Kriged Road-Traffic Maps

Hans Braxmeier[1], Volker Schmidt[2], and Evgueni Spodarev[3]

[1] University of Ulm, Department of Applied Information Processing, D–89069 Ulm, Germany hans.braxmeier@mathematik.uni-ulm.de
[2] University of Ulm, Department of Stochastics, D–89069 Ulm, Germany volker.schmidt@mathematik.uni-ulm.de
[3] University of Ulm, Department of Stochastics, D–89069 Ulm, Germany evgueni.spodarev@mathematik.uni-ulm.de

Summary. A kriging based on residuals is employed for spatial extrapolation of anisotropic directional road–traffic data. The set of data considered in the present paper includes, among others, the actual geographic positions and velocities of approximately 300 test vehicles in downtown Berlin. They are transmitted to a central station within regular time intervals. The main idea of the extrapolation technique is to interpret the recorded velocities as realizations of a random velocity field, which is sampled at selected points only. Structural properties of the resulting road–traffic maps are discussed combined with a statistical space–time analysis of polygonal traffic trajectories extracted from the original traffic data. Finally, a brief outlook to simulation and prediction of future traffic states is provided.

Key words: Spatial extrapolation, kriging, random field, anisotropy, moving neighborhood, statistical inference.
Mathematics Subject Classification 2000: Primary 62M40; Secondary 62H11, 60K30.

1 Introduction

A common difficult problem of large cities with heavy traffic is the prediction of traffic jams. In this paper, a first step towards mathematical traffic forecasting, namely the spatial reconstruction of the present traffic state from pointwise measurements is briefly described. For details, we refer to [1], where models of stochastic geometry and geostatistics are used to spatially represent the traffic state by means of velocity maps. A corresponding Java software that implements efficient algorithms of spatial extrapolation is developed; see [5].

To illustrate our extrapolation method, we use real traffic data originating from downtown Berlin. It was provided to us by the Institute of Transport Research of the German Aerospace Center (DLR). Approximately 300 test

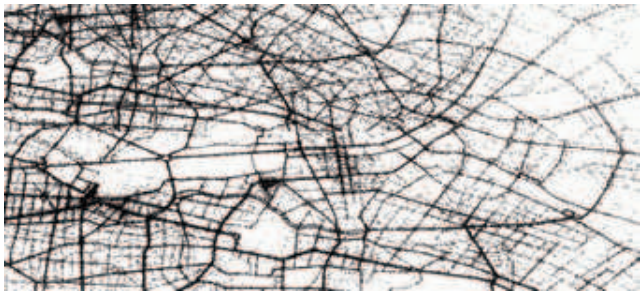


Fig. 1. Observed positions of test vehicles in downtown Berlin

vehicles (taxis) equipped with GPS sensors transmit their geographic coordinates and velocities to a central station within regular time intervals from 30 sec up to 6 min; see Fig. 2. Thus, a large data base of more than 13 million positions was formed since April 2001; see Fig. 1.

In the first stage of our research, only a smaller data set (taxi positions on all working days from 30.09.2001 till 19.02.2002, 5.00–5.30 pm, moving taxis only) was considered. Furthermore, the observation window was reduced to downtown Berlin to avoid inhomogeneities in the taxi positions.

The main idea of the extrapolation technique described in Sects. 2 and 3 below is to interpret the velocities of all vehicles at given time t as a realization of a spatial random field $V(t) = \{V(t,u)\}$ where $V(t,u)$ is a traffic velocity vector at location $u \in \mathbb{R}^2$ and time instant $t \geq 0$. The goal is to analyze the spatial structure of these random fields of velocities in order to describe the geometry of traffic jams. Since $V(t,u)$ can be measured just pointwise at some observation points $u_1, \ldots, u_n$, a spatial extrapolation of the observed data is necessary. Notice that the velocities strongly depend on the location and the direction of movement, e.g. the speed limits and consequently the mean velocities are higher on highways than in downtown streets.

The classical extrapolation methods of geostatistics (see e.g. [6]) either make no use of additional directional information or provide measurements $V(t,u+u_i)$ and $V(t,u-u_i)$ with equal weights. Both these features are not adequate to the setting mentioned above. Thus, the standard extrapolation methods had to be adapted to our specific problem. In Sect. 3, an ordinary kriging with moving neighborhood is described that allows to extrapolate directed velocity fields. First, the original data set should be split into four subsets which are directionally homogeneous. A data unit $(u, V(t,u))$ belongs to the data set i $(i = 1, \ldots, 4)$ if the polar angle of the vector $V(t,u)$ lies within the directional sector $S_i = \{\alpha \in [0, 2\pi) : (i-1)\pi/2 \leq \alpha < i\pi/2\}$. By convention, the zero polar angle corresponds to the eastward direction on the city map. The above data sets should be extrapolated separately for each directional sector. This yields four velocity maps corresponding to the four sectors of directions.

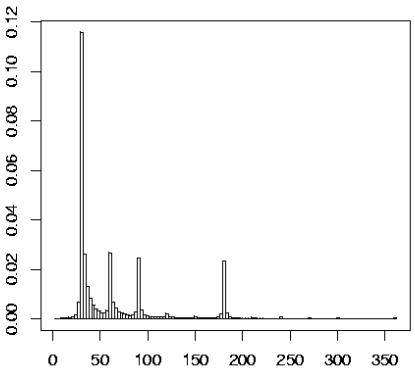


Fig. 2. Histogram of time intervals between consecutive GPS signals (in sec)

In what follows, the data from a given time interval, i.e. $[5.00, 5.30]$ pm, will be taken for extrapolation. Keeping this in mind, we shall omit the time parameter t in further notation.

The extrapolation method described in Sects. 2 and 3 has been implemented in Java, where a software library has been developed comprising the estimation and fitting of variograms as well as the ordinary kriging with moving neighborhood; see [5]. As far as it is known to the authors, this is the first complete implementation of such kriging methods in Java. Much 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; see [1] for details.

In Sect. 4, a numerical example is discussed which shows how the developed extrapolation technique can be applied to directional traffic data. Some structural features of the resulting velocity maps (see Figs. 5 and 6) are also discussed. In Sect. 5, this is combined with a statistical space–time analysis of polygonal road–traffic trajectories which have been extracted from the original traffic data. For example, it turns out that the distribution of the number of segments in these traffic trajectories can be fitted quite well by a geometric distribution. The directional distribution of the segments reflects the anisotropy of the street system of downtown Berlin, where the distribution of segment lengths is demonstrably non-normal. Furthermore, the distributions of velocity residuals, i.e. the deviations from their means, show interesting skewness properties which depend on the considered classes of low, medium, and high mean velocities, respectively. A short outlook to simulation and prediction of future traffic states is given in Sect. 6.

2 Random fields

To model traffic maps, non-stationary random fields composed of a deterministic drift and an intrinsically stationary random field of order two (residual) are used. See e.g. the monographs [4] and [6] for details.

Let $X = \{X(u),\, u \in \mathbb{R}^2\}$ be a non-stationary random field with finite second moment $EX^2(u) < \infty$, $u \in \mathbb{R}^2$. Then, $X(u)$ can be decomposed into a sum $X(u) = m(u) + Y(u)$, where $m(u) = EX(u)$ is the mean field (*drift*) and $Y(u) = X(u) - m(u)$ is the deviation field from the mean or *residual*. Assume that $\{Y(u)\}$ is intrinsically stationary of order two. Denote by

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

its variogram function. In practice, the field X can be observed in a compact (mostly 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 Sect. 3 yields an "optimal" estimator $\widehat{X}(u)$ of the value of $X(u)$ for any $u \in W$ based on the sample variables $X(u_1), \ldots, X(u_n)$.

3 Kriging based on residuals

Among the variety of extrapolation techniques for non-stationary random fields, our approach is similar to the so-called *kriging based on residuals*; see [4], p. 190. 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

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

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

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

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

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

Let $\{y(u_i) = x(u_i) - m(u_i), \quad i = 1, \ldots, n\}$ be a realization of the sample variables $Y(u_1), \ldots, Y(u_n)$. The extrapolation of $Y(u)$ can be performed by *ordinary kriging* making use of the variogram $\gamma(h)$; see [4], [6].

3.1 The kriging estimator

A simpler version of the following *ordinary kriging with moving neighborhood* can be found in [3], pp. 201–210 and [6], pp. 101–102. Consider 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 $Y(u)$ for $u \in W$ as a linear combination of the sample variables $Y(u_i)$ with unknown weights $\lambda_i = \lambda_i(u)$ by

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

The estimation involves only those sample random variables $Y(u_i)$ that are positioned in the "neighborhood" $A(u)$ of u, i.e. if $u_i \in A(u)$. 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; see Sect. 4.

Unbiasedness of the estimator introduced in (4) and minimizing its variance lead to the following conditions on the weights λ_i. For all $i = 1, \ldots, n$ with $u_i \in A(u)$ it holds

$$\begin{aligned} &\sum_{j=1}^{n} \lambda_j \gamma(u_j - u_i) \mathbf{1}\{u_j \in A(u)\} + \mu = \gamma(u - u_i) \; , \\ &\sum_{j=1}^{n} \lambda_j \mathbf{1}\{u_j \in A(u)\} = 1 \; . \end{aligned} \tag{5}$$

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 $y(u_1), \ldots, y(u_n)$.

3.2 Variograms

In this paper, the most simple and popular variogram estimator of Matheron is used (cf. [3], [6]). It is defined by

$$\widehat{\gamma}(h) = \frac{1}{2N(h)} \sum_{i,j:u_i - u_j \approx h} \left(Y(u_i) - Y(u_j)\right)^2 \tag{6}$$

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 $i, j = 1, \ldots, n$.

As shown in Fig. 4, the traffic data lead to empirical variograms that are clearly zonally anisotropic. Below, we consider zonally anisotropic variogram models constructed from isotropic ones (cf. [4], [6]). Put

$$\gamma(h) = \gamma_1(h) + \gamma_2(h) \; , \tag{7}$$

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

$$\gamma_2(h) = b_2(1 - e^{-\sqrt{h^\top Ch}/c_2}) \tag{8}$$

is a geometrically anisotropic exponential variogram model with sill $b_2 > 0$ and a further parameter $c_2 > 0$. For a vector $h = (h_1, h_2)$, the quadratic form

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

depends on two scale parameters λ_1, λ_2 and a rotation parameter $\alpha \in [0, 2\pi)$.

Let $\widehat{\gamma}(h)$ be an empirical variogram estimated from the observed data $\{y(u_i)\}$ for the field Y and $\gamma_\beta(h)$ the theoretical parametric variogram model considered in (7) with parameter vector $\beta = (a_1, b_1, c_1, b_2, \lambda_1/c_2^2, \lambda_2/c_2^2, \alpha)$. In practice, only a finite number m of values $\widehat{\gamma}(h_1), \ldots, \widehat{\gamma}(h_m)$ can be computed. In the case of traffic data, the classical least squares method is employed to fit γ_β to $\widehat{\gamma}$. Since traffic data is substantially anisotropic, the variogram model (7) has to be fitted to the data on the whole grid as well as in two directions with polar angles α and $\alpha + \pi/2$.

3.3 Drift estimation

The mean field $\{m(u)\}$ can be estimated from the data by various methods ranging from radial extrapolation to smoothing techniques such as moving average and edge preserving smoothing. 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

$$\widehat{m}(u) = \frac{1}{N_u} \sum_{u_i \in W(u)} X(u_i) \tag{9}$$

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

3.4 Residuals formed with estimated drift

In the previous sections, we supposed that the drift $m(u)$ is explicitly known. However, if it has to be estimated from the data, the theoretical background for the application of the kriging method breaks down (cf. [3], pp. 122–125, [4] p. 72, [6], p. 214). Nevertheless, 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.

4 Extrapolation of the velocity field

In what follows, the extrapolation method of Sect. 3 is applied to real traffic data of the directional sector $S_2 = \{\alpha : \pi/2 \leq \alpha < \pi\}$. This partial data set contains 19699 entries of taxis moving northwest collected over 90 days.

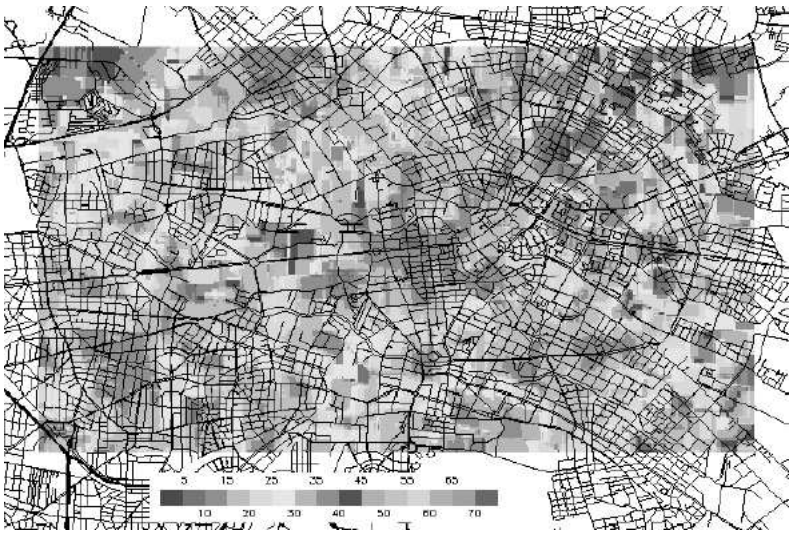


Fig. 3. Mean field $\widehat{m}(u)$ of data set 2

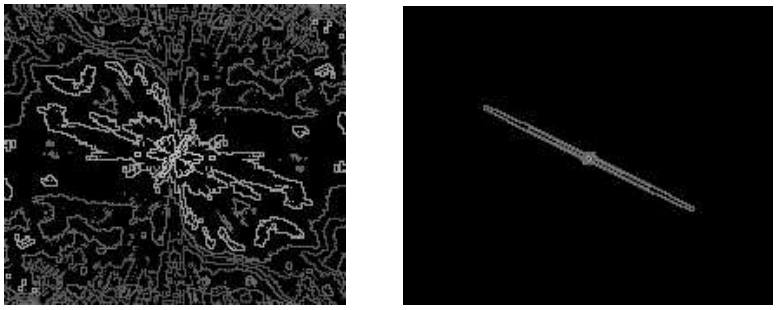


Fig. 4. Empirical variogram $\widehat{\gamma}^*(h)$ and fitted variogram model $\gamma^*(h)$ (level curves)

In Fig. 3, the northwest movement direction of the taxis can be clearly recognized in the mean velocity field $\{\widehat{m}(u)\}$. Grey tones reflect speed variation. It clearly shows that the estimator $\widehat{m}$ preserves the spatial velocity structure. To estimate the variogram γ^* of Y^*, the empirical variogram $\widehat{\gamma}_i^*$ is computed for each day $i = 1, \ldots, 90$ and then averaged over all days, i.e. $\widehat{\gamma}^*(h) = \big(\sum_{i=1}^{90} \widehat{\gamma}_i^*(h)\big)/90$. The empirical variogram $\widehat{\gamma}^*(h)$ with "maximum range" in northwest direction and "minimum range" in orthogonal direction is zonally anisotropic; see Fig. 4. The main directions of anisotropy are closely connected to the road directions in downtown Berlin. See Sect. 5 and especially Fig. 10(a) for details.

The zonally anisotropic variogram model (7) with two fixed parameters $\alpha = 170°$, $\lambda_1/c_2^2 = 1000$ has been fitted to the empirical one. The classical least squares fitting method applied to one-dimensional vertical slices of the

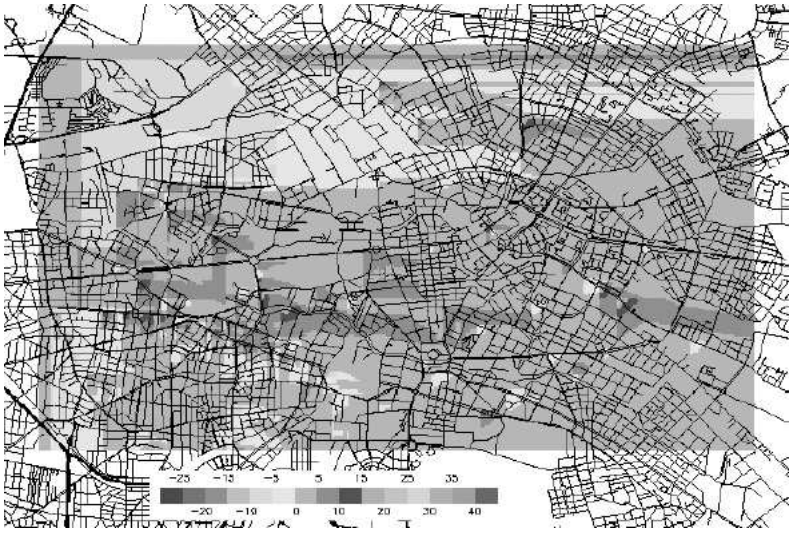


Fig. 5. Residual field $\widehat{Y}^*(u)$

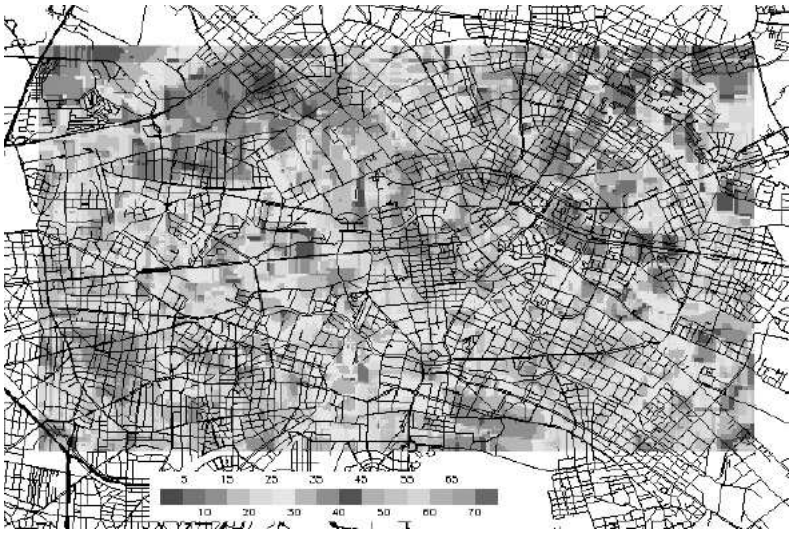


Fig. 6. Velocity field $\widehat{X}(u)$

empirical variogram in orthogonal directions $\alpha = 80°$ and $\alpha = 170°$ yields the remaining parameter values $a_1 = 31.772$, $b_1 = 116.211$, $c_1 = 245388.671$, $b_2 = 22.634$, $\lambda_2/c_2^2 = 683964.794$. Thus, the range values in directions 170° and 80° are $r_1 = 270$ m and $r_2 = 162$ m, respectively. It means that the velocities of two vehicles separated by distances $3r_1 = 810$ m in horizontal direction and $3r_2 = 486$ m in vertical direction are almost independent. These

Fig. 7. Traffic jams: $\widehat{X}(u) \leq 15$ kph

range values are conform with the results stated in Sect. 5 for the typical distance between two subsequent positions of the same test vehicle.

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 by kriging with moving neighborhood (4) using the 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 at locations u_i lying "ahead" of the current position u can influence its velocity value.

The extrapolated residuals $\widehat{Y}^*(u)$ and the resulting velocity map $\{\widehat{X}(u)\}$ are shown in Figs. 5 and 6, respectively. Due to the particular asymmetric form of the indicators, the extrapolated field of residuals is strongly discontinuous. Discontinuities of the realizations of $\{\widehat{X}(u)\}$ caused by the kriging with moving neighborhood are essential for precise localization of traffic–jam areas. In Fig. 5, most of the deviation values are zero. The routes of taxis driving in the streets are marked by peaks of the field $\widehat{Y}^*$ with subsequent tails of non–zero residual velocity values lying behind. Thus, one can distinguish separate routes of different test vehicles. See also the extracted taxi routes in Fig. 8, which are similar to those shown in Fig. 5.

In Fig. 7, areas with velocities $\widehat{X}(u) \leq 15$ kph are marked grey. Some of these regions might be caused by traffic jams, others are regions with low average velocities. Indeed, the most likely velocity value in downtown Berlin is about 20 kph as it can be seen in Fig. 12.

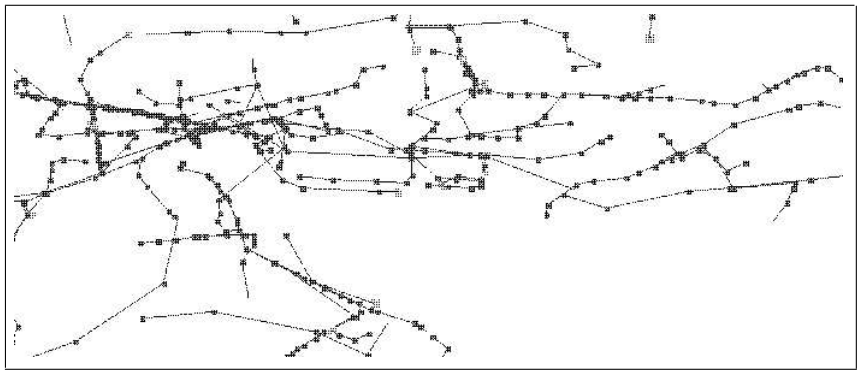


Fig. 8. Taxi routes on 18.02.2002, from 5.15 to 5.30 pm

5 Statistical analysis of traffic data

In addition to the spatial statistical inference performed above, we now discuss the histograms of velocity residuals and further traffic characteristics which bring an extra insight into the structure of traffic data. In particular, they help us to explain some features of anisotropy and spatial correlation which we already mentioned in Sect. 4. For a more detailed treatment of the subject, see [2].

5.1 Distributional properties of polygonal taxi routes

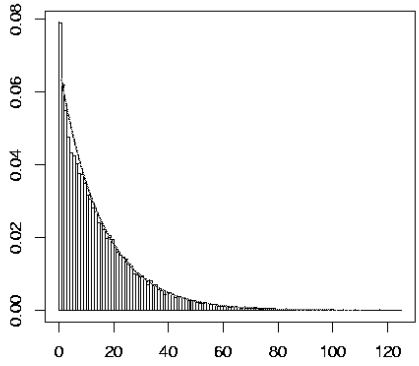
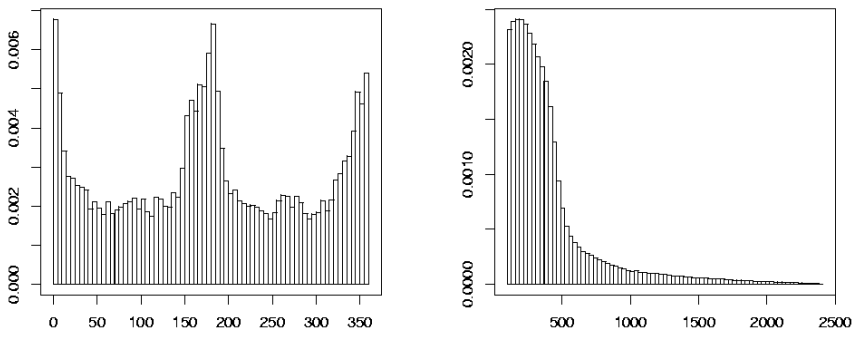


Fig. 9. Histogram of the number of segments in the taxi routes

If we think about the way the traffic data are collected we understand that the locations where the velocities are measured can not be deterministic. Moreover, they are stochastically dependent. In fact, each test vehicle follows a route that consists of a random number of segments. Each segment connects two locations where consecutive GPS signals were sent; see Fig. 8. The histogramm of the number of segments in the taxi routes is shown in Fig. 9. It turns out that this histogram can be well approximated by a geometric distribution with parameter $p = 0.9365064$ being the probability of enlarging a route by a new segment.

Furthermore, the geometry of the taxi routes explains the form of the variogram anisotropy mentioned in Sect. 4. In particular, the distribution of the angles between the movement direction of a vehicle and the eastward direction in Fig. 10(a) reflects the distribution of typical street directions with heavy traffic in downtown Berlin. The majority of main roads goes east or west which corresponds to the angles of 0°, 180°, and 360°, respectively. This is certainly the reason for the character of zonal anisotropy of the variograms in Fig. 4. Fig. 10(b) shows that the distribution of segment lengths is demonstrably



(a) Directions of route segments (b) Lengths of route segments

Fig. 10. Histograms of segment directions (in degrees) and lengths (in m)

non-normal. Furthermore, with probability of ca. 0.9, the distances between two subsequent GPS signals in the taxi routes do not exceed 1000 m. It is clear that the velocities at two positions within this distance are correlated. The opposite statement is also true. As it has been already mentioned in Sect. 4, the velocities of two cars at a distance of more than $3\sqrt{r_1^2 + r_2^2} \approx 945$ m from each other are almost independent.

5.2 Distribution of velocity residuals

The histogram in Fig. 11 shows that the distribution of velocity residuals can be well fitted by some normal distribution. Nevertheless, a more detailed statistical inference shows that the distribution of velocity residuals depends on the value of mean velocity. One reason for this is that the sum of the residual and the mean has to be non–negative. Figs. 13, 14 and 15 show the histograms

of the velocity residuals measured at locations with mean velocities (in kph) belonging to three disjoint classes: $[15, 20)$, $[25, 30)$ and $[40, 45)$, respectively. If

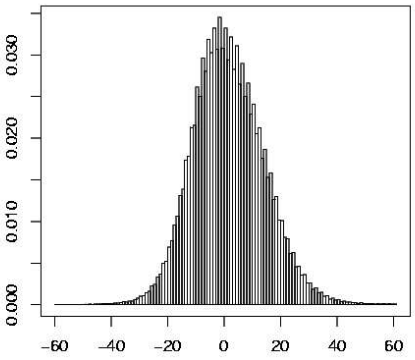
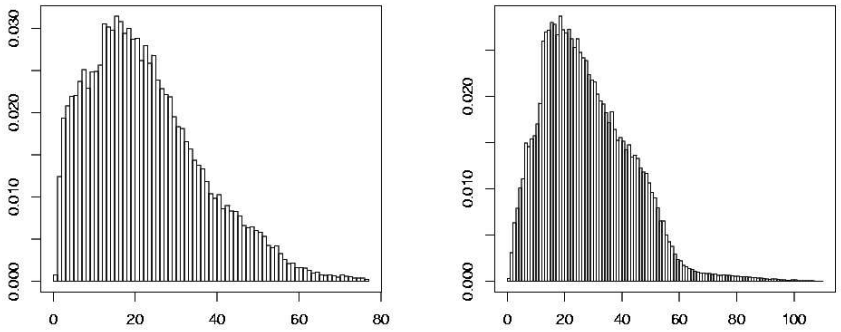


Fig. 11. Histogram of the velocity residuals (in kph)



(a) for the first route segments (b) for the remaining route segments

Fig. 12. Histogram of velocities (in kph)

we add the mean velocity values to their residuals we see that most velocities in downtown Berlin do not exceed 60 kph. The histogram of the velocities themselves is given in Fig. 12 which shows that the most likely velocity value

in downtown Berlin (i.e., the modus of the empirical velocity distribution) is about 20 kph. This explains the dominance of low velocity values in the mean field $\widehat{m}$ and the threshold maps in Figs. 3 and 7.

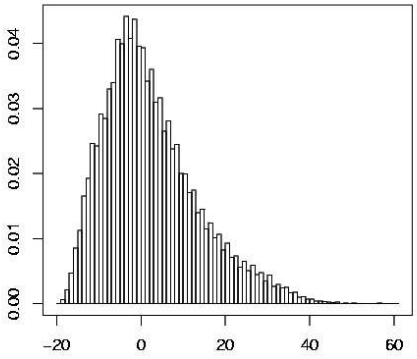


Fig. 13. Histogram of residuals given that $\widehat{m} \in [15, 20)$

The right skewness of this histogram means that large positive deviations from small mean values are more likely than negative ones.

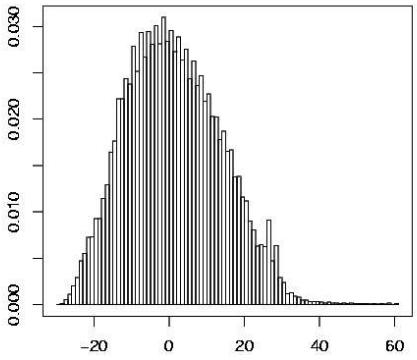


Fig. 14. Histogram of residuals given that $\widehat{m} \in [25, 30)$

This histogram is almost symmetric. Thus, both positive and negative residuals of equal size occur nearly with the same probability.

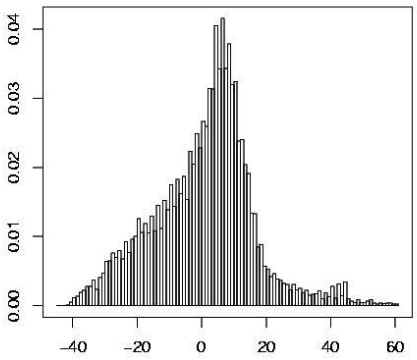


Fig. 15. Histogram of residuals given that $\widehat{m} \in [40, 45)$

It is clear from this histogram that at positions with large mean values large negative residuals are more likely than positive ones.

6 Outlook

The spatial extrapolation and statistical space–time analysis of traffic data considered in the present paper is an important step towards stochastic modelling, simulation and prediction of future road–traffic states. Our results can be used to construct a Markov–type simulator by means of which future routes of test vehicles can be generated, where the choice of the starting configuration depends on the actually measured traffic situation. In particular, when sampling the velocity residuals from histograms as given in Figs. 13, 14 and 15, the mean velocity field $\{\widehat{m}(u)\}$ will be actualized by the recently, say at the given day, observed velocities. For example, suppose that significantly larger velocities than usually have been observed at the considered day in a certain neighborhood of location u. In this case, the velocity residual at location u will be sampled from a histogram which corresponds to a larger class of mean velocities than the "historical" value $\widehat{m}(u)$. Further details concerning our simulation algorithms can be found in [2].

Then, using the extrapolation technique described in Sects. 2 and 3, velocity maps based on both the measured and simulated traffic data can be computed. To evaluate the quality of these maps, they are compared with corresponding velocity maps computed exclusively from measured traffic data. The comparison is based on morphological distance measures for digital image data. These issues will be discussed in a forthcoming paper.

Acknowledgement

This research was supported by the German Aerospace Center (DLR) through research grant 931/69175067. The authors are grateful to Reinhart Kühne, Peter Wagner and their co-workers from the DLR Institute of Transport Research for suggesting the problem as well as for stimulating and fruitful discussions on the subject.

References

1. Braxmeier H, Schmidt V, Spodarev E (2004) Spatial extrapolation of anisotropic road traffic data. *Image Analysis and Stereology* 23 (to appear)
2. Braxmeier H, Spodarev E, Schmidt V, Schweiggert F, Wagner P (2004) Statistical analysis and space–time simulation of city–road traffic. *Preprint*
3. Chilès JP, Delfiner P (1999) *Geostatistics: modelling spatial uncertainty.* Wiley, New York
4. Cressie NAC (1993) *Statistics for spatial data.* Wiley, New York
5. GeoStoch (2004) Java library. University of Ulm, Department of Applied Information Processing and Department of Stochastics, http://www.geostoch.de
6. Wackernagel H (1998) *Multivariate geostatistics.* 2nd ed., Springer, Berlin