# Approximately Linear INGARCH Models for Spatio-Temporal Counts

Christian H. Weiß [1], Malte Jahn [1], Hee-Young Kim [2]

[1]Department of Mathematics and Statistics, Helmut Schmidt University, Hamburg, Germany

[2]Division of Big Data Science, Korea University, Sejong, South Korea

May 23, 2021

# Outline

# 1.Introduction

## On the non-negative integers, $\mathbb{N}_0 = \{0, 1, \ldots\}$

- The most common ARMA-like models for count time series are the so-called INARMA models ("IN" like "integer"), which imitate the ARMA recursion by using types of "thinning operators",
- the integer-valued generalized autoregressive conditional heteroskedasticity (INGARCH) models
- Both model families lead to an autocorrelation structure similar to those of the ordinary ARMA models, i. e., their autocorrelation function (ACF) satisfies some kind of Yule–Walker equations.

## INGARCH models for unbounded counts, $\mathbb{N}_0 = \{0, 1, \ldots\}$

- INGARCH models are defined with respect to the conditional mean $M_t = E[X_t|\mathcal{F}_{t-1}]$ with $\mathcal{F}_{t-1}$ being the $\sigma$-field generated by $\{(X_{t-1}, M_{t-1}), (X_{t-2}, M_{t-2}), \ldots\}$.
- Ferland, R., Latour, A., Oraichi, D. (2006) Integer-valued GARCH processes. *Journal of Time Series Analysis* **27**(6), 923–942.
- The (exactly linear) INGARCH model of order $(p, q) \in \mathbb{N} \times \mathbb{N}_0$ is defined by the recursive scheme

$$M_t = a_0 + \sum_{i=1}^{p} a_i \, X_{t-i} + \sum_{j=1}^{q} b_j \, M_{t-j}, \tag{1}$$

where the constraints $a_0 > 0$ and $a_1, \ldots, a_p, b_1, \ldots, b_q \geq 0$ have to be satisfied to ensure a truly positive conditional mean.

# INGARCH models for unbounded counts, $\mathbb{N}_0 = \{0, 1, \ldots\}$

- Zhu, F. (2011) A negative binomial integer-valued GARCH model. *Journal of Time Series Analysis* **32**(1), 54–67.
- negative-binomial distribution $\text{NB}(\nu, \pi_t)$ with $\pi_t^{-1} = 1 + M_t/\nu$, the additional parameter $\nu > 0$ serves as a dispersion parameter.

## BINGARCH model for bounded counts, $\{0, \ldots, n\}$

- Ristić, M.M., Weiß, C.H., Janjić, A.D. (2016) A binomial integer-valued ARCH model. *International Journal of Biostatistics* **12**(2), 20150051.
- the normalized conditional mean $P_t = \frac{1}{n} M_t$, where $M_t = E[X_t | \mathcal{F}_{t-1}]$

$$P_t = a_0 + \sum_{i=1}^{p} a_i X_{t-i}/n + \sum_{j=1}^{q} b_j P_{t-j} \qquad (2)$$

  with the additional constraint $a_0 + \sum_{i=1}^{p} a_i + \sum_{j=1}^{q} b_j < 1$.
- Then the count $X_t | \mathcal{F}_{t-1} \sim \text{Bin}(n, P_t)$.

# 2. Linear Spatio-Temporal INGARCH Models

- Held, L., Höhle, M., Hofmann, M. (2005) A statistical framework for the analysis of multivariate infectious disease surveillance counts. *Statistical Modelling* **5**(3), 187–199.
- Paul, M., Held, L., Toschke, A.M. (2008) Multivariate modelling of infectious disease surveillance data. *Statistics in Medicine* **27**(29), 6250–6267.
- The starting point for the modeling of spatio-temporal counts is the extension of INGARCH(1) model.
- The $i$th component of the count vector $\boldsymbol{X}_t = (X_{t,1}, \ldots, X_{t,m})^\top \in \mathbb{N}_0^m$ expresses the number of (new) cases at time $t$ in unit $i$.
- The conditional mean $M_{t,i}$ of the $i$th unit, $E(X_{t,i}|\mathcal{F}_{t-1})$, is assumed to satisfy

$$M_{t,i} \;=\; \lambda_i\, X_{t-1,i} + \phi_i \textstyle\sum_{j\neq i} w_{ji}\, X_{t-1,j} \;+\; \gamma_i, \qquad (3)$$

where the AR-part of (3) is interpreted as the epidemic component (driven by $\boldsymbol{X}_{t-1}$), and the intercept term $\gamma_i$ as the endemic component.

## Spatial weight matrix W for spatio-temporal INGARCH model

- In Paul et al.(2008)
  - $w_{ji} = 0$ if unit $j$ is not a neighbor of region $i$, and otherwise $w_{ji} = 1/n_j$ (normalized weight), where $n_j$ expresses the total number of neighbors of unit $j$.
- Meyer, S., Held, L. (2014) Power-law models for infectious disease spread. *Annals of Applied Statistics* **8**(3), 1612–1639.
  - power-law weighting based on the path distance $o_{ji}$.
  - Here, $o_{ji} = k$ if the shortest route from $i$ to $j$ is of length $k$
  - $o_{ii} = 0$
- Bracher, J., Held, L. (2020) Endemic-epidemic models with discrete-time serial interval distributions for infectious disease prediction. *International Journal of Forecasting*, forthcoming.
  - $w_{ji} = (1 + o_{ji})^{-d} / \sum_{k=1}^{m} (1 + o_{jk})^{-d}$ with decay parameter $d > 0$.

# 3. Approximately Linear (B)INGARCH Models

## softplus INGARCH

- Weiß, C.H., Zhu, F., Hoshiyar, A. (2022) Softplus INGARCH models. *Statistica Sinica* **32**(3), forthcoming.
- $$\mathrm{sp}_c(x) \;=\; c\,\ln\big(1+\exp(x/c)\big) \quad \text{with adjustment parameter } c>0, \tag{4}$$
  which becomes piecewise linear for $c \to 0$, see Figure 1 (a)
- $$M_t \;=\; sp_c\Big(\alpha_0 + \textstyle\sum_{i=1}^{p} \alpha_i\, X_{t-i} + \sum_{j=1}^{q} \beta_j\, M_{t-j}\Big), \tag{5}$$
  where now, $\alpha_1, \ldots, \alpha_p, \beta_1, \ldots, \beta_q$ might also become negative (possibly leading to negative ACF values).
- Possible constraints to ensure the existence of a stationary solution are $\sum_{i=1}^{p} \max\{0, \alpha_i\} + \sum_{j=1}^{q} \max\{0, \beta_j\} < 1$ and $\sum_{j=1}^{q} |\beta_j| < 1$,
- For the adjustment parameter, the default choice is $c = 1$, but to achieve a closer approximation to linearity, smaller values such as $c = 0.5$ or $c = 0.25$ are sometimes preferable.

## Softplus function and softclipping function

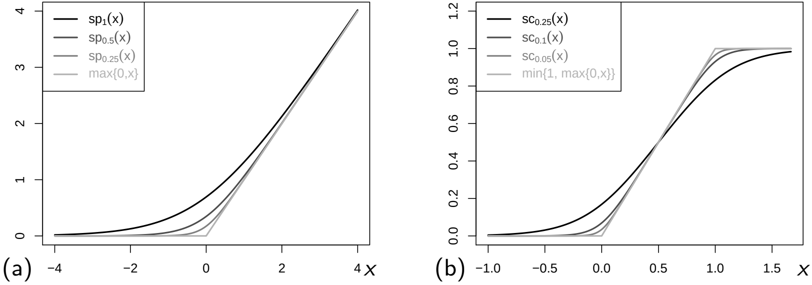



**Figure:** Plots of (a) softplus function and (b) soft clipping function against $x$.

# soft-clipping BINGARCH

- Weiß, C.H., Jahn, M. (2021) Soft-clipping INGARCH models for time series of bounded counts. *Working Paper*.
-
$$\mathrm{sc}_c(x) \; = \; c \, \ln\left(\frac{1+\exp(\frac{x}{c})}{1+\exp(\frac{x-1}{c})}\right) \quad \text{with adjustment parameter } c>0 \tag{6}$$
-
$$P_t \; = \; \mathrm{sc}_c\Big(\alpha_0 + \textstyle\sum_{i=1}^{p} \alpha_i\, X_{t-i}/n + \sum_{j=1}^{q} \beta_j\, P_{t-j}\Big). \tag{7}$$
- It is reasonable to require $\alpha_0 \in (0; 1+p+q)$, while the remaining constraints are as before.
- The value of $c$ should be chosen close to zero to achieve a good approximation to linearity, such as $c = 0.05$, see Figure 1 (b).

# 4. Approximately Linear Spatio-Temporal (B)INGARCH Models

## Approximately linear spatio-temporal (B)INGARCH models

- Combining the softclipping approach with the spatio-temporal, we obtain an approximately linear spatio-temporal INGARCH model

$$\begin{aligned} M_{t,i} \;=\; \mathrm{sp}_c \Big(\alpha_0 + \textstyle\sum_{k=1}^{p} \alpha_k\, X_{t-k,i} + \sum_{r=1}^{m} \lambda_r \sum_{j\neq i} w_{ij}\, X_{t-r,j} \\ + \textstyle\sum_{l=1}^{q} \beta_l\, M_{t-l,i} + \sum_{h=1}^{s} \phi_h \sum_{j\neq i} w_{ji}\, M_{t-h,j} \Big) \end{aligned} \tag{8}$$

- Combining the softclipping approach with the the spatio-temporal model(normalized mean), we obtain an approximately linear spatio-temporal INGARCH model for bounded counts by defining:

$$\begin{aligned} P_{t,i} \;=\; \mathrm{sc}_c \Big(\alpha_0 + \textstyle\sum_{k=1}^{p} \alpha_k\, X_{t-k,i}/n + \sum_{r=1}^{m} \lambda_r \sum_{j\neq i} w_{ij}\, X_{t-r,j}/n \\ + \textstyle\sum_{l=1}^{q} \beta_l\, P_{t-l,i} + \sum_{h=1}^{s} \phi_h \sum_{j\neq i} w_{ji}\, P_{t-h,j} \Big) \end{aligned} \tag{9}$$

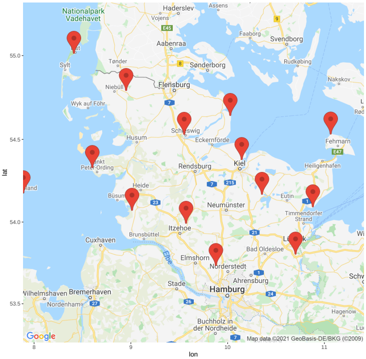
- In equation 8 (9) the conidtional mean $M_{t,i}$ (success probability $P_{t,i}$) depends approximately linear on a unit's own and its neighbors' past.

# 5. Data Analysis

## All 14 stations in northern Germany

- The present sample consists of 14 stations in northern Germany which report hourly cloud coverage data in terms of the share of the visible sky that is covered by clouds in the time period 2009-2019.
- The observations take integer values from 0 (no clouds) to 8 (sky fully overcast).
- Irregular observations are marked with flag "9" (no data) and "-1" (sky not observable).
- These flags appear relatively often and we treat them as "NA". We identify a time period of 650 consecutive hours where all of the stations report regular data in January/February 2017.

## All 14 stations in northern Germany



## Spatial Weight Matrix, W

- The locations of the weathers is given by latitude and longitude coordinates.
- We use those to calculate the great circle distances $d_{ij}$ according to the spherical law of the cosine.
- The nearest neighbor approach, in which we define the spatial weight matrix $w_{ij} = 1/K$ if j is among the K nearest neighbors of i and $w_{ij} = 0$, otherwise.

## Some results

- The results for the estimation of softclipping INGARCH(1,1,1,1) model
- A spatio-temporal soft clipping INGARCH(1,1,1,1) model

$$
\begin{aligned}
P_{t,i} \;=\; \mathrm{sc}_c\Big(\alpha_0 \quad & +\alpha X_{t-1,i}/n + \lambda \textstyle\sum_{j\neq i} w_{ij}\, X_{t-1,j}/n \\
& +\beta P_{t-1,i} + \phi \textstyle\sum_{j\neq i} w_{ji}\, P_{t-1,j}\Big)
\end{aligned}
$$

- 

| | $\alpha_0$ | $\alpha$ | $\lambda$ | $\beta$ | $\phi$ |
|---|---|---|---|---|---|
| Est. coeff. | 0.0693 | 0.5382 | 0.4086 | 0.0006 | 0.0015 |
| Approx. SE | 0.0051 | 0.0080 | 0.0096 | 0.0164 | 0.0167 |

## Discussion

- What we did
  - propose approximately linear spatio-temporal INGARCH models
  - Data analysis, which consists of 14 stations in northern Germany, hourly cloud coverage data, in the time period of 2009-2019.
- What we are doing now is
  - Method imputation that utilizes the spatio-temporal nature of the data to accurately and efficiently impute missing values.
  - Network Analysis
    Mirko Armillotta, Konstantinos Fokianos (2021) Poisson Network Autoregression *arXiv:2104.06296*
  - Moran's I is the most popular spatial test statistic, but its inability to incorporate heterogeneous populations has been long recognized.

# References1

Aldor-Noiman, S., Brown, L.D., Fox, E.B., Stine, R.A. (2016) Spatio-temporal low count processes with application to violent crime events. *Statistica Sinica* **26**(4), 1587–1610.

Bracher, J., Held, L. (2020) Endemic-epidemic models with discrete-time serial interval distributions for infectious disease prediction. *International Journal of Forecasting*, forthcoming.

Clark, N.J., Dixon, P.M. (2021) A class of spatially correlated self-exciting models. *arXiv:1805.08323v3*.

Ferland, R., Latour, A., Oraichi, D. (2006) Integer-valued GARCH processes. *Journal of Time Series Analysis* **27**(6), 923–942.

Fokianos, K. (2011) Some recent progress in count time series. *Statistics* **45**(1), 49–58.

Fokianos, K., Støve, B., Tjøstheim, D., Doukhan, P. (2020) Multivariate count autoregression. *Bernoulli* **26**(1), 471–499.

Ghodsi, A. (2015) Conditional maximum likelihood estimation of the first-order spatial integer-valued autoregressive (SINAR(1,1)) model. *Journal of the Iranian Statistical Society* **14**(2), 15–36.

## References2

Glaser, S. (2017) A review of spatial econometric models for count data. *Hohenheim Discussion Papers* 19-2017, University of Hohenheim, Germany.

Heinen, A., Rengifo, E. (2003) Multivariate modelling of time series count data: an autoregressive conditional Poisson model. *CORE Discussion Paper* 2003/25, University of Louvain, Belgium.

Held, L., Höhle, M., Hofmann, M. (2005) A statistical framework for the analysis of multivariate infectious disease surveillance counts. *Statistical Modelling* **5**(3), 187–199.

Held, L., Paul, M. (2012) Modeling seasonality in space-time infectious disease surveillance data. *Biometrical Journal* **54**(6), 824–843.

Huda, N.M., Mukhaiyar, U., Pasaribu, U.S. (2021) The approximation of GSTAR model for discrete cases through INAR model. *Journal of Physics: Conference Series* **1722**, 012100.

Kharin, Y., Zhurak, M. (2015) Statistical analysis of spatio-temporal data based on Poisson conditional autoregressive model. *Informatica* **26**(1), 67–87.

Klimek, M.D., Perelstein, M. (2020) Neural network-based approach to phase space integration. *SciPost Physics* **9**(4), 053.

# References3


Li, L., Zhu, L., Sui, D.Z. (2007) A GIS-based Bayesian approach for analyzing spatial-temporal patterns of intra-city motor vehicle crashes. *Journal of Transport Geography* **15**(4), 274–285.

Mei, H., Eisner, J. (2017) The neural Hawkes process: a neurally self-modulating multivariate point process. In von Luxburg et al. (eds): *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*, Curran Associates Inc., 6757–6767.

Meyer, S., Held, L. (2014) Power-law models for infectious disease spread. *Annals of Applied Statistics* **8**(3), 1612–1639.

Paul, M., Held, L., Toschke, A.M. (2008) Multivariate modelling of infectious disease surveillance data. *Statistics in Medicine* **27**(29), 6250–6267.

Raftery, A.E. (1985) A model for high-order Markov chains. *Journal of the Royal Statistical Society, Series B* **47**(3), 528–539.

Ristić, M.M., Weiß, C.H., Janjić, A.D. (2016) A binomial integer-valued ARCH model. *International Journal of Biostatistics* **12**(2), 20150051.

Weiß, C.H. (2018) *An Introduction to Discrete-Valued Time Series*. John Wiley & Sons, Inc., Chichester.

## References4

Weiß, C.H. (2021) Stationary count time series models. *WIREs Computational Statistics* **13**(1), e1502.

Weiß, C.H., Jahn, M. (2021) Soft-clipping INGARCH models for time series of bounded counts. *Working Paper*.

Weiß, C.H., Zhu, F., Hoshiyar, A. (2022) Softplus INGARCH models. *Statistica Sinica* **32**(3), forthcoming.

Zhu, F. (2011) A negative binomial integer-valued GARCH model. *Journal of Time Series Analysis* **32**(1), 54–67.

Thank you for your attention!

Thank you for your time!