Modeling Multivariate Surveillance Data

Xanthi Pedeli and Dimitris Karlis
Athens University of Economics & Business

January 15, 2021

Public health surveillance

- *Main purpose of public health surveillance systems:* effective and timely detection of disease outbreaks with the aim of rapidly taking control measures for the elimination of disease transmission.
- Increased availability of health surveillance data; in most cases several variables are monitored and events of different types are reported.
- Public health surveillance typically uses univariate data for monitoring disease occurrence at a local level $\longrightarrow$ correlation between series is ignored.

Additional features of health surveillance data

- Health data are typically autocorrelated over time.
- Non-negative count data which are more likely Poisson or negative binomial rather than normally distributed.
- Only shifts in a positive direction are of interest.

Available statistical methods for multivariate surveillance

- Dimensionality reduction (principal components, sufficient reduction techniques)
- Parallel surveillance (each series is monitored separately)
- Joint modeling (with alarm functions based on the LR statistic)
- Scalar accumulation (Hotelling's T^2 charts)
- Vector accumulation methods (MCUSUM and MEWMA charts)

Motivation

So, we have a fairly wide range of statistical tools in our hands to handle multivariate surveillance data.

Why another one?
Most of these approaches ignore the integer-valued property of the data and/ or its correlation structure.

Suggested approach:
Based on a modification of the multivariate integer-valued autoregressive model (PK, 2013, CSDA)

Integer-valued autoregressive model: the general idea

- Introduced by McKenzie (1985) and Al-Osh and Alzaid (1987) as a convenient way to transfer the usual autoregressive structure to discrete valued time series.
- Main concept is the notion binomial thinning.

Binomial thinning

Suppose that X is a non-negative integer-valued random variable and let $\alpha \in [0, 1)$. The binomial thinning operator "$\circ$" is defined by (Steutel and van Harn, 1979)

$$\alpha \circ X = \begin{cases} \sum_{j=1}^{X} Y_j, & X > 0 \\ 0, & \text{otherwise} \end{cases}$$

where Y_j are i.i.d. Bernoulli random variables, independent of X, with $P(Y_j = 1) = 1 - P(Y_j = 0) = \alpha$.

The integer-valued autoregressive process of order one

INAR(1) process:

$$X_t = \alpha \circ X_{t-1} + \epsilon_t,$$

where $\alpha \in [0, 1)$ and $\{\epsilon_t, t \in \mathbb{N}\}$ is a sequence of independent identically distributed non–negative integer–valued random variables with mean μ_ϵ and finite variance σ^2_ϵ .

The multivariate INAR(1) process

MINAR(1) process (PK, 2013, CSDA):

$$\mathbf{X}_t = \mathbf{A} \circ \mathbf{X}_{t-1} + \boldsymbol{\epsilon}_t, \quad t \in \mathbb{Z}$$

where,
$\mathbf{X}_t$: random vector with values in $\mathbb{N}^n$
$\mathbf{A}$: $n \times n$ matrix with independent elements $\{\alpha_{i,j}\}_{i,j=1}^n$
$\mathbf{A} \circ \mathbf{X}$: n-dimensional random vector with i-th component $[\mathbf{A} \circ \mathbf{X}]_i = \sum_{j=1}^n \alpha_{ij} \circ X_j, \quad i = 1, \ldots, n$, where the counting series in all $\alpha_{ij} \circ X_j$ are assumed to be independent.
$\{\boldsymbol{\epsilon}_t\}_{t \in \mathbb{Z}}$: a sequence of non-negative integer-valued random vectors with mean $\boldsymbol{\mu}_{\boldsymbol{\epsilon}}$ and variance-covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ independent of $\mathbf{A} \circ \mathbf{X}_{t-1}$.

The multivariate INAR(1) process

Conditional maximum likelihood estimator:

$$\hat{\theta} = \text{argmax}_{\theta} \ell(\theta),$$

where

$$\ell(\theta) = \sum_{t=2}^{T} \log f(\mathbf{x}_t | \mathbf{x}_{t-1}, \theta)$$

and $f(\mathbf{x}_t | \mathbf{x}_{t-1}, \theta)$ is the convolution of n sums of binomials and the joint distribution of ϵ_t, i.e.

$$f(\mathbf{x}_t | \mathbf{x}_{t-1}, \theta) = \sum_{k_1=0}^{m_1} \cdots \sum_{k_n=0}^{m_n} f_1(x_{1t} - k_1 | \mathbf{x}_{t-1}) \cdots f_n(x_{nt} - k_n | \mathbf{x}_{t-1}) g(k_1, \ldots, k_n),$$

where $m_i = \min(x_{it}, x_{i;t-1})$, $i = 1, \ldots, n$.

Constrained multivariate INAR(1) process

- *Motivation*: the numerical difficulty of the maximum likelihood approach increases sharply with dimensional increase.
- PK (2013, SMij) consider a constrained multivariate INAR(1) model by assuming that $\mathbf{A}$ is a $n \times n$ diagonal matrix with independent elements $\alpha_i = [\mathbf{A}]_{ii}$, $i = 1, \ldots, n$.
- Estimation of the constrained model is performed through a composite (pairwise) likelihood approach that reduces the multivariate estimation problem to a set of bivariate problems.

Linking with the multivariate health surveillance problem

- *Aim of statistical models for health surveillance data*: to effectively capture the endemic and epidemic dynamics of disease risk.
- Endemic component: explains a baseline rate of cases with stable temporal pattern - independent of the history of the epidemic process.
- Epidemic component: aims to introduce infectiousness, that is explicit dependence between events - driven by the observed past and identified with the autoregressive part of the model.

Motivation for a new model specification

- The additive decomposition of disease risk is well embodied in the multivariate INAR(1) model.
- But remember that inference becomes difficult as the dimension increase.
- The constrained version of the model,
    1. ignores the relationship with time lag between series that is typical in disease transmission;
    2. is estimated through a pairwise likelihood approach which is not appropriate for prediction purposes.

Suggested simplification

- Assume that the correlation matrix $\mathbf{A}$ is non-diagonal and relax the degree of complexity of the model by assuming that the innovation series ϵ_t, i.e. the endemic components, are uncorrelated.
- The resulting model admits a realistic epidemiological interpretation and is extremely advantageous in terms of practical implementation since the distribution of the innovations becomes a product of univariate mass functions.
- Overdispersion that is a typical characteristic of health surveillance data, can be easily accommodated even under the simplest parametric assumption of Poisson innovations.

Outbreak detection statistical process

- Assumption: the set-up phase is free or cleaned of outbreaks.
- Steps:
  1. Fit a multivariate INAR(1) model to the available series of data in the set-up phase (historical data) to obtain a parameter vector of maximum likelihood estimates $\hat{\theta}$.
  2. Use the model obtained from the set-up phase for successive monitoring of incoming observations in the operational phase (surveillance data).

Outbreak detection statistical process

Details on the second step:

- For each multivariate observation $\mathbf{x}_{t+1}$ in the operational phase, we estimate the one-step-ahead predictive distribution $\hat{P}(\mathbf{X}_{t+1} = \mathbf{x}_{t+1} | \mathbf{x}_t, \hat{\theta}), \mathbf{x} \in \mathbb{N}_0^n$ and obtain the marginal predictive probabilities $\hat{P}(X_{i,t+1} = x_{i,t+1} | \mathbf{x}_t, \hat{\theta}), i = 1, \ldots, n$.
- For each observation $x_{i,t+1}$, we construct an $(1 - \alpha)\%$ prediction interval with upper bound $x_{i,t+1}^{UB}$ equal to the $(1 - \alpha)$-quantile of the corresponding marginal predictive distribution, where α is a prespecified significance level.
- The lower bound of the prediction interval is set equal to 0 since we are only interested in detecting positive deviations from the in-control model.

Outbreak detection statistical process

Details on the second step (cont.):

- Each series flags an alarm at time $t+1$ if the corresponding observation lies outside the prediction interval, i.e. if

$$x_{i,t+1} > x_{i,t+1}^{UB}.$$

- For the overall alarm, a majority rule can be defined, i.e. flagging an alarm if a certain percentage of the series signals an alarm at the same point in time.

Simulation study

Set-up

- Time series data of length $n = 200$ simulated from a trivariate INAR(1) model with independent Poisson innovations.
- First 150 observations assumed to consist the set-up phase (that is a clean process without outbreaks) and the last 50 observations assumed to consist the monitoring phase.
- For each series i, $i = 1, 2, 3$, an outbreak of expected size κ_i at time $t = 170$ was simulated from a Poisson distribution with mean equal to κ_i.

Simulation study

Set-up (cont.)

Model:

$$\begin{pmatrix} X_{1t} \\ X_{2t} \\ X_{3t} \end{pmatrix} = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} \end{bmatrix} \circ \begin{pmatrix} X_{1,t-1} \\ X_{2,t-1} \\ X_{3,t-1} \end{pmatrix} + \begin{pmatrix} \epsilon_{1t} \\ \epsilon_{2t} \\ \epsilon_{3t} \end{pmatrix},$$

where ϵ_{it} are independent Poisson random variables with mean $E(\epsilon_{it}) = \lambda_i + \kappa_i I(t = 170)$ and $I(A)$ is an indicator function.

Simulation study

Set-up (cont.)

True parameter values:

$$\begin{bmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} \end{bmatrix} = \begin{bmatrix} 0.3 & 0.1 & 0.2 \\ 0.2 & 0.4 & 0.2 \\ 0.3 & 0.2 & 0.2 \end{bmatrix},$$

$\lambda_1 = \lambda_2 = \lambda_3 = 1$

$\kappa_1 = \kappa_2 = \kappa_3 = \kappa$, where $\kappa =$5, 8 or 10.

1000 simulation replicates per scenario ($\kappa =$5, 8 or 10).

Simulation study

Evaluation measures

- Detection rate and weekly false alarm rate based on a rule of $2/3$ i.e. assuming that an alarm is triggered if at least two out of the three series flagged an alarm at the same point in time.
- Detection rate: proportion of the 1000 replicates in which an alarm was triggered at time $t = 170$.
- False alarm rate: number of cases in which an alarm was flagged at time $t \neq 170$ divided by 1000×49.

Simulation study

Results: Detection rates (DR) and false alarm rates (FAR) for different outbreak sizes κ and different significance levels α. The reported numbers have been multiplied by 100.

	Outbreak size					
	$\kappa = 5$		$\kappa = 8$		$\kappa = 10$	
Sign. level	DR	FAR	DR	FAR	DR	FAR
$\alpha = 10\%$	89.0	1.33	99.4	1.30	99.8	1.44
$\alpha = 5\%$	80.1	0.34	98.7	0.32	99.8	0.40
$\alpha = 1\%$	55.1	0.01	93.4	0.01	98.5	0.03

Application

Data

- Syndromic surveillance data collected during Athens 2004 Olympic Games.
- The full database consists of 11 different syndromes recorded since July 2002 in emergency departments of major hospitals in the Greater Athens area (drop-in syndromic surveillance).
- We consider 3 distinct syndromes recorded in a specific hospital that are significantly correlated to each other (cross-correlations ranging from 0.31 to 0.48): respiratory infection with fever, febrile illness with rash, other syndrome with potential interest for public health.

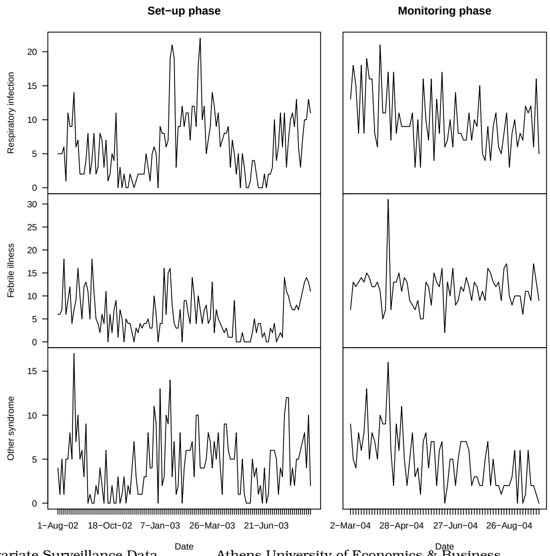
Application

Monitoring phase & set-up phase

- Monitoring period: March 2, 2004 - September 28, 2004
- Set-up phase: August 1, 2002 - August 29, 2003 is considered as the set-up phase.
- During both periods syndromes were recorded every three days so that the historical and surveillance data consist of $t_0 = 127$ and $t_1 = 71$ observations respectively.

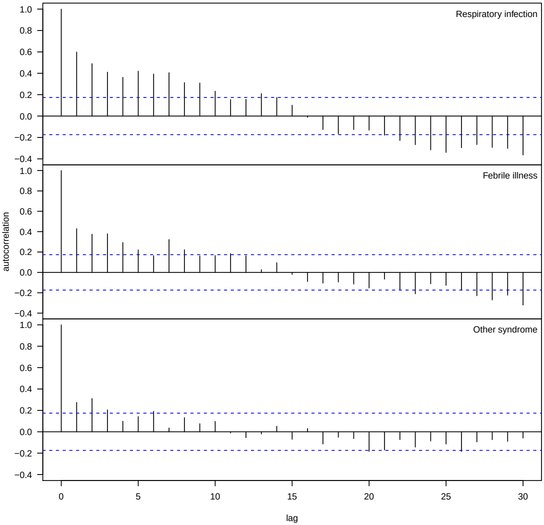
Application

Time series plot of the data



Application

Plots of the autocorrelations of the historical data



Application

Statistical surveillance approach

A trivariate INAR(1) regression model with indepedent Poisson innovations fitted to the historical syndromic surveillance data: each marginal series is modeled as $X_{it} = \sum_{j=1}^{3} \alpha_{ij} \circ X_{j,t-1} + \epsilon_{it}$, $i = 1, 2, 3$, where ϵ_{it} are independent Poisson random variables with mean

$$\begin{aligned} E(\epsilon_{it}) &= \exp\left\{\beta_{i0} + \beta_{i1}\text{Weekday} + \beta_{i2}\cos\left(\frac{2\pi t}{122}\right)\right. \\ &\quad \left. + \beta_{i3}\sin\left(\frac{2\pi t}{122}\right)\right\} \end{aligned}$$

for $t = 1, \ldots, t_0$.

Application

Statistical surveillance approach (cont.)

- A univariate surveillance approach based on fitting three indepedent INAR(1) regression models with Poisson innovations also employed for comparison purposes.
- We assume a type I error of $\alpha = 0.01$ and for the overall alarm we set a rule of $2/3$ that is an alarm is triggered if at least two out of the three series flag an alarm at the same point in time.

Application

Results: Maximum likelihood estimates (standard errors) of the correlation parameters obtained from fitting three independent Poisson INAR(1) or a trivariate INAR(1) regression model with independent Poisson innovations to the historical data.

correlation parameters	trivariate INAR(1)	independent INAR(1)
$\hat{\alpha}_{11}$	0.329 (0.044)	0.393 (0.039)
$\hat{\alpha}_{12}$	0.126 (0.043)	-
$\hat{\alpha}_{13}$	0.134 (0.054)	-
$\hat{\alpha}_{21}$	0.160 (0.040)	-
$\hat{\alpha}_{22}$	0.177 (0.045)	0.263 (0.041)
$\hat{\alpha}_{23}$	0.141 (0.048)	-
$\hat{\alpha}_{31}$	0.062 (0.039)	-
$\hat{\alpha}_{32}$	0.108 (0.039)	-
$\hat{\alpha}_{33}$	0.131 (0.047)	0.179 (0.045)

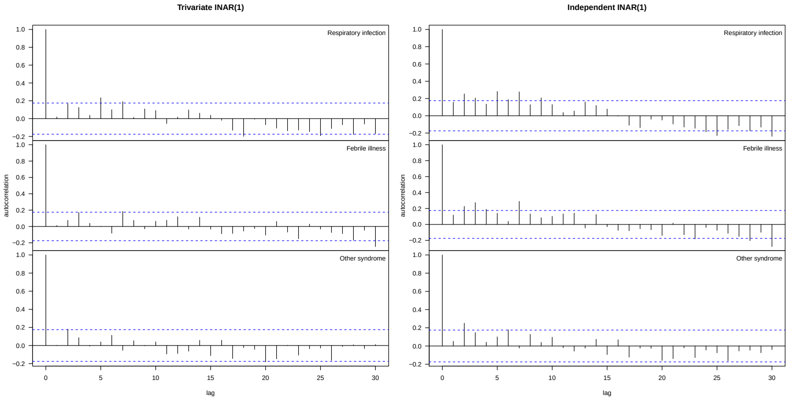
Application

Results: Maximum likelihood estimates (standard errors) of the regression parameters obtained from fitting three independent Poisson INAR(1) or a trivariate INAR(1) regression model with independent Poisson innovations to the historical data.

regression parameters	trivariate INAR(1)	indepedent INAR(1)
$\hat{\beta}_{10}$	1.190 (0.153)	1.506 (0.099)
$\hat{\beta}_{11}$	-0.255 (0.145)	-0.278 (0.110)
$\hat{\beta}_{12}$	-0.359 (0.118)	-0.222 (0.078)
$\hat{\beta}_{13}$	-0.218 (0.098)	-0.140 (0.073)
$\hat{\beta}_{20}$	1.197 (0.135)	1.496 (0.096)
$\hat{\beta}_{21}$	-0.267 (0.133)	-0.118 (0.102)
$\hat{\beta}_{22}$	0.411 (0.121)	0.156 (0.070)
$\hat{\beta}_{23}$	0.548 (0.110)	0.296 (0.068)
$\hat{\beta}_{30}$	0.990 (0.155)	1.246 (0.109)
$\hat{\beta}_{31}$	0.047 (0.142)	0.046 (0.113)
$\hat{\beta}_{32}$	-0.174 (0.099)	-0.112 (0.072)
$\hat{\beta}_{33}$	-0.198 (0.090)	-0.146 (0.071)

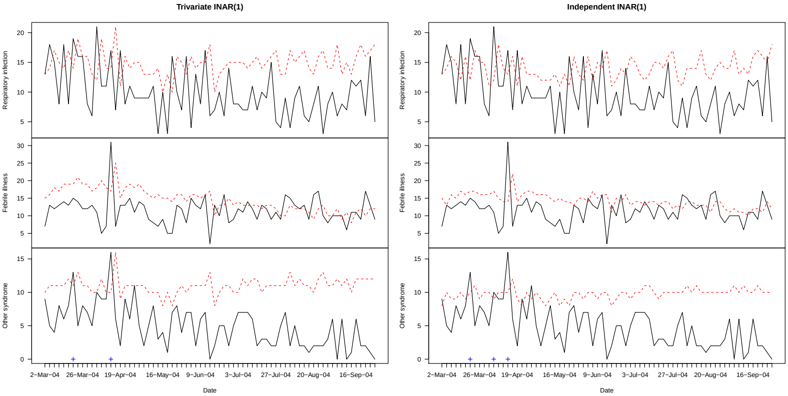
Application

Results: Plots of the autocorrelations of the residuals obtained by the trivariate INAR(1) (left panel) and the independent INAR(1) (right panel) regression models.



Application

Results: Surveillance plots. Statistical alarms (blue crosses) are raised when at least two series exceed the upper bounds of the corresponding 99% prediction intervals (red dashed lines).



Final remarks

- We suggest a multivariate INAR(1) approach suitable for joint modeling of multivariate surveillance data. The introduced model admits a realistic epidemiological interpretation and accounts for overdispersion that is typical with surveillance data.
- Emphasis has been put on the case of independent Poisson innovations but other discrete distributions, as e.g. the negative binomial, can also be considered instead.
- A series of interesting points should be further exploited, as e.g. updating the data basis for the model fit in a regular basis and keep the newest obs. only for building the model or downweight past outbreaks by suitable adjustments (Noufaily et al, 2013, SIM).