Multivariate integer-valued autoregressive models applied to earthquake counts

Mathieu Boudreault, Arthur Charpentier

UQAM

February 3rd, 2012

Motivation

- Major earthquakes (and/or tsunamis) can cause very significant damage (e.g. Japan in March 2011));
- Earthquake risk is usually assumed by reinsurance companies that diversify their risk across the planet;
- It is known in geophysics and seismology that strong earthquakes can trigger smaller earthquakes at **large** distances;
  - These **smaller** earthquakes that occur very remotely *generally* do not create damage;
  - Better diversification;
- What if very large earthquakes increased the global seismic risk?
  - In other words, if a magnitude (M) 8 earthquake occurs in Japan, can it increase the risk of a M$>$6 earthquake in California or in the Middle-East ?
- That would be very bad news for reinsurance companies and the insurance industry as a whole
  - Systematic (undiversifiable) seismic risk;

Motivation

- It is only very recently that a study confirmed that the global seismic risk is not increased after a very large earthquake;
  - "*Thus, we conclude that the regional hazard of larger earthquakes is increased after a mainshock, but the global hazard is not*"
  - Citation from Tom Parsons & Aaron Velasco (2011) in Nature Geoscience;
- Good news for risk management: reinsurers can still diversify earthquake risk across various regions across the Earth !
- **Goals of the paper**:
  - Further investigate the multivariate INAR model;
  - Extend the bivariate INAR model of Pedeli & Karlis (2011);
  - Perform an empirical study on the dynamics of earthquake counts with various applications;
    - Size and direction of interactions among tectonic plates, and among different magnitudes;

Background on integer-valued autoregression

- Time series of counts occur in various situations in insurance, finance, epidemiology, sports statistics, etc.;
  - Gourieroux and Jasiak (2004), Boucher, Denuit, Guillen (2008) in car insurance;
- Cannot extend a basic AR(1) process to discrete variables for obvious reasons;
- **Solutions**:
  - Model the mean as a latent process (see Ferland, Latour, Oraichi (2006) for example);
  - Thinning operators (Steutel and van Harn (1979));

Background on integer-valued autoregression

- Integer-valued autoregression (INAR)

$$p \circ N = Y_1 + \cdots + Y_N \text{ if } N \neq 0, \text{ and } 0 \text{ otherwise}$$

  - N is a random variable with values in $\mathbb{N}$ and $p \in [0, 1]$
  - $Y_1, Y_2, \cdots$ are i.i.d. Bernoulli variables, independent of N, with $\mathbb{P}(Y_i = 1) = p$.
  - $p \circ N$ has a binomial distribution with parameters N and p;

- INAR(1) (Al-Osh and Alzaid (1987) or McKenzie (1985))

$$N_t = p \circ N_{t-1} + \varepsilon_t = \sum_{i=1}^{N_{t-1}} Y_i + \varepsilon_t$$

- INAR(q) (Du and Li (1991)):

$$N_t = \sum_{j=1}^{q} (p_j \circ N_{t-j}) + \varepsilon_t$$

Multivariate integer-valued autoregression

- Not much on the issue: Franke & Subba Rao (1993) and Latour (1997);
  - Have developed some theoretical results on multivariate INAR based on binomial thinning or Steutel and van Harn operators;
- $d-$variate INAR of order 1 (binomial thinning in the line of Franke & Subba Rao (1993)):

$$\begin{bmatrix} N_t^{(1)} \\ N_t^{(2)} \\ \\ N_t^{(d)} \end{bmatrix} = \begin{bmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,d} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ p_{d,1} & p_{d,2} & \cdots & p_{d,d} \end{bmatrix} \circ \begin{bmatrix} N_{t-1}^{(1)} & N_{t-1}^{(2)} & \cdots & N_{t-1}^{(d)} \\ N_{t-1}^{(1)} & N_{t-1}^{(2)} & \cdots & N_{t-1}^{(d)} \\ \vdots & \vdots & \ddots & \vdots \\ N_{t-1}^{(1)} & N_{t-1}^{(2)} & \cdots & N_{t-1}^{(d)} \end{bmatrix} + \begin{bmatrix} \varepsilon_t^{(1)} \\ \varepsilon_t^{(2)} \\ \vdots \\ \varepsilon_t^{(d)} \end{bmatrix}$$

  where the thinning operator is applied element-by-element;
- **Further assumptions**:
  - All Bernoulli r.v. are i.i.d.;
  - ε_t follows a d-variate discrete distribution;

Multivariate integer-valued autoregression

- Franke & Subba Rao (1993) have derived for a d-variate INAR:
  - Conditions for stationarity: largest eigenvalue of the $\mathbf{P}$ matrix is strictly smaller than 1;
  - Conditional maximum likelihood estimates are asymptotically normal (and unbiaised), likelihood ratio tests exist;
- In the line of Franke & Subba Rao (1993), i.e. d-variate INAR with binomial thinning, we have derived the following results:
  - Unconditional mean;
  - Autocorrelation matrices;
  - Forecasting results;

Some properties of the model

- **Definitions**:
  - $\mathbf{P}$ matrix $= \left[p_{i,j}\right], i,j = 1,2,...,d$;
  - $\mathbf{N}_t$ (column) vector $= \left[N_t^{(j)}\right], j = 1,2,...,d$
  - $\boldsymbol{\lambda} \equiv \mathrm{E}\left[\varepsilon_t\right]$ and $\boldsymbol{\Lambda} \equiv \mathrm{Var}\left[\varepsilon_t\right]$;
  - $\boldsymbol{\mu} \equiv \mathrm{E}\left[\mathbf{N}_t\right]$ and $\gamma\left(h\right) \equiv \mathrm{Cov}(\mathbf{N}_t, \mathbf{N}_{t-h})$;
  - $\boldsymbol{\Delta} \equiv$diag$(\mathbf{V}\boldsymbol{\mu})$ and $\mathbb{I}$ is the identity matrix;
- Then, we obtain

$$\begin{aligned} \boldsymbol{\mu} &= \left[\mathbb{I} - \mathbf{P}\right]^{-1} \boldsymbol{\lambda} \\ \gamma(h) &= \mathbf{P}^h \gamma(0) \end{aligned}$$

where $\gamma\left(0\right)$ is the solution of

$$\gamma(0) = \mathbf{P}\gamma(0)\mathbf{P}' + (\boldsymbol{\Delta} + \boldsymbol{\Lambda})$$

which can be solved numerically using a fixed-point algorithm.

Forecasting

- **Definitions**:
  - $\mathbf{V}$ is a matrix with entries $p_{i,j}\left(1-p_{i,j}\right)$, $i,j=1,2,...,d$;
  - $V_h(\mathbf{N}_t) \equiv \mathrm{Var}\left[\mathbf{N}_{t+h}|\mathbf{N}_t\right]$;
- **Conditional moments**:
  - Expectation: by recursion

$$\mathrm{E}\left[\mathbf{N}_{t+h}|\mathbf{N}_t\right] = \mathbf{P}^h\mathbf{N}_t + \left(\mathbb{I}+\mathbf{P}+\cdots+\mathbf{P}^{h-1}\right)\boldsymbol{\lambda}$$

  - Variance: recursion and decomposition of the variance. Then $V_h(\mathbf{N})$ is defined recursively by

$$V_1(\mathbf{N}) = \mathrm{diag}(\mathbf{VN}) + \boldsymbol{\Lambda}$$

and

$$V_h(\mathbf{N}) = \mathbb{E}[V_{h-1}(\mathbf{P}\circ\mathbf{N}+\varepsilon)|\mathbf{N}] + \mathbf{P}^{h-1}[\mathrm{diag}(\mathbf{VN})+\boldsymbol{\Lambda}](\mathbf{P}^{h-1})'.$$

  - Closed-form expressions when $\mathbf{P}$ is diagonal;

Bivariate INAR

- Proposed model:

$$\begin{aligned} N_t^{(1)} &= p_{1,1} \circ N_{t-1}^{(1)} + p_{1,2} \circ N_{t-1}^{(2)} + \varepsilon_t^{(1)} \\ N_t^{(2)} &= p_{2,1} \circ N_{t-1}^{(1)} + p_{2,2} \circ N_{t-1}^{(2)} + \varepsilon_t^{(2)} \end{aligned}$$

- Pedeli & Karlis (2011a, b): diagonal model (i.e. $p_{2,1} = p_{1,2} = 0$)
  - They have derived various properties of the model and applied the model to daytime and nighttime road accidents in Netherlands;
  - Generalization to non-diagonal **P** matrix is straightforward
  - Earthquake counts: adds spatial contagion of order 1;
- We derive:
  - Granger causality tests;
  - Closed-form expressions for likelihood function, unconditional mean and variance, autocorrelations and cross autocorrelations with Poisson errors;

Some results

- Granger causality tests:
  - Form of likelihood ratio test over the diagonal model;
  - Example: model with $(p_{1,2} = 0$ and $p_{2,1} \neq 0)$ over diagonal model checks if $N_t^{(1)}$ causes $N_t^{(2)}$;
- Likelihood function:
  - Based upon the convolution of two independent but **not** identically distributed binomial r.v.;
- Unconditional mean:

$$\mu_1 = \frac{(1-p_{2,2})\lambda_1 + p_{1,2}\lambda_2}{(1-p_{1,1})(1-p_{2,2}) - p_{2,1}p_{1,2}}$$
$$\mu_2 = \frac{(1-p_{1,1})\lambda_2 + p_{2,1}\lambda_1}{(1-p_{1,1})(1-p_{2,2}) - p_{2,1}p_{1,2}}$$

- Auto and cross autocorrelations (see paper);

Background on earthquakes

- Fault: fracture or discontinuity in a volume of rock located underground;
  - Example: tectonic faults, San Andreas fault, Laurentian fault;
- Rocks tend to very slowly move because layers of the Earth underneath are softer;
- This creates an accumulation of energy along a geological fault: earthquakes occur after the sudden release of this energy;
- We are sitting on a layer of rocks that is divided in many sections: tectonic plates;
- Depending on the source, there are 15-20 major tectonic plates;
- Earthquakes occur at the limits of tectonic plates (interplate) or within a plate (intraplate) (due to other faults);
- When an earthquake occurs, seismic wave travels on a long distance, fragilizes stability on other faults, creating aftershocks;
- Foreshocks and aftershocks are smaller size earthquakes that precede or follow a mainshock (main event);

Tectonic plates

- Mapping of tectonic plates

- 16 (17) tectonic plates:
  - Japan is at the limit of 4 tectonic plates (Pacific, Okhotsk, Philippine and Amur);
  - California is at the limit of the Pacific, North American and Juan de Fuca (not shown) plates, and is close to the San Andreas fault;

Data

- Map of tectonic plates (ArcGIS shapefiles): Dept of Geography, U Colorado (Boulder);
- List of earthquakes: Advanced National Seismic System (ANSS) Composite Earthquake Catalog;
  - Issues: databases added, improvements of seismic instruments, addition of seismic monitoring stations;
  - Statistical tests of changes of structure to determine cutoff dates;
- 1965-2011 for magnitude (M) $>$5 earthquakes (70 000 events);
- 1992-2011 for M$>$6 earthquakes (3000 events);
- To count the number of earthquakes, we used time ranges of 3, 6, 12, 24, 36 and 48 hours;
- Approximately 8500 to 135 000 periods of observation;

Quality of fit

- **Remarks**:
  - Earthquakes that occur within h hours (i.e. in $[0, h]$) count toward dependence in the noise;
  - Earthquakes that occur in $[h, 2h]$ count toward first order space-time contagion;
  - Earthquakes that occur in $[kh, (k+1)\, h]$ count toward the k-th order of space-time contagion;
- Five models investigated:
  - Poisson with independent noise;
  - Poisson with dependent noise;
  - Independent INARs (time contagion of order 1)
  - Diagonal BINAR of Pedeli & Karlis (2011a, b) (time contagion of order $1 +$ dependent noise);
  - Proposed model (space-time contagion of order $1 +$ dependent noise);
- 136 possible pairs of tectonic plates, 6 time frequencies $= 816$ estimations for each model;

Quality of fit - Results

- For all pairs of tectonic plates, at all frequencies, autoregression **in time** is important (very high statistical significance);
  - Long sequence of zeros, then mainshocks and aftershocks;
  - Rate of aftershocks decreases exponentially over time (Omori's law);
- For 7-13% of pairs of tectonic plates, diagonal BINAR has significant better fit than independent INARs;
  - Contribution of dependence in noise;
  - Spatial contagion of order 0 (within h hours);
  - Contiguous tectonic plates;
- For 7-9% of pairs of tectonic plates, proposed BINAR has significant better fit than diagonal BINAR;
  - Contribution of spatial contagion of order 1 (in time interval $[h, 2h]$);
  - Contiguous tectonic plates;
- **Conclusion**: for approximately 90%, there is no significant spatial contagion for M$>$5 earthquakes;
  - When there is, in most cases, plates are contiguous (small distance);
  - Add evidence to Parsons & Velasco (2011);

Analysis of pairs of tectonic plates (Interpretation)

- Okhotsk (#1) and West Pacific (#2) tectonic plates (underneath part of Japan): at 24-hour frequency we have

$$\begin{bmatrix} N_t^{(1)} \\ N_t^{(2)} \end{bmatrix} = \begin{bmatrix} 0.0817 & 0.0280 \\ 0.1060 & 0.1552 \end{bmatrix} \circ \begin{bmatrix} N_{t-1}^{(1)} & N_{t-1}^{(2)} \\ N_{t-1}^{(1)} & N_{t-1}^{(2)} \end{bmatrix} + \begin{bmatrix} \varepsilon_t^{(1)} & (0.1620) \\ \varepsilon_t^{(2)} & (0.4261) \end{bmatrix}$$

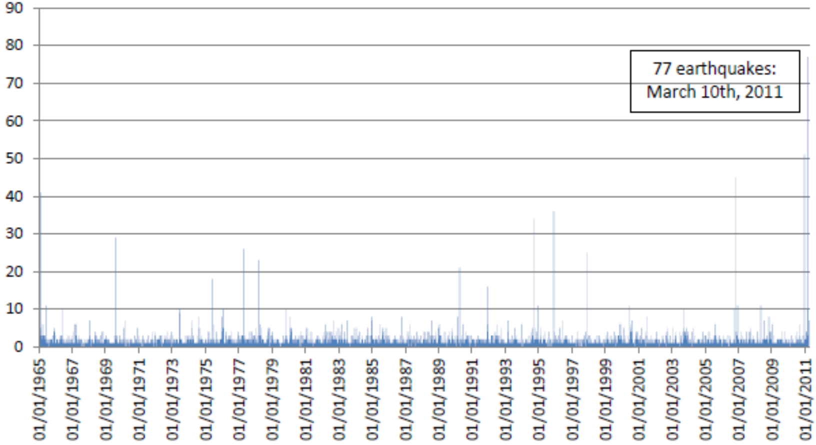
- Unconditional mean of earthquakes: Okhotsk $= 0.1926$ / day, West Pacific $= 0.5285$ / day;
- Suppose n (m) earthquakes observed on plate # 1 (#2);
- Mean number of earthquakes on Okhotsk plate:

$$0.0817n + 0.028m + 0.162 \text{ (proposed model)}$$
$$0.0922n + 0.1748 \text{ (diagonal model)}$$

- Given that the West Pacific plate is very active (a lot of days with zeros, a few days with tens of earthquakes), diagonal model (ignoring cross autocorrelation) can mean severe underestimation of the number of earthquakes;

Analysis of pairs of tectonic plates (Interpretation)

- Daily number of earthquakes (M>5) on the West Pacific tectonic plate



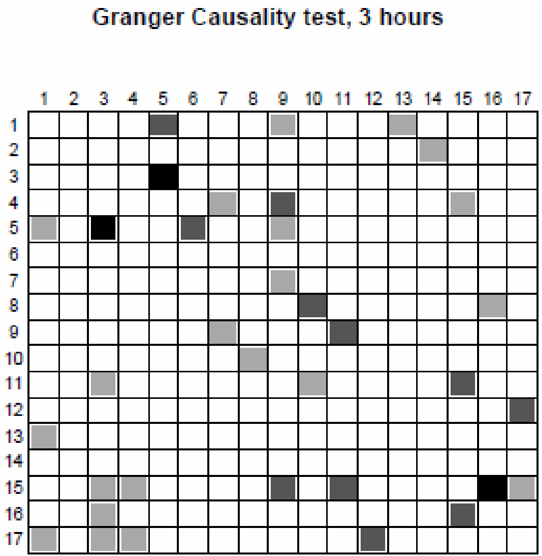
Foreshocks and aftershocks

- For each tectonic plate, 2 sets of data: medium-size earthquakes ($5<M<6$) and large-size earthquakes ($M>6$);
- Space-time contagion: "space" is now the variation in magnitude;
- Want to know if $M>6$ earthquakes help explain number of ($5<M<6$) earthquakes and vice-versa;
- Same five models, over the 17 tectonic plates, and 6 observation frequencies;
- Diagonal BINAR has significant better fit over all three simpler models;
- Due to the (very largely documented) presence of foreshocks and aftershocks, should expect proposed BINAR to be (statistically) significant;
  - It is indeed the case for the very large majority of plates and frequencies;

Granger causality tests

- Investigate direction of relationship (which one causes the other, or both);
- Pairs of tectonic plates:
  - Uni-directional causality: most common for contiguous plates (North American causes West Pacific, Okhotsk causes Amur);
  - Bi-directional causality: Okhotsk and West Pacific, South American and Nasca for example;
- Foreshocks and aftershocks:
  - Aftershocks much more significant than foreshocks (as expected);
  - Foreshocks announce arrival of larger-size earthquakes;
  - Foreshocks significant for Okhotsk, West Pacific, Indo-Australian, Indo-Chinese, Philippine, South American;

Granger causality matrix



Risk management

- Interested in computing $\mathbb{P}\left(\sum_{k=1}^{T}\left(N_k^{(1)}+N_k^{(2)}\right)\geq n \middle| \mathcal{F}_0\right)$ for various values of T (time horizons) and n (tail risk measure);
  - Total number of earthquakes on a set of two tectonic plates;
- 100 000 simulated paths of diagonal and proposed BINAR models;
  - Use estimated parameters of both models;
  - Pair: Okhotsk and West Pacific;
- Scenario: on a 12-hour period, 23 earthquakes on Okhotsk and 46 earthquakes on West Pacific (second half of March 10th, 2011);
- Results on next slide

Risk management

Diagonal model				
n / days	**1 day**	**3 days**	**7 days**	**14 days**
5	0.9680	0.9869	0.9978	0.9999
10	0.5650	0.7207	0.8972	0.9884
15	0.1027	0.2270	0.4978	0.8548
20	0.0067	0.0277	0.1308	0.4997
Proposed model				
n / days	**1 day**	**3 days**	**7 days**	**14 days**
5	0.9946	0.9977	0.9997	1.0000
10	0.8344	0.9064	0.9712	0.9970
15	0.3638	0.5288	0.7548	0.9479
20	0.0671	0.1573	0.3616	0.7256

Conclusion

- We have presented additional results regarding the multivariate INAR;
- We have extended the diagonal BINAR of Pedeli & Karlis (2011a,b) and derived other results;
- Application to earthquakes:
  - Pairs of tectonic plates: spatial contagion of order 1 is important for contiguous plates.
  - Foreshocks and aftershocks: cross autocorrelation of order 1 is significant (aftershocks especially)
- Risk management:
  - For periods following an active day, lack of spatial contagion may seriously understate number of events;
  - Impacts over the long-run are unclear for the moment;