MATH STAT

HELMUT SCHMIDT UNIVERSITÄT
Universität der Bundeswehr Hamburg

INGARCH Models for Spatio-Temporal Counts or Ordinal Data

Malte Jahn[1], Christian H. Weiß[2], and Hee-Young Kim[3]

[2] Department of Mathematics and Statistics, Helmut Schmidt University, Hamburg, Germany; **weissc@hsu-hh.de**.

[1] Merseburg University of Applied Sciences, Merseburg, Germany [3] Korea University, Sejong, South Korea

Introduction

During the last forty years, (univariate) count processes $(X_t)_{\mathbb{N}}$, i.e., where the X_t have a range contained in $\mathbb{N}_0 = \{0,1,\dots\}$, received a lot of interest in the research literature [W18]. Many model families have been proposed for $(X_t)_{\mathbb{N}}$, where (among others) the integer-valued generalized autoregressive conditional heteroscedasticity (INGARCH) models are quite popular in applications. The exactly linear INGARCH model of [FLO06] has a conditional Poisson distribution, $X_t|X_{t-1},\dots \sim \text{Poi}(M_t)$, with

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

while for bounded counts with range $\{0,\dots,n\}$, [RWJ16] propose a conditional binomial distribution, $X_t|X_{t-1},\dots \sim \text{Bin}(n, P_t)$, with

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

In both cases, strict parameter constraints necessary to ensure $M_t > 0$ or $0 < P_t < 1$, respectively. In particular, only positive autocorrelation is possible as $a_1,\dots,b_q \geq 0$. If using a log- or logit-link, respectively, negative parameters are possible, but then the models are highly non-linear.

To enable both an approximately linear autocorrelation structure and negative dependencies, [WZH22] propose to use an additional softplus response function for unbounded counts,

$$\text{sp}_c(x) = c\,\ln(1+\exp(x/c)) \text{ with } c > 0,$$

while [WJ24] use a soft-clipping response for bounded counts:

$$\text{sc}_c(x) = c\,\ln\left(\left(1+\exp\left(\frac{x}{c}\right)\right)/\left(1+\exp\left(\frac{x-1}{c}\right)\right)\right) \text{ with } c > 0.$$

These nearly piecewise-linear functions are shown in Figure 1.

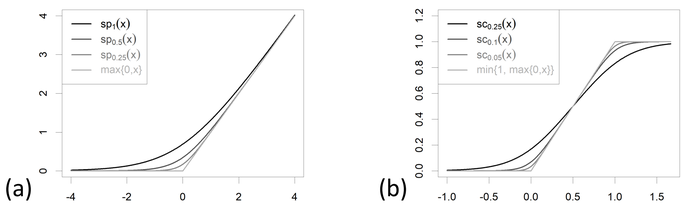

Fig. 1: Plots of (a) softplus and (b) soft-clipping function.

Spatio-temporal INGARCH models

There have been various proposals for spatio-temporal extensions of the above INGARCH models, for spatial count vectors $\boldsymbol{X}_t = (X_{t,1},\dots,X_{t,m})$ with $X_{t,i}$ referring to the i^{th} location. For example, [PHT08, AF24, MFF24] propose exactly linear ST-INGARCH models for unbounded counts (but limited to positive dependence), and [AF24, MFF24] also provide corresponding log-linear versions (allowing for negative dependence, but being highly non-linear). To overcome these limitations, [JWK23] make use of the softplus resp. soft-clipping function. The ST-spINGARCH$(p,r;q,s)$ model for unbounded counts assumes the conditional means

$$M_{t,i} = \text{sp}_c\Big(\alpha_0 + \sum_{k=1}^{p}\alpha_k X_{t-k,i} + \sum_{g=1}^{r}\lambda_g \sum_{j\neq i} w_{ij} X_{t-g,j} + \sum_{l=1}^{q}\beta_l M_{t-l,i} + \sum_{h=1}^{s}\phi_h \sum_{j\neq i} w_{ij} M_{t-h,j}\Big),$$

with the w_{ij} being specified spatial weights. Similarly, the ST-scBINGARCH $(p,r;q,s)$ model for bounded counts has the normalized conditional means

$$P_{t,i} = \text{sc}_c\big(\alpha_0 + \sum_{k=1}^{p}\alpha_k X_{t-k,i}/n + \sum_{g=1}^{r}\lambda_g \sum_{j\neq i} w_{ij} X_{t-g,j}/n + \sum_{l=1}^{q}\beta_l P_{t-l,i} + \sum_{h=1}^{s}\phi_h \sum_{j\neq i} w_{ij} P_{t-h,j}\big).$$

Possible model specifications and practical issues are illustrated for the weather data (DWD Climate Data Center) in Figures 2–3.

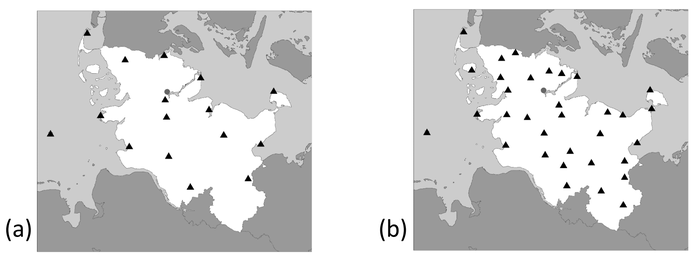

Fig. 2: Map of Schleswig–Holstein (Germany) with weather stations for (a) cloud coverage ($m = 17$) and (b) precipitation ($m = 34$). Grey circle = weather station 'Schleswig'.

Model Specifications and Practical Issues

Dealing with Missing Values

Consider the hourly cloud-coverage data (2009–2019), which express the share of visible sky covered by clouds via bounded counts ("okta") from 0 (no clouds) to $n = 8$ (sky fully overcast). Then 57.2% of the vectors $\boldsymbol{x}_t$ have at least one missing component $x_{t,i}$ (while altogether 5.5% of all $x_{t,i}$ are missing).

Our solution: If $x_{t,i}$ is missing, then we impute an integer value by choosing the value $x_{t,j}$ from most correlated non-missing station at time t. So the selected $x_{t,j}$ is a kind of "nearest neighbor" in terms of cross-correlation. Note that the average maximum correlation between stations over time is 0.709.

Choice of Spatial Weights

We considered two extreme cases for choosing $\mathbf{W} = (w_{ij})$:

- sparse $\mathbf{W}^{(1)}$: includes only nearest neighbor, $w_{ij}^{(1)} = 1$ iff j geographically closest to i;
- dense $\mathbf{W}^{(2)}$: all neighbors considered via $w_{ij}^{(2)} = 1/d_{ij}$ together with row-normalization.

During model fitting, $\mathbf{W}^{(2)}$ produced lower AIC values.

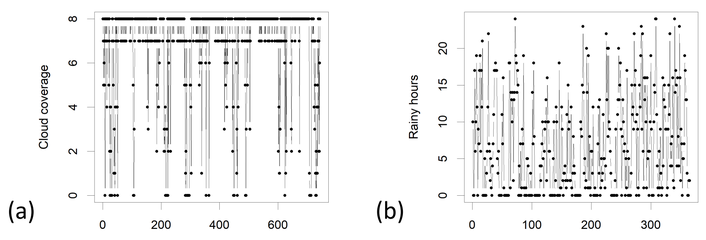

Fig. 3: (a) Hourly cloud coverage at station 'Schleswig' in Dec. 2019, and (b) rainy hours per day at station 'Schleswig' in 2019.

Accounting for Zero Inflation

Consider $x_{t,i}$ to be the bounded counts of rainy hours per day (2011–2019), so $n = 24$. After imputing missing values (0.36% of total hourly observations), we recognized strong zero inflation (38% of zeros). But ST-scBINGARCH model easily adapted by choosing a conditional ZIB distribution with parametrization

$$P(X_{t,i} = x|\boldsymbol{X}_{t-1},\dots) = \omega(1-P_{t,i})\,\delta_{x,0} + \big(1-\omega(1-P_{t,i})\big) \cdot p_{\text{Bin}}\left(x \mid n, {}^{P_{t,i}}\!/_{1-\omega(1-P_{t,i})}\right).$$

Accounting for Cross-Correlation

Cloud coverages show strong cross-correlation (average ≈ 0.52). Our default model assumes conditionally independent components, i.e., $X_{t,i}|\boldsymbol{X}_{t-1},\dots$ independently $\text{Bin}(n, P_{t,i})$, but this only leads to cross-correlation ≈ 0.25. Therefore, we use additional Gaussian copula, where we try two versions:

- "exchangeable correlation structure", where correlation matrix $\mathbf{R} = (1-\rho)\mathbf{I} + \rho\mathbf{E}$ with $\rho \in \left(\frac{-1}{m-1}, 1\right)$;
- "spatial error model (SEM) copula" with covariance matrix $\mathbf{B}\mathbf{B}^{\top}$, where $\mathbf{B} = (\mathbf{I} - \rho\mathbf{W})^{-1}$ with $\rho \in (-1,1)$.

During model fitting, SEM copula produced lower AIC values. Note that exact likelihood computation is not feasible as 2^m discrete differences would have to be computed. Therefore, we use mid-point approximation proposed by [KP10], although it induces bias into parameter estimation [JWK23, JW25].

Modeling spatio-temporal ordinal data

From now on, the $X_{t,i}$ are ordinal random variables, with ordered categories labeled as $0,\dots,K$. Let $\boldsymbol{Y}_{t,i}$ denote nominal binarization of $X_{t,i}$, i.e. its k^{th} component equals $Y_{t,i}^{(k)} = \delta_{X_{t,i},k}$. Then, $\boldsymbol{P}_{t,i} = E[\boldsymbol{Y}_{t,i}|\boldsymbol{X}_{t-1},\dots]$ is conditional PMF vector. To account for natural ordering of the range and to achieve parsimony, define the "expected categories" $C_{t,i} := e(\boldsymbol{P}_{t,i})$, where $e(\boldsymbol{p}) = \sum_{k=0}^{K} k\,p_k$.

[JW25] assume a conditional multinomial distrib. for locations i, $\boldsymbol{Y}_{t,i}|\boldsymbol{X}_{t-1},\dots \sim \text{Mult}(1, \boldsymbol{P}_{t,i})$, with conditional independence between the locations in the default model, or with additional Gaussian copula for intensified cross-dependence as before.

The most general setup uses the softmax response $\boldsymbol{\sigma}$ with $\sigma_k(\boldsymbol{x}) = \exp(x_k)/\sum_j \exp(x_j)$ and assumes condit. PMF $\boldsymbol{P}_{t,i}$ as

$$P_{t,i}^{(k)} = \sigma_k\Big(\big(\alpha_{i,u}^0 + \textstyle\sum_{g=1}^{p}\alpha_{i,u}^g X_{t-g,i}/K + \sum_{h=1}^{r}\lambda_{i,u}^h \sum_{j=1,j\neq i}^{m} w_{i,j} X_{t-h,i}/K + \sum_{a=1}^{q}\beta_{i,u}^a C_{t-a,i}/K + \sum_{b=1}^{s}\nu_{i,u}^b \sum_{j=1,j\neq i}^{m} w_{i,j} C_{t-b,i}/K\big)_{u=0,\dots,K}\Big).$$

It has location- and category-specific parameters, but the number of parameters can be reduced by assuming that some parameters are constant over at least one dimension. For example, our most parsimonious model assumes $\alpha_{i,u}^g = \alpha^g$, $\lambda_{i,u}^h = \lambda^h$, etc., so only the intercept $\alpha_{i,u}^0$ is still both location- and category-specific.

To fully eliminate the category-specific dimension, one can use a parametric conditional PMF such as $X_{t,i}|\boldsymbol{X}_{t-1},\dots \sim \text{Bin}(K, P_{t,i})$, where $P_{t,i}$ is defined like for the ST-scBINGARCH$(p,r;q,s)$ model. Here, [JW25] use a logistic response function instead of the soft-clipping one.

Some Notes on Statistical Analyses

By contrast to the above count models, the $X_{t,i}$ are now of qualitative nature, so moments and autocorrelations are not defined. Thus, following [W20], we express location by the median, and dispersion by the index of ordinal variation, $\text{IOV} = \frac{4}{K}\sum_{k=0}^{K-1} F^{(k)}(1-F^{(k)})$, where $F^{(k)} = P(X \leq k)$ for the ordinal r.v. X.

For serial dependence, we use serial Cohen's κ, see [W20]. Let $F^{(k,l)}(h) = P(X_t \leq k, X_{t-h} \leq l)$, then

$$\kappa_{\text{ord}}(h) = \frac{\sum_{k=0}^{K-1}\left(F^{(k,k)}(h) - \left(F^{(k)}\right)^2\right)}{\sum_{k=0}^{K-1} F^{(k)}(1-F^{(k)})}.$$

Similarly, for cross-dependence,

$$\kappa_{\text{ord}}(i,j) = \frac{2\sum_{k=0}^{K-1}\left(F_{ij}^{(k,k)} - F_i^{(k)}F_j^{(k)}\right)}{\sum_{k=0}^{K-1}\left(F_i^{(k)}\left(1-F_j^{(k)}\right) + F_j^{(k)}\left(1-F_i^{(k)}\right)\right)},$$

where $F_{ij}^{(k,l)} = P(X_i \leq k, X_j \leq l)$ and $F_i^{(k)} = P(X_i \leq k)$.

Acknowledgements

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Projektnummer 516522977.

References:

[AF24] Armillotta M, Fokianos K (2024) Count network autoregression. *J Time Series Anal* **45**, 584–612.

[FLO06] Ferland R, Latour A, Oraichi D (2006) Integer-valued GARCH processes. *J Time Series Anal* **27**, 923–942.

[JW25] Jahn M, Weiß CH (2025) Modeling multivariate ordinal time series. *J Appl Stat*, in press (DOI: 10.1080/02664763.2025.2575034).

[JWK23] Jahn M, Weiß CH, Kim HY (2023) Approximately linear INGARCH models for spatio-temporal counts. *J Royal Stat Soc C* **72**, 476–497.

[KP10] Kazianka H, Pilz J (2010) Copula-based geostatistical modeling of continuous and discrete data … *Stoch Environ Res Risk Ass* **24**, 661–673.

[MFF24] Maletz S, Fokianos K, Fried, R (2024) Spatio-temporal count autoregression. *Data Sci Sci* **3**, 2425171. R package "glmSTARMA" at CRAN.

[PHT08] Paul M, Held L, Toschke AM (2008) Multivariate modelling of infectious disease surveillance data. *Stat Med* **27**, 6250–6267.

[RWJ16] Ristić MM, Weiß CH, Janjić AD (2016) A binomial integer-valued ARCH model. *Int J Biostat* **12**, 20150051.

[W18] Weiß CH (2018) *An Introduction to Discrete-valued Time Series*. John Wiley & Sons, Inc, Chichester.

[W20] Weiß CH (2020) Distance-based analysis of ordinal data and ordinal time series, *J Am Stat Assoc* **115**, 1189–1200.

[WJ24] Weiß CH, Jahn M (2024) Soft-clipping INGARCH models for time series of bounded counts. *Stat Mod* **24**, 295–319.

[WZH22] Weiß CH, Zhu F, Hoshiyar A (2022) Softplus INGARCH models. *Stat Sinica* **32**, 1099–1120.