Multivariate modelling of infectious disease surveillance data

Leonhard Held Michaela Paul

Biostatistics Unit
Institute of Social and Preventive Medicine
University of Zurich

21. May 2008

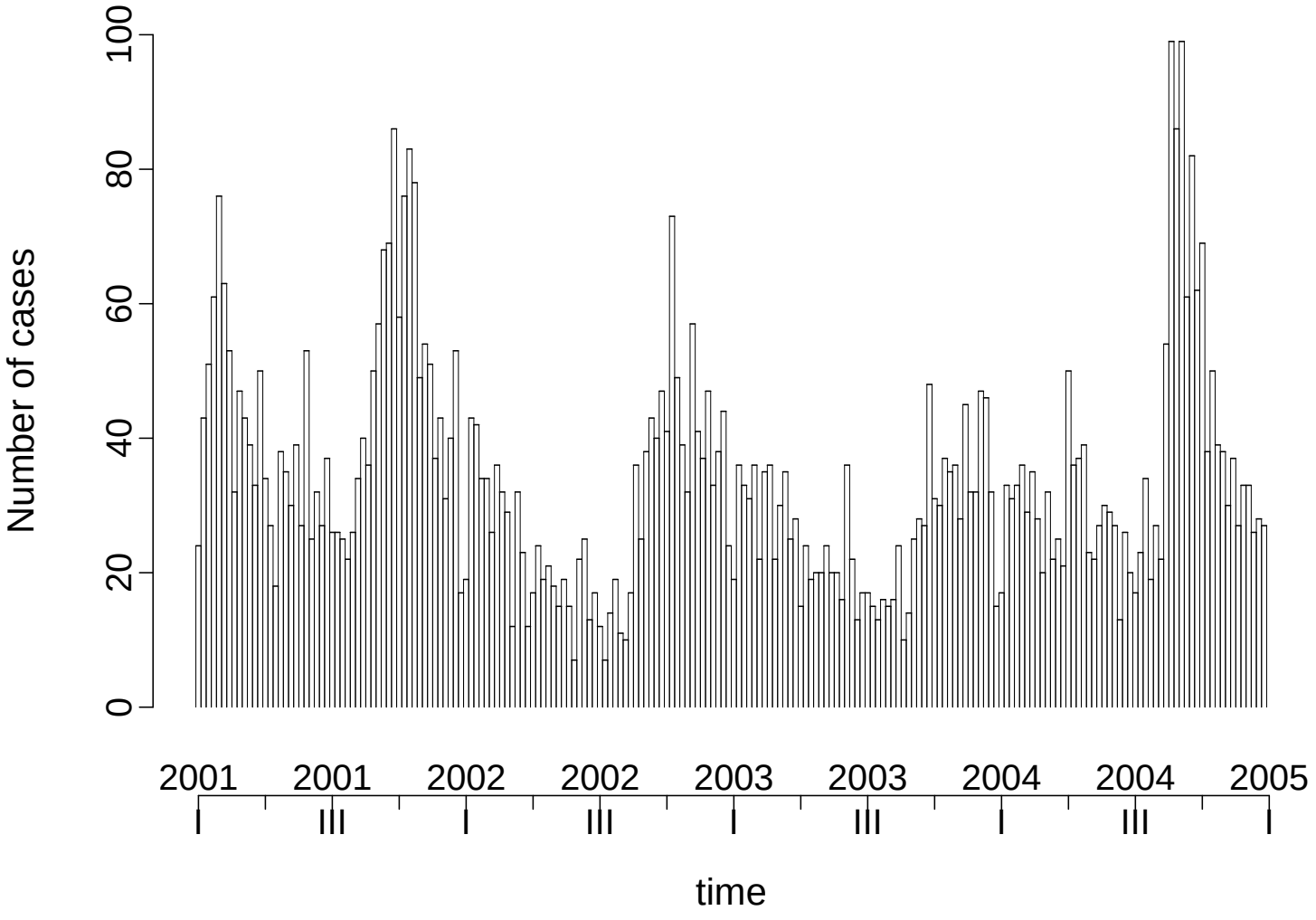
Financial support by the German Science Foundation (DFG, 2003-2006)
and the Swiss National Science Foundation (SNF, since 2007)

Outline

1. Introduction

- This talk is about the statistical analysis of routinely collected surveillance data seen as multiple time series of counts
- The statistical methods of this talk are implemented in the R-package `surveillance` available from the Comprehensive R Archive Network (CRAN)
- Details in Held et al. (2005) and Paul et al. (2008)

Example: Hepatitis A in Germany 2001-2005



Characteristics

- Seasonality
- Occasional outbreaks
- Non-stationarity, for example caused by increased vaccination
- Non-availability of information on susceptibles

Different setting as in classical infectious disease epidemiology

Pure mechanistic modelling impossible!

Previous modelling approaches

- Inclusion of past disease counts as covariates in log-linear Poisson model, i.e. counts act multiplicatively on disease incidence

Causes severe problems as it only allows for negative association

- Modifications have been proposed (Zeger and Qaqish, 1988), which avoid this problem but which are difficult to interpret

Objective

Development of a realistic stochastic model for the statistical analysis of surveillance data of infectious disease counts

- A compromise is needed between mechanistic and empirical modelling
- Model based on a generalized branching process with immigration (Held et al., 2005)

Past counts act additively on disease incidence

- Explicit decomposition of the incidence in an endemic and epidemic component

$\longrightarrow$ model is not a GLM

- Note: Branching process is the common approximation of SIR-models in the absence of information on susceptibles

Model

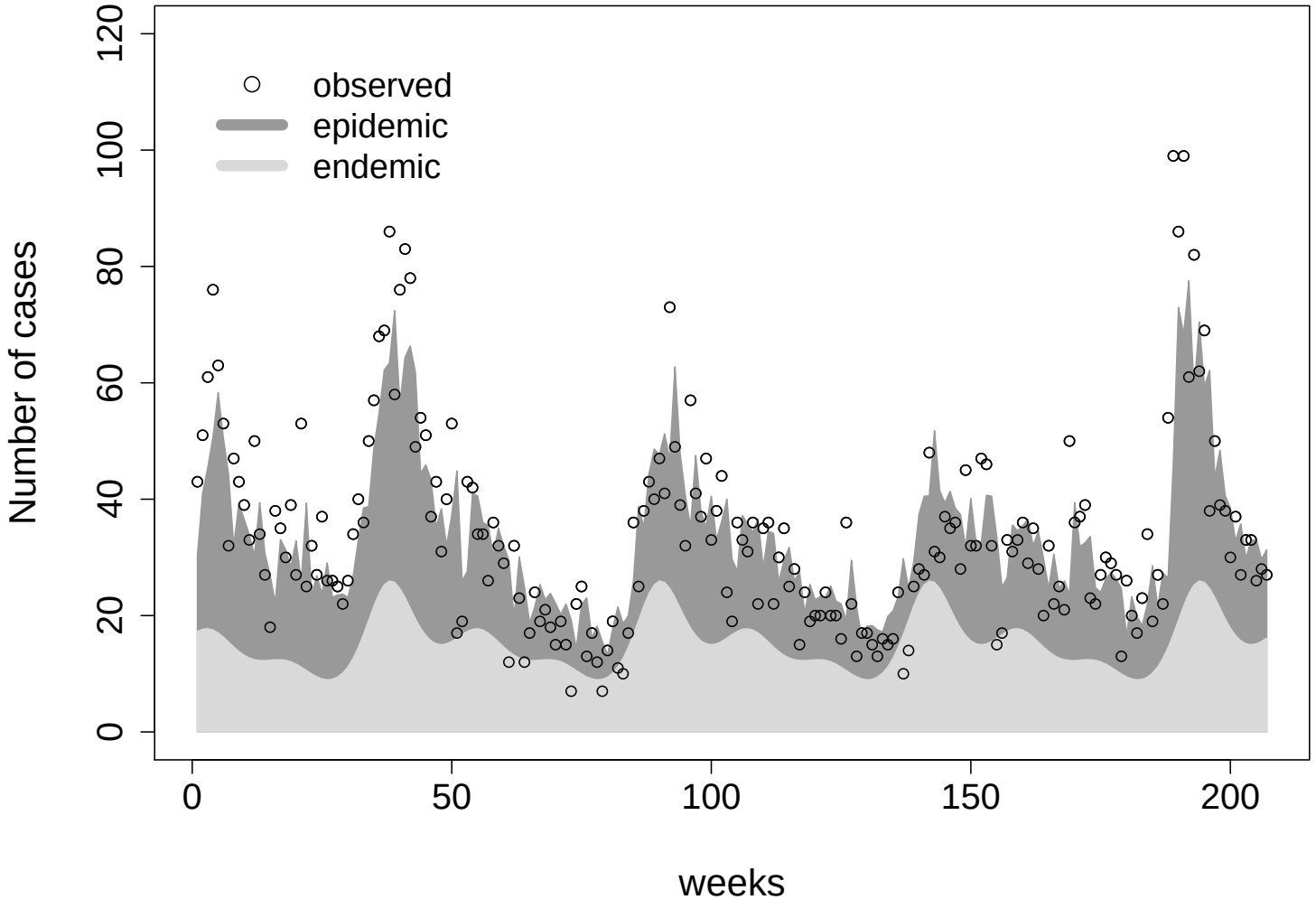
$$
\begin{aligned}
y_t | y_{t-1} &\sim \text{Po}(\mu_t) \\
\mu_t &= \nu_t + \lambda y_{t-1} \\
\log(\nu_t) &= \alpha + \sum_{s=1}^{S} (\gamma_s \sin(\omega_s t) + \delta_s \cos(\omega_s t))
\end{aligned}
$$

- Autoregressive coefficient $\lambda < 1$ determines stationarity of y_t, can be interpreted as epidemic proportion
- $\log \nu_t$ is modelled parametrically as in log-linear Poisson regression; includes terms for seasonality
- Adjustments for overdispersion straightforward: Replace $\text{Po}(\mu_t)$ by $\text{NegBin}(\mu_t, \psi)$-Likelihood
- Model can be fitted by Maximum-Likelihood in `surveillance`

Example: Hepatitis A in Germany 2001-2005

S	$\hat{\lambda}_{ML}$ (se)	$\hat{\psi}_{ML}$ (se)	$\log L$	p	AIC
3	-	-	-1024.5	7	2063.1
3	0.57 (0.03)	-	-870.7	8	1757.5
3	-	9.45 (1.19)	-799.0	8	1614.0
3	0.54 (0.06)	15.36 (2.26)	-763.8	9	1545.6

Fitted values



Multivariate modelling

- Suppose now multiple time series $i = 1, \ldots, n$ are available over the same time horizon $t = 1, \ldots, T$
- Notation: $y_{i,t}$ is the number of disease cases made from the i-th time series at time t
- Examples:
  - Incidence in different age groups
  - Incidence of related diseases
  - Incidence in different geographical regions
- Idea: Include now also the number of counts from other time series as autoregressive covariates
  $\longrightarrow$ multi-type branching process

Bivariate modelling

Joint analysis of two time series $i = 1, 2$

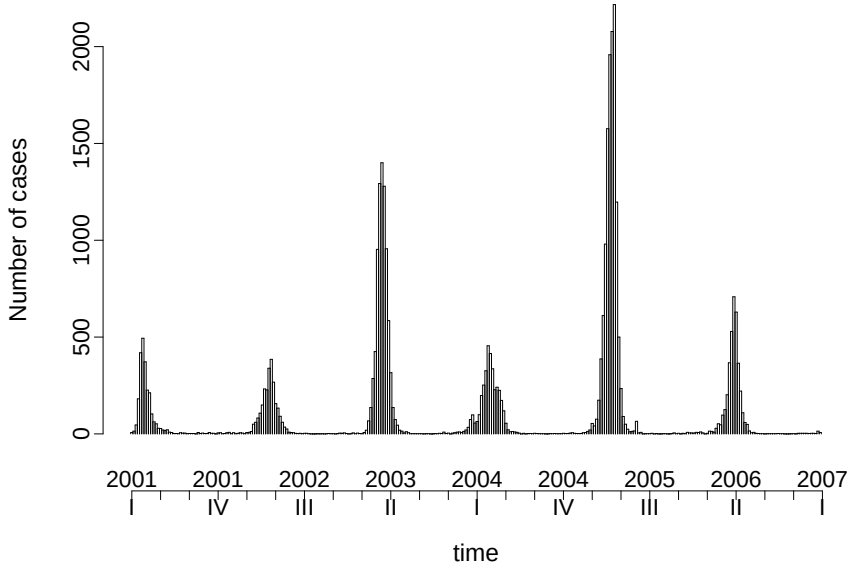
$$\begin{aligned} y_{i,t}|\mathbf{y}_{t-1} &\sim \text{NegBin}(\mu_{i,t}, \psi) \\ \mu_{i,t} &= \nu_t + \lambda y_{i,t-1} + \phi y_{j,t-1} \qquad \text{where } j \neq i \end{aligned}$$

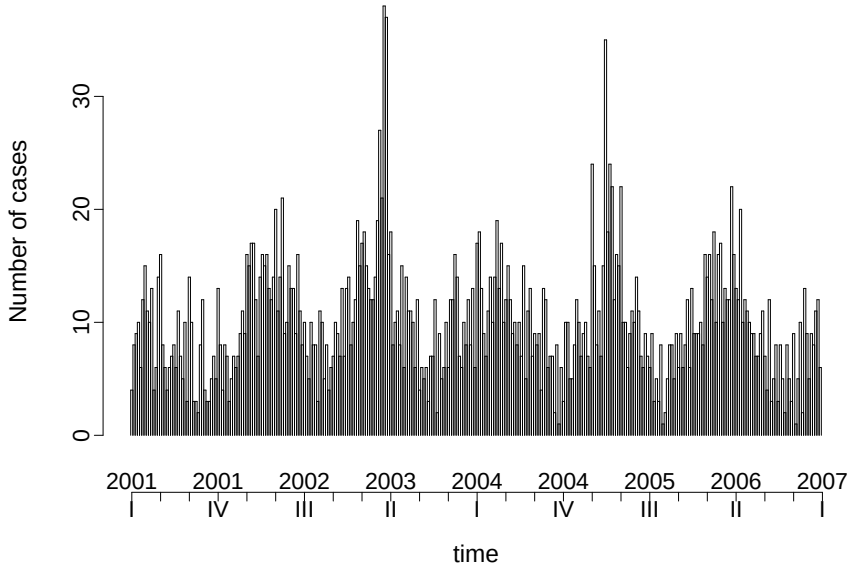
Note: ψ, ν_t, λ and ϕ may also depend on i

Example: Influenza and meningococcal disease

- Inter-dependencies between disease cases caused by different pathogens might be of particular interest to further understand the dynamics of such diseases
- For example, several studies describe an association between influenza and meningococcal disease (Cartwright et al., 1991; Hubert et al., 1992)
- We analyse routinely collected surveillance data from Germany, 2001-2006

Data


Number of cases
0
500
1000
1500
2000
2001 I
2001 IV
2002 III
2003 II
2004 I
2004 IV
2005 III
2006 II
2007 I
time


Number of cases
0
10
20
30
2001 I
2001 IV
2002 III
2003 II
2004 I
2004 IV
2005 III
2006 II
2007 I
time

Univariate analysis of influenza infections

S	$\hat{\lambda}_{ML}$ (se)	$\hat{\psi}_{ML}$ (se)	log L	p	AIC
0	0.99 (0.01)	-	-4050.9	2	8105.9
0	0.98 (0.05)	2.41 (0.27)	-1080.2	3	2166.5
1	0.86 (0.05)	2.74 (0.31)	-1064.1	5	2138.2
2	0.76 (0.05)	3.12 (0.37)	-1053.3	7	2120.6
3	0.74 (0.05)	3.39 (0.41)	-1044.1	9	2106.3
4	0.74 (0.05)	3.44 (0.42)	-1042.2	11	2106.3

Univariate analysis of meningococcal infections

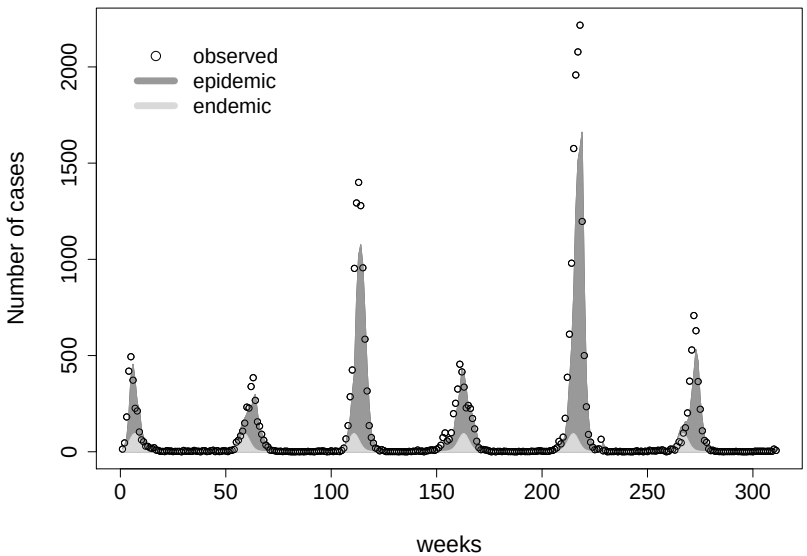
S	$\hat{\lambda}_{ML}$ (se)	$\hat{\psi}_{ML}$ (se)	log L	p	AIC
0	0.50 (0.04)	-	-919.2	2	1842.4
0	0.48 (0.05)	11.80 (2.09)	-880.5	3	1767.0
1	0.16 (0.06)	20.34 (4.83)	-845.6	5	1701.2
2	0.16 (0.06)	20.41 (4.86)	-845.5	7	1705.0

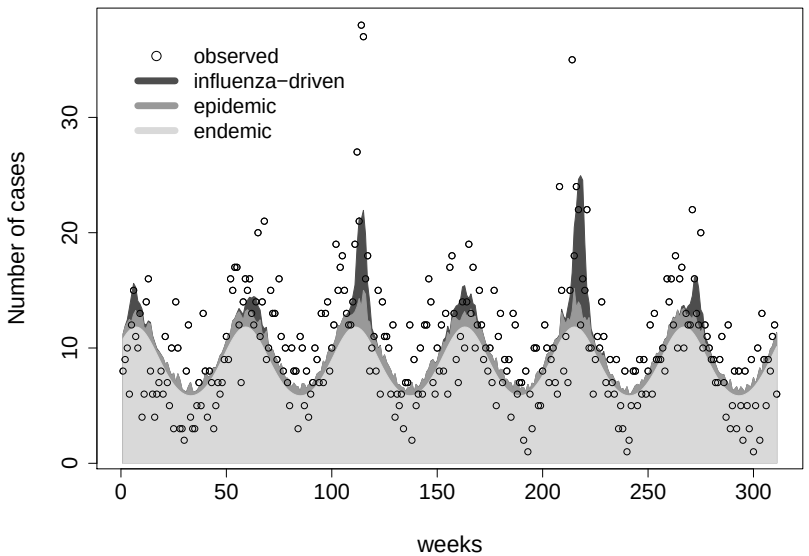
Multivariate analyses

Model	S		$\hat{\lambda}_{ML}$ (se)		$\hat{\phi}_{ML}$ (se)	
	flu	men	flu	men	men $\longrightarrow$ flu	flu $\longrightarrow$ men
1	3	1	0.74 (0.05)	0.16 (0.06)	-	
2	3	1	0.74 (0.05)	0.16 (0.06)	0.000 (0.000)	-
3	3	1	0.74 (0.05)	0.10 (0.06)	-	0.005 (0.001)
4	3	1	0.74 (0.05)	0.10 (0.06)	0.000 (0.000)	0.005 (0.001)

Model	$\hat{\psi}_{ML}$ (se)		log L	p	AIC
	flu	men			
1	3.39 (0.41)	20.34 (4.83)	-1889.7	14	3807.5
2	3.39 (0.41)	20.34 (4.83)	-1889.7	15	3809.5
3	3.39 (0.41)	25.32 (6.98)	-1881.0	15	3791.9
4	3.40 (0.41)	25.32 (6.98)	-1881.0	16	3793.9

Fitted time series





Is a one-week lag correct?

lag	$\hat{\phi}_{ML} \times 10^3$ (se $\times 10^3$)
3	2.92 (1.30)
2	4.54 (1.41)
1	5.32 (1.42)
0	5.30 (1.39)
-1	4.68 (1.31)
-2	3.73 (1.26)
-3	2.30 (1.22)

Spatio-temporal models

- Suppose surveillance data on the same pathogen are available for several geographical locations $i = 1, \ldots, n$
- A possible model extension is:

$$\mu_{i,t} \quad = \quad \nu_t + \lambda y_{i,t-1} + \phi \sum_{j \neq i} w_{ji} y_{j,t-1}$$

- A possible choice for the weights w_{ji} is $w_{ji} = \mathbb{1}(j \sim i)$, i.e. only regions adjacent to region i are taken into account
- Perhaps more natural is $w_{ji} = 1/n_j \cdot \mathbb{1}(j \sim i)$, where n_j denotes the number of neighbours of region j
- Note: λ and ϕ may also depend on i

Incorporating travel information

- Linking of parallel time series based on adjacencies $w_{ji} = \mathbb{1}(j \sim i)$ or $w_{ji} = 1/n_j \cdot \mathbb{1}(j \sim i)$ may be unrealistic in a globalized world
- Alternative: Include (air) travel information, if available
- Convincing example: SARS epidemic, as analysed in Hufnagel et al. (2004)
- Our example: Influenza in USA, as analysed in Brownstein et al. (2006)

Multi-type branching process with immigration

Mean model can be written as

$$\boldsymbol{\mu}_t = \boldsymbol{\Lambda}\mathbf{y}_{t-1} + \boldsymbol{\nu}_t$$

where

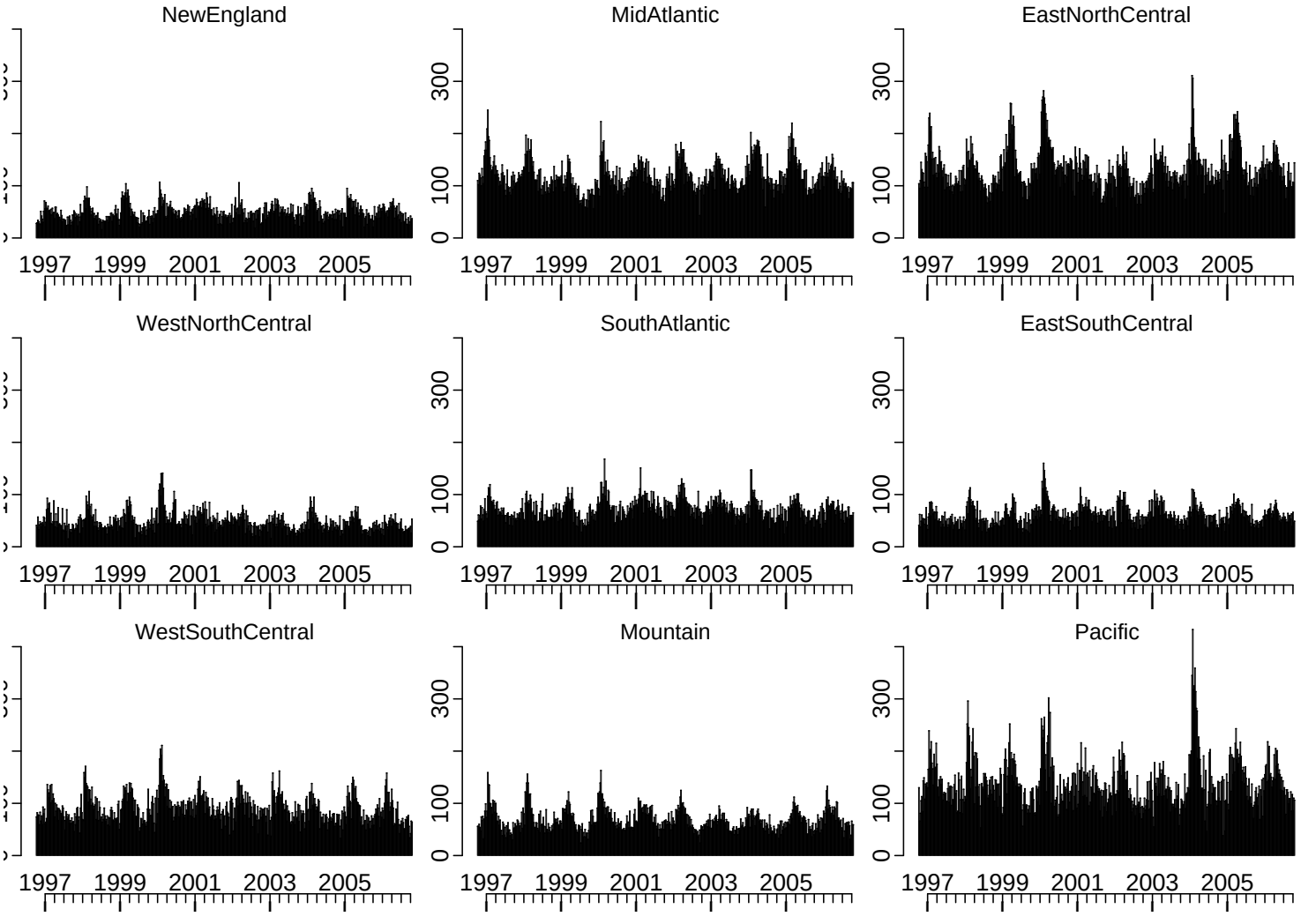
$$\boldsymbol{\Lambda}_{ij} = \begin{cases} \lambda_i & \text{for } i = j \\ \phi_i w_{ji} & \text{for } i \neq j \end{cases}$$

Largest eigenvalue of $\boldsymbol{\Lambda}$ determines stationarity, can be seen as multivariate analogue of λ

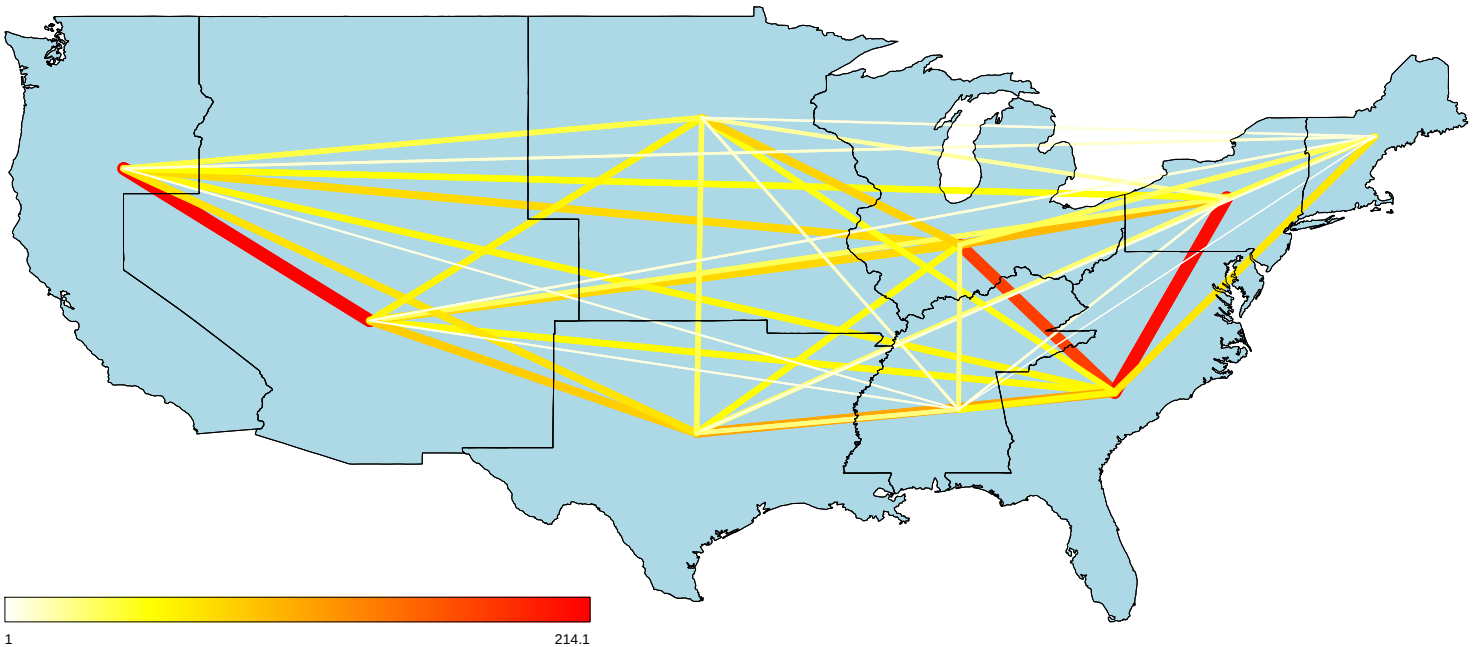
Example: Influenza in USA, 1997-2007

- Data on weekly mortality from pneumonia and influenza obtained from the CDC 121 Cities Mortality Reporting System
- These reports summarize the total number of deaths due to pneumonia and influenza in 9 geographical regions
- Data on the average/yearly number of passengers travelling by air obtained from TranStats database, U.S. Department of Transportation

Data


NewEngland
MidAtlantic
EastNorthCentral
WestNorthCentral
SouthAtlantic
EastSouthCentral
WestSouthCentral
Mountain
Pacific
0
100
300
1997
1999
2001
2003
2005

Air travel data, 1997-2007



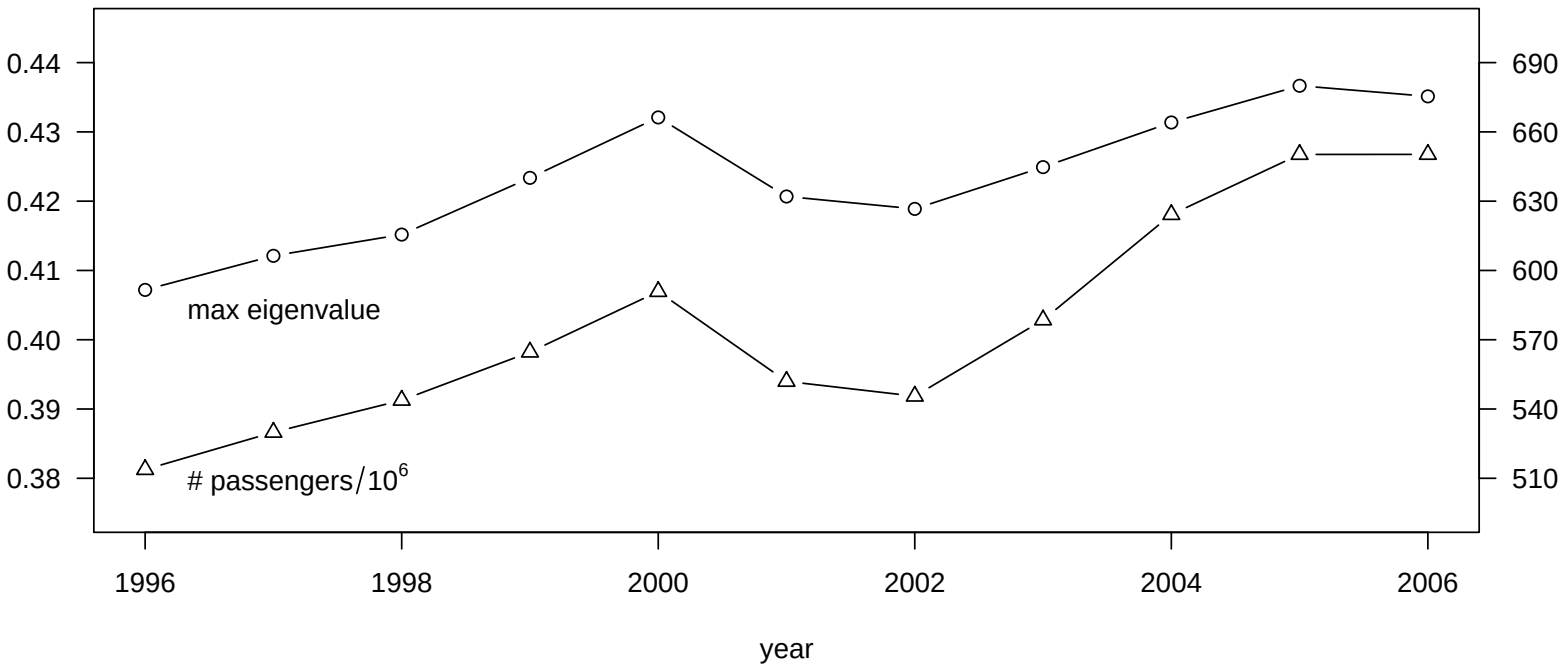
Shown is the average yearly number of passengers per 100,000

Parameter estimates (NegBin, $S = 4$)

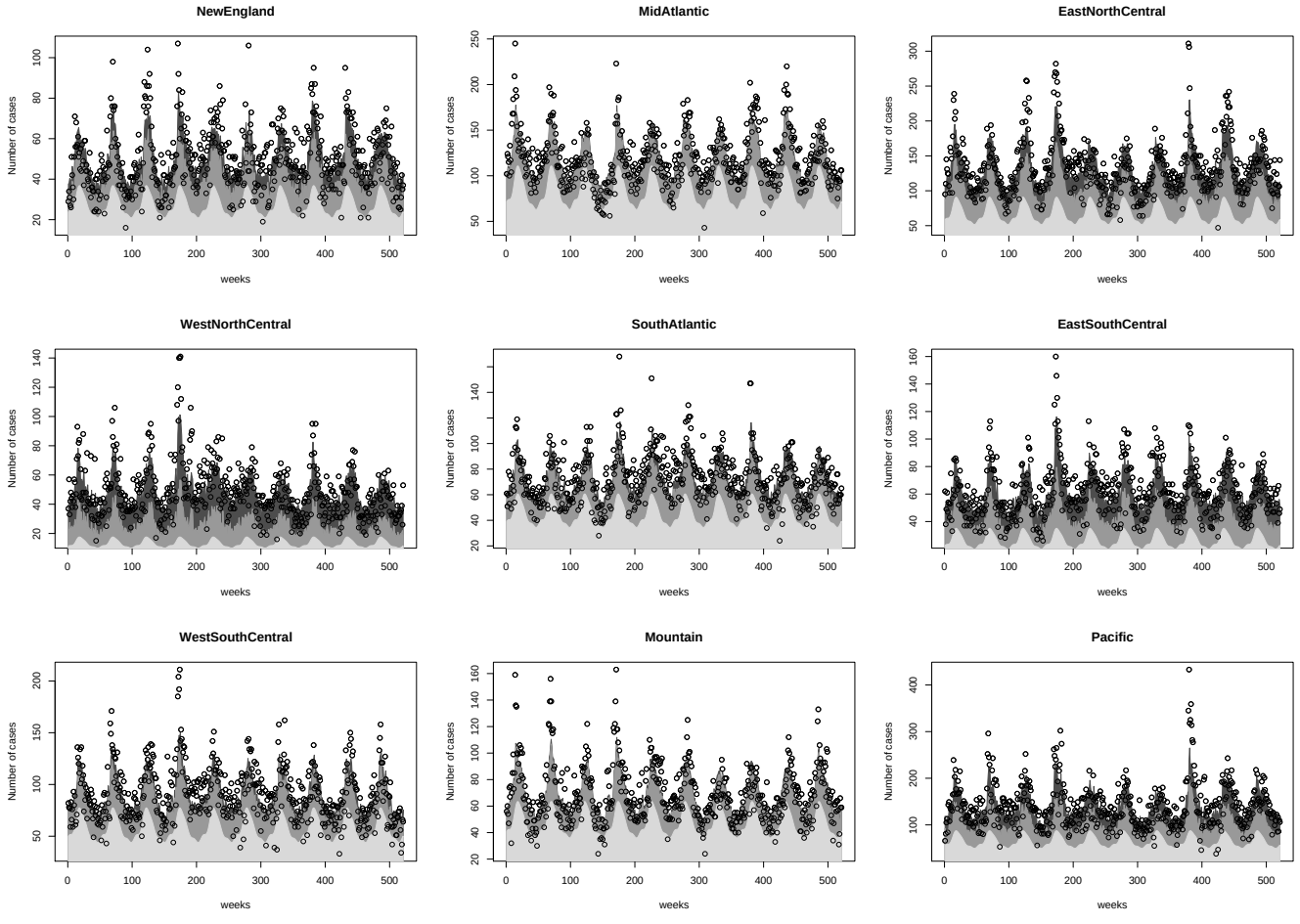
weights	$\hat{\lambda}_{ML}$ (se)	$\hat{\phi}_{i,ML}$ (se)	AIC	max EV
–	–	–	40300.5	–
–	0.34 (0.01)	–	39693.6	0.34
$\mathbb{1}(j \sim i)$	0.30 (0.01)	0.01 (0.01) - 0.23 (0.08)	39632.2	0.45
$\mathbb{1}(j \sim i)/n_j$	0.30 (0.01)	0.01 (0.02) - 0.68 (0.25)	39631.6	0.44
p_{ji}	0.28 (0.01)	0.89 (3.13) - 31.58 (6.04)	39617.0	0.45
p_{ji}(yearly)	0.28 (0.01)	0.84 (1.09) - 28.68 (5.02)	39593.5	*

Here p_{ji} denotes the relative proportion of persons travelling from region j to region i

Max eigenvalues in the best-fitting model



Fitted values



Model validation

- We validate the models based on probabilistic one-step-ahead predictions
- Mean squared prediction error score does not incorporate prediction uncertainty
- We use proper scoring rules (Gneiting and Raftery, 2007), which address calibration and sharpness simultaneously:
  - Logarithmic score
  - Ranked probability score
- Calibration alone is assessed using PIT histograms for count data (Czado et al., 2007)

Proper scoring rules

- The logarithmic score is strictly proper and defined as

$$\text{LogS}(Y, y_{obs}) = -\log f(y_{obs}),$$

the log predictive density ordinate at the observed value y_{obs}.

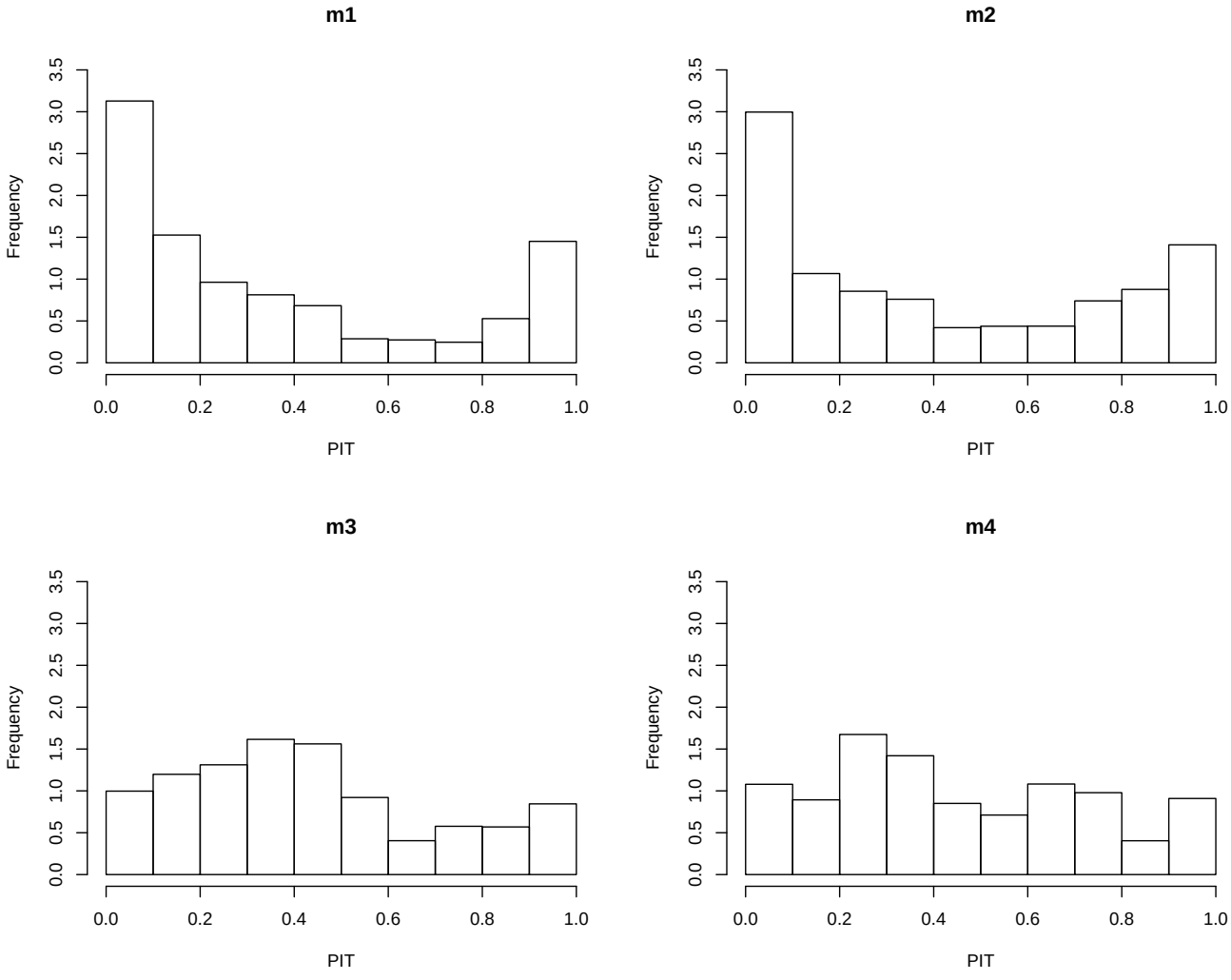
- A popular strictly proper score which is less sensitive to outliers but sensitive to distance is the so-called ranked probability score

$$\text{RPS}(Y, y_{obs}) = \sum_{t=0}^{\infty} (P(Y \leq t) - \mathbf{1}(y_{obs} \leq t))^2 dt,$$

the sum of the Brier scores for binary predictions at all possible thresholds t.

Hepatitis A in Germany: PIT histograms

Based on 100 one-step-ahead predictions



Hepatitis A in Germany: Scoring rules

distr	S	autoreg	logs	rps
Poisson	3	$-$	5.483	8.198
Poisson	3	$+$	4.357	6.265
NegBin	3	$-$	3.909	7.420
NegBin	3	$+$	3.691	5.851

Hepatitis A in Germany: Scoring rules

distr	S	autoreg	logs	(p-value)	rps	(p-value)
Poisson	3	$-$	5.483	($<$0.001)	8.198	($<$0.001)
Poisson	3	$+$	4.357	($<$0.001)	6.265	(0.0019)
NegBin	3	$-$	3.909	(0.0015)	7.420	($<$0.001)
NegBin	3	$+$	3.691		5.851	

p-values are based on Monte-Carlo permutation tests for paired individual scores

Meningococcal infections: Scoring rules

Based on 156 one-step-ahead predictions (3 years)

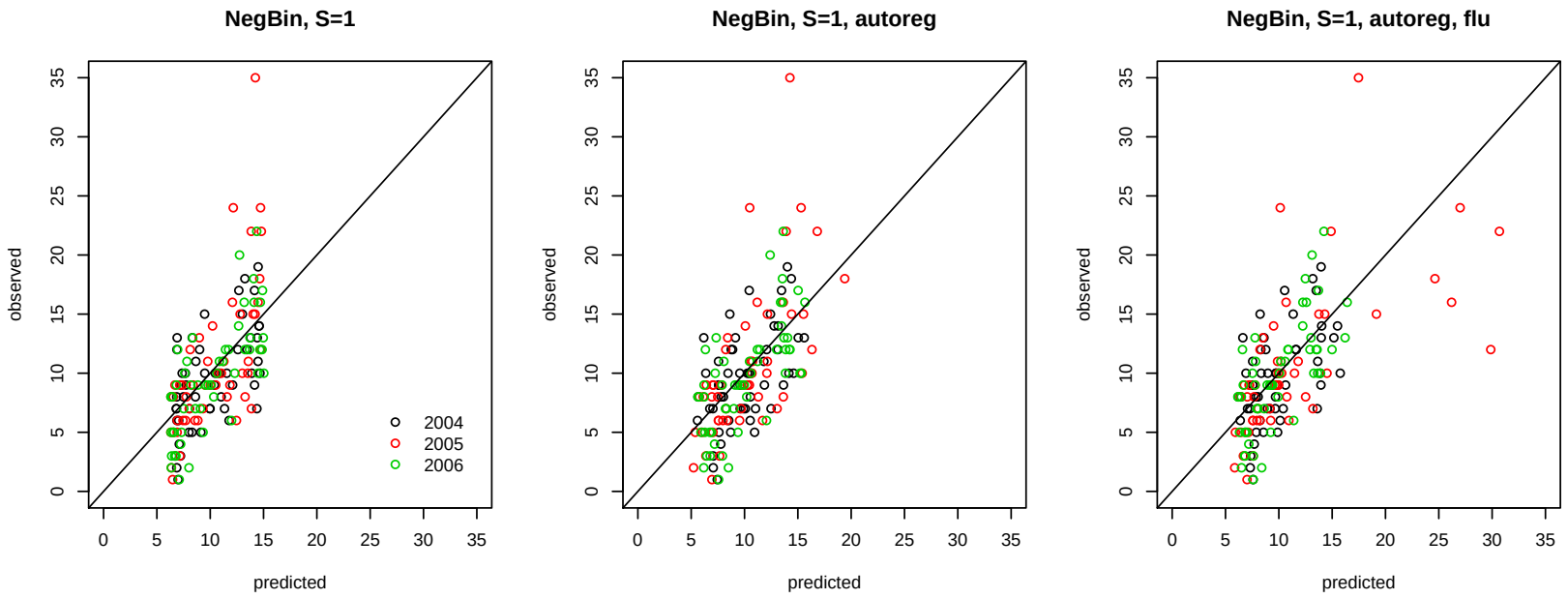
distr	S	autoreg	flu	logs	rps
NegBin	1	$-$	$-$	2.679	2.023
NegBin	1	$+$	$-$	2.709	2.080
NegBin	1	$+$	$+$	2.708	2.128

Meningococcal infections: Scoring rules

Based on 156 one-step-ahead predictions (3 years)

distr	S	autoreg	flu	logs	(p-value)	rps	(p-value)
NegBin	1	$-$	$-$	2.679		2.023	
NegBin	1	$+$	$-$	2.709	(0.104)	2.080	(0.138)
NegBin	1	$+$	$+$	2.708	(0.390)	2.128	(0.323)

Observed versus predicted and shrinkage



Shrinkage (Copas, 1997) applied to flu coefficient did not improve the predictions much.

Influenza in USA: Scoring rules

Based on 260 one-step-ahead predictions (5 years)

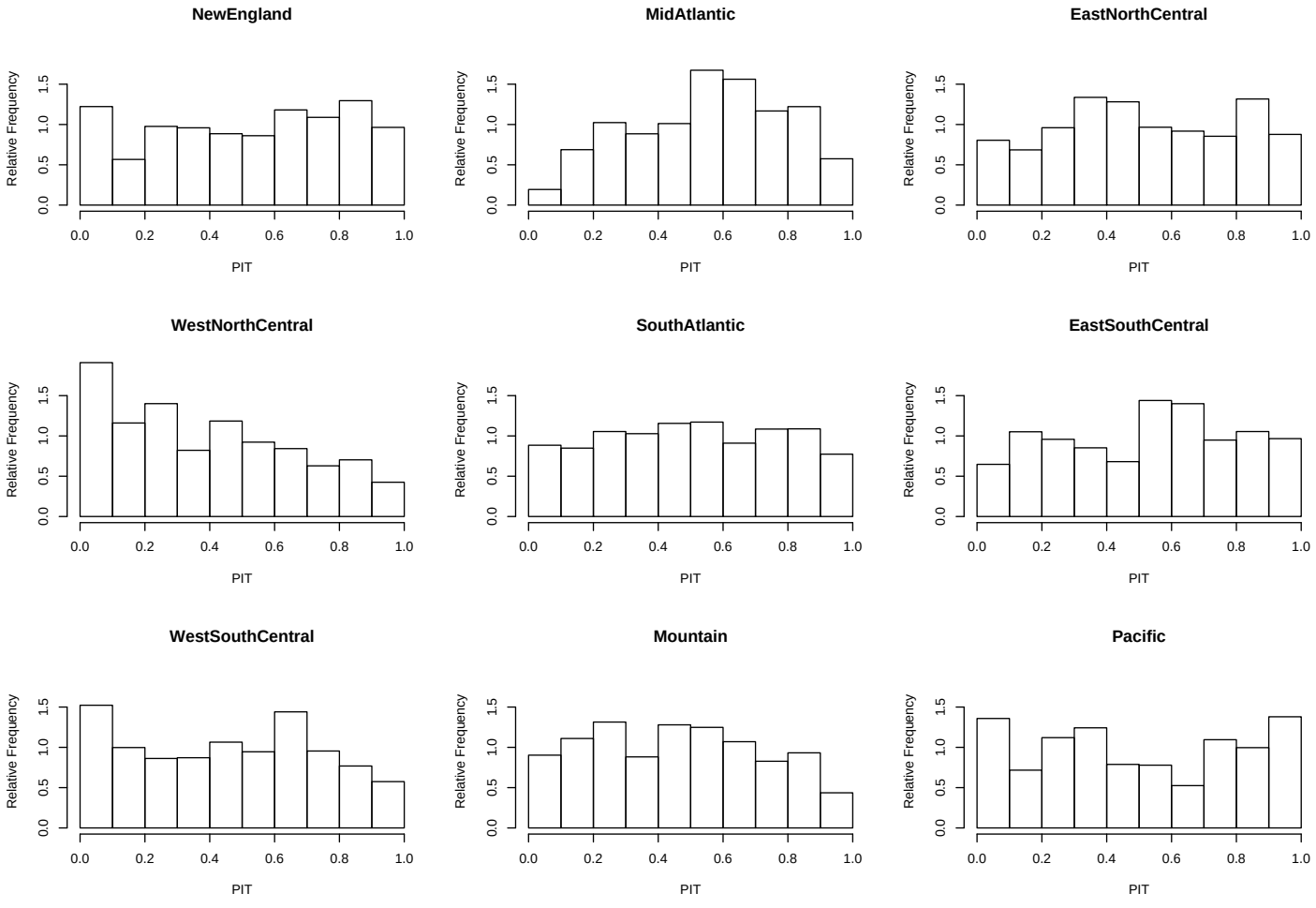
autoreg	weights	logs	rps
$-$		4.2816	10.671
$+$		4.2249	9.8478
$+$	$\mathbb{1}(j \sim i)$	4.2253	9.8582
$+$	$\mathbb{1}(j \sim i)/n_j$	4.2248	9.8531
$+$	p_{ji}	4.2247	9.8493
$+$	p_{ji}(yearly)	4.2278	9.8689

Influenza in USA: Scoring rules

Based on 260 one-step-ahead predictions (5 years)

autoreg	weights	logs	(p-value)	rps	(p-value)
$-$		4.2816	($<$0.001)	10.671	($<$0.001)
$+$		4.2249	(0.964)	9.8478	(0.971)
$+$	$\mathbb{1}(j \sim i)$	4.2253	(0.764)	9.8582	(0.672)
$+$	$\mathbb{1}(j \sim i)/n_j$	4.2248	(0.939)	9.8531	(0.856)
$+$	p_{ji}	4.2247		9.8493	
$+$	p_{ji}(yearly)	4.2278	(0.326)	9.8689	(0.571)

Influenza in USA: PIT histograms



Discussion

- Useful tool for the analysis of multivariate time series of counts of disease
- Can be used to detect inter-dependencies between time series
- Predictive model validation through proper scoring rules
- Next steps:
  - Dependence of autoregressive components on covariates, e.g. vaccination levels or hygiene interventions
  - Inclusion of area-level random-effects

References

Brownstein, J. S., C. J. Wolfe, and K. D. Mandl (2006). Empirical evidence for the effect of airline travel on inter-regional influenza spread in the United States. *PLOS Medicine 3(10)*, 1826–1835.

Cartwright, K., D. Jones, A. Smith, J. Stuart, E. Kaczmarski, and S. Palmer (1991). Influenza A and meningococcal disease. *Lancet 338(8766)*, 554–557.

Copas, J. P. (1997). Using regression models for prediction: shrinkage and regression to the mean. *Statistical Methods in Medical Research 6*, 167–183.

Czado, C., T. Gneiting, and L. Held (2007). Predictive model assessment for count data. Technical report. In revision for *Biometrics*.

Gneiting, T. and A. E. Raftery (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* **102**, 359–378.

Held, L., M. Höhle, and M. Hofmann (2005). A statistical framework for the analysis of multivariate infectious disease surveillance data. *Statistical Modelling 5*, 187–199.

Hubert, B., L. Watier, P. Garnerin, and S. Richardson (1992). Meningococcal disease and influenza-like syndrome: a new approach to an old question. *Journal of Infectious Diseases 166*, 542–545.

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

Paul, M., L. Held, and A. M. Toschke (2008). Multivariate modelling of infectious disease surveillance data. Technical report.

Zeger, S. L. and B. Qaqish (1988). Markov regression models for time series: a quasi-likelihood approach. *Biometrics 44*(4), 1019–1031.