DRAFT: Statistical Deep Learning for Dependent Establishment Data

Qi Wang, Paul A. Parker, and Robert Lund

Abstract

The recent boom of data science has lead to a surge in popularity of deep-learning methods. These modeling approaches are powerful, with the ability to learn highly flexible nonlinear functions. However, in many cases, deep-learning approaches are cast as optimization problems, with no straightforward mechanism to assess the uncertainty around any generated estimates. This limits their use in situations where uncertainty is important, such as production of establishment statistics. In this work, we develop a statistical deep learning approach that allows for prediction of establishment data, while accounting for the uncertainty around these model-based estimates. Importantly, the proposed approach can also leverage various dependence structures (i.e., spatial, temporal, etc.) in order to improve the precision of relevant estimates. We illustrate the approach through a motivating example using data from the National Center for Science and Engineering Statistics.

Keywords: Counts, Echo State Network, Poisson, NCSES.

1 Introduction

The field of establishment statistics is a growing and important sub-field of survey research. It is distinguished by the study of population units other than individual people or households. For example, some common common examples of establishments are businesses, farms, and schools, among other institutions (Snijkers et al., 2023). Establishment statistics play a critical role in informing policy-makers as well as improving decision making for the underlying establishments themselves. However, due to the complexity of many establishment datasets, there is a need to develop new methodology in order to improve the quality of any statistics produced.

One prominent issue that arises when working with establishment data is the need to model non-Gaussian response variables. For example, many variables are count-valued, such as the number of employees in a given business or the number of students in a given school (see Savitsky and Toth (2016) for example). Traditional linear models are often inappropriate due to the discrete nature of these counts. Additionally, various dependence structures often present themselves in the analysis of establishment data. For example, data might be collected sequentially over a number of years, resulting in temporal correlation, or units may be nested in larger entities such as states, with differing policies. Together, these challenges highlight the need for new statistical techniques that are tailored to the specific challenges presented by establishment data.

The primary goal of our research is to analyze and interpret counts of graduate students provided by the Survey of Graduate Students and Postdoctorates in Science and Engineering (GSS). The GSS is an annual census of U.S. academic institutions sponsored by the National Center for Science and Engineering Statistics, a subsidiary of the National Science Foundation. Data is available from 1972 to 2021 and includes the total number of full-time graduate students during each year for each school. The data provides valuable insight into the demographics and distribution of graduate students across different fields and institutions, and may be used to understand past trends and predict future dynamics in graduate education.

To model GSS graduate student counts, we develop statistical deep learning methods for sequential count data based on the echo state network (Jaeger, 2002). An important advantage of echo state networks in statistical settings is their ability to handle non-linear and non-stationary time series data. Traditional time series models often assume linearity and stationarity. These assumptions can be limiting in many real-world scenarios where data exhibit complex dependencies and non-linear interactions. Echo state networks, which use reservoir dynamics, can capture such complexities effectively (Lukoševičius and Jaeger, 2009). Researchers have successfully applied statistical echo state networks to various domains, including financial time series prediction (Parker et al., 2021) and environmental or climate modeling (McDermott and Wikle, 2017; Huang et al., 2022; Bonas et al., 2024).

The remainder of this paper is outlined as follows. In Section 2, we introduce our count echo state network after covering appropriate background information on deep learning for sequential data. Section 3 illustrates an analysis of GSS graduate student count data using the count echo state network, including an out-of-sample prediction study. Finally, Section 4 provides some discussion and concluding remarks.

2 Methodology

In order to predict graduate student counts based on data from the GSS, we develop deep learning methodology for count time series. In particular, we build upon a type of recurrent neural network known as an echo state network, by adapting it to handle counts.

2.1 Recurrent Neural Networks

The most common type of neural network is a feed-forward neural network. Although these are powerful tools for estimating complex nonlinear functions, they are not equipped to handle dependent data, such as time series data. In contrast to this, recurrent neural networks (RNNs) have been developed specifically for prediction of sequential data.

As an example, consider the GSS graduate student count data. Let, $\{Y_t\}_{t=1}^{T}$, denote the observations from a single school, from time $t = 1, \ldots, T$. Further suppose that a length r vector of covariates is also given for each time point, $\boldsymbol{x}_t$, with a one in the first element, corresponding to in intercept term. An RNN can be constructed by first modeling a length-n_h vector $\boldsymbol{h}_1 = g\left(\boldsymbol{U}\boldsymbol{x}_1\right)$, for some pre-specified function $g(\cdot)$, and $n_h \times r$ matrix of coefficients, $\boldsymbol{U}$. The vector $\boldsymbol{h}_1$ is termed the hidden layer for time $t = 1$, and n_h represents the number of hidden nodes. Subsequently, for each $t = 2, 3, \ldots, T$ the hidden layer $\boldsymbol{h}_t$ is calculated as:

$$\boldsymbol{h}_t = g(\boldsymbol{W}\boldsymbol{h}_{t-1} + \boldsymbol{U}\boldsymbol{x}_t).$$

Thus, the hidden layer at time t is constructed by first taking an affine transformation of the current inputs ($\boldsymbol{x}_t$) and the previous hidden state ($\boldsymbol{h}_{t-1}$), and then taking a nonlinear transformation. Importantly, the hidden layer at time t is connected to the hidden layer at time $t-1$, which accounts for the sequential nature of the data. The nonlinear function $g(\cdot)$ is known as the activation function and is usually chosen to be bounded in order to avoid overflow issues. Common choices include the hyperbolic tangent function and the Sigmoid function. The calculation of $\boldsymbol{h}_t$ in a recurrent layer is graphically illustrated in Figure 1.

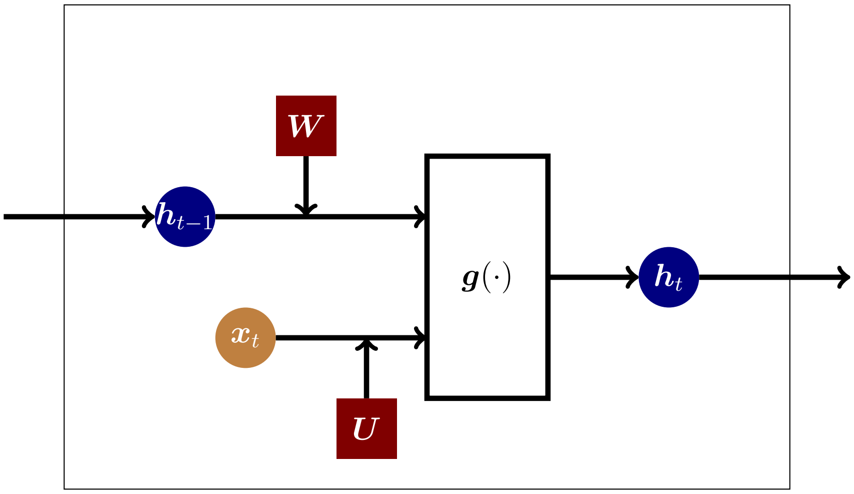

Figure 1: Graphical depiction of a recurrent layer.

Finally, an output layer is included in the model,

$$\hat{y}_t = g_o(\boldsymbol{h}_t'\boldsymbol{\eta}),$$

where $\boldsymbol{\eta}$ is a length n_h vector of coefficients, and $g_o(\cdot)$ is known as the output layer activation function. The choice of activation function for the output layer typically depends on the support of the data. For example, with unbounded continuous data, one would use the identity function, whereas for binary data, one would use the sigmoid function. Under this setup, $\hat{y}_t$ can serve as a prediction for the true response Y_t. Note that the standard RNN is not a statistical model, as no distributional assumptions have been made. The parameters $\boldsymbol{U}$, $\boldsymbol{W}$, and $\boldsymbol{\eta}$ must be estimated. This can be done through the minimization of an appropriate loss function. For example, when working with continuous data, one might minimize the sum of squares loss,

$$\sum_{t=1}^{T}(\hat{y}_t - Y_t)^2.$$

Gradient descent techniques are usually employed for the optimization problem, and in particular, variants of stochastic gradient descent have become the norm in the field of deep learning.

2.2 Echo State Networks

One alternative to traditional RNNs is the Echo State Network (ESN). Instead of estimating all parameters in the model, an ESN randomly samples and fixes the hidden layer weights before model fitting (Jaeger, 2007). The ESN for a univariate time series has the same general structure as an RNN,

$$\begin{aligned} y_t &= g_o(\boldsymbol{h}_t'\boldsymbol{\eta}) \\ \boldsymbol{h}_t &= g\left(\frac{\nu}{|\lambda_W|}\boldsymbol{W}\boldsymbol{h}_{t-1} + \boldsymbol{U}\boldsymbol{x}_t\right), \end{aligned} \tag{1}$$

Importantly, the weight matrices $\boldsymbol{W}$ and $\boldsymbol{U}$ are randomly chosen and fixed. Thus, the only parameters that must be learned are $\boldsymbol{\eta}$. This results in a much easier optimization problem. For example, with the identity activation function in the output layer, $\boldsymbol{\eta}$ can be estimated via ordinary least squares. Alternatively, it may be desirable to add regularization in order to prevent overfitting, through the use of penalized regression approaches. Note here that λ_W represents the largest eigenvalue of the randomly generated matrix $\boldsymbol{W}$, while ν is a fixed scaling factor between zero and one. This scaling is recommended by McDermott and Wikle (2019a) to limit the spectral radius of $\boldsymbol{W}$ and prevent instability in model fitting. For illustration purposes, we fix $\nu = 0.35$. Although we have not found the results to be sensitive to this choice for our application of interest, with other datasets, it may be desirable to tune this parameter.

More recently, McDermott and Wikle (2017) used the ESN in a statistical context by linking the hidden layer outputs to a Gaussian likelihood. They quantify uncertainty through the use of an ensemble of ESN models. McDermott and Wikle (2019a) provide a natural alternative for uncertainty quantification through a Bayesian ESN by placing a prior distribution on the output layer weights. Finally, McDermott and Wikle (2019b) consider deeper model hierarchies through the use of multiple hidden layers.

2.3 Count Echo State Network

Existing echo state networks are not adapted to handle count data, however, as McDermott and Wikle (2017) show, it is natural to link an echo state network to a likelihood. With this in mind, we propose the count echo state network (CESN), which links an echo state network to a Poisson likelihood,

$$\begin{aligned} Y_t|\lambda_t &\overset{ind}{\sim} \text{Poisson}(\lambda_t) \\ \log(\lambda_t) &= \boldsymbol{h}_t'\boldsymbol{\eta} \\ \boldsymbol{h}_t &= g\left(\frac{\nu}{|\lambda_W|}\boldsymbol{W}\boldsymbol{h}_{t-1} + \boldsymbol{U}\boldsymbol{x}_t\right),\ t = 2, \ldots, T \\ \boldsymbol{h}_1 &= g\left(\boldsymbol{U}\boldsymbol{x}_1\right). \end{aligned} \tag{2}$$

Here, we use a hyperbolic tangent activation function $g(\cdot)$ and the elements of $\boldsymbol{U}$ and $\boldsymbol{W}$ are randomly generated and fixed before model fitting. Specifically, we generate each element independently from a two component mixture with 10% weight from a uniform distribution between -0.1 and 0.1, while the other 90% weight is given to a point mass at zero. Although the two component approach with a point mass at zero is not necessary, it can lead to sparse matrices which may be computationally advantageous. Other distributions may work as well, although in general we recommend distributions that are symmetric and centered around zero. Finally, after generating and fixing $\boldsymbol{U}$ and $\boldsymbol{W}$, the only parameter that must be estimated is $\boldsymbol{\eta}$.

The model (2) can be estimated via penalized maximum likelihood. For example, with an L1 penalty, to estimate $\boldsymbol{\eta}$, we would maximize the loss function,

$$\mathcal{L} = \sum_{t=1}^{T} \exp\left(Y_t \boldsymbol{h}_t' \boldsymbol{\eta} - \exp(\boldsymbol{h}_t' \boldsymbol{\eta})\right) - \tau \sum_{j=1}^{n_h} |\eta_j|.$$

This can be done via standard software such as the `glmnet` package (Friedman et al., 2010). Note that for the purposes of illustration, we fix $\tau = 2$, although this parameter could be tuned via cross validation to further improve predictive performance.

Fitting model (2) a single time results in a point estimate but no uncertainty quantification. As an alternative, one may consider an ensemble approach, where the model is refit $M > 1$ times, resampling the matrices $\boldsymbol{U}$ and $\boldsymbol{W}$ each time, as suggested by McDermott and Wikle (2017). This allows one to use the ensemble average as a point estimate, which may be more robust than a single CESN. Additionally, one can consider ensemble quantiles or standard deviations as a measure of uncertainty. Herein, when fitting ensemble CESNs, we let $M = 100$.

2.4 Bayesian CESN

Another approach to fitting the CESN is through the Bayesian paradigm. That is, after putting an appropriate prior on $\boldsymbol{\eta}$, we can sample from the posterior distribution. A standard Gaussian prior is not conjugate with the Poisson likelihood, and thus would require the use of rejection sampling techniques such as the Metropolis-Hastings algorithm. These can be difficult to tune, especially in high-dimensional settings, such as the many hidden nodes used in a CESN. Instead, we turn to the multivariate log-Gamma (MLG) distribution, developed by Bradley et al. (2018) and Bradley et al. (2020).

The density for the MLG distribution is given as

$$\det(\boldsymbol{V}^{-1})\left\{\prod_{i=1}^{n}\frac{\kappa_i^{\alpha_i}}{\Gamma(\alpha_i)}\right\}\exp\left[\boldsymbol{\alpha}'\boldsymbol{V}^{-1}(\boldsymbol{Y}-\boldsymbol{\mu})-\boldsymbol{\kappa}'\exp\left\{\boldsymbol{V}^{-1}(\boldsymbol{Y}-\boldsymbol{\mu})\right\}\right], \tag{3}$$

denoted by $\text{MLG}(\boldsymbol{\mu}, \mathbf{V}, \boldsymbol{\alpha}, \boldsymbol{\kappa})$. Sampling from $\boldsymbol{Y} \sim \text{MLG}(\boldsymbol{\mu}, \mathbf{V}, \boldsymbol{\alpha}, \boldsymbol{\kappa})$ can be done efficiently through the following steps:

1. Generate a vector $\mathbf{g}$ as n independent Gamma random variables with shape α_i and rate κ_i, for $i = 1, \ldots, n$
2. Let $\mathbf{g}^* = \log(\mathbf{g})$
3. Let $\mathbf{Y} = \mathbf{V}\mathbf{g}^* + \boldsymbol{\mu}$.

Bayesian inference using the MLG prior distribution will also require simulation from the related conditional multivariate log-Gamma distribution (cMLG). Letting $\mathbf{Y} \sim \text{MLG}(\boldsymbol{\mu}, \mathbf{V}, \boldsymbol{\alpha}, \boldsymbol{\kappa})$, Bradley et al. (2018) show that $\mathbf{Y}$ can be partitioned into $(\mathbf{Y_1}', \mathbf{Y_2}')'$, where $\mathbf{Y_1}$ is r-dimensional and $\mathbf{Y_2}$ is $(n-r)$-dimensional. The matrix $\mathbf{V}^{-1}$ is also partitioned into $[\mathbf{H}\ \mathbf{B}]$, where $\mathbf{H}$ is an $n \times r$ matrix and $\mathbf{B}$ is an $n \times (n-r)$ matrix. Then

$$\boldsymbol{Y_1}|\boldsymbol{Y_2} = \boldsymbol{d}, \boldsymbol{\mu}^*, \boldsymbol{H}, \boldsymbol{\alpha}, \boldsymbol{\kappa} \sim \text{cMLG}(\boldsymbol{\mu}^*, \boldsymbol{H}, \boldsymbol{\alpha}, \boldsymbol{\kappa})$$

with density

$$M\exp\left\{\boldsymbol{\alpha}'\boldsymbol{H}\boldsymbol{Y_1} - \boldsymbol{\kappa}'\exp(\boldsymbol{H}\boldsymbol{Y_1} - \boldsymbol{\mu}^*)\right\}, \tag{4}$$

where $\boldsymbol{\mu}^* = \mathbf{V}^{-1}\boldsymbol{\mu} - \mathbf{B}\mathbf{d}$. Bradley et al. (2018) show that it is also straightforward to sample from the cMLG distribution using $(\mathbf{H}'\mathbf{H})^{-1}\mathbf{H}'\mathbf{Y}$, where $\mathbf{Y}$ is sampled from $\text{MLG}(\boldsymbol{\mu}, \mathbf{I}, \boldsymbol{\alpha}, \boldsymbol{\kappa})$.

Another important result given by Bradley et al. (2018) is that $\text{MLG}(\mathbf{c}, \alpha^{1/2}\mathbf{V}, \alpha\mathbf{1}, \alpha\mathbf{1})$ converges in distribution to a multivariate normal distribution with mean $\mathbf{c}$ and covariance matrix $\mathbf{V}\mathbf{V}'$ as the value of α approaches infinity. This allows for the use of MLG priors in place of Gaussian priors, in situations where it is computationally preferable, while still achieving the same posterior in the limit of α.

With this in mind, we place an MLG prior on $\boldsymbol{\eta}$ in the CESN, $\boldsymbol{\eta} \sim \text{MLG}(\mathbf{0}, \alpha^{1/2}\kappa\boldsymbol{I}, \alpha\mathbf{1}, \alpha\mathbf{1})$, which is asymptotically equivalent to a Gaussian prior. We have found letting $\alpha = 1000$ is enough for the prior to be indistinguishable from the Normal distribution. Here, κ acts as a ridge or L2 penalty. We fix $\kappa = 10$ for illustration purposes, although a further prior could be placed on this parameter if desired. Finally, letting $\boldsymbol{H}_T = [\boldsymbol{h}_1, \dots, \boldsymbol{h}_T]'$, this results in a cMLG posterior distribution for $\boldsymbol{\eta}$,

$$
\begin{aligned}
\boldsymbol{\eta}|\boldsymbol{Y} \propto \prod_{t=1}^{T} & \exp\left\{\boldsymbol{h}_t'\boldsymbol{\eta} - Y_t \text{exp}(\boldsymbol{h}_t'\boldsymbol{\eta})\right\} \\
& \times \exp\left\{\alpha\mathbf{1}_{n_h}'\alpha^{-1/2}\frac{1}{\kappa}\boldsymbol{I}_{n_h}\boldsymbol{\eta} - \alpha\mathbf{1}_{n_h}'\text{exp}\left(\alpha^{-1/2}\frac{1}{\kappa}\boldsymbol{I}_{n_h}\boldsymbol{\eta}\right)\right\} \\
& = \exp\left\{\boldsymbol{\alpha}_\eta'\boldsymbol{H}_\eta\boldsymbol{\eta} - \boldsymbol{\kappa}_\eta'\text{exp}(\boldsymbol{H}_\eta\boldsymbol{\eta})\right\} \\
& \boldsymbol{H}_\eta = \begin{bmatrix} \boldsymbol{H}_T \\ \alpha^{-1/2}\frac{1}{\kappa}\boldsymbol{I}_{n_h} \end{bmatrix}, \quad \boldsymbol{\alpha}_\eta = \left(\mathbf{1}_T', \alpha\mathbf{1}_{n_h}'\right)', \\
& \boldsymbol{\kappa}_\eta = \left(Y_1, \dots, Y_T, \alpha\mathbf{1}_{n_h}'\right)' \\
\boldsymbol{\eta}|\boldsymbol{Y}\cdot \sim & \text{ cMLG}(\boldsymbol{H}_\eta, \boldsymbol{\alpha}_\eta, \boldsymbol{\kappa}_\eta),
\end{aligned}
$$

which can be sampled from directly and efficiently.

3 Analysis of GSS Data

Our data of interest originates 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. This comprehensive dataset spans nearly five decades, from 1972 to 2021. The GSS is a key source of information on the demographics, fields of study, sources of support, and post-graduation plans of graduate students and postdoctoral researchers in science, engineering, and selected health fields. The dataset is rich in temporal frequency, providing a valuable resource for examining trends and patterns over time, as well as differences across institutions.

The data collected in the GSS includes variables such as the number of graduate students and postdoctoral researchers by field of study, gender, citizenship status, and race/ethnicity. Additionally, the survey captures 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 within each school. Note that schools (i.e. colleges) may be nested within institutions (i.e. universities).

Given the longitudinal nature of the dataset, it is possible to explore changes over time in the composition of graduate students. For instance, model fitting can allow for trend filtering on already observed data, as well as prediction for future years. The cross-sectional component of the data also enables comparisons between institutions, providing potential insight into how different universities and research institutions contribute to the training and support of the next generation of scientists and engineers.

As a case study, we focus on schools in the state of California that have complete data for all years 1972-2021. This results in a dataset of 109 schools, each with graduate student counts at 50 time points. The time series plots for four randomly selected schools are shown in Figure 2. The scale of graduate student counts can be quite different across schools and time, with some counts being in the hundreds, while others are less than ten. These small counts in particular allude to the need for a count data model rather than assuming approximate normality.

Auto-correlation function (ACF) plots are also presented for the same four schools in Figure 3. Note that the degree and strength of auto-correlation varies across schools. Thus, a standard ARMA model developed for one school may not be appropriate for another school. Thus, we wish to develop a more flexible approach that can adapt to the unique temporal structure present within each specific school.

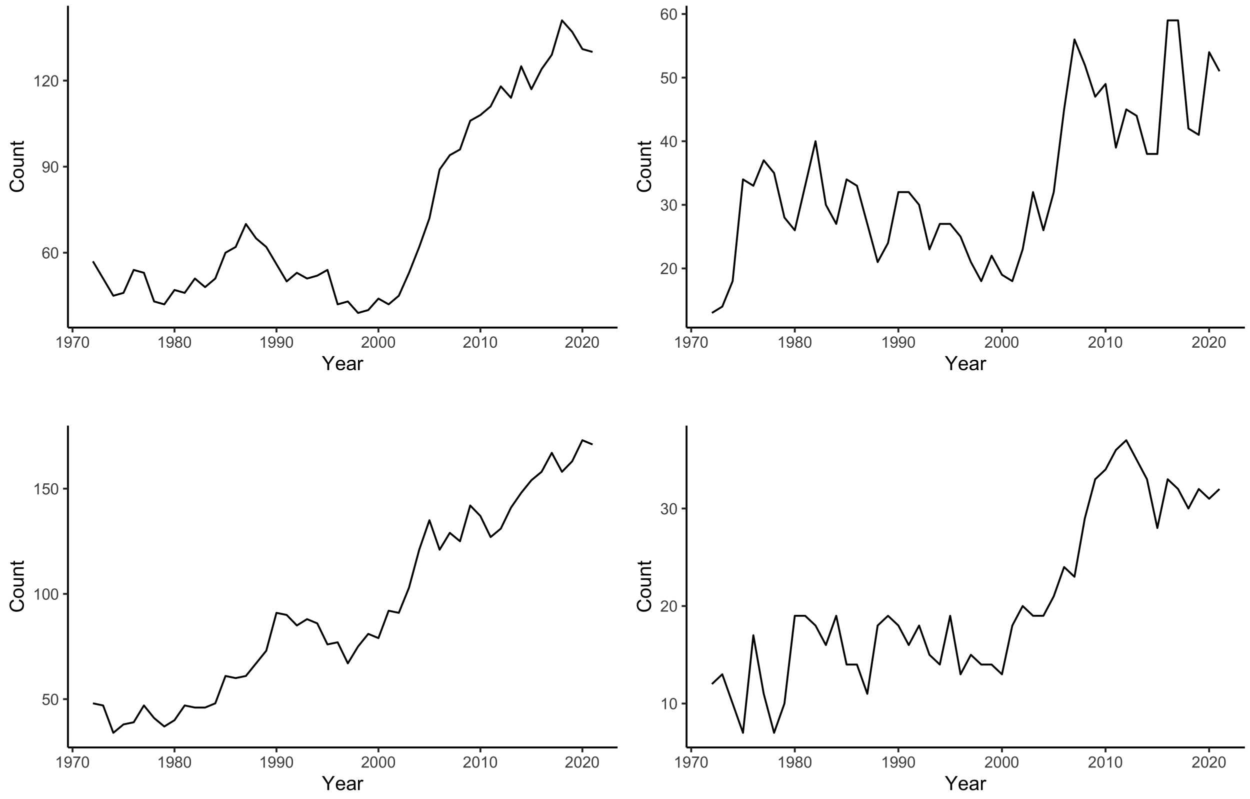

Figure 2: Time series plot of graduate student counts for four randomly selected schools in California.

3.1 Model Comparison

We compare a variety of models to evaluate their effectiveness on prediction and modeling of GSS count data. As a baseline, we fit an intercept only model for each school. This serves as a measuring stick for which any proposed models should improve upon. Next, we fit an INGARCH(1,1) model (Ferland et al., 2006; Fokianos et al., 2009), fit via the `tscount` package in `R` (Liboschik et al., 2017). This results in comparison to a standard approach that has code readily available. The remaining three models are variants of the proposed CESN. First, a single CESN is fit. Note that this variant results in a point estimate, but no uncertainty quantification. Next, an ensemble CESN is fit, which allows for both a point estimate and uncertainty quantification. Finally, we fit the Bayesian variance of the CESN,

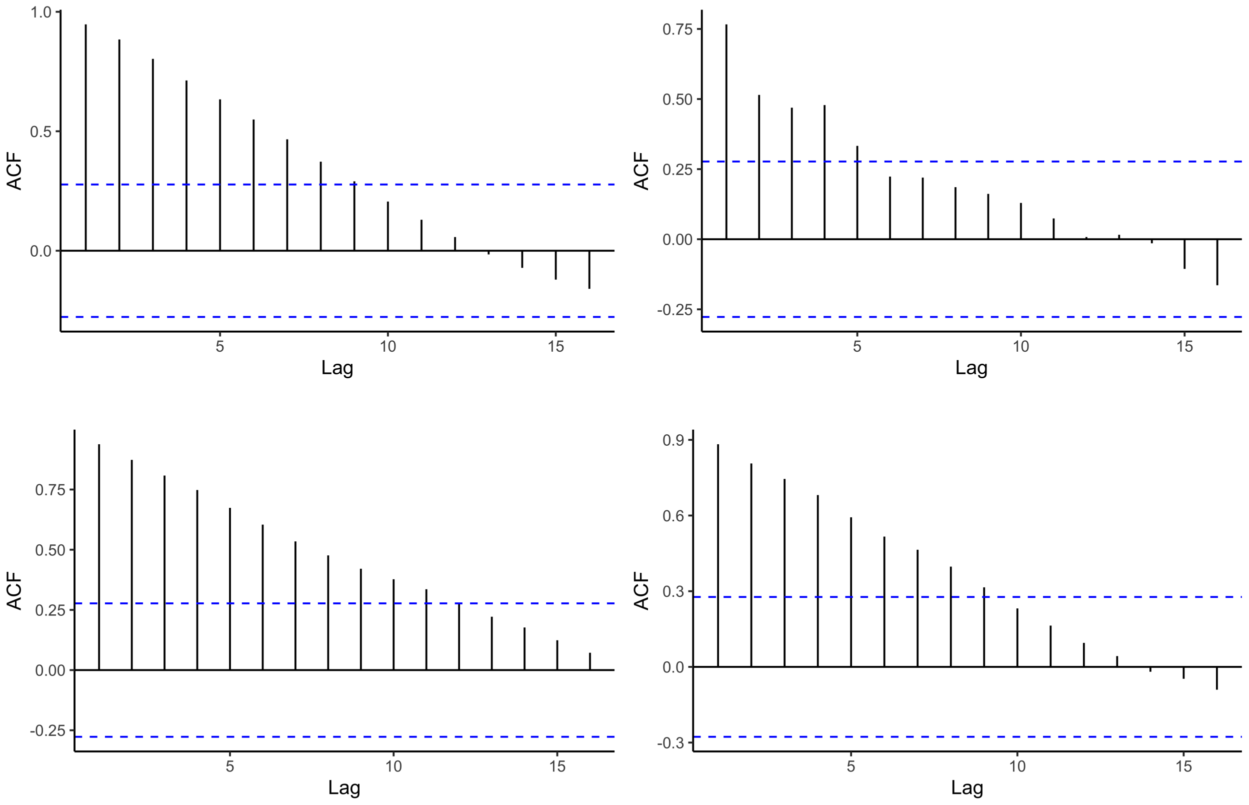

Figure 3: ACF plot of graduate student counts for four randomly selected schools in California.

which again allows for uncertainty quantification. For all CESN models, we let $\boldsymbol{x}_t = (1, Y_{t-1})'$, similar to McDermott and Wikle (2017).

We are interested in the quality of our point predictions as well as predictive uncertainty. To assess the quality of point predictions, we compare the point predictions to the held-out test data. First, we compute the mean square prediction error (MSPE) across schools,

$$\text{MSPE}_t = \frac{1}{S}\sum_{s=1}^{S}(\hat{Y}_{st} - Y_{st})^2,$$

where Y_{st} is the actual graduate student count of schools s at time t and $\hat{Y}_{st}$ is the point prediction. In some cases, for a given year, some schools see a dramatic shift in graduate student counts, and MSPE is not robust to these outliers. Thus, we also look at the mean square logarithmic prediction error (MSLPE),

$$\text{MSLPE}_t = \frac{1}{S}\sum_{s=1}^{S}\Big(\log(\hat{Y}_{st}+1) - \log(Y_{st}+1)\Big)^2 .$$

Finally, we evaluate the quality of our uncertainty estimates through the interval score (IS),

$$\frac{1}{S}\sum_{s=1}^{S}\left\{(u_{st}-\ell_{st}) + \frac{2}{\alpha}(\ell_{st}-Y_{st})I(Y_{st}<\ell_{st}) + \frac{2}{\alpha}(Y_{st}-u_{st})I(Y_{st}>u_{st})\right\},$$

where ℓ_{st} and u_{st} are the lower and upper bounds of the prediction interval for school s at time t. We let $\alpha = 0.05$ to correspond to a 95% prediction interval. Note that similar to MSPE and MSLPE, a lower interval score indicates a better model fit.

The MSE results are summarized in Table 1. All four models that account for temporal dependence provide substantial improvement over the intercept only model. All three CESN approaches have similar MSPE, beating out the alternatives. Although the INGARCH model can occasionally have competitive performance, such as the year 2019, on average, the CESN models outperform the INGARCH approach. Also note that the MSE in 2019 was heavily driven by an outlier.

Table 1: Mean square prediction error for one step ahead predictions of graduate student counts from 2017-2021.

Model	2017	2018	2019	2020	2021	5 Year Avg.
Intercept Only	3103	4908	3451	2454	3413	3466
INGARCH(1,1)	1116	1605	1033	1735	1245	1347
Single CESN	489	699	2483	1968	311	1190
Ensemble CESN	495	677	2404	1845	324	**1149**
Bayesian CESN	512	651	2645	1766	247	1164

Table 2 summarizes the results in terms of MSLPE. Here we see that after diminishing the weight given to outliers, the CESN models outperform the INGARCH model on the 2019 test set, and on average as well. Also, here the Bayesian CESN has substantially lower error than the single and ensemble CESNs.

Table 2: Mean square logarithmic prediction error ($\times 10^2$) for one step ahead predictions of graduate student counts from 2017-2021.

Model	2017	2018	2019	2020	2021	5 Year Avg.
Intercept Only	17.97	38.32	36.31	42.03	45.07	35.94
INGARCH(1,1)	8.33	19.77	15.11	8.02	15.58	13.30
Single CESN	8.23	20.41	7.17	9.90	8.02	10.75
Ensemble CESN	8.28	20.33	7.21	10.59	7.93	10.87
Bayesian CESN	8.13	17.80	5.02	9.53	2.84	**8.66**

Finally, Table 3 summarizes the prediction interval score results. Note that intervals are only constructed for the INGARCH, ensemble CESN, and Bayesian CESN approaches. Here, both CESN approaches substantially outperform the INGARCH model, which tends to have the widest intervals and poor coverage rate. Among the two CESN approaches, the Bayesian CESN model has slightly more favorable results. Due to this, and the superior point estimates in terms of MSLPE, we recommend the Bayesian CESN model for prediction of graduate student counts.

Table 3: 95% prediction interval score for one step ahead predictions of graduate student counts from 2017-2021.

Model	2017	2018	2019	2020	2021	5 Year Avg.
INGARCH(1,1)	307	285	280	315	314	300
Ensemble CESN	188	174	292	282	109	209
Bayesian CESN	179	154	303	290	86	**202**

4 Discussion

This work introduces the Count Echo State Network (CESN) as a novel methodological approach for analyzing and predicting sequential count data, specifically targeting the complex dataset of graduate student counts from the Survey of Graduate Students and Postdoctorates in Science and Engineering (GSS). The effectiveness of the CESN was demonstrated through comparison with existing count time series methods, showing significant improvements in predictive performance.

The CESN models, particularly the Bayesian variant, consistently outperformed the traditional INGARCH model and the baseline intercept model. This superiority was evident across multiple metrics, including mean square prediction error (MSPE) and mean square logarithmic prediction error (MSLPE). The Bayesian CESN, with its capability to provide uncertainty quantification, further demonstrated lower prediction interval scores, indicating more reliable and narrower predictive intervals, alongside the superior point estimates.

One of the important features of the CESN approach is its ability to handle non-linear and non-stationary time series data, which are common characteristics of establishment datasets like the GSS. By linking the echo state network to a Poisson likelihood, the CESN was able to effectively capture the count nature of the data.

The random generation of weight matrices in the CESN significantly reduces the computational burden compared to traditional recurrent neural networks (RNNs), which require gradient descent techniques. This efficiency makes the CESN a scalable solution for large datasets, enabling real-time analysis and predictions without sacrificing accuracy.

Finally, the application of the CESN to the GSS dataset could yield actionable insights for policymakers and educational institutions. By accurately modeling and predicting trends in graduate student populations, institutions can better plan for resource allocation, funding, and program development. The ability to forecast future trends also aids in anticipating changes in the workforce and preparing for shifts in educational demands.

While the CESN model shows considerable promise, there are areas for further refinement and exploration. For example, future research may focus on including more detailed covariates, such as economic indicators or state/local policy changes. This could enhance the model's predictive power and provide deeper insights into the factors influencing graduate student populations. Investigating other prior distributions and regularization methods within the Bayesian framework is another potential avenue of research. There is a large literature on Bayesian variable selection and shrinkage methods that could further improve model robustness.

The CESN represents an advancement in the modeling of count time series data, offering a powerful tool for statisticians and data scientists working with complex establishment datasets. Its application to the GSS dataset not only demonstrates its efficacy but also sets the stage for future innovations in this vital field of statistical research. The CESN may also be of broader interest in the field of establishment statistics. For example, applications such as business or agricultural statistics, could be used to test test the versatility of the method.

References

Bonas, M., Wikle, C. K., and Castruccio, S. (2024). "Calibrated forecasts of quasi-periodic climate processes with deep echo state networks and penalized quantile regression." *Environmetrics*, 35, 1, e2833.

Bradley, J. R., Holan, S. H., and Wikle, C. K. (2018). "Computationally efficient multivariate spatio-temporal models for high-dimensional count-valued data (with discussion)."

— (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.

Ferland, R., Latour, A., and Oraichi, D. (2006). "Integer-valued GARCH process." *Journal of time series analysis*, 27, 6, 923–942.

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

Friedman, J., Hastie, T., and Tibshirani, R. (2010). "Regularization paths for generalized linear models via coordinate descent." *Journal of statistical software*, 33, 1, 1.

Huang, H., Castruccio, S., and Genton, M. G. (2022). “Forecasting high-frequency spatio-temporal wind power with dimensionally reduced echo state networks.” *Journal of the Royal Statistical Society Series C: Applied Statistics*, 71, 2, 449–466.

Jaeger, H. (2002). “Adaptive nonlinear system identification with echo state networks.” *Advances in neural information processing systems*, 15.

— (2007). “Echo state network.” *scholarpedia*, 2, 9, 2330.

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

Lukoševičius, M. and Jaeger, H. (2009). “Reservoir computing approaches to recurrent neural network training.” *Computer science review*, 3, 3, 127–149.

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

— (2019a). “Bayesian recurrent neural network models for forecasting and quantifying uncertainty in spatial-temporal data.” *Entropy*, 21, 2, 184.

— (2019b). “Deep echo state networks with uncertainty quantification for spatio-temporal forecasting.” *Environmetrics*, 30, 3, e2553.

Parker, P. A., Holan, S. H., and Wills, S. A. (2021). “A general Bayesian model for heteroskedastic data with fully conjugate full-conditional distributions.” *Journal of Statistical Computation and Simulation*, 91, 15, 3207–3227.

Savitsky, T. D. and Toth, D. (2016). “Bayesian Estimation Under Informative Sampling.” *Electronic Journal of Statistics*, 10, 1, 1677 – 1708.

Snijkers, G., Bavdaž, M., Bender, S., Jones, J., MacFeely, S., Sakshaug, J. W., Thompson, K. J., and Van Delden, A. (2023). “Advances in business statistics, methods and data collection: introduction.” *Advances in Business Statistics, Methods and Data Collection*, 1–22.