we put $W(u)$ to be a square with side length τ centered in u.

The estimator (16) yields arbitrarily smooth results for large moving neighborhoods $W(u)$. Thus, an optimal side length τ should be found to fit the problem. In the traffic problem, τ must be small because edges of the surface $\{m(u), u \in W\}$ are intrinsic to the image structure and have to be preserved by smoothing.

In all large cities, there are areas D of parks, forests, building blocks, etc. where no road–traffic data is available. By (16), this implies $\widehat{m}(u) = 0$ for all points u with $W(u) \subset D$. Consequently, such points u would automatically belong to traffic–jam regions and so contaminate traffic–jam maps with artefacts. To avoid this, the neighborhood $W(u)$ of points u with $N_u = 0$ has to be enlarged till it contains at least one observation point. In this way, meaningful average velocity maps are obtained that allow the correct analysis of traffic jams.

Since X is not stationary and, consequently, $m(u)$ is not constant the estimator (16) is biased. Nevertheless, in practical applications, the bias $E\,\widehat{m}(u) - m(u)$ is small provided that the area $|W(u)|$ is small and the net of observation points is spatially dense enough.

RESIDUALS FORMED WITH ESTIMATED DRIFT

In previous sections, it has been assumed that the drift $m(u)$ was explicitly known. If it has to be estimated from the data, the theoretical background for the application of the kriging method breaks down.

Indeed, kriging requires intrinsic stationarity of the field of residuals $Y^*(u)$ introduced in (2). This requirement is clearly not satisfied even in the case of an unbiased estimator $\widehat{m}(u)$ since the variogram

$$\gamma^*(h) = \frac{1}{2} E[Y^*(u) - Y^*(u+h)]^2$$

is not equal any more to the variogram $\gamma(h)$ of Y (see Chilès and Delfiner, 1999, pp. 122–125, Cressie,

1993 p. 72, Wackernagel, 1998, p. 214) and depends clearly on u.

Despite these theoretical obstacles, practitioners continue to use the ordinary kriging of residuals with estimated drift based on the data $y^*(u_i) = x(u_i) - \widehat{m}(u_i),\ i = 1, \ldots, n$ legitimized by its ease and satisfactory results.

ALGORITHMS AND IMPLEMENTATION IN JAVA

In the following, some efficient algorithms for spatial extrapolation are discussed. Their implementation in Java was integrated into the GeoStoch library GeoStoch, 2004 as a separate package. The software is supplied with detailed comments generated by Java–Doc complying with the Sun standards; see Niemeyer and Peck, 1996, pp. 80–81.

Fast estimation of variograms and drifts

Matheron's estimator (8) requires all pairs of positions u_i and u_j with $u_i - u_j \in U(h)$ to be found for each lattice vector h. For k lattice vectors and n positions, it costs $k * \mathcal{O}(n^2)$ operations. By means of the binary search tree structure DTree, this complexity can be significantly reduced.

Such algorithm tessellates the searching space into rectangles and saves positions of actual measurements in a binary tree. Thus, searching k points from p costs in average $r + \log(p)$ operations; see Segewick, 1992. Since $r \ll p$ always holds, the average complexity of the search is $\mathcal{O}(\log(p))$. Additionally, the complexity of filling the tree with values is $\mathcal{O}(p * \log(p))$. For variogram estimation, one stores $p = \frac{n*(n-1)}{2}$ polar coordinates of the vectors between any two measurement points in a DTree. Thus, the overall complexity for the variogram computation is $\mathcal{O}(p * \log(p)) + k * \mathcal{O}(\log(p))$. For large square lattices with side length $m > 200$ ($k = m^2$), the difference in run times is significant!

The DTree structures can be used also for the fast computation of the moving average. There, measurement points lying in a certain square neighborhood should be found. The complexity of such computation can be estimated as mentioned above.

Variogram fitting

In variogram fitting, one employs essentially known algorithms for the minimization of functions. The idea of all stochastic algorithms lies in cleverly modifying parameters of the variogram model at random till the maximal quadratic distance to the empirical curve becomes smaller than a critical value ε. This can be done for instance by means of *genetic algorithms* (see Goldberg, 1989) or the method of *simulated annealing*; see e.g. Press *et al.*, 2002, pp. 448–460. Genetic algorithms were implemented in Java and integrated in the GeoStoch library. The simulated

annealing Java package `JSimul` is available from Mégnin, 2001. Additionally, one–dimensional variogram fitting by slices was implemented in Java by Faulkner, 2002. This Java package provides good GUI but poor runtime performance for large data samples.

TEST EXAMPLE: BOOLEAN MODEL

To test the performance of the above extrapolation method, one needs to generate synthetic data whose theoretical properties are known. In other words, one has to find a random field $\{X(u)\}$ with known structure of distribution, variogram and shape of realizations that is easy to simulate. In the following, we construct such a random field on the basis of the so–called *Boolean random field*, a model that is classical in stochastic geometry.

Definition and properties of the simulation model

In what follows, basic properties of the Boolean model in $\mathbb{R}^2$ are described. For more details, see e.g. Stoyan *et al.*, 1995. Let

$$\Phi = \{X_1, X_2, X_3, \ldots\}$$

be a stationary Poisson point process in $\mathbb{R}^2$ with intensity λ; see Fig. 3.

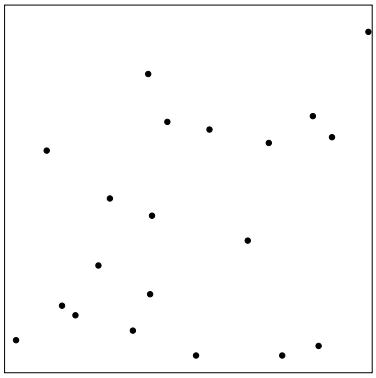

Fig. 3: Realization of Φ

For simulation of Poisson processes, see e.g. Lantuéjoul, 2002, Stoyan *et al.*, 1995. A *Boolean random set* Ξ with deterministic grains can be introduced as

$$\Xi = \bigcup_{i=1}^{\infty} (\Xi_0 + X_i)$$

where Ξ_0 is the so–called *primary grain* and $\Xi_0 + X_i$ a grain translated to the germ position X_i.

Fig. 4: Realization Ξ

The primary grain Ξ_0 can be an arbitrary compact set in $\mathbb{R}^2$. In the present paper, a rectangle

$$\Xi_0 = [0, a] \times [0, b] \qquad (17)$$

with width $a > 0$ and height $b > 0$ is considered; see Fig. 4. On the basis of Ξ, one constructs a stationary random field $Y = \{Y(u), u \in \mathbb{R}^2\}$ by setting

$$Y(u) = \mathbf{1}\{u \in \Xi\} - p$$

where the constant $p = 1 - e^{-\lambda ab}$ is the *volume fraction* of Ξ. This random field is a special case of a *Boolean random function* considered e.g. in Serra, 1988. It can take only values $-p$ or $1-p$. The field Y is stationary of order two and it holds $E\,Y(u) = 0$. So it can be used to model the "deviations from the mean".

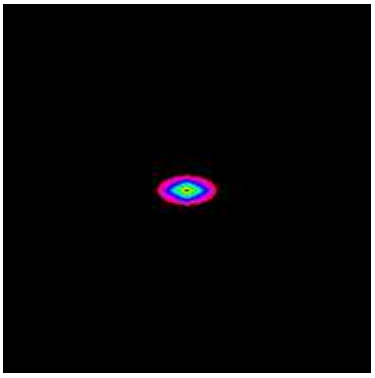

Fig. 5: Variogram γ (level curves)

The variogram $\gamma(h)$ of Y is given by

$$\gamma(h) = e^{-\lambda ab}\left(1 - e^{-\lambda(ab - |\Xi_0 \cap (\Xi_0 - h)|)}\right).$$

For a vector $h = (h_1, h_2)$, the area $|\Xi_0 \cap (\Xi_0 - h)|$ is equal to $(a - |h_1|)(b - |h_2|)$ for $|h_1| \leqslant a$, $|h_2| \leqslant b$, and zero, otherwise. This variogram is clearly anisotropic as shown in Fig. 5 for parameter values $a = 40$, $b = 20$ and $\lambda = 0.0006$ in the observation window $W = [0, 200]^2$.

In order to model a non–stationary field X, one adds a deterministic drift variable $m(u)$ to the field $Y(u)$. As a toy example, $m(u)$ is chosen here to be the indicator function

$$m(u) = \mathbf{1}\{u \in B_r(u_0)\} \qquad (18)$$

of a deterministic circle $B_r(u_0)$ with center u_0 and radius $r > 0$; see Fig. 6. The resulting field

$$X(u) = m(u) + Y(u), \quad u \in \mathbb{R}^2$$

attains only three values $-p$, $1 - p$, $2 - p$.

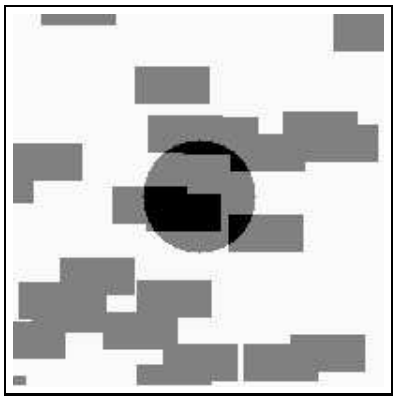

Fig. 6: Realization of X

To test the extrapolation quality on synthetic data, one simulates X and measures its realization $x(u)$ at a finite number of points u_i. Then one extrapolates X from the data $x(u_i)$ and compares the result with the original realization $x(u)$. Measurement points u_i are generated by an independent Poisson process Φ_1 with intensity $\lambda_1 = 0.01$; see Fig. 7.

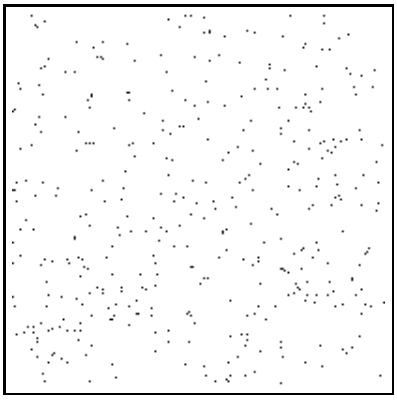

Fig. 7: Realization of Φ_1

The intensity of Φ_1 is substantially higher than that of Φ since otherwise the information contained in the data is insufficient to reconstruct the original image.

Synthetic data

Practically, the experiment described above should be repeated many times in order to reduce the randomness in the quality of results. In this paper, 90 realizations of X have been sampled. They yield 90 data sets each of them containing ca. 300 pairs $(u_i, x(u_i))$. These data sets correspond to the traffic data of a half an hour. The intensity of Φ_1 is chosen to produce in average about 300 measurement points to comply with the real traffic situation.

For simulations, we used the following parameter values:

$$W = [0, 200]^2,\ u_0 = (100, 100),$$
$$r = 30,\ a = 40,\ b = 20,$$
$$\lambda = 0.0006.$$

The mean area fraction is then

$$p = 0.38121662.$$

In Fig. 4, a realization of Ξ with these parameter values is shown. By adding a circle in the middle of the picture and subtracting p, one obtains a realization of the random field X; see Fig. 6.

Numerical results; reconstruction of simulated images

To estimate the drift, moving average with the side length $\tau = 3$ of the square neighborhood was used.

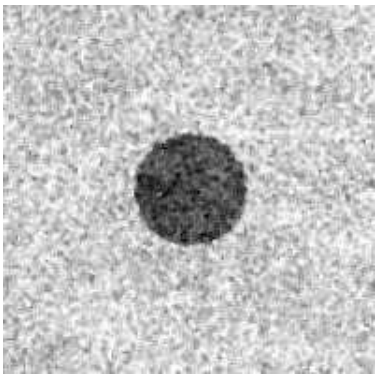

Fig. 8: Estimated drift $\widehat{m}(u)$

As seen in Fig. 8, the estimated drift preserves the original drift structure.

After subtracting the estimated drift $\widehat{m}(u)$ from the data in each data set j, $j = 1, \ldots, 90$, the empirical variogram $\widehat{\gamma}_j^*$ of Y^* is computed; see Fig. 9. The parameter values of the circular segment (9) are $\delta = 2$, $\varepsilon = 3°$.

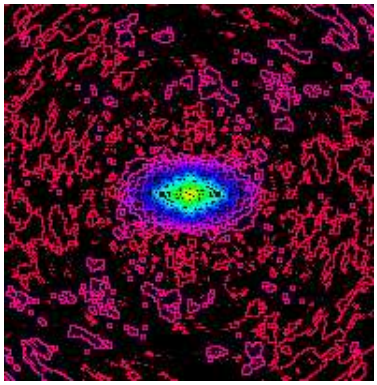

Fig. 9: Estimated variogram $\widehat{\gamma}^*(h)$ (level curves)

Then, one averages the variogram over all 90 estimated copies $\widehat{\gamma}_j^*$ by arithmetic mean:

$$\widehat{\gamma}^*(h) = \frac{1}{90}\sum_{j=1}^{90}\widehat{\gamma}_j^*(h).$$

This mean variogram can be well–fitted by the true variogram of the Boolean model. For fitting, simulated annealing was used to minimize the target function (15) in the least squares method.

The parameters of the simulated annealing are chosen as follows: maximal temperature 10^6, annealing rate 20, number of iterations 10, tolerance value 10^{-5}; see Press *et al.*, 2002 for their meaning. The starting values of the variogram parameters were $a_0 = 20$, $b_0 = 10$, $\lambda_0 = 0.006$. The fitting yields parameter values

$$\widehat{a} = 39.7605124, \quad \widehat{b} = 20.7768498,$$
$$\widehat{\lambda} = 0.001193$$

lying quite close to the original ones. The maximal (mean) deviation of $\widehat{\gamma}^*$ from γ^* is 0.03684976 ($6.337388{\cdot}10^{-5}$, respectively); see Fig. 11.

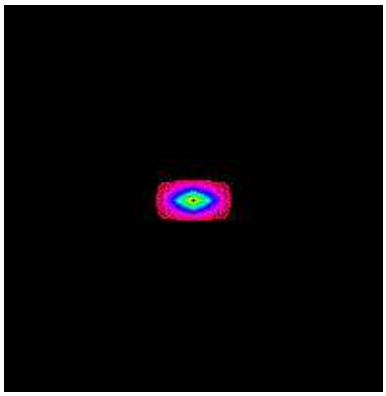

Fig. 10: Fitted variogram model $\gamma^*(h)$ (level curves)

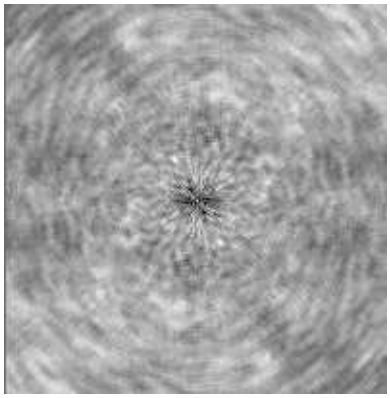

Fig. 11: Difference between the fitted theoretical model $\gamma^*(h)$ and the empirical variogram $\widehat{\gamma}^*(h)$

The knowledge of the grain shape (17) can be integrated in the indicator functions of the kriging with moving neighborhood. Put the set $\{u_i \in A(u)\}$ in (4) to be equal to

$$\{|x - x_i| \leqslant \widehat{a}, |y - y_i| \leqslant \widehat{b}\}$$

where $u_i = (x_i, y_i)$ and $u = (x, y)$ denote the Euclidean coordinates of points u_i and u.

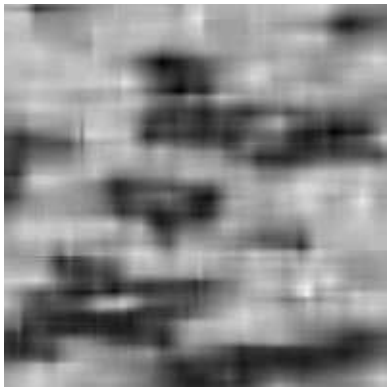

Fig. 12: Residual $\widehat{Y}^*(u)$

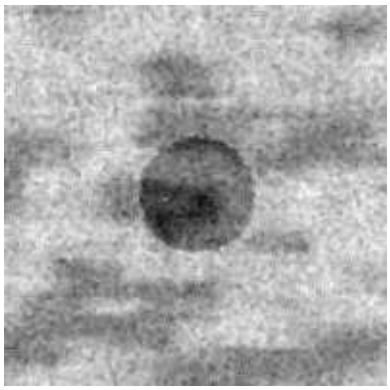

Fig. 13: Extrapolated field $\widehat{X}(u)$

Thus, the extrapolation method will use only those points u_i that can potentially affect the value $Y^*(u)$. The extrapolation results $\widehat{Y}^*(u)$ and $\widehat{X}(u)$ are shown in Figs. 12 and 13. The striking similarity of the images for $X(u)$ and $\widehat{X}(u)$ in Figs. 6 and 13, respectively is a clear evidence for the high quality of the extrapolation method.

Statistical tests for the area fraction

Fig. 14: The threshold image $\widehat{\Xi}$ of $\widehat{Y}^*$

The threshold image of $\widehat{Y}^*$ in Fig. 14 is a binary image that can be compared with the original image of Y in Fig. 4. Written in terms of functions, it is equal to $\mathbf{1}\{u \in \widehat{\Xi}\}$ where $\widehat{\Xi} = \{u : \widehat{Y}^*(u) \geqslant 1/2 - p\}$ and $1/2 - p = 0.11878339181$. To quantify visual similarities in both images, statistical tests for the area fraction can be used, see Böhm *et al.*, 2004. For each of 90 threshold images, the null hypothesis $H_0 : \widehat{p} = p$ is tested vs. its alternative $H_1 : \widehat{p} \neq p$ where

$$\widehat{p} = \frac{|\widehat{\Xi} \cap W|}{|W|}$$

is an estimator of the area fraction of the threshold image and

$$p = 0.38121660819385905$$

the area fraction of the original Boolean model. If the threshold image is a realization of a Boolean model and the null hypothesis H_0 is true the corresponding test statistic

$$T = \frac{\sqrt{|W|}(\widehat{p} - p)}{\sqrt{\sum_{|h| \leqslant b} |W \cap (W-h)| \widehat{C}_1(h)}} \sim N(0,1)$$

is asymptotically $N(0,1)$ – distributed as $|W| \to \infty$ where

$$\widehat{C}_1(h) = \frac{|\widehat{\Xi} \cap (\widehat{\Xi}-h) \cap W \cap (W-h)|}{|W \cap (W-h)|} - \widehat{p}^2$$

is a consistent estimator of the covariance function

$$C_1(h) = P(o \in \widehat{\Xi},\, h \in \widehat{\Xi}) - P^2(o \in \widehat{\Xi})$$

of the random set $\widehat{\Xi}$.

Thus, the null hypothesis H_0 is rejected at the asymptotic significance level $1-\theta$ if

$$|T| > z_{1-\theta/2}$$

where $z_{1-\theta/2}$ is the $(1-\theta/2)$–quantile of the standard normal distribution.

For $\theta = 0.04$ and $z_{1-\theta/2} = 2.054$, the null hypothesis H_0 was rejected in 6 % to 10 % of realizations depending on the series of the 90 images. It attests statistically the visual similarity of the images of Ξ and $\widehat{\Xi}$. The test results can be improved by choosing larger observation windows (e.g. 400×400 pixels), smaller grains (e.g. $a = 20$, $b = 10$) and more measurement points per image (say, 2000). The reason for that is the asymptotic nature of the test. The significance level is approximately equal to $1-\theta$ if W is large enough, i.e. beginning from a particular relation between the sizes of grains and the observation window. Additionally, we suppose that increasing the number of measurement points would improve the extrapolation quality and consequently reduce the rejection rate of H_0 to 4 %. However, in our experiments we kept the number of approx. 300 measurement points constant in order to preserve analogies to the traffic problem setting.

ANALYSIS OF TRAFFIC DATA

In what follows, the above extrapolation method is applied to real traffic data.

The original data set contains entries with spatial positions scattered not only over Berlin but also over a wide region with radius of approx. 100 km from the city center. To avoid too large inhomogeneities, the observation window is reduced to downtown Berlin with geographic coordinates

$$13.3 \leqslant x \leqslant 13.46666,$$
$$52.48333 \leqslant y \leqslant 52.55\,.$$

Then, the data analysis is performed for the directional sector $S_2 = \{\alpha : \pi/2 \leqslant \alpha < \pi\}$ including data of taxis moving northwest. This partial data set contains 19699 entries collected over 90 days (see Fig. 15).

To calculate the mean velocity field $\widehat{m}(u)$, the moving average in (16) is applied to the data of sector 2 (see Fig. 16). The side length of the square moving neighborhood is $\tau = 0.005$. In Figs. 15 and 16, the northwest movement direction of the taxis can be clearly recognized. Color gradations reflect speed variation from green and blue for high values through yellow for the middle ones up to red for the low ones. Figure 16 shows the corresponding mean field $\widehat{m}(u)$.

The comparison of both maps confirms that the estimator $\widehat{m}$ preserves the spatial velocity structure of the data. To estimate the variogram γ^* of Y^*, the mean values $\widehat{m}(u)$ have to be subtracted from the actual velocity values. Then, the empirical variogram $\widehat{\gamma}_i^*$ is calculated for each day $i = 1, \ldots, 90$.

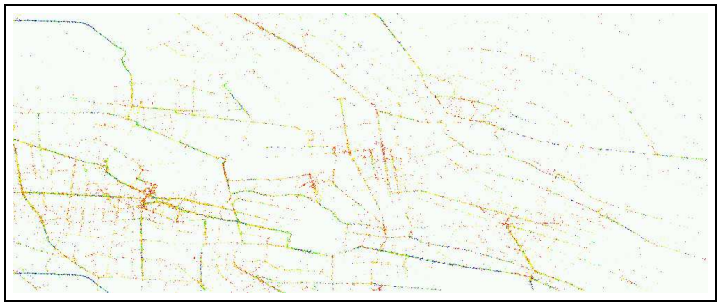

Fig. 15: Positions of taxis moving northwest

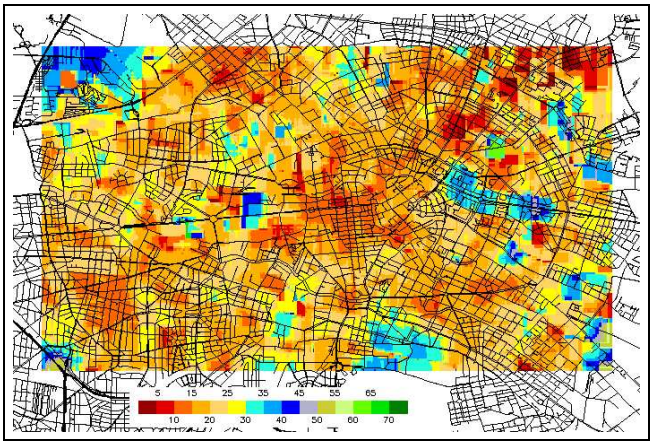

Fig. 16: Mean field $\widehat{m}(u)$ of data set 2

Averaging on all days, one obtains the following variogram estimator for Y^*

$$\widehat{\gamma}^*(h) = \frac{1}{90}\sum_{i=1}^{90}\widehat{\gamma}_i^*(h)\,;$$

see Fig. 17.

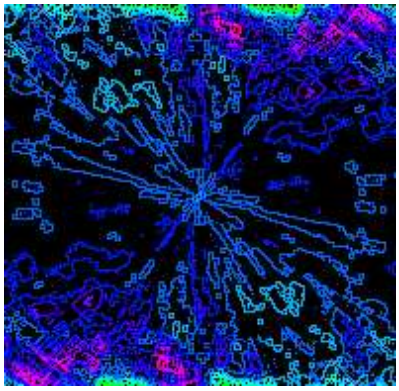

Fig. 17: Empirical variogram $\widehat{\gamma}^*(h)$ (level curves)

The parameters of the segment in (9) used for variogram calculation are $\delta = 0.006$ and $\varepsilon = 3°$ with maximal distance $h = 0.07$ being approximately a half diameter of W. The empirical variogram $\widehat{\gamma}^*(h)$ with maxima in northwest direction and minima in orthogonal direction is zonally anisotropic showing substantial northwest correlation in the data.

In Fig. 17, level curves are colored in accordance with the increasing variogram values from green, blue and yellow to red, where the zonally anisotropic behavior of $\widehat{\gamma}^*(h)$ near the origin becomes clear. The variogram values are low in a narrow sector at the polar angle of approximately 170°, i.e. traffic velocities are highly correlated in this direction.

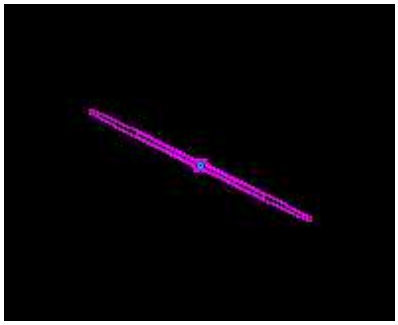

Fig. 18: Fitted variogram model $\gamma^*(h)$ (level curves)

The zonally anisotropic variogram model (10) with two fixed parameters $\alpha = 170°$, $\lambda_1/c_2^2 = 1000$ taken from Fig. 17 has been fitted to the empirical one; see Fig. 18. The classical least squares fitting method applied to one–dimensional vertical slices of the empirical variogram in orthogonal directions $\alpha = 80°$ and $\alpha = 170°$ yields other parameter values:

$$\begin{aligned} a_1 &= 31.77189640437076,\\ b_1 &= 116.21092322,\\ c_1 &= 245388.67081,\\ b_2 &= 22.6344102,\\ \lambda_2/c_2^2 &= 683964.79366. \end{aligned}$$

Thus, the range values in directions 170° and 80° are $r_1 = 0.27$ km and $r_2 = 0.162$ km, respectively. So a vehicle driving in direction $\alpha = 170°$ influences only those vehicles driving behind it in the same direction at a

maximal distance $3r_1 = 810$ m. Vehicles driving behind it in the orthogonal direction 80° are influenced up to a distance of $3r_2 = 648$ m.

Additionally, the two–dimensional surface of the above variogram model has been fitted by least squares to the empirical one using genetic minimizing algorithms. Resulting parameter values are very close to those obtained above:

$$\begin{aligned} a_1 &= 19.379745108968454, \\ b_1 &= 95.3944270699768, \\ c_1 &= 245867.97491680854, \\ b_2 &= 9.486514644862856, \\ \lambda_2/c_2^2 &= 684317.2022809463, \\ \lambda_1/c_2^2 &= 1023.8357320907359, \\ \alpha &= 146,84^\circ. \end{aligned}$$

For extrapolation, the sample of velocities $x(u_1), \ldots, x(u_n)$ $(n = 223)$ observed on Monday, 18.02.2002 is used. Compared to the whole data set 2 representing the "past", it is interpreted as "actual" data. The random field Y^* of deviations from mean velocities is extrapolated using kriging with moving neighborhood (4) with the following indicator function

$$\mathbf{1}\{u_i \in A(u)\} = \mathbf{1}\{\varphi(u_i - u) \in S_2\}$$

where $\varphi(u_i - u)$ is the polar angle of the vector $u_i - u$. This assumption is rather intuitive since only those measurements with positions u_i lying "ahead" of the current position u can influence its velocity value.

Extrapolated residuals $\widehat{Y}^*(u)$ and the resulting velocity map $\widehat{X}(u)$ are shown in Figs. 19 and 20, respectively. Due to the particular asymmetric form of the indicators, the extrapolated field of residuals is strongly discontinuous. This obviously affects the geometric characteristics of $\widehat{X}(u)$. Discontinuities of the realizations of X caused by the kriging with moving neighborhood are essential for precise localization of traffic–jam areas. In Fig. 21, areas with velocities $\widehat{X}(u) \leqslant 15$ kph are marked yellow. Some of these regions might be caused by traffic jams.

ACKNOWLEDGEMENT

This research has been supported by the German Aerospace Center (DLR) through research grant 931/69175067. The authors are grateful to Prof. Reinhard Kühne and his co-workers from the DLR Institute of Transport Research for suggesting the problem and fruitful discussions on the subject.

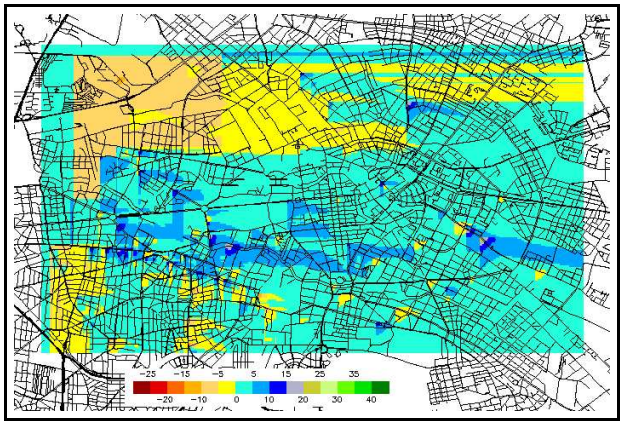

Fig. 19: Residual field $\widehat{Y}^*(u)$

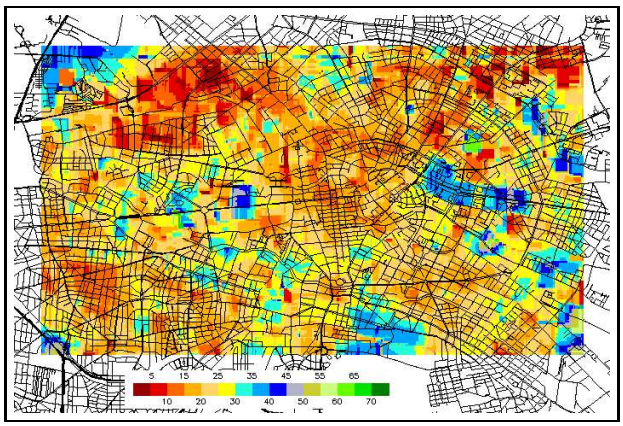

Fig. 20: Velocity field $\widehat{X}(u)$

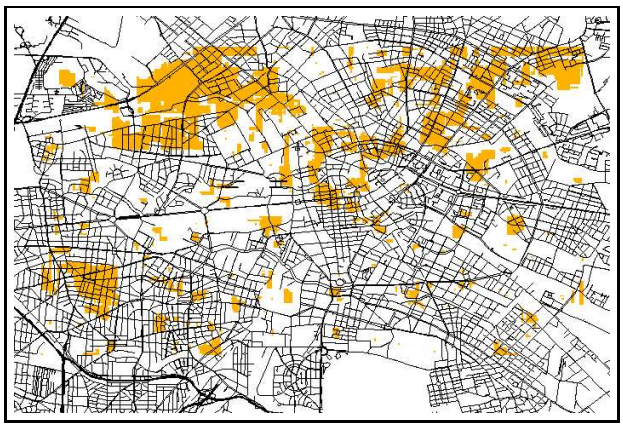

Fig. 21: Traffic jams: $\widehat{X}(u) \leqslant 15$ kph

REFERENCES

Böhm S, Heinrich L, Schmidt V (2004). Asymptotic properties of estimators for the volume fraction of jointly stationary random sets. Statistica Neerlandica (to appear).

Chilès JP, Delfiner P (1999). Geostatistics: modelling spatial uncertainty. Wiley, New York.

Cressie NAC (1993). Statistics for spatial data. Wiley, New York.

Faulkner B (2002). Java classes for nonprocedural variogram modelling. Computers and Geosciences, 28 (3):387–397.

Genton MG (1998a). Highly robust variogram estimation. Mathematical Geology, 30:213–221.

Genton MG (1998b). Variogram fitting by generalized least squares using an explicit formula for the covariance structure. Mathematical Geology, 30:323–345.

Genton MG (2001). Robustness problems in the analysis of spatial data. In: Moore M, ed. Spatial Statistics: Methodological Aspects and Applications. Lecture Notes in Statistics, Springer, New York, 21–37.

GeoStoch (2004). Java library. University of Ulm, Department of Applied Information Processing and Department of Stochastics, http://www.geostoch.de .

Goldberg DE (1989). Genetic algorithms in search, optimization, and machine learning. Addison-Wesley, Massachussetts.

Heinrich L, Schmidt H, Schmidt V (2004). Limit theorems for stationary tessellations with random inner cell structures. Preprint (submitted).

Kitanidis PK (1997). Introduction to geostatistics: applications to hydrogeology. Cambridge University Press, New York.

Klenk S, Schmidt V, Spodarev E (2004). A new algorithmic approach to the computation of Minkowski functionals of polyconvex sets. Preprint (submitted).

Lantuéjoul C (2002). Geostatistical simulation: models and algorithms. Springer, Berlin.

Lehmann E, Casella G (1998). Theory of point estimation. Springer, New York, 2nd edition.

Mayer J, Schmidt V, Schweiggert F (2004). A unified simulation framework for spatial stochastic models. Simulation Modelling Practice and Theory (to appear).

Mégnin C (2001). Jsimul: a Java–based simulated annealing package. http://www.theblueplanet.org/JSimul/html/JSimul_readme.html.

Niemeyer P, Peck J (1996). Exploring Java. O'Reilly, Bonn.

Press WH, Teukolsky SA, Vetterling WT, Flannery BP (2002). Numerical recipes in C++. The art of scientific computing. Cambridge University Press, Cambridge, 2nd edition.

Schmidt V, Spodarev E (2004). Joint estimators for the specific intrinsic volumes of stationary random sets. Stochastic Processes and Their Applications (to appear).

Segewick R (1992). Algorithmen in C. Addison-Wesley, Bonn.

Serra J (1988). Image analysis and mathematical morphology: theoretical advances, volume 2. Academic Press, London.

Stoyan D, Stoyan H, Jansen U (1997). Umweltstatistik. Teubner, Stuttgart.

Stoyan D, Kendall WS, Mecke J (1995). Stochastic geometry and its applications. Wiley, Chichester, 2nd edition.

Tomasi C, Manduchi R (1998). Bilateral filtering for gray and color images. In: Proceedings of the 1998 IEEE International Conference on Computer Vision, Bombay, India, 839–846.

Wackernagel H (1998). Multivariate geostatistics. Springer, Berlin, 2nd edition.

zu Castell W, Weller U, Zipprich M, Sommer M, Wehrhan M (2002). Kriging considered from the deterministic point of view. In: Bayer U, Burger H, and Skala W, eds. Terra Nostra. Schriften der Alfred–Wegener–Stiftung, volume 3, Berlin, 249–254.