Hierarchical Count Echo State Network Models with Application to Graduate Student Enrollments

Qi Wang, Paul A. Parker, and Robert Lund
Department of Statistics
University of California, Santa Cruz

ARTICLE HISTORY
Compiled June 25, 2025

ABSTRACT
Poisson autoregressive count models have evolved into a time series staple for correlated count data. This paper proposes an alternative to Poisson autoregressions: count echo state networks. Echo state networks can be statistically analyzed in frequentist manners via optimizing penalized likelihoods, or in Bayesian manners via MCMC sampling. This paper develops Poisson echo state techniques for count data and applies them to a count data set containing the number of graduate students from 1,758 United States universities during the years 1972-2021 inclusive. Negative binomial models are also implemented and generally better handle the overdispersion in the counts. Performance of the proposed models are compared via their forecasting performance as judged by several methods. In the end, a hierarchical negative binomial based echo state network is judged as the superior model.

KEYWORDS
Count Time Series, Echo State Network, High-dimensional Counts, Multivariate Log-Gamma Prior, Spatio-temporal Data.

1. Introduction

Correlated count models are often used to describe positive integer-valued data, such as infectious disease counts (Agosto and Giudici 2020), bank failures (Schoenmaker 1996), storm frequencies (Robbins et al. 2011), hospital visits (Neelon et al. 2013; Matteson et al. 2011), and crime incidents (Kim et al. 2021). Standard Gaussian time series models may not describe discrete counts well, especially when the counts are small (Davis et al. 2016). As such, a discrete distribution is often adopted for the marginal distribution of the counts; Poisson, generalized Poisson, Conway-Maxwell Poisson, binomial, and negative binomial are common choices. The objective then becomes building correlated but non-Gaussian models for the counts.

Classical approaches to the count problem include integer autoregression and general thinning approaches (Jin-Guan and Yuan 1991; Jung and Tremayne 2006; Weiß 2008; Zhu and Joe 2010) and Joe (2016), generalized linear autoregressive moving-average (GLARMA) techniques where model parameters evolve in a stationary but random fashion (Dunsmuir 2015), Poisson autoregressive methods (Fokianos et al. 2009), generalized linear mixed models (McCulloch and Searle 2004), and Gaussian transformations (Jia et al. 2023). Recently, Kong and Lund (2024) have written exclusively on count series having marginal Poisson distributions. The above papers mostly consider univariate series; literature considering multivariate counts is sparser, but includes Ord et al. (1993); Heinen and Rengifo (2007); Brandt and Sandler (2012); Serhiyenko et al. (2015); Karlis (2016), and Fokianos (2021). Our setting, which contains over 1,700 count time series, lies in the high-dimensional realm where work is almost non-existent. The only papers we are aware of that study high-dimensional counts are Bradley et al. (2018); Pan and Pan (2024), and Düker et al. (2024).

This paper studies modeling of high-dimensional count time series data with echo state networks (ESNs) (see McDermott and Wikle (2017) for more on ESNs). The weights in our recurrent neural network are "pre-generated" and fixed throughout training. Our count modeling tactics are similar to the GLARMA methods of Dunsmuir (2015). Penalized likelihood methods in frequentist settings and regression models with multivariate log-gamma priors in the Bayesian setting (Bradley et al. 2018) are developed for estimation. For the negative binomial case, a Pólya-gamma augmentation method is used to improve MCMC computational efficiency. Literature on count ESNs is sparse except for Schafer (2020). Another innovation of this paper includes a count ESN model with a Bayesian hierarchical structure. Computationally efficient MCMC routines for Bayesian estimation are developed for both the Poisson and negative binomial models.

Our interest is motivated by graduate student enrollment counts from the Survey of Graduate Students and Postdoctorates in Science and Engineering, which is sponsored by the National Center for Science and Engineering Statistics, a federal statistical agency tasked with the measurement and reporting of the U.S. science and engineering enterprise. Modeling graduate student counts at individual schools across the United States helps us understand higher education data, facilitating a deeper understanding of enrollment dynamics and providing insight into evidence-based policy making, resource allocation, and institutional evaluation. This aligns with general official statistics objectives, which strive to ensure accurate, transparent, and actionable information for various stakeholders. Our methodology may also interest other statistical agencies that deal with count data over time; for example, the American Community Survey examines counts containing the number of residents in a household (Parker et al. 2020).

The rest of this paper proceeds as follows. Section 2 describes the graduate student counts motivating this study. Section 3 narrates the count time series and spatio-temporal modeling background needed to develop our methods. Our ESN approach is developed in Section 4 and some competing models and scoring rules are given in Section 5. Section 6 fits the ESN model to our graduate student counts, comparing to several other modelling techniques. Section 7 summarizes our findings and proposes several directions for future work.

2. Graduation Count Series

The data in this paper were extracted from the Survey of Graduate Students and Postdocs in Science and Engineering (GSS), an annual census of all academic institutions in the United States that grant research-based graduate degrees. The data is available at `https://ncses.nsf.gov/surveys/graduate-students-postdoctorates-s-e/2023#data`. This comprehensive dataset spans 1972-2021 inclusive and contains 1,758 schools. The GSS is a key source of information on demographics, study fields, support sources, and post-graduate plans of graduate students and postdoctoral researchers in selected fields of science, engineering, and health. The data are annual and allow us to examine time trends, patterns, and differences among institutions.

The data collected in the GSS includes the number of graduate students and postdoctoral researchers by field of study, gender, citizenship status, and race/ethnicity. The survey also contains information on the primary sources of financial support for these individuals, such as federal agencies, universities, and private industry. Herein, we focus solely on the number of graduate students in each program. Different programs may come from the same institution/university.

From the longitudinal nature of our data, it is possible to explore time changes in graduate student compositions. Model fitting methods could allow for trends or forecast counts in future years. The cross-sectional component of our data enables comparisons between institutions, providing insight into how different universities train and support our next generation of scientists and engineers.

Figure 1 plots time count trajectories for four randomly chosen schools in our study: The University of Florida (School of Chemistry), The University of Rochester (School of Electrical, Electronics, and Communications Engineering), The Georgia Institute of Technology (School of Chemical Engineering), and Oklahoma State University (School of Geological and Earth Sciences). The counts have varying scales; some schools have hundreds of students, while others report in teens. Sample autocorrelations for these four schools are shown in Figure 2 and exhibit positive but non-negligible temporal associations. The scales for these schools are significantly different from each other. Figure 1 suggests that some schools have counts close to zero, making it inappropriate to base inferences on Gaussian-based analyses. For added feel, Figure 3 shows sample time series for 100 randomly sampled schools.

3. Background

This section reviews count time series and neural network methods. Conjugacy approaches for our Bayesian methods are also discussed.

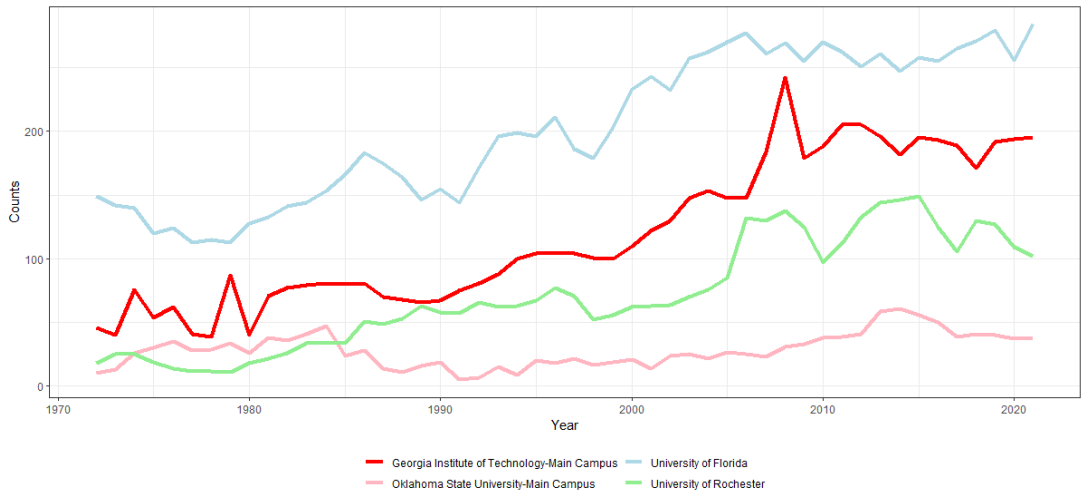

Figure 1.: Time series plots of graduate student counts for four randomly selected example schools in our study.

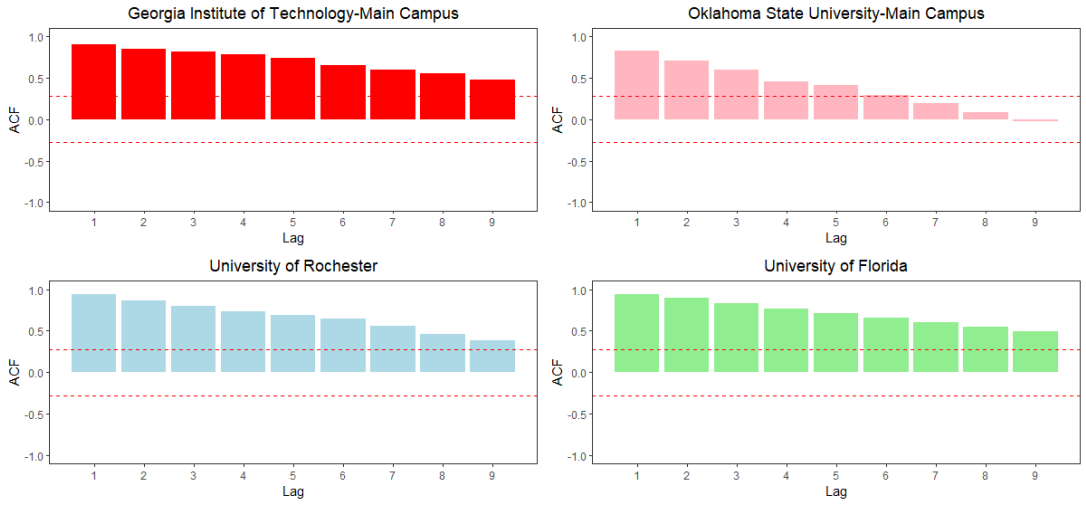

Figure 2.: Sample autocorrelations for four randomly selected schools. The dashed lines are pointwise 95% confidence thresholds for a zero correlation.

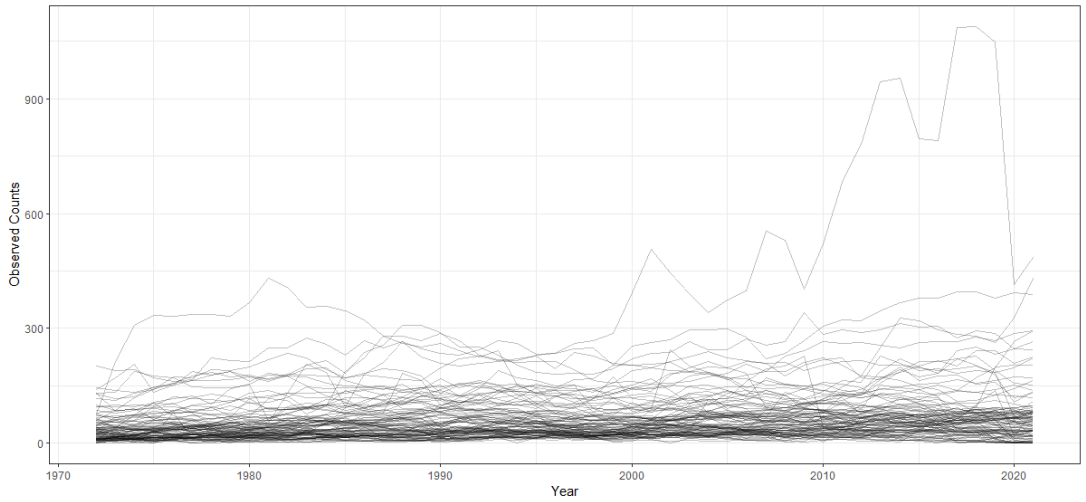

Figure 3.: Time series plots for 100 randomly sampled schools. Significant scale differences exist in these series.

3.1. Count Time Series

The classical way to model count time series is through integer autoregressions, which are based on thinning operations. An integer autoregression of order one (INAR(1)) for the counts $\{Y_t\}_{t=1}^{T}$ obeys

$$Y_t = \delta \circ Y_{t-1} + \varepsilon_t,$$

where $\{\varepsilon_t\}$ is an IID positive integer-valued random sequence. The thinning operator $\circ$ operates on the integer-valued random variable N via

$$\delta \circ N = \sum_{i=1}^{N} B_i,$$

where $\{B_i\}$ are IID Bernoulli trials having success probability $\delta \in [0, 1]$. One often selects $\{\varepsilon_t\}$ to produce a specific count marginal distribution for Y_t. For example, selecting ε_t to be Poisson with mean $\lambda(1-\delta)$ will produce a Poisson series with mean λ for every t. See Joe (2016) for quantification about the marginal distributions that can be constructed. INAR models cannot have any negative autocorrelations, which does not appear to be relevant here given the plots in Figure 2. Prominent INAR references include Jin-Guan and Yuan (1991); Jung and Tremayne (2006); Weiß (2008), and Zhu and Joe (2010).

A more general count modeling approach was recently introduced in Jia et al. (2023). This approach transforms a standardized Gaussian series into the desired count series and can accommodate any marginal distribution. Suppose that we want the marginal cumulative distribution function (CDF) $F_Y(\cdot)$ for the observation Y_t. If $\{Z_t\}$ is a standard Gaussian series with $E[Z_t] \equiv 0$, $\mathrm{Var}(Z_t) \equiv 1$, and $\mathrm{Corr}(Z_t, Z_{t+h}) = \rho_Z(h)$, then set

$$Y_t = F_Y^{-1}(\Phi(Z_t)), \tag{1}$$

where $\Phi(\cdot)$ denotes the CDF of a standard normal distribution and $F_Y^{-1}(\cdot)$ is the quantile function

$$F_Y^{-1}(u) = \inf\{x : F_Y(x) \geq u\}.$$

The probability integral transformation theorem shows that $\Phi(Z_t)$ has a uniform[0,1] distribution. A second application of the same result shows that Y_t has the desired marginal distribution $F_Y(\cdot)$. The autocovariances of $\{Y_t\}$ in (1) can be computed through Hermite expansions as in Jia et al. (2023). Kong and Lund (2024) focuses exclusively on Poisson distributed series.

State-space models are yet another popular count modeling technique. A hierarchical model with Poisson dynamics, for example, takes

$$Y_t|\theta_t \sim \text{Poisson}(e^{\theta_t}),$$

where $\{\theta_t\}$ is a process to be clarified and a log link is used to keep the Poisson parameter positive. When $\{\theta_t\}$ has autoregressive moving-average (ARMA) dynamics, these structures belong to the generalized linear autoregressive moving-average (GLARMA) model class. A Gaussian autoregressive structure is often placed on $\{\theta_t\}$:

$$\theta_t = \phi_0 + \sum_{i=1}^{p} \phi_i \theta_{t-i} + \varepsilon_t.$$

Here, $\{\varepsilon_t\}$ is an IID Gaussian noise process with zero mean and $\phi_1, \ldots, \phi_p$ are the p autoregressive coefficients that are assumed to produce a causal autoregression. While these and other models are discussed in Fokianos et al. (2009); Dunsmuir (2015) and Davis et al. (2021), a useful tactic allows θ_t to depend on the past counts. Models that recurse on past counts are called integer generalized autoregressive conditional heteroskedastic (INGARCH) models, but do not generate white noise like ordinary GARCH series. For example, in the first-order case, one could posit that

$$\theta_t = \phi_0 + a\theta_{t-1} + bY_{t-1}.$$

See Gamerman et al. (2015); Dunsmuir (2015) and Holan and Wikle (2016) for additional work.

3.2. Echo State Networks

Neural networks are combinations of linear and nonlinear transformations, which often capably describe nonlinear features in sequential data. Neural networks that are based on the common recursions governing sequential data are called recurrent neural networks (RNNs) and were introduced in (Rumelhart et al. 1986). Thereafter, Hochreiter and Schmidhuber (1997) developed long short-term memory (LSTM) networks by adding additional structure that makes dependencies decay more slowly in lag. Dey and Salem (2017) simplified LSTMs while retaining similar model performance.

Analysis of neural networks can be computationally intensive with gradient descent techniques over a large parameter space, especially with RNNs. ESNs, proposed by Jaeger and Haas (2004), are RNN variants that allow for more efficient parameter estimation. Hidden weights in the ESNs are randomly generated from a symmetric

distribution centered about zero; the only parameters requiring estimation reside in the output layer.

McDermott and Wikle (2017) introduce a quadratic ESN (QESN) for spatio-temporal data. Uncertainty quantification is done via ensembling, i.e., randomly generating the hidden layer weights multiple times. A dimension reduction layer such as a principal component analysis is added to the QESN in McDermott and Wikle (2019), making it a deep-ensembled ESN; the model is implemented in both frequentist and Bayesian frameworks. Furthering this work, Schafer (2020) generalize ESNs to exponential families. Wang et al. (2024) integrate graph convolutional neural networks and ESNs to capture areal-level spatial dependence.

The ESN in our paper resembles the structure described by McDermott and Wikle (2017). As an example, consider a single GSS school and let $\{\boldsymbol{Y}_t\}_{t=1}^{T}$ denote the observations from the school. The observation for the school at time t, Y_t, has the conditional mean

$$E[Y_t|Y_1,\ldots,Y_{t-1}] = \boldsymbol{h}_t'\boldsymbol{\eta}, \tag{2}$$

where the $n_h \times 1$ vector $\boldsymbol{h}_t$ denotes output from a fixed-weight recurrent layer with n_h hidden nodes (which may depend on previous counts for this school as elaborated upon below). Also, $\boldsymbol{\eta}$ is a length n_h vector of regression coefficients. The parameters in $\boldsymbol{\eta}$ are the only quantities that need to be estimated; this can be done by optimizing a likelihood-based loss function, or by Bayesian MCMC.

We have a length-r covariate $\boldsymbol{x}_t$, which is school specific, at time t. An ESN with n_h hidden nodes can be built to fit this data. To begin, the length-n_h vector $\boldsymbol{h}_1$ is initialized from the covariate $\boldsymbol{x}_1$. Thereafter, for each $t \in \{2,3,\ldots,T\}$, the hidden unit $\boldsymbol{h}_t$ in our echo state network hidden layer is calculated via

$$\boldsymbol{h}_t = g\left(\frac{\nu}{|\lambda_W|}\boldsymbol{W}'\boldsymbol{h}_{t-1} + \boldsymbol{U}_y'\boldsymbol{Y}_{(t-p):(t-1)} + \boldsymbol{U}_x'\boldsymbol{x}_t\right), \tag{3}$$

where the length-p vector $\boldsymbol{Y}_{(t-p):(t-1)}$ contains the past p observations for this school; i.e., $(Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p})'$. Here and in the rest of this paper, we use $p = 1$ exclusively. The scalar λ_W is the largest eigenvalue of $\boldsymbol{W}$, an $n_h \times n_h$ weight matrix, and ν is a regularization parameter, lying between zero and unity, used to ensure that $\boldsymbol{h}_t$ does not explode/overflow. Here, $g(\cdot)$ is a predetermined activation function, typically a sigmoid or hyperbolic tangent function. Moreover, $\boldsymbol{U}_Y$ is a $p \times n_h$ matrix, where p is the autoregressive order (larger p's typically induce longer temporal memory). Similarly, $\boldsymbol{U}_X$ is an $r \times n_h$ matrix. The parameters $\boldsymbol{W}$, $\boldsymbol{U}_Y$, and $\boldsymbol{U}_X$ are randomly generated from some specified distribution, but thereafter are fixed and not estimated in the model fitting procedure. An example of the calculation of $\boldsymbol{h}_t$ in a hidden layer without autoregressive component is graphically illustrated in Figure 4.

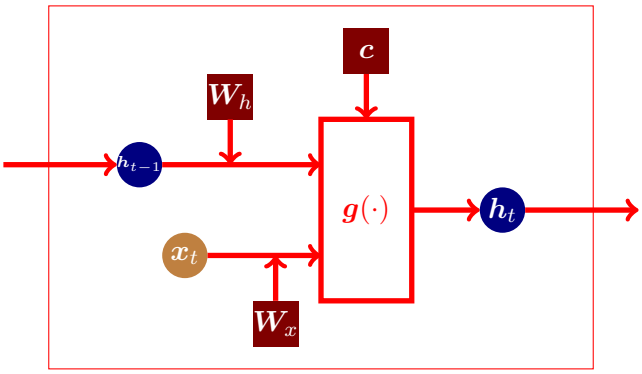

Figure 4.: An example of a recurrent layer for sequentially calculating nodes.

In further detail, $W_{i,j}$, $(U_Y)_{i,j}$, and $(U_X)_{i,j}$, for $1 \leq i, j \leq n_h$, are independently generated from the following spike and slab distributions:

$$\begin{aligned}
(W)_{i,j} &= \gamma_{i,j}^{W} \text{Unif}(-a_w, a_w) + (1 - \gamma_{i,j}^{W})\delta_0, \\
(U_Y)_{i,j} &= \gamma_{i,j}^{U_Y} \text{Unif}(-a_{u_Y}, a_{u_Y}) + (1 - \gamma_{i,j}^{U_Y})\delta_0, \\
(U_X)_{i,j} &= \gamma_{i,j}^{U_X} \text{Unif}(-a_{u_X}, a_{u_X}) + (1 - \gamma_{i,j}^{U_X})\delta_0, \\
\gamma_{i,j}^{W} \sim \text{Bern}(\pi_w), \gamma_{i,j}^{U_Y} &\sim \text{Bern}(\pi_{u_Y}), \gamma_{i,j}^{U_X} \sim \text{Bern}(\pi_{u_X}).
\end{aligned} \tag{4}$$

These choices were made following McDermott and Wikle (2017); however, other zero-centered weight distributions could also be used. Here, δ_0 denotes a unit point mass at zero. The elements $\gamma_{i,j}^{W}$ are IID, as are $\gamma_{i,j}^{U_Y}$ and $\gamma_{i,j}^{U_X}$. This model contains hyperparameters that can be chosen via cross-validation. In other words, the uniform distribution's parameters $a_{(\cdot)}$ and the Bernoulli distribution parameter $\pi_{(\cdot)}$ in (4) are not estimated. Here, the notation $a_{(\cdot)}$ denotes a generic a_w, a_{u_Y}, or a_{u_X}, and $\pi_{(\cdot)}$ denotes a generic π_w, π_{u_Y}, or π_{u_X}. After fixing these hyperparameters, the weight matrices $\boldsymbol{W}$, $\boldsymbol{U}_Y$, and $\boldsymbol{U}_X$ are generated.

The only parameter estimated in (2) is $\boldsymbol{\eta}$. This setup can be interpreted as first feeding information into the ESN and using the hidden layer output as explanatory variables in a regression fit. McDermott and Wikle (2017) use a ridge penalty to avoid overfitting when estimating $\boldsymbol{\eta}$. One may also add empirical orthogonal functions (EOFs) or other spatial basis functions as covariates in (3), enabling one to capture cross-sample or spatial dependence as in McDermott and Wikle (2017). The ESN parameters were chosen via cross-validation: we select $a \in \{0.01, 0.1, 1\}$ and $\pi \in \{0.1, 0.3, 0.5\}$. Candidate values for n_h are $\{30, 50, 100, 120\}$; values of ν considered lie in $\{0.1, 0.5, 0.7, 0.9\}$.

3.3. Multivariate log-Gamma Priors

This subsection introduces the multivariate log-Gamma (MLG) distribution from Bradley et al. (2018) and Bradley et al. (2020). This distribution's conjugacy properties make Bayesian Poisson count modeling computationally convenient. To simulate from this distribution, an m-dimensional random vector $\boldsymbol{w} = (w_1, \ldots, w_m)'$ is first generated with mutually independent components: $w_i \sim \text{LG}(\alpha_i, \kappa_i)$. Here, $\text{LG}(\alpha, \kappa)$ denotes the log-Gamma distribution, which is the logarithm of a Gamma draw with

shape parameter α and scale parameter κ. Then set

$$\boldsymbol{q} = \boldsymbol{\mu} + \boldsymbol{V}\boldsymbol{w},$$

where $\boldsymbol{V} \in \mathcal{R}^m \times \mathcal{R}^m$ and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)'$ are deterministic (hierarchical structures will be placed on these parameters later). We call $\boldsymbol{q}$ a multivariate log-gamma random vector and write $\boldsymbol{q} \sim \text{MLG}(\boldsymbol{\mu}, \boldsymbol{V}, \boldsymbol{\alpha}, \boldsymbol{\kappa})$. Here, $\boldsymbol{q}$ can be shown to have the probability density

$$f(\boldsymbol{q}|\boldsymbol{\mu}, \boldsymbol{V}, \boldsymbol{\alpha}, \boldsymbol{\kappa}) = \frac{1}{\det(\boldsymbol{V}\boldsymbol{V}')^{1/2}} \left(\prod_{i=1}^{m} \frac{\kappa_i^{\alpha_i}}{\Gamma(\alpha_i)} \right) \exp[\boldsymbol{\alpha}'\boldsymbol{V}^{-1}(\boldsymbol{q}-\boldsymbol{\mu}) - \boldsymbol{\kappa}' \exp\{\boldsymbol{V}^{-1}(\boldsymbol{q}-\boldsymbol{\mu})\}]$$

over $\boldsymbol{q} \in \mathcal{R}^m$.

Bradley et al. (2018) show that a multivariate log-Gamma random variable with parameters $(\boldsymbol{c}, \alpha^{1/2}\boldsymbol{V}, \alpha\boldsymbol{1}, \alpha\boldsymbol{1})$ converges in distribution to a multivariate normal distribution with mean $\boldsymbol{c}$ and covariance matrix $\boldsymbol{V}\boldsymbol{V}'$ as $\alpha \to \infty$, noting also that $\alpha = 1,000$ is typically sufficiently large for this convergence to "kick in". Accordingly, we set $\alpha = 1,000$ throughout.

One can also partition a MLG variate $\boldsymbol{q}$ with parameters $(\boldsymbol{\mu}, \boldsymbol{V}, \boldsymbol{\alpha}, \boldsymbol{\kappa})$ into an r-dimensional vector $\boldsymbol{q}_1$ and an $(n-r)$-dimensional vector $\boldsymbol{q}_2$ via $\boldsymbol{q} = (\boldsymbol{q}_1', \boldsymbol{q}_2')'$. Here, $\boldsymbol{V}^{-1}$ can also be partitioned into $[\boldsymbol{H}|\boldsymbol{B}]$, where $\boldsymbol{H}$ is an $n \times r$ dimensional matrix and $\boldsymbol{B}$ is an $n \times n - r$ matrix.

Our notation uses $\boldsymbol{q}_1|\boldsymbol{q}_2$ as a conditional multivariate log-gamma distribution:

$$\boldsymbol{q}_1|\boldsymbol{q}_2 \sim \text{cMLG}(\boldsymbol{H}, \boldsymbol{\alpha}, \boldsymbol{\kappa}^*),$$

which can be shown to have the probability density

$$f(\boldsymbol{q}_1|\boldsymbol{q}_2) = M \exp\{\boldsymbol{\alpha}'\boldsymbol{H}\boldsymbol{q}_2 - \boldsymbol{\kappa}^{*\prime} \exp(\boldsymbol{H}\boldsymbol{q}_1\},$$

over $\boldsymbol{q}_1 \in \mathbb{R}^r$. Here, $\boldsymbol{\kappa}^* = \exp\{\boldsymbol{B}\boldsymbol{q}_2 - \boldsymbol{V}^{-1}\boldsymbol{\mu} - \ln(\boldsymbol{\kappa})\}$ and M is some normalizing constant.

Bradley et al. (2018) and Bradley et al. (2020) describe an efficient data augmentation strategy to sample from this distribution, which is an important step when building Bayesian hierarchical models with a Poisson conditional distribution. This will be discussed in more detail later.

4. Methods

This section proposes several ESN models for count data having various structures. We begin by describing a general count ESN model. For school i, we posit the model

$$Y_{i,t}|\boldsymbol{\theta}_{i,t} \overset{ind}{\sim} F_{\boldsymbol{\theta}_{i,t}}, \tag{5}$$

where F is some count marginal distribution that depends on the parameter $\boldsymbol{\theta}_{i,t}$. Depending on the choice of F, $\boldsymbol{\theta}_{i,t}$ may be multivariate. For example, $\boldsymbol{\theta}_{i,t}$ is univariate when F is Poisson and bivariate when F is negative binomial. Our models allow $\boldsymbol{\theta}_{i,t}$ to evolve recursively in time t according to an ESN. One example has $\boldsymbol{\theta}_{i,t} = \boldsymbol{h}_{i,t}'\boldsymbol{\eta}$,

where $h_{i,t}$ evolves in t as in (3). Negative binomial based models are also discussed to handle overdispersion.

4.1. Poisson Echo State Networks

Suppose that $Y_{i,t}|\theta_{i,t}$ is Poisson with mean $\exp(\theta_{i,t})$, and $\theta_{i,t}$ is a linear combination of the output from an ESN:

$$\begin{aligned} Y_{i,t}|\theta_{i,t} &\sim \text{Poisson}(e^{\theta_{i,t}}), \\ \theta_{i,t} &= \boldsymbol{h}'_{i,t}\boldsymbol{\eta}_i, \end{aligned}$$

where $\{\boldsymbol{h}_{i,t}\}$ evolves in time via an ESN. A first-order autoregressive example employing the logarithm of past observations posits

$$\begin{aligned} \boldsymbol{h}_{i,t} &= g\left(\frac{\nu}{|\lambda_W|}\boldsymbol{W}'\boldsymbol{h}_{i,t-1} + \ln(Y_{i,t-1}+1)\boldsymbol{U}'_y + \boldsymbol{U}'_x\boldsymbol{x}_{i,t}\right), \quad t = 2,\ldots,T, \\ \boldsymbol{h}_{i,1} &= g\big(\boldsymbol{U}'_x\boldsymbol{x}_{i,1}\big). \end{aligned} \tag{6}$$

In the above, the activation function $g(\cdot)$ is applied coordinate-wise to the quantities in parentheses, and unity is added to all observations to avoid taking a logarithm of zero. Here, $g(\cdot)$ is the hyperbolic tangent function, chosen to avoid overflow and introduce nonlinearity. Also, the elements of $\boldsymbol{U}_x$, $\boldsymbol{U}_y$, and $\boldsymbol{W}$ are randomly generated from (4) and fixed throughout. The only unknown parameter for the ith school is $\boldsymbol{\eta}_i$.

Assuming conditional independence of $Y_{i,t}|\theta_{i,t}$ in t, the likelihood for the ith school, denoted by $\mathcal{L}_i(\boldsymbol{\eta}_i)$, is

$$\mathcal{L}_i(\boldsymbol{\eta}_i) = \prod_{t=1}^{T} \frac{\exp(-e^{\theta_{i,t}})e^{\theta_{i,t}Y_{i,t}}}{Y_{i,t}!}.$$

Based on this likelihood, two frequentist models and two Bayesian models are now developed.

4.1.1. Single and Ensemble Poisson ESN

The parameter $\boldsymbol{\eta}_i$ can be estimated by maximizing the LASSO-based penalized log likelihood $\mathcal{L}_i^*$:

$$\mathcal{L}_i^*(\boldsymbol{\eta}_i) = \sum_{t=1}^{T}\left[Y_{i,t}\theta_{i,t} - e^{\theta_{i,t}}\right] - \tau\sum_{j=1}^{n_h}|\eta_{i,j}|,$$

where τ is a penalty parameter typically chosen by cross-validation. The weights in the ESN need only be generated once to fit the model, which we call a "Single Poisson ESN." Alternatively, the ESN parameters can be generated multiple times, and the model fitted to each realization used to construct an "Ensemble ESN." In penalized models like LASSO regression, uncertainty quantification usually follows from a bootstrap. However, ensembling provides another source of uncertainty quantification. The regularization parameter τ is chosen as 1.0 via cross-validation among $\{0.5, 1.0, 1.5\}$.

Note, similar to McDermott and Wikle (2017), a small grid of candidate values was used here and predictive performance was generally strong; however, in some cases, it may be worthwhile to use a finer grid of candidates to further optimize prediction accuracy.

4.1.2. Bayesian Poisson ESN

This model can also be implemented using a Bayesian framework. Conjugacy is ensured with the multivariate log-gamma (MLG) prior above for $\boldsymbol{\eta}_i$:

$$\boldsymbol{\eta}_i \sim \text{MLG}(\mathbf{0}_{n_h}, \alpha^{1/2}\sigma_\eta \boldsymbol{I}_{n_h}, \alpha\mathbf{1}_{n_h}, \alpha\mathbf{1}_{n_h}),$$

where the distributional notation follows Bradley et al. (2018), $\boldsymbol{I}_k$ denotes the $k \times k$ identity matrix, $\mathbf{1}_k$ denotes the k-dimensional vector containing all unit entries, and $\mathbf{0}_k$ is the k-dimensional vector containing all zeros. The variance parameter σ_η is chosen to be 0.1 to induce shrinkage. Note that σ_η can also be random, an aspect discussed later. In this case, the joint density can be shown to have form

$$\begin{aligned} p(\boldsymbol{Y}_i, \boldsymbol{\eta}_i) \quad \propto \quad & \prod_{t=1}^{T} [\exp(\theta_{i,t} Y_{i,t} - \exp(\theta_{i,t}))] \times \\ & \exp\left(\alpha \mathbf{1}'_{n_h} \alpha^{-1/2} \frac{1}{\sigma_\eta} \boldsymbol{I}_{n_h} \boldsymbol{\eta}_i - \alpha \mathbf{1}'_{n_h} \exp\left[\alpha^{-1/2} \frac{1}{\sigma_\eta} \boldsymbol{I}_{n_h} \boldsymbol{\eta}_i \right] \right). \end{aligned}$$

Consequently, the posterior distribution $p(\boldsymbol{\eta}_i | \boldsymbol{Y}_i)$ has the conditional MLG form

$$\begin{aligned} p(\boldsymbol{\eta}_i | \boldsymbol{Y}_i) \quad \propto \quad & \exp\left\{ \left(\sum_{t=1}^{T} Y_{i,t} \boldsymbol{h}'_{i,t} + \alpha \mathbf{1}'_{n_h} \alpha^{-1/2} \frac{1}{\sigma_\eta} \boldsymbol{I}_{n_h} \right) \boldsymbol{\eta}_i - \right. \\ & \left. \exp\left[\sum_{t=1}^{T} \boldsymbol{h}'_{i,t} \boldsymbol{\eta}_i \right] - \alpha \mathbf{1}'_{n_h} \exp\left[\alpha^{-1/2} \frac{1}{\sigma_\eta} \boldsymbol{I}_{n_h} \boldsymbol{\eta}_i \right] \right\}. \end{aligned}$$

To induce simplicity, define the notations $\boldsymbol{H}_i = (\boldsymbol{h}'_{i,1}, \ldots, \boldsymbol{h}'_{i,T})'$ and

$$\boldsymbol{L}_i = \begin{bmatrix} \boldsymbol{H}_i \\ \alpha^{-1/2} \frac{1}{\sigma_\eta} \boldsymbol{I}_{n_h} \end{bmatrix}, \quad \boldsymbol{\xi}_i = \begin{bmatrix} \boldsymbol{Y}'_i & \alpha \mathbf{1}'_{n_h} \end{bmatrix}', \qquad \boldsymbol{\psi}_i = \begin{bmatrix} \mathbf{1}'_T & \alpha \mathbf{1}'_{n_h} \end{bmatrix}'.$$

Under this formulation, the posterior distribution is

$$\boldsymbol{\eta}_i | \boldsymbol{Y}_i \sim \text{cMLG}(\boldsymbol{L}_i, \boldsymbol{\xi}_i, \boldsymbol{\psi}_i).$$

Bradley et al. (2018) and Bradley et al. (2020) develop a computationally efficient procedure to sample from a cMLG distribution via a "data augmentation" approach. The full model is a latent conjugate multivariate process (LCM) model (Bradley et al. 2020). In this case, the model can be partitioned into two stages:

- Data stage:

$$Y_{i,t}|\boldsymbol{\eta}_i, \boldsymbol{q}_i \overset{ind}{\sim} \text{Poisson}(e^{\boldsymbol{h}'_{i,t}\boldsymbol{\eta}_i + \boldsymbol{b}'_i\boldsymbol{q}_i})$$

- Parameter stage:

$$\boldsymbol{\eta}_i|\boldsymbol{V}, \boldsymbol{\alpha}_\eta, \boldsymbol{\kappa}_\eta, \boldsymbol{q}_i \sim \text{MLG}(-\boldsymbol{V}\boldsymbol{B}\boldsymbol{q}_i, \boldsymbol{V}, \boldsymbol{\alpha}_\eta, \boldsymbol{\kappa}_\eta)$$
$$f(\boldsymbol{q}_i) \propto 1$$

  where

$$\boldsymbol{V} = \alpha^{1/2}\sigma_\eta \boldsymbol{I}_{n_h},$$
$$\boldsymbol{\alpha}_\eta = \alpha \mathbf{1}_{n_h}, \boldsymbol{\kappa}_\eta = \alpha \mathbf{1}_{n_h}.$$

  Here, $\boldsymbol{b}_i$ is a pre-specified T-dimensional vector, and the $n_h \times T$ matrix $\boldsymbol{B}$ is also pre-specified. The T-dimensional $\boldsymbol{q}_i$ has an improper flat prior. Conditional on $\boldsymbol{q}_i = \mathbf{0}_T$, the likelihood is proportional to the original model.

Therefore, to sample $\boldsymbol{\eta}$ from the posterior distribution, one can perform the following two steps:

- First, sample $\tilde{\boldsymbol{\eta}}_i$ as a draw from the $\text{MLG}(\mathbf{0}, \boldsymbol{I}_{n_h}, \boldsymbol{\xi}_i, \boldsymbol{\psi}_i)$ distribution.
- Next, affinely transform $\tilde{\boldsymbol{\eta}}_i$ via

$$(\boldsymbol{L}'_i\boldsymbol{L}_i)^{-1}\boldsymbol{L}'_i\tilde{\boldsymbol{\eta}}_i \tag{7}$$

  as a data augmented sample of $\boldsymbol{\eta}_i|\boldsymbol{Y}_i$.

Sampling via this approach will not require a Metropolis-Hastings step, making the scheme computationally efficient.

4.1.3. Bayesian Hierarchical Poisson ESN

We now develop a Bayesian hierarchical model that incorporates similarities among schools within the same geographic state, which are likely to be subject to the same regulations and hence behave similarly. Our model specification is

$$Y_{i,t}|\theta_{i,t} \overset{ind}{\sim} \text{Poisson}(e^{\theta_{i,t}}),$$
$$\theta_{i,t} = \boldsymbol{h}'_{i,t}\boldsymbol{\eta}_{s(i)} + \delta_{s(i)},$$

where the $\boldsymbol{h}_{i,t}$ evolves via (6).

Here, the subscript $s(i)$ denotes which state (within the United States) contains the ith school. The shared parameters $\boldsymbol{\eta}$ and δ for schools within the same state account inject our hierarchical dependence. Accordingly, a hierarchical model structure is imposed on both $\boldsymbol{\eta}_{s(i)}$ and $\delta_{s(i)}$ as follows:

$$\boldsymbol{\eta}_j \sim \text{MLG}(\mathbf{0}_{n_h}, \alpha^{1/2}\sigma_\eta \boldsymbol{I}_{n_h}, \alpha \mathbf{1}_{n_h}, \alpha \mathbf{1}_{n_h}),\ j = 1, \ldots, n_s$$
$$\boldsymbol{\delta} \sim \text{MLG}(\mathbf{0}_{n_s}, \alpha^{1/2}\sigma_\delta \boldsymbol{I}_{n_s}, \alpha \mathbf{1}_{n_s}, \alpha \mathbf{1}_{n_s}),$$

where $n_s = 49$ is the number of states with schools in the GSS data, and the hyperparameters σ_η and σ_δ follow half-Cauchy priors with the fixed scale parameter υ:

$$\sigma_\eta \sim \text{Half-Cauchy}(0, \upsilon), \quad \sigma_\delta \sim \text{Half-Cauchy}(0, \upsilon).$$

A vague prior distribution is implemented by choosing $\upsilon = 100$. Let $\boldsymbol{S}_i'$ denote a length-n_s vector containing only zeroes and ones, where the $s(i)$th element is unity (and all other entries are zero). Then $\theta_{i,t} = \tilde{\boldsymbol{h}}_{i,t}'\tilde{\boldsymbol{\eta}}$, where

$$\tilde{\boldsymbol{h}}_{i,t} := (\boldsymbol{0}'_{(s(i)-1)\cdot n_h}, \boldsymbol{h}_{i,t}', \boldsymbol{0}'_{(n_s - s(i)-1)\cdot n_h}, \boldsymbol{S}_i'), \quad \text{and,} \quad \tilde{\boldsymbol{\eta}} = (\boldsymbol{\eta}_1', \ldots, \boldsymbol{\eta}_{n_s}', \boldsymbol{\delta}')'. \tag{8}$$

The prior distribution of $\tilde{\boldsymbol{\eta}}$ is

$$\tilde{\boldsymbol{\eta}} \sim \text{MLG}\left(\boldsymbol{0}_{(n_h+1)\times n_s}, \begin{bmatrix} \sigma_\eta \boldsymbol{I}_{(n_h \cdot n_s)} & \boldsymbol{0}_{(n_h\cdot n_s)\times n_s} \\ \boldsymbol{0}_{n_s \times (n_h \cdot n_s)} & \sigma_\delta \boldsymbol{I}_{n_s} \end{bmatrix}, \alpha \boldsymbol{1}_{(n_h+1)\cdot n_s}, \alpha \boldsymbol{1}_{(n_h+1)\cdot n_s}\right).$$

The joint density $\pi(\boldsymbol{Y}, \tilde{\boldsymbol{\eta}}, \sigma_\eta, \sigma_\delta)$ can be expressed as

$$\begin{aligned}
\pi(\boldsymbol{Y}, \tilde{\boldsymbol{\eta}}, &\sigma_\eta, \sigma_\delta) \propto \\
&\prod_{i=1}^{N}\prod_{t=1}^{T} \exp\left(\tilde{\boldsymbol{h}}_{i,t}'\tilde{\boldsymbol{\eta}} i y_{i,t} - \exp\left(\tilde{\boldsymbol{h}}_{i,t}'\tilde{\boldsymbol{\eta}}_i\right)\right) \times \frac{1_{\{\sigma_\delta > 0\}}}{1 + (\sigma_\delta/\upsilon)^2} \times \frac{1_{\{\sigma_\eta > 0\}}}{1 + (\sigma_\eta/\upsilon)^2} \\
&\times \left| \begin{matrix} \sigma_\eta^{-1}\boldsymbol{I}_{(n_h\cdot n_s)} & \boldsymbol{0}_{(n_h\cdot n_s)\times n_s} \\ \boldsymbol{0}_{n_s\times(n_h\cdot n_s)} & \sigma_\delta^{-1}\boldsymbol{I}_{n_s} \end{matrix} \right| \times \exp\left(\alpha \boldsymbol{1}'_{(n_h+1)\cdot n_s} \cdot \alpha^{-1/2} \begin{bmatrix} \sigma_\eta^{-1}\boldsymbol{I}_{(n_h\cdot n_s)} & \boldsymbol{0}_{(n_h\cdot n_s)\times n_s} \\ \boldsymbol{0}_{n_s\times(n_h\cdot n_s)} & \sigma_\delta^{-1}\boldsymbol{I}_{n_s} \end{bmatrix} \tilde{\boldsymbol{\eta}}\right) \\
&\times \exp\left(-\alpha \boldsymbol{1}'_{(n_h+1)\cdot n_s} \exp\left(\alpha^{-1/2} \begin{bmatrix} \sigma_\eta^{-1}\boldsymbol{I}_{(n_h\cdot n_s)} & \boldsymbol{0}_{(n_h\cdot n_s)\times n_s} \\ \boldsymbol{0}_{n_s\times(n_h\cdot n_s)} & \sigma_\delta^{-1}\boldsymbol{I}_{n_s} \end{bmatrix} \tilde{\boldsymbol{\eta}}\right)\right).
\end{aligned}$$

With the definitions

$$\begin{aligned}
\boldsymbol{H} &= (\tilde{\boldsymbol{h}}'_{1,1}, \ldots, \tilde{\boldsymbol{h}}'_{1,T}, \ldots, \tilde{\boldsymbol{h}}'_{S,1}, \ldots, \tilde{\boldsymbol{h}}'_{S,T})', \\
\boldsymbol{Y} &= (y_{1,1}, \ldots, y_{1,T}, \ldots, y_{S,1}, y_{S,T})',
\end{aligned}$$

the full conditional distribution of $\tilde{\boldsymbol{\eta}}$ is still a conditional MLG:

$$\tilde{\boldsymbol{\eta}}|\cdot \sim \text{cMLG}(\text{L}, \boldsymbol{\xi}, \boldsymbol{\psi}),$$

where

$$\begin{aligned}
\boldsymbol{L} &= \begin{bmatrix} \boldsymbol{H} \\ \alpha^{-1/2} \begin{bmatrix} \sigma_\eta^{-1}\boldsymbol{I}_{(n_h\cdot n_s)} & \boldsymbol{0}_{(n_h\cdot n_s)\times n_s} \\ \boldsymbol{0}_{n_s\times(n_h\cdot n_s)} & \sigma_\delta^{-1}\boldsymbol{I}_{n_s} \end{bmatrix} \end{bmatrix}, \\
\boldsymbol{\xi} &= (\boldsymbol{Y}', \alpha \boldsymbol{1}'_{(n_h+1)\cdot n_s}), \\
\boldsymbol{\psi}_\eta &= (\boldsymbol{1}'_{NT}, \alpha \boldsymbol{1}_{(n_h+1)\cdot n_s}).
\end{aligned}$$

Note that $\boldsymbol{H}$ is an $(n_S \cdot n_T) \times ((n_h + 1) \cdot n_s)$ matrix; therefore, direct inversion of the matrices in (7) can be computationally intensive when sampling from the posterior

distribution. One way to increase posterior sampling efficiency exploits the sparsity in $\boldsymbol{L}$.

For σ_η and σ_δ, the full conditional distributions are

$$p(\sigma_\eta|\cdot) \propto \frac{1_{\{\sigma_\eta>0\}}}{1+(\sigma_\eta/\upsilon)^2} \times \sigma_\eta^{-(n_h \times n_s)} \times \exp(\alpha^{1/2}\sigma_\eta^{-1}\mathbf{1}'_{n_h \times n_s}\tilde{\boldsymbol{\eta}}_{[1:(n_h \times n_s)]}) \times \exp(-\alpha \mathbf{1}'_{n_h \cdot n_s} \exp(\alpha^{-1/2}\sigma_\eta^{-1}\tilde{\boldsymbol{\eta}}_{[1:(n_h \times n_s)]}))$$

and

$$p(\sigma_\delta|\cdot) \propto \frac{1_{\{\sigma_\delta>0\}}}{1+(\sigma_\delta/\upsilon)^2} \times \sigma_\delta^{-n_s} \times \exp(\alpha^{1/2}\mathbf{1}'_{n_s}\sigma_\delta^{-1}\boldsymbol{I}_{n_s}\boldsymbol{\xi} - \alpha \mathbf{1}'_{n_s} \exp(\alpha^{-1/2}\sigma_\delta^{-1}\boldsymbol{\xi})),$$

where $\boldsymbol{\xi}$ contains the last n_s elements in $\tilde{\boldsymbol{\eta}}$.

4.2. Negative Binomial Echo State Networks

This subsection proposes a negative binomial model as a Poisson alternative, accounting for possible over-dispersion in our conditional marginal distributions. Our setup utilizes the negative binomial distributional specification

$$Y_{i,t}|r_i, p_{i,t} \overset{ind}{\sim} \text{NB}(r_i, p_{i,t}),$$

with the probability mass function

$$P(Y_{i,t} = y_{i,t}) = \frac{\Gamma(y_{i,t}+r_i)}{\Gamma(r_i)\Gamma(y_{i,t}+1)} (p_{i,t})^{r_i} (1-p_{i,t})^{y_{i,t}}.$$

We model $p_{i,t}$ via the logit transformation

$$\ln\left(\frac{p_{i,t}}{1-p_{i,t}}\right) = \boldsymbol{h}'_{i,t}\boldsymbol{\eta}_{s(i)} + \boldsymbol{\delta}_{s(i)},$$

where $\boldsymbol{h}_{i,t}$ again generated by Equation 6. While an alternative parametrization of the negative binomial distribution would allow one to model the mean directly rather than indirectly through the probabilities, however, our parametrization choice here permits a computationally convenient data augmentation approach to model fitting. Lastly, although one could fit a negative Binomial model without the hierarchical structure (similar to the Poisson model from Section 4.1.2), we do not explore this here. The hierarchical structure and the negative binomial likelihood are two different aspects (dependence structure vs. over-dispersion); as we demonstrate in Section 6, both are important.

For estimation, the likelihood function for this model is

$$\mathcal{L}(\tilde{\boldsymbol{Y}}_{i,t}|\tilde{\boldsymbol{\eta}}, r_i) \propto \frac{\Gamma(y_{i,t}+r_i)}{\Gamma(r_i)\Gamma(y_{i,t}+1)} \frac{\exp(\tilde{\boldsymbol{h}}'_{i,t}\tilde{\boldsymbol{\eta}})^{r_i}}{\left(1+\exp(t\tilde{\boldsymbol{h}}'_{i,t}\tilde{\boldsymbol{\eta}})\right)^{y_{i,t}+r_i}}, \tag{9}$$

where $\tilde{\boldsymbol{h}}_{i,t}$ and $\tilde{\boldsymbol{\eta}}$ are as in (8). In a Bayesian framework, Pólya-Gamma data augmentation can be employed to aid in posterior sampling. By Equation (2) in Polson et al. (2013), the second fraction in (9) is

$$2^{-b} e^{\varkappa_{i,t}\psi_{i,t}} \int_0^\infty e^{-\omega_{i,t}\psi_{i,t}^2/2} p(\omega_{i,t}) d\omega_{i,t},$$

where $b_{i,t} = Y_{i,t} + r_i$, $\varkappa_{i,t} = r_i - (y_{i,t} + r_i)/2$, and ω follows a Pólya-Gamma$(b_{i,t}, 0)$ distribution with $\psi = \tilde{\boldsymbol{h}}_{i,t}'\tilde{\boldsymbol{\eta}}$. Therefore, the likelihood is

$$\frac{\Gamma(y_{i,t} + r_i)}{\Gamma(r_i)\Gamma(y_{i,t}+1)} 2^{-(y_{i,t}+r_i)} \exp(\varkappa_{i,t}\tilde{\boldsymbol{h}}_{i,t}'\tilde{\boldsymbol{\eta}}) \int_0^\infty \exp(-\omega_{i,t}\psi_{i,t}^2/2) p(\omega_{i,t}) d\omega_{i,t}.$$

Given a prior $\pi(\tilde{\boldsymbol{\eta}})$, the full conditional for $\tilde{\boldsymbol{\eta}}$ can be calculated via

$$p(\tilde{\boldsymbol{\eta}}|\cdot) \propto \pi(\tilde{\boldsymbol{\eta}}) \exp\left\{-\frac{1}{2}(\boldsymbol{H}\tilde{\boldsymbol{\eta}} - \boldsymbol{\zeta})'\Omega(\boldsymbol{H}\tilde{\boldsymbol{\eta}} - \boldsymbol{\zeta})\right\},$$

where $\boldsymbol{H}$ is the combined $\tilde{\boldsymbol{h}}_{i,t}$ by row, $\boldsymbol{\Omega}$ is a diagonal matrix with diagonal elements $(\omega_{1,1}, \ldots, \omega_{1,T}; \ldots; \omega_{N,1}, \ldots, \omega_{N,T})$, and

$$\boldsymbol{\zeta} = \left(\frac{\varkappa_{1,1}}{\omega_{1,1}}, \ldots, \frac{\varkappa_{N,T}}{\omega_{N,T}}\right).$$

If a multivariate normal prior with mean $\boldsymbol{\mu}_\eta$ and covariance matrix $\boldsymbol{\Sigma}_\eta$ is assumed for $\tilde{\boldsymbol{\eta}}$, then

$$\tilde{\boldsymbol{\eta}}|\cdot \sim \text{MVN}(\boldsymbol{\mu}_\eta^*, \boldsymbol{\Sigma}_\eta^*),$$

where

$$\begin{aligned}\boldsymbol{\Sigma}_\eta^* &= (\boldsymbol{H}'\boldsymbol{\Omega}\boldsymbol{H} + \boldsymbol{\Sigma}_\eta^{-1})^{-1}, \\ \boldsymbol{\mu}_\eta^* &= \boldsymbol{\Sigma}_\eta^*(\boldsymbol{H}'\boldsymbol{\varkappa} + \boldsymbol{\Sigma}_\eta^{-1}\boldsymbol{\mu}_\eta).\end{aligned}$$

On the other hand,

$$\omega_{i,t}|\cdot \sim \text{PG}(b_{i,t}, \tilde{\boldsymbol{h}}_{i,t}'\tilde{\boldsymbol{\eta}}),$$

where PG denotes a Pólya-Gamma distribution, and $\boldsymbol{\varkappa}$ is a vector of stacked $\varkappa_{i,t}$. A zero vector is chosen for $\boldsymbol{\mu}_\eta$ to shrink the parameters towards zero, and $\boldsymbol{\Sigma}_\eta$ is the diagonal matrix

$$\begin{bmatrix} \sigma_\eta^2 \boldsymbol{I}_{(n_h \cdot n_s)} & \boldsymbol{0}_{(n_h \cdot n_s)\times n_s} \\ \boldsymbol{0}_{n_s \times (n_h \cdot n_s)} & \sigma_\delta^2 \boldsymbol{I}_{n_s} \end{bmatrix}.$$

The prior distributions for σ_η^2 and σ_δ^2 are inverse Gammas using the rate parameterizations $\sigma_\eta^2 \sim \text{IG}(\alpha_\eta, \beta_\eta)$ and $\sigma_\delta^2 \sim \text{IG}(\alpha_\delta, \beta_\delta)$. In this paper, $\alpha_\eta = \alpha_\delta = 0.001$ and $\beta_\eta = \beta_\delta = 0.001$ are chosen to yield vague priors. Finally, the inverse of the dispersion

parameter r_i follows the half-Cauchy distribution

$$\frac{1}{r_i} \sim \text{Half-Cauchy}(0, 1).$$

The joint distribution now follows as

$$\begin{aligned}
&p(\boldsymbol{Y}, \tilde{\boldsymbol{\eta}}, \boldsymbol{\omega}, \sigma_\eta, \sigma_\delta, r_i) \propto \\
&\prod_{i=1}^{N}\prod_{t=1}^{T}\left\{\frac{\Gamma(y_{i,t}+r_i)}{\Gamma(r_i)\Gamma(y_{i,t}+1)}2^{-(y_{i,t}+r_i)}\right\}\exp\left\{-\frac{1}{2}(\boldsymbol{H}\tilde{\boldsymbol{\eta}}-\boldsymbol{\zeta})'\Omega(\boldsymbol{H}\tilde{\boldsymbol{\eta}}-\boldsymbol{\zeta})\right\}\times \\
&|\boldsymbol{\Sigma}_\eta|^{-1/2}\exp\left(-\frac{1}{2}\tilde{\boldsymbol{\eta}}'\boldsymbol{\Sigma}_\eta^{-1}\tilde{\boldsymbol{\eta}}\right)\times\prod_{i=1}^{N}\prod_{t=1}^{T}\text{PG}(\omega_{i,t}|b_{i,t},0)\times \\
&\text{IG}(\sigma_\eta|\alpha_\eta,\beta_\eta)\times\text{IG}(\sigma_\delta|\alpha_\delta,\beta_\delta)\times\text{Half-Cauchy}(1/r_i|0,1).
\end{aligned}$$

When sampling r_i, Metropolis-Hastings is used. The proposal distribution is chosen as uniform having a mean of the current value and range $2(\min\{10, r_i\})$.

5. Model Comparisons and Scoring Procedures

This section discusses how we evaluate model fits (scoring criteria). One of the most commonly used scoring criteria involves the one-step-ahead mean squared prediction errors (MSPE), defined at time t by

$$\text{MSPE}_t = \frac{1}{n_S}\sum_{i=1}^{n_S}(\hat{Y}_{i,t} - Y_{i,t})^2,$$

where $\hat{Y}_{i,t}$ denotes the one-step-ahead prediction of the number of graduate students in school i at time t. As discussed above, the type of one-step-ahead prediction used depends on the model type fitted. In a few cases, schools experience a large shift in their year-to-year graduate student counts. The MSPE is not robust to such outliers. To address this limitation, the mean squared logarithmic prediction error (MSLPE), defined as

$$\text{MSLPE}_t = \frac{1}{N}\sum_{i=1}^{N}(\ln(\hat{Y}_{i,t}+1) - \ln(Y_{i,t}+1))^2,$$

can be used. Here, unity is added to observed and predicted counts to avoid taking a logarithm of zero. Smaller MSPEs and MSLPEs indicate better-fitting models.

A measurement of the quality of the uncertainty quantification is the interval score (IS), defined as

$$\text{IS}_t(\alpha) = \frac{1}{N}\sum_{i=1}^{N}\left\{(u_{i,t}-l_{i,t}) + \frac{2}{1-\alpha}(l_{i,t}-Y_{i,t})I_{[Y_{i,t}<l_{i,t}]} + \frac{2}{1-\alpha}(Y_{i,t}-u_{i,t})I_{[Y_{i,t}>l_{i,t}]}\right\},$$

where $l_{i,t}$ and $u_{i,t}$ are the lower and upper bounds of the α prediction interval for school i at time t, respectively. Here, the 95% prediction interval $\alpha = 0.95$ is used. A lower interval score indicates a more favorable predictive distribution.

Another uncertainty metric is the interval coverage rate (ICR), which measures the proportion of observations that fall in the $\alpha \times 100\%$ prediction interval:

$$\text{ICR}(\alpha) = \frac{1}{N}\sum_{i=1}^{N} I_{[l_{i,t}<Y_{i,t}<u_{i,t}]}.$$

ICR values close to α are indicative of a well-calibrated predictive distribution.

6. Analysis of the GSS Series

This section fits the proposed models to the GSS data and compares their performance. As a baseline, an intercept-only model for each school, which describes the scenario where the counts are modeled through mean effects only, is fitted. We also fit a "persistence model" where the observation for the previous year is used as the prediction for the current year for each school (separately). Although simple, this prediction is often competitive with state of the art statistical and machine learning models (Bonas et al. 2025) and makes for a natural point of comparison. Next, a Poisson INGARCH model (Ferland et al. 2006; Fokianos et al. 2009) is fitted via the `R` package `tscount` (Liboschik et al. 2017). The remaining fitted models are variants of ESNs. First, a Single Poisson ESN is fitted. Only point estimate for this model are considered (no bootstrapping); hence, uncertainty quantification is not considered. Thereafter, an Ensemble Poisson ESN is fitted, which allows for both point estimation and uncertainty quantification. A Bayesian Poisson ESN is also fitted. Finally, Bayesian hierarchical ESNs, which take into account dependence between schools residing within the same US state, are considered for both Poisson and negative binomial marginal distributions. For all ESN models, a lag-one autoregressive component is added to the output layer.

The ESN hyperparameters n_h, ν, $a_{(\cdot)}$, τ, and $\pi_{(\cdot)}$ were chosen via cross validation. For the Bayesian Poisson ESN, no burn-in or thinning was used since sampling was done directly from the posterior without MCMC. For the Bayesian Hierarchical NB ESN, 1,000 samples were used for burn in and every other sample in the generated Markov chain was used therefter. For the Bayesian Hierarchical Poisson ESN, the first 500 samples are burned in, and every other sample is again saved. Since the posterior distribution of σ_η and σ_δ do not have a closed form, a Metropolis-Hastings step is needed. Due to the positivity of σ_η and σ_δ, the proposal distribution is taken as uniform centered at the current value and having range $2\min\{0.5, \sigma_\delta^{(t-1)}\}$ and $2\min\{0.5, \sigma_\eta^{(t-1)}\}$, respectively. Here, $\sigma_\delta^{(t-1)}$ and $\sigma_\eta^{(t)}$ represents the t-th posterior sample for $\sigma_\delta^{(t-1)}$ and $\sigma_\eta^{(t)}$, respectively. Each Bayesian MCMC was run until 1,000 prediction samples were generated.

One-step-ahead predictions for the number of graduate students from 2017 to 2021 were made. These predictions use all data up to the previous year analyzed. Specifically, data from 1972-2016 is used to predict the 2017 counts; data from 1972-2020 are used to predict the 2021 counts. The models are refit each year for a new prediction year.

The GSS data are generally overdispersed. To see this, the mean of the sample mean counts over all 1,758 schools is 63.48 and the mean of the sample variances (a denom-

	2017	2018	2019	2020	2021	5 Year Mean(SD)
Intercept	3351	3110	3014	1886	2431	2758(594)
Persistence	887	**223**	**241**	1192	290	566(445)
INGARCH(1,1)	1129	612	512	960	587	760(269)
Single ESN	846	261	422	1222	295	609(414)
Ensemble ESN	843	286	415	1232	291	614(414)
Bayesian Poisson ESN	915	279	278	1240	263	595(455)
Bayesian Hierarchical NB ESN	**822**	250	243	**968**	289	**514**(352)
Bayesian Hierarchical Poisson ESN	878	224	275	1021	**246**	529(388)

Table 1.: Mean squared one-step-ahead prediction errors for all fitted models. The best score in each column is bolded.

	2017	2018	2019	2020	2021	5 Year Mean(SD)
Intercept	0.405	0.393	0.401	0.457	0.490	0.429(0.042)
Persistence	0.209	0.079	0.085	**0.159**	0.099	0.126(0.056)
INGARCH(1,1)	0.215	0.109	0.110	0.193	0.144	0.154(0.048)
Single ESN	0.217	0.099	0.122	0.194	0.138	0.154(0.050)
Ensemble ESN	0.217	0.106	0.122	0.200	0.139	0.157(0.049)
Bayesian Poisson ESN	0.210	**0.073**	**0.085**	0.161	**0.097**	**0.125**(0.058)
Bayesian Hierarchical NB ESN	**0.205**	0.090	0.099	0.172	0.122	0.138(0.049)
Bayesian Hierarchical Poisson ESN	0.208	0.073	0.088	0.169	0.100	0.128(0.058)

Table 2.: Mean squared log prediction errors for all fitted models to three significant digits. The best score in each column is bolded.

inator of $n-1$ is used) is 1320.28, significantly more than the mean. Hence, Poisson marginal distributions, which have a unit dispersion, will be inadequate. Because of this, a hierarchical negative binomial setup, which permits more overdispersion than a hierarchical Poisson setup, is considered.

6.1. One-step-ahead Predictions

MSPE and MSLPE scores for various model fits are shown in Table 1 and 2. For MSPE scores, the Bayesian Hierarchical NB ESN performs best on average, followed by the Bayesian Hierarchical Poisson ESN, which also has a relatively small MSE. The INGARCH(1,1) model and the frequentist ESNs, including the Single ESN and Ensemble ESN, exhibit similar performance. For MSEs, the NB ESN model does not perform uniformly best. However, it was able to better capture 2020's volatility (while other models suffer), producing a smaller average.

For MSLPE scores, the Bayesian Poisson ESN performs best, although the MSLPE is very similar to that of the Bayesian Hierarchical Poisson ESN. Note the similarity between the frequentist ESNs and INGARCH scores: on a logarithmic scale, frequentist ESNs and INGARCH models perform similarly.

For model fits with uncertainty quantification, Tables 3 and 4 present IS(0.95) and ICR(0.95) scores. The INGARCH(1,1) model has the best five-year average IS(0.95). For ICR(0.95) scores, the Bayesian Hierarchical NB ESN model has an overall ICR closer to 0.95, suggesting that the Bayesian Hierarchical NB and the Bayesian Hierarchical Poisson ESNs intervals are wider. This could be due to the ESN's sensitivity to changepoint-type shifts in some of the series. Although the prediction intervals are capable of covering such shifts, this comes at the cost of widening the intervals in

other regions. It is useful to note here that the ensemble ESN and the Bayesian ESNs take two different views of uncertainty quantification. The ensemble ESN allows for the construction of an interval that reflects variability in the model estimates based on the underlying distribution of the hidden layer weights. However, for the Bayesian ESNs, intervals are constructed based on the posterior predictive distribution, which is intended to capture uncertainty in the prediction itself. This is reflected in the results, where the Bayesian Poisson ESN has better average interval score and coverage rate compared to the ensemble ESN. As the data are overdispersed, the superior uncertainty quantification for the negative Binomial model is expected.

	2017	2018	2019	2020	2021	5 Year Mean(SD)
INGARCH(1,1)	**309**	**165**	**148**	**237**	**176**	**207**(66)
Ensemble ESN	499	305	304	461	308	376(96)
Bayesian Poisson ESN	471	238	245	410	246	322(110)
Bayesian Hierarchical NB ESN	401	189	195	351	226	272(97)
Bayesian Hierarchical Poisson ESN	519	282	278	432	286	360(110)

Table 3.: Rounded interval scores (0.95) for models with uncertainty quantification. The best score in each column is bolded.

	2017	2018	2019	2020	2021	5 Year Mean(SD)
INGARCH(1,1)	0.720	0.811	0.795	0.769	0.766	0.772(0.035)
Ensemble ESN	0.742	0.815	0.821	0.779	0.820	0.796(0.034)
Bayesian Poisson ESN	0.771	0.874	0.865	0.825	0.862	0.839(0.044)
Bayesian Hierarchical NB ESN	**0.879**	**0.958**	**0.956**	**0.909**	**0.938**	**0.928**(0.034)
Bayesian Hierarchical Poisson ESN	0.761	0.880	0.856	0.825	0.849	0.834(0.046)

Table 4.: Rounded interval coverage rates (0.95) for models with uncertainty quantification. The best score in each column is bolded.

6.2. In-sample Model Adequacy Checking

To further assess the model fits, the conditional standardized Pearson residuals introduced by Weiß et al. (2020) were computed. These residuals are

$$R_t(\hat{\boldsymbol{\theta}}) := \frac{Y_t - E(Y_t|Y_{t-1}, Y_{t-2}, \ldots, Y_1; \hat{\boldsymbol{\theta}})}{\sqrt{\text{Var}(Y_t|Y_{t-1}, Y_{t-2}, \ldots, Y_1; \hat{\boldsymbol{\theta}})}}$$

at time t. Since these residuals are centered and scaled, if the fitted model is adequate, the computed R_t's should have a sample mean close to zero and a unit variance. In other words, if the residuals have a sample mean far from zero, this is an indication of bias. A sample variance that deviates from unity indicates a lack of fit for the variance of data distribution. Boxplots of the sample variance of the residuals from each of the 1,728 schools is shown in Figure 5 for the Bayesian hierarchical NB ESN and Bayesian Hierarchical Poisson ESN fits. The solid lines in the boxes depict the sample median. The Bayesian Hierarchical NB ESN has a sample variance that is close to unity for most schools, and is seen to significantly outperform the Poisson model. This said, two major outliers exist (sample variances above 50) for the Bayesian Hierarchical NB

model. Series for these two schools are plotted in Figure 6, where significant "jumps" during several years are evident. These changepoint-type features may merit further "localized" investigation.

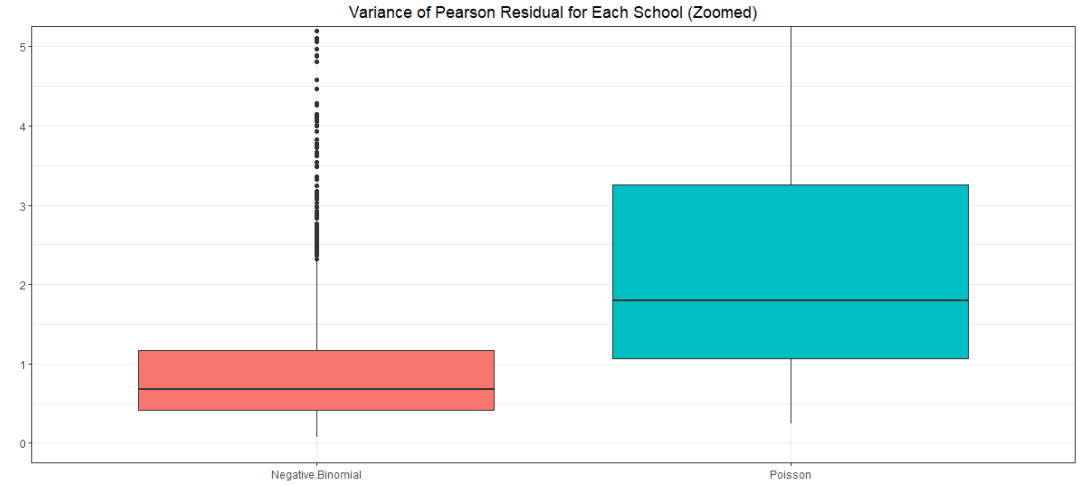

Figure 5.: Boxplot of sample variances of the residuals over all 1,728 schools. Due to two extreme outlying series, the y-axis is truncated to 5. The sample mean residual variance for the Poisson and negative binomial models are 3.39 and 1.18, respectively. Sample medians are indicated by a solid horizontal line in the box.

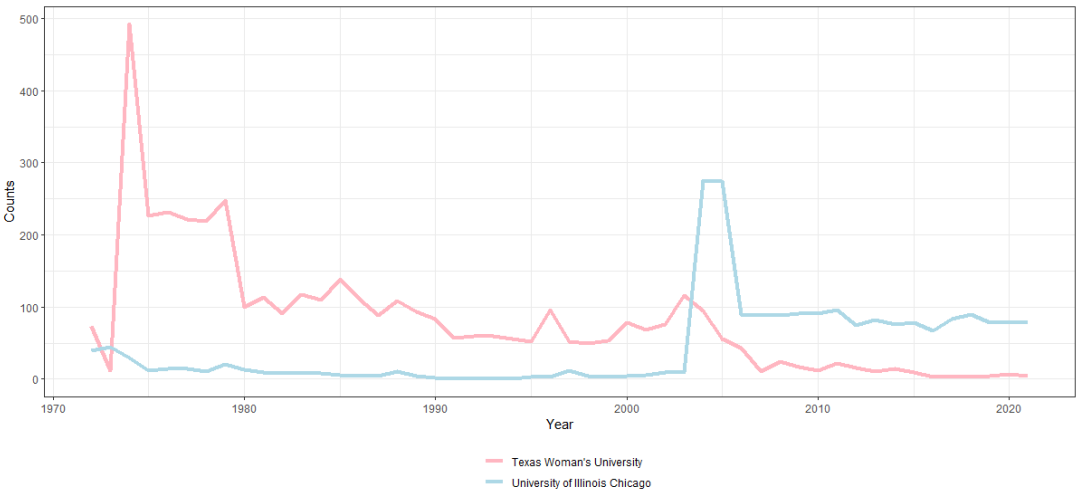

Figure 6.: Time series plots for our two outlying series.

These residuals were further scrutinized in Figures 7 and 8, which show time series plots and sample autocorrelations for the same four schools in Figure 1. The residual series have minimal autocorrelation overall, suggesting that the ESNs effectively capture temporal autocorrelations. However, for the University of Rochester, a few autocorrelations appear significantly different from zero. Given the number of schools and lags considered, this anomaly is not particularly concerning. Moreover, a residual trace plot including 100 randomly selected schools is shown in Figure 9 for reference.

Again, there is no clear autocorrelation present in these residuals.

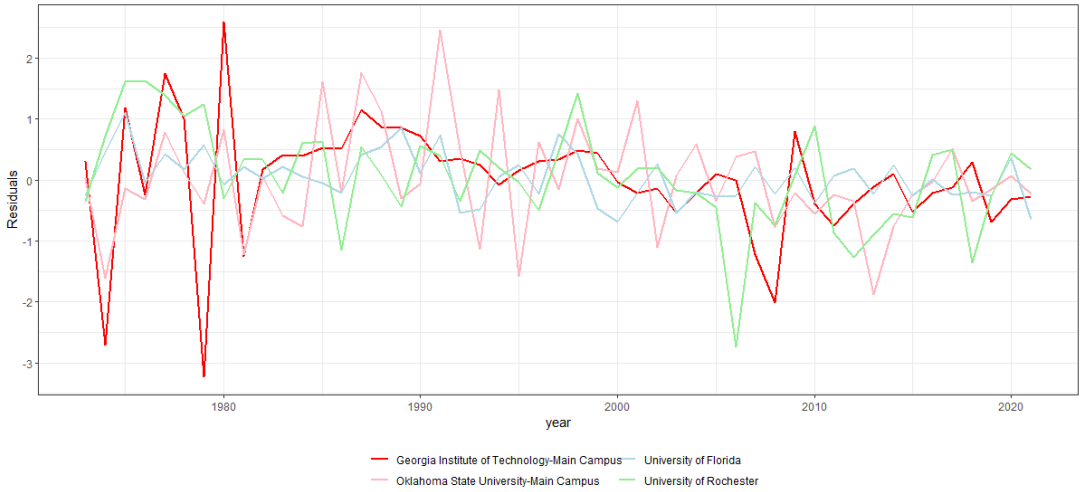

Figure 7.: Model residuals of four randomly selected schools over the period 1973-2021.

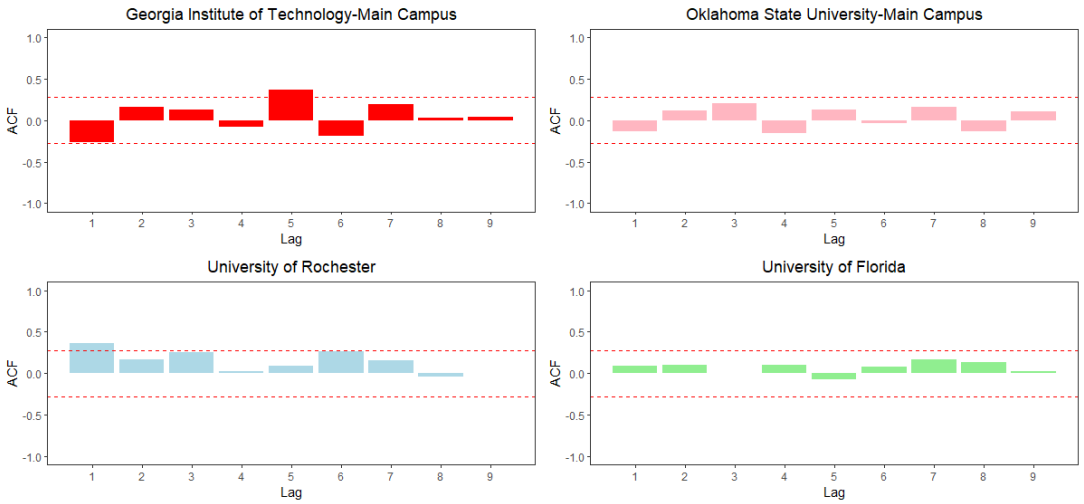

Figure 8.: Residual autocorrelations. The dashed lines are pointwise 95% confidence thresholds for a zero autocorrelation.

7. Discussion

This work developed extensions of ESNs to model graduate student enrollment counts. Parameters were estimated in both frequentist and Bayesian frameworks. Hierarchical structures were introduced to account for correlations from schools within the same geographic state of the US. Both Poisson and negative Binomial data models were considered. One-step-ahead prediction residuals were assessed from 2017 to 2021 using various scoring metrics to compare methods. Our enrollment count series were best modeled by a Bayesian hierarchical model with negative binomial dynamics. In

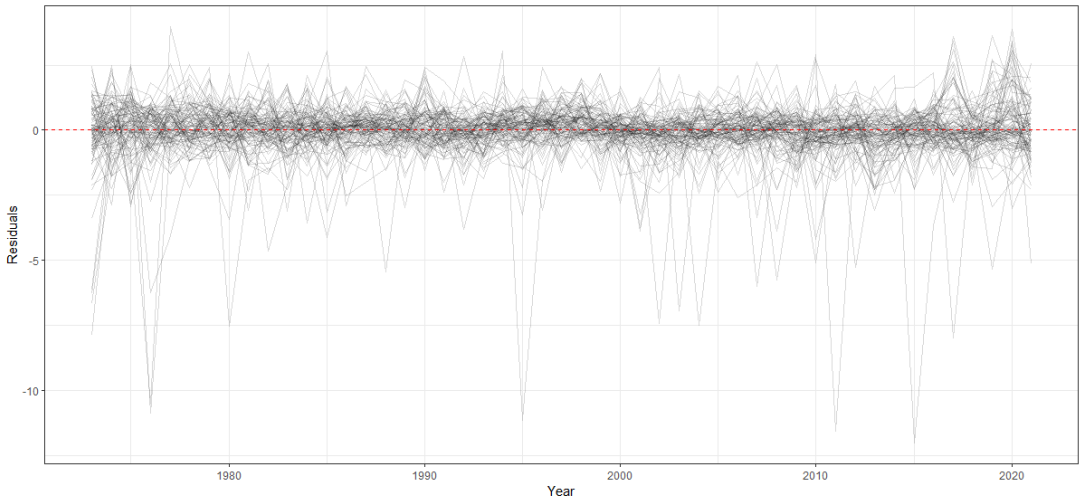

Figure 9.: In-sample standardized residuals ($\{R_t(\hat{\boldsymbol{\theta}})\}$) for the Bayesian hierarchical negative binomial ESN across 100 randomly selected schools.

particular, the use of the negative binomial data model accounts for overdispersion in the data, while the hierarchical structure accounts for dependencies across schools within the same state.

Although our goal here was forecasting and not inference, ESNs inherit the general criticisms of neural networks as being uninterpretable "black-box" models. A growing body of work attempts to address this gap (e.g., see Wikle et al. (2023) for an overview). For example, one straightforward approach is to "feature shuffle" or randomly perturb features and then compare model errors. In policy relevant situations where models need to be interpretable, such methods may allow the use of machine learning approaches such as ESNs.

Future work aims to explore methods for capturing spatial effects in ways not considered here. Wang et al. (2025) propose using spatial radial basis functions in a convolutional neural network to incorporate "spatial effects" at different scales. One potential approach is to apply a spatial deep convolutional neural network (SDCNN) in an extreme learning machine framework (Huang et al. 2006). Here, the outputs from the SDCNN could be used as covariates. By combining SDCNN with ESNs, non-separable and non-stationary spatial covariances (in either count or continuous cases) could possibly be handled. Another area where research is needed lies with marginal distribution types. Although only Poisson and negative binomial structures were considered here, research into an unspecified count family via nonparametric techniques could prove useful. One more avenue of future work involves modeling dependence structures within institutions and across disciplines. One primary challenge here is that universities often group departments within colleges or divisions in very different ways; for example, statistics departments can reside in Engineering, Natural Sciences, or Mathematics clusters. Finally, the strong predictive performance of the persistence approach indicates that it may be worth pursuing models on the difference in counts $\{Y_t - Y_{t-1}\}$ rather than the counts themselves. Although this would induce challenges, recent literature involving distributions supported on all integers are being developed (Kang et al. 2025). Model in this way may result in improved predictions (similar to the persistence approach), while still being model-based and allowing for uncertainty

quantification.

Acknowledgement

This research was partially supported by the U.S. National Science Foundation (NSF) under Grant NCSE-2215169. Robert Lund was partially supported by the NSF grant DMS-2113592. This article is released to inform interested parties of ongoing research and to encourage discussion. The views expressed on statistical issues are those of the authors and not those of the NSF.

References

Agosto, A. and P. Giudici (2020). A Poisson autoregressive model to understand covid-19 contagion dynamics. *Risks 8*(3), 77.

Bonas, M., A. Datta, C. K. Wikle, E. L. Boone, F. S. Alamri, B. V. Hari, I. Kavila, S. J. Simmons, S. M. Jarvis, W. S. Burr, et al. (2025). Assessing predictability of environmental time series with statistical and machine learning models. *Environmetrics 36*(1), e2864.

Bradley, J. R., S. H. Holan, and C. K. Wikle (2018). Computationally efficient multivariate spatio-temporal models for high-dimensional count-valued data (with discussion). *Bayesian Analysis 13*, 253–310.

Bradley, J. R., S. H. Holan, and C. K. Wikle (2020). Bayesian hierarchical models with conjugate full-conditional distributions for dependent data from the natural exponential family. *Journal of the American Statistical Association 115*(532), 2037–2052.

Brandt, P. T. and T. Sandler (2012). A Bayesian Poisson vector autoregression model. *Political Analysis 20*(3), 292–315.

Davis, R. A., K. Fokianos, S. H. Holan, H. Joe, J. Livsey, R. Lund, V. Pipiras, and N. Ravishanker (2021). Count time series: A methodological review. *Journal of the American Statistical Association 116*(535), 1533–1547.

Davis, R. A., S. H. Holan, R. B. Lund, and N. Ravishanker (2016). *Handbook of Discrete-valued Time Series*. Taylor & Francis / CRC Press.

Dey, R. and F. M. Salem (2017). Gate-variants of gated recurrent unit (gru) neural networks. In *2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS)*, pp. 1597–1600. IEEE.

Düker, M.-C., R. Lund, and V. Pipiras (2024). High-dimensional latent Gaussian count time series: Concentration results for autocovariances and applications. *Electronic Journal of Statistics 18*(2), 5484–5562.

Dunsmuir, W. T. (2015). Generalized linear autoregressive moving average models. In R. Davis, S. Holan, R. Lund, and N. Ravishanker (Eds.), *Handbook of Discrete-Valued Time Series. CRC Monographs*. Taylor & Francis / CRC Press.

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

Fokianos, K. (2021). Multivariate count time series modelling. *Econometrics and Statistics 31*, 100–116.

Fokianos, K., A. Rahbek, and D. Tjøstheim (2009). Poisson autoregression. *Journal of the American Statistical Association 104*(488), 1430–1439.

Gamerman, D., C. A. Abanto-Valle, R. S. Silva, T. G. Martins, R. Davis, S. Holan, R. Lund, and N. Ravishanker (2015). Dynamic Bayesian models for discrete-valued time series. In R. Davis, S. Holan, R. Lund, and N. Ravishanker (Eds.), *Handbook of Discrete-valued Time Series*, pp. 165–186. Chapman & Hall / CRC Press.

Heinen, A. and E. Rengifo (2007). Multivariate autoregressive modeling of time series count data using copulas. *Journal of Empirical Finance 14*(4), 564–583.

Hochreiter, S. and J. Schmidhuber (1997). Long short-term memory. *Neural Computation 9*(8), 1735–1780.

Holan, S. H. and C. K. Wikle (2016). Hierarchical dynamic generalized linear mixed models for discrete-valued spatio-temporal data. In R. Davis, S. Holan, R. Lund, and N. Ravishanker (Eds.), *Handbook of Discrete-Valued Time Series*, pp. 327–348. Taylor & Francis / CRC Press.

Huang, G.-B., Q.-Y. Zhu, and C.-K. Siew (2006). Extreme learning machine: theory and applications. *Neurocomputing 70*(1-3), 489–501.

Jaeger, H. and H. Haas (2004). Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. *science 304*(5667), 78–80.

Jia, Y., S. Kechagias, J. Livsey, R. Lund, and V. Pipiras (2023). Latent Gaussian count time series. *Journal of the American Statistical Association 118*(541), 596–606.

Jin-Guan, D. and L. Yuan (1991). The integer-valued autoregressive (inar(p)) model. *Journal*

of Time Series Analysis 12(2), 129–142.

Joe, H. (2016). Markov models for count time series. In R. Davis, S. Holan, R. Lund, and N. Ravishanker (Eds.), *Handbook of Discrete-valued Time Series*, pp. 49–70. Taylor & Francis / CRC Press.

Jung, R. C. and A. Tremayne (2006). Binomial thinning models for integer time series. *Statistical Modelling 6*(2), 81–96.

Kang, Y., Y. Zhang, S. Wang, and Z. Zhao (2025). A new class of z-valued inar (1) models with application to mutual fund flows. *Economics Letters*, 112339.

Karlis, D. (2016). Models for multivariate count time series. In R. Davis, S. Holan, R. Lund, and N. Ravishanker (Eds.), *Handbook of Discrete-valued Time Series*, pp. 407–424. Taylor & Francis / CRC Press.

Kim, B., S. Lee, and D. Kim (2021). Robust estimation for bivariate Poisson INGARCH models. *Entropy 23*(3), 367.

Kong, J. and R. B. Lund (2024). Poisson count time series. *Journal of Time Series Analysis 45*, doi.org/10.1111/jtsa.12799.

Liboschik, T., K. Fokianos, and R. Fried (2017). tscount: An r package for analysis of count time series following generalized linear models. *Journal of Statistical Software 82*, 1–51.

Matteson, D. S., M. W. McLean, D. B. Woodard, and S. G. Henderson (2011). Forecasting emergency medical service call arrival rates. *Annals of Applied Statistics 5*, 1379–1406.

McCulloch, C. E. and S. R. Searle (2004). *Generalized, linear, and mixed models*. John Wiley & Sons.

McDermott, P. L. and C. K. Wikle (2017). An ensemble quadratic echo state network for non-linear spatio-temporal forecasting. *Stat 6*(1), 315–330.

McDermott, P. L. and C. K. Wikle (2019). Deep echo state networks with uncertainty quantification for spatio-temporal forecasting. *Environmetrics 30*(3), e2553.

Neelon, B., P. Ghosh, and P. F. Loebs (2013). A spatial Poisson hurdle model for exploring geographic variation in emergency department visits. *Journal of the Royal Statistical Society Series A: Statistics in Society 176*(2), 389–413.

Ord, K., C. Fernandes, and A. Harvey (1993). Time series models for multivariate series of count data. In T. S. Rao (Ed.), *Developments in Time Series Analysis*, pp. 295–309. Taylor & Francis.

Pan, Y. and J. Pan (2024). Modelling volatilities of high-dimensional count time series with network structure and asymmetry. *arXiv preprint arXiv:2409.01521*.

Parker, P. A., S. H. Holan, and R. Janicki (2020). Conjugate bayesian unit-level modelling of count data under informative sampling designs. *Stat 9*(1), e267.

Polson, N. G., J. G. Scott, and J. Windle (2013). Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American Statistical Association 108*(504), 1339–1349.

Robbins, M. W., R. B. Lund, C. M. Gallagher, and Q. Lu (2011). Changepoints in the North Atlantic tropical cyclone record. *Journal of the American Statistical Association 106*(493), 89–99.

Rumelhart, D. E., G. E. Hinton, and R. J. Williams (1986). Learning representations by back-propagating errors. *Nature 323*(6088), 533–536.

Schafer, T. L. (2020). *Alternative learning strategies for spatio-temporal processes of complex animal behavior*. Ph. D. thesis, University of Missouri–Columbia.

Schoenmaker, D. (1996). *Contagion Risk in Banking*. Citeseer.

Serhiyenko, V., N. Ravishanker, and R. Venkatesan (2015). Approximate Bayesian estimation for multivariate count time series models. In *Ordered Data Analysis, Modeling and Health Research Methods: In Honor of HN Nagaraja's 60th Birthday*, pp. 155–167. Springer.

Wang, Q., P. A. Parker, and R. Lund (2025). Spatial deep convolutional neural networks. *Spatial Statistics*, 100883.

Wang, Z., S. H. Holan, and C. K. Wikle (2024). Echo state networks for spatio-temporal area-level data. *arXiv preprint arXiv:2410.10641*.

Weiß, C., L. Scherer, B. Aleksandrov, and M. Feld (2020). Checking model adequacy for

count time series by using pearson residuals. *Journal of Time Series Econometrics 12*(1), 20180018.

Weiß, C. H. (2008). Thinning operations for modeling time series of counts—a survey. *AStA Advances in Statistical Analysis 92*, 319–341.

Wikle, C. K., A. Datta, B. V. Hari, E. L. Boone, I. Sahoo, I. Kavila, S. Castruccio, S. J. Simmons, W. S. Burr, and W. Chang (2023). An illustration of model agnostic explainability methods applied to environmental data. *Environmetrics 34*(1), e2772.

Zhu, R. and H. Joe (2010). Negative binomial time series models based on expectation thinning operators. *Journal of Statistical Planning and Inference 140*(7), 1874–1888.