Dargatz, Georgescu, Held:

Stochastic modelling of the spatial spread of influenza in Germany

Sonderforschungsbereich 386, Paper 450 (2005)

Online unter: http://epub.ub.uni-muenchen.de/

Projektpartner

Stochastic Modelling of the Spatial Spread of Influenza in Germany

Christiane Dargatz, Vera Georgescu, Leonhard Held
Ludwig-Maximilians-University, Munich

September 2, 2005

Abstract: In geographical epidemiology, disease counts are typically available in discrete spatial units and at discrete time-points. For example, surveillance data on infectious diseases usually consists of weekly counts of new infections in pre-defined geographical areas. Similarly, but on a different time-scale, cancer registries typically report yearly incidence or mortality counts in administrative regions.
A major methodological challenge lies in building realistic models for space-time interactions on discrete irregular spatial graphs. In this paper, we will discuss an observation-driven approach, where past observed counts in neighbouring areas enter directly as explanatory variables, in contrast to the parameter-driven approach through latent Gaussian Markov random fields (Rue & Held, 2005) with spatio-temporal structure. The main focus will lie on the demonstration of the spread of influenza in Germany, obtained through the design and simulation of a spatial extension of the classical SIR model (Hufnagel et al., 2004).

Zusammenfassung: In der räumlichen Epidemiologie liegen Fallzahlen typischerweise für diskrete Gebiete und diskrete Zeitpunkte vor. Bei der Erfassung infektiöser Krankheiten beispielsweise zählt man die wöchentlichen Inzidenzen in vorgegebenen Regionen. Ähnlich, aber auf einer anderen Zeitskala, werden Krebsfälle jährlich registriert.
Eine Herausforderung liegt darin, eine Methode zur realistischen Modellierung von räumlich-zeitlichen Zusammenhängen auf diskreten, unregelmäßigen räumlichen Graphen zu entwickeln. In diesem Artikel beschäftigen wir uns mit einem Ansatz, der bereits erfasste Fälle in angrenzenden Gebieten direkt als erklärende Variablen einbezieht, im Gegensatz zur Modellierung durch latente Gauß-Markov-Zufallsfelder (Rue & Held, 2005) mit räumlich-zeitlicher Struktur. Dazu stellen wir die Ausbreitung von Influenza in Deutschland mittels einer räumlichen Erweiterung des klassichen SIR-Modells (Hufnagel et al., 2004) in Computersimulationen nach.

Keywords: Space-time interaction; Gaussian Markov random fields; Epidemic modelling; Stochastic differential equations; Global SIR model; Influenza.

1 Introduction

There has been much recent interest in space-time models for disease counts collected in discrete spatial units and discrete time points. While most of the work has mainly focused on non-infectious diseases, in particular on cancer, recently models for infectious disease data have been developed. For non-infectious diseases, hierarchical Bayesian approaches have been proposed, where latent parameters follow Gaussian Markov random field (GMRF) models (Waller et al., 1997, Knorr-Held & Besag, 1998, Knorr-Held, 2000a, Lagazio et al., 2001, Lagazio et al., 2003, Schmid & Held, 2004). Common to these models is the assumption that the observed counts are conditionally independent, given the latent parameters.

However, the allowance for realistic space-time interaction in GMRFs is non-trivial, one approach that dates back to Clayton (1996) is to use Kronecker product structures (see also Rue & Held, 2005, Sec-

tion 3.4.3) for interaction parameters while keeping main effects for overall spatial and temporal trends.

In this paper we will focus on a different modelling strategy, where past counts enter explicitly in the disease rate and hence the conditional independence assumption is lost. This class of models, called *observation-driven* (Cox, 1981), is motivated by the fact that *parameter-driven* models, such as the GMRF models mentioned above, are not able to capture the epidemic trends observable in data on infectious diseases. Indeed, epidemic models have used such *observation-driven* models for decades; in particular the class of SIR models (susceptible–infected–removed) has been extensively studied. However, this has been done mainly in a purely temporal and simplistic context, ignoring the fact that global epidemics spread in a spatio-temporal fashion.

A recent approach described in Hufnagel et al. (2004) fills this gap, proposing a spatio-temporal model on two scales (local and global) to describe the spread of the SARS epidemic based on stochastic differential equation models. Watts et al. (2005) developed a metapopulation model which incorporates mixing even at multiple scales. We adopt and extend the model of Hufnagel et al. (2004) and use it to investigate if it is able to describe an influenza epidemic in Germany 2005.

A major requirement in SIR models is knowledge of the number of susceptibles. In surveillance applications, often the whole population is considered as susceptible, due to the lack of available data (e.g. Knorr-Held & Richardson, 2003). An alternative approach is to use a branching process model as an approximation to the SIR model. This class of models has the advantage that it does not require knowledge of the number of susceptibles, however, some form of stationarity is needed to ensure that the stochastic process, describing the number of counts at each time point, does not explode to infinity. Held, Höhle, and Hofmann (2005) have used an extended version of this model in a series of applications from surveillance data. In particular, they showed that maximum likelihood estimation is straightforward and extended the model to the space-time domain using a multivariate branching process formulation. However, the application of this class of model to highly infectious diseases such as influenza is perhaps not suitable, due to the underlying assumption of stationarity.

In the next section, we will start with the classical SIR model and then describe the approach by Hufnagel et al. (2004) to model the spatio-temporal spread of infectious diseases. In Section 3, we develop an algorithm for simulating this model. A central feature of the formulation is that the dispersal of infected cases in space is not necessarily solely local but also global, if necessary. For example, infected cases might travel through air traffic large distances in a small amount of time. Based on data on air and train traffic in Germany, we define such a dispersal rate matrix for administrative regions in Germany and investigate whether simulations from such a model show similar patterns as an influenza epidemic in Germany in 2005. Model parameters are chosen based on external knowledge.

2 From the Standard Deterministic to a Global Stochastic SIR Model

2.1 Standard SIR Model

In the SIR model, we divide a population into three categories: Those who are susceptible to the disease (S), those who are infected and infectious (I), and those who are removed from the system because they are recovered and immune, or quarantined, or dead (R). With s, j, and r we denote the fractions of susceptible, infectious, and removed individuals of the total population N. Transitions from one category

to another happen according to

$$S + I \xrightarrow{\alpha} 2I, \quad I \xrightarrow{\beta} R,$$

where α is the rate of an individual's contacts per day which are sufficient to spread the disease, and β^{-1} is the average infectious period. The infection dynamics in the standard deterministic SIR model is given by the set of differential equations

$$ds/dt \;=\; -\alpha sj, \quad dj/dt \;=\; \alpha sj - \beta j. \tag{1}$$

Hence, while recovery follows a linear process, infections occur on high rate only when both the numbers of susceptibles *and* infectives are sufficiently large. Since we assume a closed population, i.e. ignoring births, non-related deaths, and migration during the relatively short duration of an influenza epidemic, we expect the size of the population to be constant. The fraction of recovered individuals thus reads $r = 1 - s - j$. The ratio $\rho = \alpha/\beta$ is called the basic reproduction number and states a decisive parameter for the course of the epidemic: When ρ^{-1} is greater than the initial fraction of susceptibles s_0, no epidemic will develop. Otherwise, the epidemic will fall off as soon as the decreasing function $s(t)$ drops below ρ^{-1}.

In case of influenza, an infected individual aquires immunity to the strain he was affected by and can hence not become susceptible during the same wave of flu again. Therefore, there is no need for a transition from state R back to S. However, there are steadily new antigen mutants of the influenza virus coming up, which is why at the beginning of the next epidemic the whole population will be susceptible again.

2.2 Stochastic SIR Model

Bearing in mind that the infection and recovery processes are of rather stochastic than deterministic character, we write (1) in terms of stochastic Langevin equations:

$$\begin{aligned}
\frac{ds}{dt} &= -\alpha sj + \frac{1}{\sqrt{N}}\sqrt{\alpha sj}\;\xi_1(t) \\
\frac{dj}{dt} &= \alpha sj - \beta j - \frac{1}{\sqrt{N}}\sqrt{\alpha sj}\;\xi_1(t) + \frac{1}{\sqrt{N}}\sqrt{\beta j}\;\xi_2(t),
\end{aligned}$$

where $\xi_1(t)$ and $\xi_2(t)$ are independent Gaussian white noise forces, modelling fluctuations in transmission and recovery matters. These are of particular importance during the initial phase when the number of infected individuals is relatively small.

2.3 Excursus: SLIR Model

It is possible to also incorporate a latent status in our considerations, which yields the following transitions, the so-called SLIR model:

$$S + I \xrightarrow{\alpha} L + I, \quad L \xrightarrow{\varepsilon} I, \quad I \xrightarrow{\beta} R,$$

where ε^{-1} is the average latent period. Let l denote the fraction of latent individuals of the total population. The differential equations then read

$$ds/dt \;=\; -\alpha sj, \quad dl/dt \;=\; \alpha sj - \varepsilon l, \quad dj/dt \;=\; \varepsilon l - \beta j$$

in the deterministic case and

$$\begin{aligned}
\frac{ds}{dt} &= -\alpha sj + \frac{1}{\sqrt{N}}\sqrt{\alpha sj}\,\xi_1(t) \\
\frac{dl}{dt} &= \alpha sj - \varepsilon l - \frac{1}{\sqrt{N}}\sqrt{\alpha sj}\,\xi_1(t) + \frac{1}{\sqrt{N}}\sqrt{\varepsilon l}\,\xi_3(t) \\
\frac{dj}{dt} &= \varepsilon l - \beta j - \frac{1}{\sqrt{N}}\sqrt{\varepsilon l}\,\xi_3(t) + \frac{1}{\sqrt{N}}\sqrt{\beta j}\,\xi_2(t)
\end{aligned}$$

in the stochastic model, where $\xi_3(t)$ accounts for noise in the duration of the latent period.

Since our objective is the modelling of the spread of influenza, where an individual can normally pass on the virus from the moment of infection, we from now on suppress the consideration of latency. Nevertheless, the following observations can easily be adjusted to the SLIR model (cf. supporting material at `http://www.statistik.lmu.de/~dargatz/publications`).

2.4 Global SIR Model

So far, our model describes the spread of a disease in a single closed population under the assumption of homogeneous mixing. But this condition applies only as long as individuals cover relatively short distances—an assumption that is not given in our fully connected world anymore, even if we restrict our focus to a comparatively small area like Germany. As suggested in Hufnagel et al. (2004), we introduce a network of subregions $1, \ldots, n$ of the primarily observed area, each region i having a population size N_i being composed of S_i, I_i, and R_i susceptible, infectious and removed individuals. Whilst the local infection dynamics within a subregion is given by the stochastic SIR model as introduced above, the global dispersal between the knots of the network is rated in a connectivity matrix $\gamma = (\gamma_{ij})_{ij}$:

$$S_i + I_i \xrightarrow{\alpha} 2I_i, \quad I_i \xrightarrow{\beta} R_i, \quad S_i \xrightarrow{\gamma_{ij}} S_j, \quad I_i \xrightarrow{\gamma_{ij}} I_j.$$

The system of stochastic differential equations now changes to

$$\begin{aligned}
\frac{ds_i}{dt} &= -\alpha s_i j_i - \sum_k \gamma_{ik} s_i + \sum_k \gamma_{ki} s_k + \frac{1}{\sqrt{N_i}}\sqrt{\alpha s_i j_i}\,\xi_1^{(i)}(t) \\
&\quad + \frac{1}{\sqrt{N_i}}\sqrt{\sum_k \gamma_{ik} s_i}\,\xi_4^{(i)}(t) - \frac{1}{\sqrt{N_i}}\sqrt{\sum_k \gamma_{ki} s_k}\,\xi_5^{(i)}(t) \\
\frac{dj_i}{dt} &= \alpha s_i j_i - \beta j_i - \sum_k \gamma_{ik} j_i + \sum_k \gamma_{ki} j_k - \frac{1}{\sqrt{N_i}}\sqrt{\alpha s_i j_i}\,\xi_1^{(i)}(t) + \frac{1}{\sqrt{N_i}}\sqrt{\beta j_i}\,\xi_2^{(i)}(t) \qquad (2) \\
&\quad + \frac{1}{\sqrt{N_i}}\sqrt{\sum_k \gamma_{ik} j_i}\,\xi_4^{(i)}(t) - \frac{1}{\sqrt{N_i}}\sqrt{\sum_k \gamma_{ki} j_k}\,\xi_5^{(i)}(t) \\
\frac{dr_i}{dt} &= \beta j_i - \frac{1}{\sqrt{N_i}}\sqrt{\beta j_i}\,\xi_2^{(i)}(t).
\end{aligned}$$

for $i = 1, \ldots, n$. Here, $\boldsymbol{\xi_1(t)} = (\xi_1^{(1)}(t), \ldots, \xi_1^{(n)}(t))$, $\boldsymbol{\xi_2(t)}$, $\boldsymbol{\xi_4(t)}$, and $\boldsymbol{\xi_5(t)}$ are independent vector-valued white noise forces which stand for fluctuations in transmission, recovery, and outbound and inbound traffic, respectively.

Since in the global model the single populations are *not* closed anymore due to migration, the property $s_i + j_i + r_i = 1$, $i = 1, \ldots, n$, does not necessarily hold. Instead, s_i, j_i, and r_i indicate the fractions of susceptible, infectious and removed individuals as measured by the original population N_i. That is why in (2) we also declared the formula for r_i.

3 Implementation

3.1 Keeping the System Closed

Let us focus on the Gaussian white noises $\boldsymbol{\xi_j}$. The components of $\boldsymbol{\xi_1(t)}$ and $\boldsymbol{\xi_2(t)}$ (and also of $\boldsymbol{\xi_3(t)}$) are all stochastically independent of each other, but we have to introduce a weak form of dependence to the components of $\boldsymbol{\xi_4(t)}$ and $\boldsymbol{\xi_5(t)}$ due to the following: Since we assume the area of our n regions to be closed, we have to require

$$\sum_{i=1}^{n} \left(\frac{ds_i}{dt} + \frac{dj_i}{dt} + \frac{dr_i}{dt} \right) = 0 \,.$$

The left hand side of this equation reads

$$\sum_i \left(-\sum_k \gamma_{ik} s_i + \sum_k \gamma_{ki} s_k \right) + \sum_i \left(-\sum_k \gamma_{ik} j_i + \sum_k \gamma_{ki} j_k \right) \tag{3}$$

$$+ \sum_i \frac{1}{\sqrt{N_i}} \left(\sqrt{\sum_k \gamma_{ik} s_i} + \sqrt{\sum_k \gamma_{ik} j_i} \right) \xi_4^{(i)}(t) \tag{4}$$

$$+ \sum_i \frac{1}{\sqrt{N_i}} \left(\sqrt{\sum_k \gamma_{ki} s_k} + \sqrt{\sum_k \gamma_{ki} j_k} \right) \xi_5^{(i)}(t) \,. \tag{5}$$

Obviously, the two sums over i in (3) both equal 0. In order to also let rows (4) and (5) disappear, we correlate the components of $\boldsymbol{\xi_4(t)}$ and those of $\boldsymbol{\xi_5(t)}$ among each other such that equality with zero holds almost surely.

For the components of $\boldsymbol{\xi_4}$, we proceed as follows (see Knorr-Held, 2000b): Define

$$x_i(t) := \frac{1}{\sqrt{N_i}} \left(\sqrt{\sum_k \gamma_{ik} s_i(t)} + \sqrt{\sum_k \gamma_{ik} j_i(t)} \right) .$$

We hence seek

$$\sum_{i=1}^{n} x_i(t)\, \xi_4^{(i)}(t) = 0 \quad \text{a.s. for all } t. \tag{6}$$

Define the $n \times n$-matrices

$$\mathbf{M} := \mathbf{I_n} - \frac{1}{n} \mathbf{1_n 1_n'} = \begin{pmatrix} \frac{n-1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} \\ -\frac{1}{n} & \frac{n-1}{n} & \cdots & -\frac{1}{n} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{n} & -\frac{1}{n} & \cdots & \frac{n-1}{n} \end{pmatrix} ,$$

where $\mathbf{I_n} \in \mathbb{R}^{n\times n}$ denotes the identity matrix and $\mathbf{1_n} = (1,1,\ldots,1)' \in \mathbb{R}^{n\times 1}$. Furthermore,

$$\mathbf{\Sigma(t)} := \operatorname{diag}\big(x_1^2(t),\ldots,x_n^2(t)\big)$$

and

$$\mathbf{Q(t)} := \mathbf{M}\cdot\mathbf{\Sigma(t)}\cdot\mathbf{M} = \big(q_{ij}(t)\big)_{ij}$$

with

$$q_{ii} = \left(\frac{1}{n^2}\sum_{k\neq i} x_k^2(t) + \Big(\frac{n-1}{n}\Big)^2 x_i^2(t)\right),$$

$$q_{ij} = \left(\frac{1}{n^2}\sum_{k\neq i,j} x_k^2(t) - \frac{n-1}{n^2}\left(x_i^2(t)+x_j^2(t)\right)\right) \qquad \text{for } i\neq j,$$

and let

$$\mathbf{u(t)} := \big(x_1(t)\,\xi_4^{(1)}(t),\ldots,x_n(t)\,\xi_4^{(n)}(t)\big)' \sim N\big(\mathbf{0},\mathbf{Q(t)}\big)\,, \tag{7}$$

i.e. $\mathbf{Q(t)}$ is the covariance matrix of $\mathbf{u(t)}$. Then, as required,

$$\text{var}\left(x_i(t)\,\xi_4^{(i)}(t)\right) = q_{ii}(t) \approx x_i^2(t) \quad \text{for } n \text{ large and } i = 1,\ldots,n$$

($x_i(t)$ remains constant for t fixed, hence var $(x_i(t)\,\xi_4^{(i)}(t)) \stackrel{!}{=} x_i^2(t)$). Moreover,

$$\mathbb{E}\left(\sum_{i=1}^n x_i(t)\,\xi_4^{(i)}(t)\right) = 0 \qquad \text{and} \qquad \text{var}\left(\sum_{i=1}^n x_i(t)\,\xi_4^{(i)}(t)\right) = \mathbf{1_n'Q(t)1_n} = 0,$$

yielding (6). Unfortunately, the desired property $\sum_{i,j} q_{ij}(t) = 0$ yields the drawback that $\mathbf{Q(t)}$ is not positive definite and hence unsuitable as covariance matrix. Instead of $\mathbf{u(t)}$, we hence consider a linear transformation $\mathbf{Lu(t)}$ with

$$\mathbf{L} := \begin{pmatrix} \mathbf{I_{n-1}} & \mathbf{-1_{n-1}} \\ \mathbf{1'_{n-1}} & 1 \end{pmatrix} \in \mathbb{R}^{n\times n}\,,$$

whose first $n-1$ components have dispersion

$$\mathbf{P(t)} := \operatorname{diag}\big(x_1^2(t),\ldots,x_{n-1}^2(t)\big) + x_n^2(t)\mathbf{1_{n-1}1'_{n-1}} \in \mathbb{R}^{(n-1)\times(n-1)}.$$

Draw $\boldsymbol{\pi}\mathbf{(t)} = (\pi_1(t),\ldots,\pi_{n-1}(t),0)'$ with

$$\big(\pi_1(t),\ldots,\pi_{n-1}(t)\big)' \sim N\big(\mathbf{0},\mathbf{P(t)}\big)$$

and retransform $\mathbf{u(t)} = (u_1(t),\ldots,u_n(t)) = \mathbf{M}\boldsymbol{\pi}\mathbf{(t)}$. We obtain

$$\xi_4^{(i)}(t) = \frac{u_i(t)}{x_i(t)} \quad , i = 1,\ldots,n.$$

Note that, for any i, we have $x_i(t) > 0$ as long as $r_i(t) < 1$, since for all $i \in \{1,\ldots,n\}$ there is a $k \in \{1,\ldots,n\}$ with $\gamma_{ik} > 0$ (i.e. each district is directly connected to at least one other). However, if $x_i(t) = 0$, the value of $\xi_4^{(i)}(t)$ does not matter since in (4) it will be multiplied by $x_i(t)$.

Obtain $\boldsymbol{\xi_5}$ in the same way as $\boldsymbol{\xi_4}$, replacing $x_i(t)$ by

$$y_i(t) := \frac{1}{\sqrt{N_i}}\left(\sqrt{\sum_k \gamma_{ik}s_k(t)} + \sqrt{\sum_k \gamma_{ik}j_k(t)}\right).$$

3.2 Numerical Scheme

Given initial conditions $s_i(0)$, $j_i(0)$, and $r_i(0)$, $i = 1, \ldots, n$, as well as fixed values for the transmission rate α and the reciprocal average infectious period β, we simulate the epidemic process at discrete, equidistant instants in the time domain $[0, t_{\text{max}}]$. Let δ be the (suitably small) time step. For the approximation of the differential equations (2), we apply the Euler-Maruyama approximation scheme

$$\begin{aligned} s_i\big(m\delta\big) &= s_i\big((m-1)\delta\big) + \big[ds_i(t)/dt\big]_{t=(m-1)\delta} \\ j_i\big(m\delta\big) &= j_i\big((m-1)\delta\big) + \big[dj_i(t)/dt\big]_{t=(m-1)\delta} \\ r_i\big(m\delta\big) &= r_i\big((m-1)\delta\big) + \big[dr_i(t)/dt\big]_{t=(m-1)\delta} \end{aligned} \tag{8}$$

for $m \geq 1$ and $i = 1, \ldots, n$ (Kloeden & Platen, 1999).

3.3 Algorithm

After having fixed the parameters α, β, and γ, the time step δ and initial values for s_i, j_i, and r_i, $i = 1, \ldots, n$, the proceeding for each instant of time now reads as follows ($m = 0, \ldots, (t_{\text{max}}/\delta) - 1$):

1. For $i = 1, \ldots, n$, calculate

   $$\mu_i := \alpha\, s_i(m\delta)\, j_i(m\delta), \quad \nu_i := \beta\, j_i(m\delta),$$

   and

   $$\eta_i := \sum_{k=1}^{n} \gamma_{ik}\, s_i(m\delta), \quad \zeta_i := \sum_{k=1}^{n} \gamma_{ki}\, s_k(m\delta), \quad \rho_i := \sum_{k=1}^{n} \gamma_{ik}\, j_i(m\delta), \quad \tau_i := \sum_{k=1}^{n} \gamma_{ki}\, j_k(m\delta).$$

2. For $i = 1, \ldots, n$, compute $x_i = m_i(\sqrt{\eta_i} + \sqrt{\rho_i}\,)$ and $y_i = m_i(\sqrt{\zeta_i} + \sqrt{\tau_i}\,)$, where $m_i := \sqrt{N_i}^{-1}$.

3. Set $\mathbf{P_4} = \text{diag}\big(x_1^2, \ldots, x_{n-1}^2\big) + x_n^2 \mathbf{1_{n-1}} \mathbf{1'_{n-1}}$ and $\mathbf{P_5} = \text{diag}\big(y_1^2, \ldots, y_{n-1}^2\big) + y_n^2 \mathbf{1_{n-1}} \mathbf{1'_{n-1}}$.

4. Generate $\boldsymbol{\pi}^{(\mathbf{j})} = \big(\pi_1^{(j)}, \ldots, \pi_n^{(j)}\big)$, $j = 4, 5$, with $\big(\pi_1^{(j)}, \ldots, \pi_{n-1}^{(j)}\big) \sim N(\mathbf{0}, \mathbf{P_j})$ and $\pi_n^{(j)} = 0$.

5. Compute $\mathbf{u} = (u_1, \ldots, u_n) = \mathbf{M}\boldsymbol{\pi_4}$ and $\mathbf{v} = (v_1, \ldots, v_n) = \mathbf{M}\boldsymbol{\pi_5}$ with $\mathbf{M} = \mathbf{I_n} - n^{-1}\mathbf{1_n}\mathbf{1'_n}$.

6. Evaluate $\boldsymbol{\xi_1}, \boldsymbol{\xi_2} \sim N(\mathbf{0}, \mathbf{I_n})$ and $\boldsymbol{\xi_4}, \boldsymbol{\xi_5}$ with $\xi_4^{(i)} = u_i/x_i$, $\xi_5^{(i)} = v_i/y_i$, $i = 1, \ldots, n$.

7. For $i = 1, \ldots, n$, calculate

   $$\begin{aligned} \big[ds_i/dt\big]_{t=m\delta} &= -\mu_i - \eta_i + \zeta_i + m_i \left(\sqrt{\mu_i}\, \xi_1^{(i)} + \sqrt{\eta_i}\, \xi_4^{(i)} - \sqrt{\zeta_i}\, \xi_5^{(i)}\right) \\ \big[dj_i/dt\big]_{t=m\delta} &= \mu_i - \nu_i - \rho_i + \tau_i + m_i \left(-\sqrt{\mu_i}\, \xi_1^{(i)} + \sqrt{\nu_i}\, \xi_2^{(i)} + \sqrt{\rho_i}\, \xi_4^{(i)} - \sqrt{\tau_i}\, \xi_5^{(i)}\right) \\ \big[dr_i/dt\big]_{t=m\delta} &= \nu_i - m_i \sqrt{\nu_i}\, \xi_2^{(i)} \end{aligned}$$

8. Approximate $s_i((m+1)\delta)$, $j_i((m+1)\delta)$, and $r_i((m+1)\delta)$, $i = 1, \ldots, n$, with the Euler-Maruyama formula (8).

9. For $i = 1, \ldots, n$, correct approximation errors by setting negative values of s_i, j_i, and r_i equal to zero.

10. (Optional step.) Rescale s_i, j_i, and r_i, $i = 1, \ldots, n$, via

$$s_i\big((m+1)\delta\big) \leftarrow s_i\big((m+1)\delta\big) \cdot \big(s_i\big((m+1)\delta\big) + j_i\big((m+1)\delta\big) + r_i\big((m+1)\delta\big)\big)^{-1}$$
$$j_i\big((m+1)\delta\big) \leftarrow j_i\big((m+1)\delta\big) \cdot \big(s_i\big((m+1)\delta\big) + j_i\big((m+1)\delta\big) + r_i\big((m+1)\delta\big)\big)^{-1}$$
$$r_i\big((m+1)\delta\big) \leftarrow r_i\big((m+1)\delta\big) \cdot \big(s_i\big((m+1)\delta\big) + j_i\big((m+1)\delta\big) + r_i\big((m+1)\delta\big)\big)^{-1}.$$

With this transformation, we constantly adjust the fractions of susceptible, infected, and recovered individuals to the current population size of the respective region.

4 Initialization

We use our simulation program for the demonstration of spread of influenza in Germany for varying resolutions: for districts ("Landkreise/Stadtkreise"), counties ("Regierungsbezirke"), and states ("Bundesländer").

4.1 Dataset

The underlying data about incidences of influenza in Germany is taken from the Robert Koch Institute (RKI): SurvStat, `http://www3.rki.de/SurvStat`, deadline: 8 July 2005. We only consider cases categorized as A or A/B (i.e. no further differentiation), since it is the influenza A virus that is most responsible for national epidemics of the flu. Unfortunately, the data suffers from underreporting. According to estimations of the Federal Ministry of Health and Women, Austria (`http://www.bmgf.gv.at`), and the Robert Koch Institute (`http://www.rki.de`), the annual number of influenza cases is approximately 4.5% of the total population. However, only one out of 500 of these cases is reported to the RKI. Moreover, the number of announced cases depends on the number of medical examinations induced and does hence not reflect the actual geographical distribution. In particular, affections will be more clustered in the dataset than in reality.

4.2 Connectivity Matrix

The connectivity matrix γ describes the strength of traffic between the subunits of Germany. For its design we take into account the dispersal between adjacent regions, caused e.g. by commuters, and the domestic train and air traffic. Each of these three components is provided with a weight regulating its influence.

At district level, we assume that the major part of the traffic between regions arises from commuters. Data from the Federal Statistical Office Germany (`http://www.destatis.de/e_home.htm`) about the lengths of ways to work lead us to the assumption that about 30% of the employees work in a different district than their home town. Adding private traffic, we obtain an estimated fraction of 16% of the total population that is migrating between districts every day, which is reflected by γ having an average row total of 0.16. We choose the weights of the train and of the flight network to be 1/20 and 1/80 of the traffic between neighboured districts according to the annual amounts of travellers, which are about 200 million in the inter urban rail services and 50 million in the domestic flight connections. Certainly, these weights depend on the kind of disease and time period under observation. For example, the influence of the flight network will be less when considering children's diseases, and during school terms an increasing national mixing rate should be considered.

Within the matrix γ, the strength of migration between two adjacent districts is measured by their densities and numbers of surrounding districts. Our rail network model consists of 57 cities which are served by ICE trains. Data about flight connections was obtained from the OAGflights database (`http://www.oagflights.com`) and composed as in Hufnagel et al. (2004).

For counties and states, we assume the migration between parts of Germany to be more uniform than in the case of districts. For more details, see the supporting material.

4.3 Transmission Rate and Infectious Period

Before being able to run the simulation, we need an estimate of the parameters α and β. Recall that α is the daily number of contacts sufficient for infection an individual has with other individuals, and β^{-1} is the average infectious period of the disease. Due to these meanings, it is easy to estimate β, but more complicated to guess α. We hence try to estimate the basic reproduction number $\rho = \alpha/\beta$. For that, we return to the standard deterministic SIR: Devide the second equation of (1) by the first one and obtain the time-independent differential equation

$$\frac{dj}{ds} = -1 + \frac{1}{\rho s}\,,$$

which has the explicit solution

$$j(t) = -s(t) + \frac{1}{\rho}\log s(t) + c$$

with a constant c. At the very beginning of an influenza epidemic, almost all individuals of the considered population are susceptible, whilst the number of infected should be about zero. With these assumptions, i.e. $s_0 = 1$ and $i_0 = 0$, we obtain $c = 1$. Consequently,

$$\rho = \frac{\log s(t)}{j(t) + s(t) - 1} \qquad \text{for all } t \geq 0. \tag{9}$$

Certainly, the term on the right is not constant for the available data. Moreover, as time goes by and safety measures like vaccination or isolation are increased, the reproduction number is going to fall. However, we assume ρ to be constant in time, but varying in space. From the application of formula (9) to our district-level data and $\lim_{t\to\infty} s(t) = 0.955$ (compare Section 4.1) and $\lim_{t\to\infty} j(t) = 0$, we set

$$\rho(d_i) = 10^{-5} \cdot d_i + 1.0179,$$

where d_i is the population density of region i. This relation reflects the intuitively clear fact that the disease is more likely to spread in areas with high population densities. Since the infectious period of influenza usually lasts for four to five days, we assume $\beta = 2/9$ and calculate α_i via $\beta \cdot \rho(d_i)$, $i = 1, \ldots, n$.

5 Simulation Results

In this section we want to present the results of our simulations. We run the program for different starting scenarios for both the deterministic and stochastic model and try the effects of the parameters on the outcomes. Although we draw comparisons between the (highly under-)reported and the simulated data, we want to emphasize that the objective of this paper is neither to predict the future nor to exactly repeat

former data, but to give an idea of the spatio-temporal spread of influenza and the effect of stochastic fluctuations on its outbreak.

Results of the simulations are returned as animated maps of Germany, which are available at `http://www.statistik.lmu.de/~dargatz/publications`.

5.1 Long-Term Simulations

We repeatedly run our simulation at district level with α and β as estimated in Section 4.3 and with an initial number of infectious individuals according to week 5/2005 in our dataset (Section 4.1). Though being probabilistic, the simulations generally yield the same pattern (see also Figure 1): From South Germany, where an increased level of prevalence was observed in week 5, the disease bounces to Bremen and at the same time moves via Frankfurt to North Rhine-Westphalia and Lower Saxony, from where it spreads to the Eastern part of Germany and finally affects the whole nation. This shows surprisingly good agreement with the actual course of the influenza epidemic in 2005 as demonstrated at `http://influenza.rki.de`. In the last graphic of Figure 1, we interpret the increased morbidity at the national borders, especially in North and East Germany, as edge effects.

As mentioned in Section 4.1, cases in our dataset appear more concentrated in one region than they probably are, which might be due to different reporting behaviour. In contrast to that, our simulation does not leave any district unaffected. The final size of the epidemic, which is the fraction of individuals that have been affected by the disease at the end of the epidemic, is about 4.5% on average. Figure 2 shows the simulated final sizes of the epidemic in each district after 100 iterations. The duration of the outbreak in our simulations is about 150 days, which is twice as long as the actual continuance of the influenza wave from week 5/2005 on. We see the reason for this in a slow onset of the epidemic—caused by too small numbers of initially infectious individuals—and the reproduction number ρ not decreasing in time but being constant, which contradicts the real situation (see Section 4.3). Investigations show that the amount of initially infected individuals hardly affects the final size of the epidemic—as long as the fraction of susceptibles at the very beginning does not fall below ρ^{-1}—but rather shifts the starting point of the major outbreak. To our surprise, changes in the time step δ do not really matter regarding the speed, course and intensity of the spread, which means that our numerical scheme already yields good results for relatively large δ.

5.2 One Week's Forecasts

We initialize the computer program with data from various weeks of our dataset and simulate the following week's spread of the epidemic repeatedly. Figure 3 compares the distributions of the proportions of infected individuals of the total population in week 7/2005 for 2000 simulations of the stochastic model with the respective deterministic result and actual data for the three considered divisions of Germany and $\delta \in \{0.1, 1\}$ (measured in days). It turns out that the deterministic outcomes are similar for all resolutions and both values of δ, but do not agree with the dataset, which is not surprising due to the high level of underreporting mentioned in Section 4.1. On county and state level, the stochastic results seem to be normally distributed with the deterministic value as mean, where the variance is smaller for $\delta = 0.1$ than for $\delta = 1$. In contrast to that, the probabilistic modelling on district level yields results that are larger than in the deterministic case (for $\delta = 1$ much more clearly than for $\delta = 0.1$), though apparently also normally distributed. We suspect the reason for this offset in the starting distribution of infectious individuals: While there are reported affections in almost all counties and states, prevalences are concen-

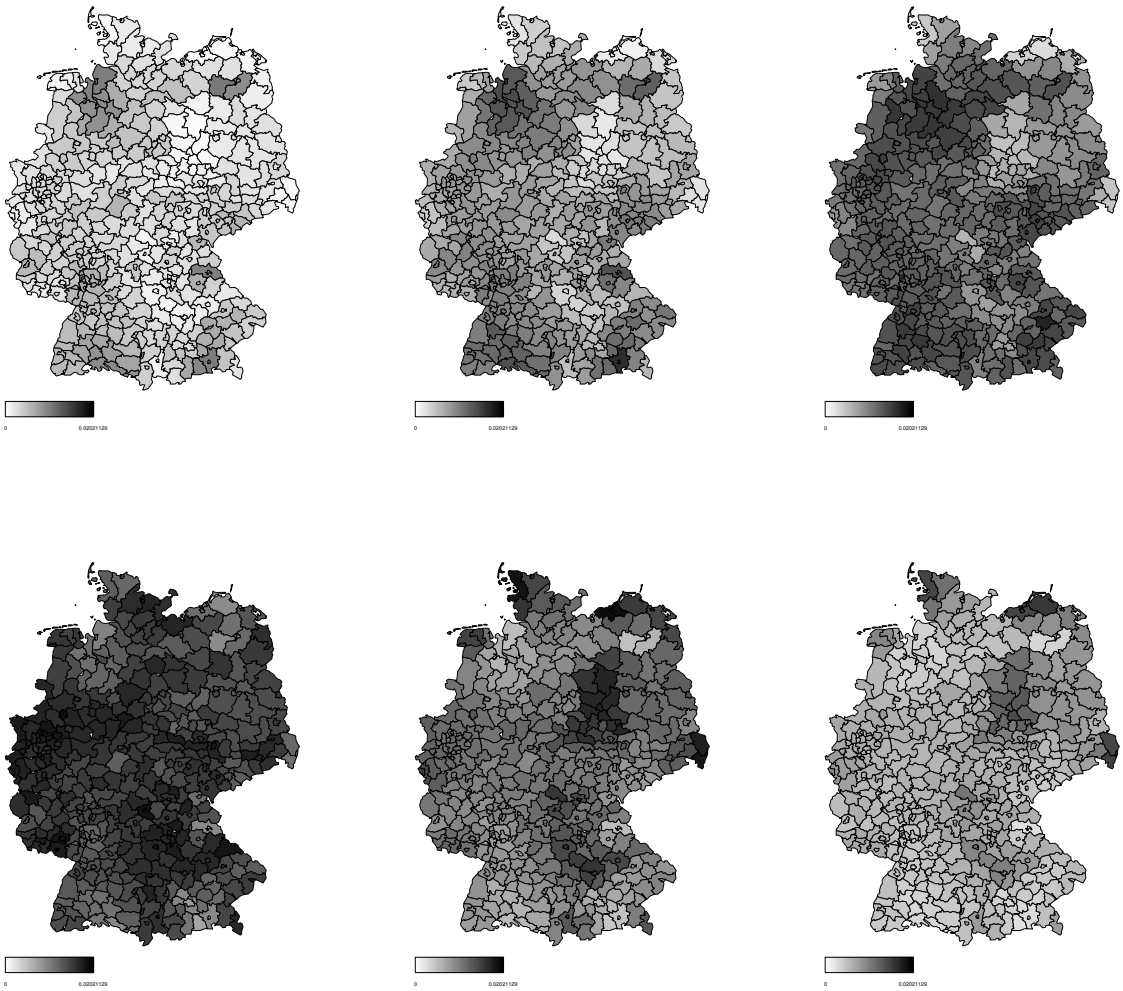

Figure 1: Stochastic simulation of the spread of influenza in Germany. The initial situation corresponds to week 5/2005 in the dataset. Displayed are the fractions of infectives at days 50, 70, 85, 110, 133, and 150 after the starting point.

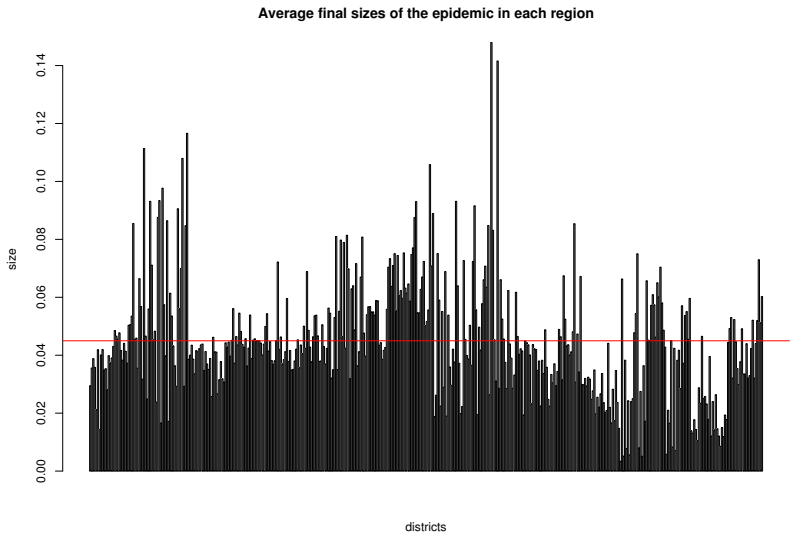

Figure 2: Average final sizes of the epidemic in the 438 German districts after 100 iterations, started at week 5/2005. The horizontal line indicates the mean of all bars.

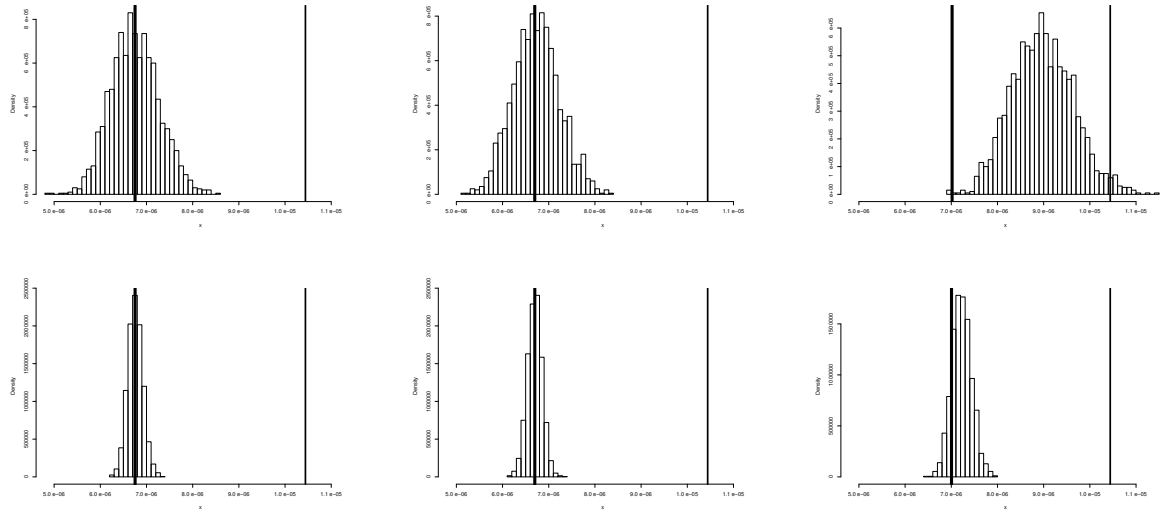

Figure 3: Distributions of the fractions of infectives after 2000 stochastic simulations of one week's spreads. The starting scenario corresponds to week 6/2005. The vertical marks display the respective deterministic (thick line) and actual (thin line) outcomes according to the dataset. Simulations were performed on state, county, and district level (from the left to the right). The first row shows the results for $\delta = 1$, the second one for $\delta = 0.1$.

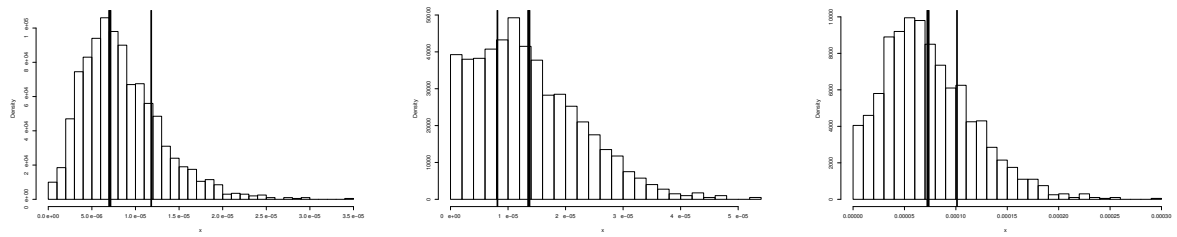

Figure 4: Distributions of the fractions of infectives in Berlin, Böblingen, and Bremerhaven (from the left to the right) after 2000 simulations with $\delta = 1$. The starting scenario corresponds to week 6/2005. The vertical marks display the respective deterministic (thick line) and actual (thin line) outcomes.

trated on relatively few districts. When in our simulation the epidemic spreads to those districts with the initial fraction of infectives being zero, the stochastic fluctuations in this dynamics are kind of bounded to one side (compare with step 9 of the algorithm in Section 3). Obviously, this effect is deeper for larger values of δ. If we focus on those few districts where morbidity was already present at the beginning of the simulation, we obtain rather satisfying results already for $\delta = 1$ (see Figure 4). For these districts, the actual data lies within the range of the stochastic results. We conclude that the stochastic simulation at district level is rather inappropriate as long as we consider relatively short terms or cannot improve the quality of the underlying data.

6 Conclusion and Outlook

In this paper, we presented a global extension of the classical SIR model as well as technical details for its implementation and initialization. Computer simulations provided quite realistic demonstrations of the spread of diseases in Germany. In future work, we will further refine the model, e.g. by involving time-dependent parameters (cf. Sections 4.3 and 5.1). Furthermore, we intend to deal with the question of finding surveillance strategies in case of a sudden outbreak of an epidemic, like specific isolation, vaccination or observation of migration. One main purpose of our research will certainly involve the application of more formal statistical inference techniques for estimating the model parameters based on available data from surveillance databases.

References

Anderson, R., & May, R. (1991). *Infectious Diseases of Humans*. Oxford: Oxford University Press.

Clayton, D. (1996). Generalized linear mixed models. In W. Gilks, S. Richardson, & D. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in Practice* (p. 275-301). London: Chapman & Hall.

Cox, D. (1981). Statistical analysis of time series. Some recent developments. *Scandinavian Journal of Statistics*, *8*, 93-115.

Held, L., Höhle, M., & Hofmann, M. (2005). A statistical framework for the analysis of multivariate infectious disease surveillance counts. *Statistical Modelling (to appear)*.

Hufnagel, L., Brockmann, D., & Geisel, T. (2004). Forecast and control of epidemics in a globalized world. *Proceedings of the National Academy of Sciences*, *101*, 15124-15129.

Kloeden, P., & Platen, E. (1999). *Numerical Solution of Stochastic Differential Equations* (3rd ed.). Berlin, Heidelberg, New York: Springer.

Knorr-Held, L. (2000a). Bayesian modelling of inseperable space-time variation in disease risk. *Statistics in Medicine*, *19*, 2555-2567.

Knorr-Held, L. (2000b). Dynamic rating of sports teams. *Journal of the Royal Statistical Society, Series D (The Statistician)*, *49*, 261-276.

Knorr-Held, L., & Besag, J. (1998). Modelling risk from a disease in time and space. *Statistics in Medicine*, *17*, 2045-2060.

Knorr-Held, L., & Richardson, S. (2003). A hierarchical model for space-time surveillance data on meningococcal disease incidence. *Applied Statistics*, *52*, 169-183.

Lagazio, C., Biggeri, A., & Dreassi, E. (2003). Age-period-cohort models and disease mapping. *Environmetrics*, *14*, 475-490.

Lagazio, C., Dreassi, E., & Biggeri, A. (2001). A hierarchical Bayesian model for space-time variation of disease risk. *Statistical Modelling*, *1*, 17-29.

Rue, H., & Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications* (Vol. 104). London: Chapman & Hall.

Schmid, H., & Held, L. (2004). Bayesian extrapolation of space-time trends in cancer registry data. *Biometrics*, *60*, 1034-1042.

Waller, L., Carlin, B., Xia, H., & Gelfand, A. (1997). Hierarchical spatio-temporal mapping of disease rates. *Journal of the American Statistical Association*, *92*, 607-617.

Watts, D., Muhamad, R., Medina, D., & Dodds, P. (2005). Multiscale, resurgent epidemics in a hierarchical metapopulation model. *Proceedings of the National Academy of Sciences*, *102*, 11157-11162.