Determining economic factors for sex trafficking in the United States using count time series regression

Yuhyeong Jang[1] · Raanju R. Sundararajan[1] · Wagner Barreto-Souza[2] · Elizabeth Wheaton-Paramo[3]

© The Author(s) 2024

Abstract

The article presents a robust quantitative approach for determining significant economic factors for sex trafficking in the United States. The aim is to study monthly counts of sex trafficking-related convictions, and use a wide range of economic variables as covariates to investigate their effect on conviction counts. A count time series model is considered along with a regression setup to include economic time series as covariates (economic factors) to explain the counts on sex trafficking-related convictions. The statistical significance of these economic factors is investigated and the significant factors are ranked based on appropriate model selection methods. The inclusion of time-lagged versions of the economic factor time series in the regression model is also explored. Our findings indicate that economic factors relating to immigration policy, consumer price index and labor market regulations are the most significant in explaining sex trafficking convictions.

✉ Raanju R. Sundararajan
  rsundararajan@smu.edu

  Yuhyeong Jang
  yuhyeongj@smu.edu

  Wagner Barreto-Souza
  wagner.barreto-souza@ucd.ie

  Elizabeth Wheaton-Paramo
  ewheaton@smu.edu

[1] Department of Statistics and Data Science, Southern Methodist University, Dallas, Texas, USA

[2] School of Mathematics and Statistics, University College Dublin, Dublin, Ireland

[3] Department of Economics, Southern Methodist University, Dallas, Texas, USA

⌀ Springer

1 Introduction

Human trafficking is a serious human rights problem that is caused by multiple factors concerning social and economic conditions. The three major types of human trafficking, sex, labor and organ, are often driven by social and economic factors. The International Organization for Migration (IOM) mentions the significance of economic factors in the victim's country of origin and destination in determining a trafficking occurrence; for instance, see IOM (2012). Push factors for human trafficking refer to the social and economic conditions in the victim's country of origin and pull factors are the same conditions observed in the destination country. The work in O'Brien et al. (2013) identifies poverty, gender inequality, and unemployment as key socio-economic variables that are push factors, and employment opportunities and demand for cheap labor as key economic pull factors. However, in the previous work, a formal statistical analysis is not undertaken to determine the significance of these push and pull factors.

Rigorous statistical research on human trafficking is extremely scarce in the literature mostly due to the lack of availability of reliable data (Kangaspunta 2003). Cho (2015) takes up an empirical analysis to determine significant push and pull factors of human trafficking based on data covering 153 countries. Their work classifies the large set of push and pull factors into four pillars namely migration, vulnerability, crime, and policy and institutional effects. They employ an additive regression type model with the response being a continuous variable measuring the level of human trafficking inflows/outflows. Significant push factors identified by their method include variables relating to GDP, fertility rates, information flows, share of food, beverage and tobacco industries in GDP, control of corruption, crime rates, and infant mortality rates. Some of the significant pull factors that were identified are GDP, the percentage of the workforce employed in the agriculture sector, refugee inflows, and crime rates. While this previous work considers a large set of push and pull factors and employs an extreme bound analysis approach to robustly detect the push and pull factors, it fails to adequately recognize the time series nature of the response and explanatory variables. More precisely, the data considered in their work involve aggregates of the response and explanatory variables over time. This can potentially limit the proper understanding of the time-varying connection between push and pull factors and human trafficking rates.

In the United States, several challenges exist that deter access to high-quality human trafficking data. In Hanson and Finklea (2022), the main difficulties in precisely measuring human trafficking are discussed. Their work lists inconsistent definitions of a human trafficker, underground actions leading to concealed trafficking activities, lack of awareness, and the absence of victims self-identifying as such as some of the main challenges to accurate data collection. In Farrell et al. (2019), the law enforcement's lack of accurate trafficking data collection in both sex and labor trafficking cases is mentioned. The Office for Victims of Crime[1] in the United States describes human trafficking as "hidden in plain sight" due to the absence of comprehensive data. These are some of the key reasons behind the absence of rigorous quantitative human traf-

---

[1] www.ovcttac.gov.
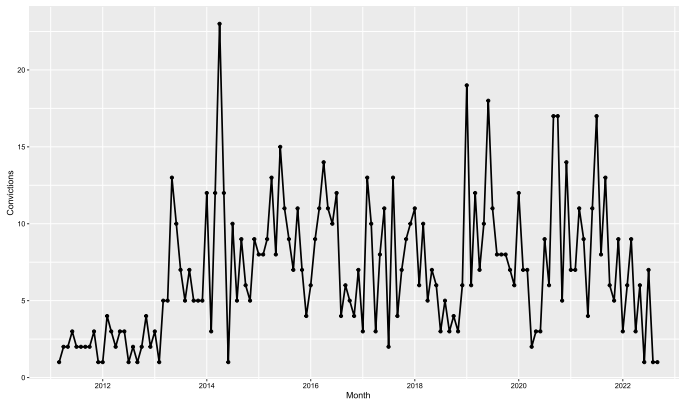
Fig. 1 Plot of monthly counts of sex trafficking-related convictions in the United States from March 2011 to September 2022
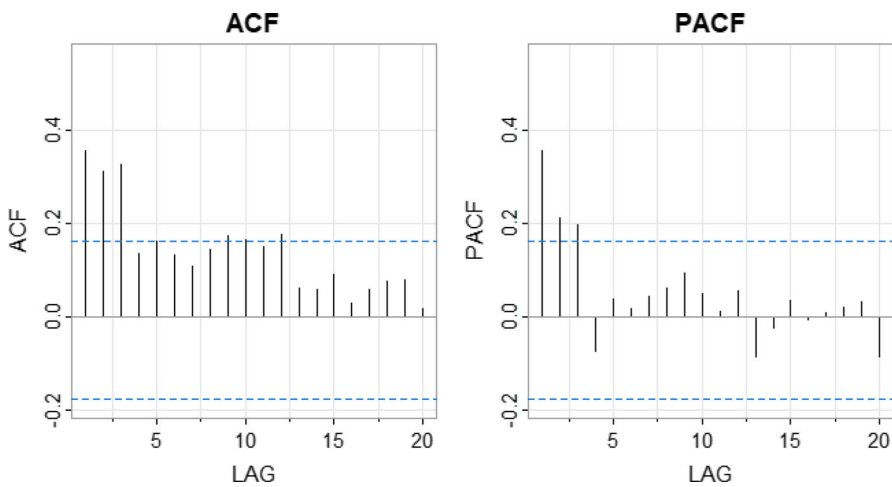
Fig. 2 Autocorrelation (ACF) and partial ACF plots for count time series of sex trafficking-related convictions in the United States

ficking research in the United States. In this article, we consider monthly counts of sex trafficking-related convictions, at the federal level, in the United States[2] during the period March 2011 to September 2022. Plots of this count time series and its autocorrelations are shown in Figs. 1 and 2, respectively. One of the objectives of this article is to properly explore and study this convictions dataset using count time series models along with necessary model adequacy checks.

Two popular methods for modeling time series of counts are the INteger-valued AutoRegressive (INAR) model (McKenzie 1985; Al-Osh and Alzaid 1987) and the INteger-valued Generalized AutoRegressive Conditional Heteroscedastic (INGARCH) model (Ferland et al. 2006). The INAR model is constructed on the basis

[2] Data source: DeliverFund https://deliverfund.org/.

of a thinning operator (Weiß 2008), while the INGARCH model utilizes the structure of generalized linear models (GLM) (Davis et al. 2016; Weiß 2018, Chapter 4, Davis et al. 2021.) The INGARCH approach is more accessible to non-statisticians because no special knowledge, except an exposure to the GLM, is required for understanding it, whereas one needs to be familiar with the thinning operator to understand the INAR model. The INGARCH model also involves the choice of a link function in order to accommodate the inclusion of covariates in a flexible manner. Two link functions that will be considered in this paper are the logarithmic (Fokianos and Tjøstheim 2011) and the recent softplus function (Weiß et al. 2022). We refer to those INGARCH models with log and softplus link functions as log-INGARCH and sp-INGARCH processes, respectively. It must be mentioned that a linear link function can also be considered but those models can only accommodate positive-valued covariates that exhibit positive associations with the count responses. In contrast, such constraints are not imposed on log-INGARCH and sp-INGARCH models. Additionally to the previous point, in our data analysis, we consider differenced or log-differenced versions of economic variables such as monthly consumer price index (CPI), monthly GDP, and monthly unemployment rate to be incorporated in the model as covariates. These transformed time series assume positive and negative values thereby making the linear INGARCH model unsuitable.

In this work, due to the aforementioned features, we adopt the log-INGARCH and sp-INGARCH setup for modeling the monthly counts of sex trafficking-related convictions, at the federal level, in the United States. These two INGARCH models are described, along with their extensions to include covariates (economic factors). Methods to test statistical significance of the regression coefficients associated with the various economic factors are provided. This enables us to detect the significant economic factors for sex trafficking. A wide range of economic variables,[3] treated as time series data, are considered. Some examples here include equity market volatility trackers on fiscal policy, immigration policy, agricultural policy, to name a few, and also other well-known macroeconomic variables such as consumer price index, gross domestic product (GDP) and unemployment rate; see Table 1 for the complete list. It must be noted that, in this paper, we will denote these economic variables as economic factors and not economic pull factors. This is mainly because the term 'pull factors' is associated with international human trafficking instances and our sex trafficking convictions dataset does not carry information on the countries of origins of the victims. The main contributions of the proposed work are listed below.

(i) To the best of our knowledge, all existing quantitative methods to analyze human trafficking data do not treat the response variable (counts of federal sex trafficking convictions) and the covariates (economic factors) as time series data. This is critical in uncovering the true time-dependent relationship between economic factors and sex trafficking-related convictions. Our approach treats the response and covariates as time series data, and statistical significance of the economic factors is investigated using appropriate methods.

(ii) We consider time-lagged versions of the economic factors in the regression model and study their relationship with sex trafficking. This allows for a lead-lag-type

---

[3] Data source: Federal Reserve Economic Data https://fred.stlouisfed.org/.

relationship between federal sex trafficking-related convictions and economic factors. This is highly significant since there is always a lag between the time of the initial arrest of the perpetrator in a sex trafficking case, and the eventual conviction/sentencing outcome on that case.

(iii) Going forward, with the availability of additional data, straightforward extensions of the proposed modeling approach can detect other social and economic factors for sex trafficking convictions, and can also uncover the (joint) time-dependent impact of these factors on trafficking convictions.

The paper unfolds as follows. In Sect. 2, we describe INGARCH models with log and softplus link functions, followed by their extensions to the regression setup to include covariates. Section 3 includes a comprehensive and complete statistical analysis of the sex trafficking convictions dataset, including model order and variable selection, multiple regression analysis, model diagnostics and forecasting. More precisely, we first discuss a model order and variable selection procedure, within the INGARCH framework, to select the economic factors that will be included in the multiple regression analysis. Second, we apply a multiple regression analysis and discuss the significant economic factors. Model adequacy checks are performed using probability integral transform (PIT) plots and residual analysis. Finally, an out-of-sample forecasting exercise is performed and presented. Concluding remarks are provided in Sect. 4.

2 INGARCH models

In this section, we begin by introducing the log-INGARCH and sp-INGARCH models by Fokianos and Tjøstheim (2011) and Weiß et al. (2022), respectively. The log-INGARCH model assumes a multiplicative effect in terms of the previous conditional means and observations, while the sp-INGARCH model leads to an approximately additive effect. When it comes to the distributional assumption (conditioned on the past), the Poisson and negative binomial (NB) distributions will be considered. Finally, extensions of the INGARCH model to include covariates are discussed, which will be crucial for our study on the sex trafficking conviction counts explained by economic factors.

2.1 Log-INGARCH model

Let $\{Y_t\}_{t\in\mathbb{Z}}$ be our count time series of interest. The Poisson log-INGARCH(p, q) model assumes that

$$\begin{cases} Y_t|\mathcal{F}_{t-1} \sim \text{Poisson}(\lambda_t), & (1) \\ \log\lambda_t = d + \sum_{i=1}^{p} a_i \log\lambda_{t-i} + \sum_{j=1}^{q} b_j \log(Y_{t-j} + 1), & (2) \end{cases}$$

for $t \in \mathbb{Z}$, where $\mathcal{F}_{t-1} \equiv \sigma(Y_{t-1}, Y_{t-2}, \ldots)$ denotes this count process' history, and d, a_i's and b_j's are real-valued parameters with $|a_i| < 1$, $|b_j| < 1$, for

each $i = 1, 2, \ldots, p$, and $j = 1, 2, \ldots, q$. It is further assumed that $|\sum\limits_{i=1}^{p} a_i + \sum\limits_{j=1}^{q} b_j| < 1$; see Fokianos and Tjøstheim (2011) and Liboschik et al. (2017) for details on this stationarity condition. It must be noted that (2) can be rewritten as $\lambda_t = \exp\left(d + \sum\limits_{i=1}^{p} a_i \log \lambda_{t-i} + \sum\limits_{j=1}^{q} b_j \log(Y_{t-j} + 1)\right)$, which shows that the conditional mean proceeds in a multiplicative way.

The Poisson log-INGARCH model above can be seen as a type of Poisson regression, where (1) corresponds to the random component, while (2) can be regarded as the regression component with the logarithm of the past mean and past observations serving as explanatory variables (Kedem and Fokianos 2002; Tjøstheim 2012). It must however be noted that in contrast to the Poisson generalized linear model, (2) also poses randomness and thereby makes the marginal distribution of Y_t different from the typical Poisson distribution.

2.2 sp-INGARCH model

In the aforementioned log-INGARCH model, the conditional mean λ_t involves a multiplicative effect in terms of the previous conditional means and observations. However, in some applications, additive effects can be more appropriate as discussed by Weiß et al. (2022). In such cases, the softplus link function can be an alternative because it nearly preserves a linear structure and simultaneously allows for negative autocorrelation. The softplus function is defined as

$$s_c(x) = c \log\left(1 + \exp\left(x/c\right)\right),$$

for $x \in \mathbb{R}$, where $c > 0$ is a tuning parameter controlling the degree of linearity in $s_c(x)$. In this paper, we follow the default setup $c = 1$ suggested in Weiß et al. (2022). In the sp-INGARCH model, the regression component (2) is replaced by the softplus link as follows:

$$\lambda_t = s_c\left(d + \sum_{i=1}^{p} a_i \lambda_{t-i} + \sum_{j=1}^{q} b_j Y_{t-j}\right), \tag{3}$$

for $t \in \mathbb{Z}$, where d, a_i's and b_j's are real-valued parameters with stationarity constraints $\sum_{i=1}^{p} \max\{0, a_i\} + \sum_{j=1}^{q} \max\{0, b_j\} < 1$, and $\sum_{i=1}^{p} |a_i| < 1$. Then, the count Y_t is assumed to be conditionally generated from a Poisson(λ_t) distribution as in (1), for $t \in \mathbb{Z}$.

2.3 Negative binomial (NB) INGARCH model

The conditional Poisson distribution in (1) restricts the conditional variance of Y_t to be equal to λ_t. Although the unconditional model is overdispersed, it still could be insuf-

ficient in accommodating all of the overdispersion involved in our count time series of interest. To overcome this problem, a conditional negative binomial distribution can be imposed instead of the Poisson, which will have an additional parameter to control the dispersion. Under this assumption, the conditional probability density function of Y_t given $\mathcal{F}_{t-1}$ assumes the form

$$\Pr(Y_t = y|\mathcal{F}_{t-1}) = \frac{\Gamma(y+\phi)}{\Gamma(y+1)\Gamma(\phi)}\left(\frac{\phi}{\lambda_t + \phi}\right)^{\phi}\left(\frac{\lambda_t}{\lambda_t + \phi}\right)^{y}, y = 0, 1, 2, \cdots, \quad (4)$$

where $\phi > 0$ is a dispersion parameter and λ_t follows the same temporal structure as in (2) or (3). The conditional variance $\mathrm{Var}(Y_t|\lambda_t) = \lambda_t(1 + \phi^{-1}\lambda_t)$ readily indicates that ϕ controls the degree of overdispersion and having the Poisson distribution as a limiting case when $\phi \to \infty$. In Sects. 3.1 and 3.2, we contrast the suitability of the Poisson and NB distributions for the sex trafficking convictions time series.

2.4 Count time series regression model

Here we discuss INGARCH regression models where effects of external covariates can be additionally incorporated into the temporal structure. Suppose that $\mathbf{Z}_t = (Z_{1,t}, \cdots, Z_{r,t})^{\top}$, for $t \in \mathbb{Z}$, denotes a vector of time-dependent covariates. The inclusion of $\mathbf{Z}_t$ into the systematic (regression) components in (2) and (3) is done in the following manner:

$$\log \lambda_t = d + \sum_{i=1}^{p} a_i \log \lambda_{t-i} + \sum_{j=1}^{q} b_j \log(Y_{t-j} + 1) + \sum_{k=1}^{r} \beta_k Z_{k,t}, \quad (5)$$

$$\lambda_t = s_c\left(d + \sum_{i=1}^{p} a_i \lambda_{t-i} + \sum_{j=1}^{q} b_j Y_{t-j} + \sum_{k=1}^{r} \beta_k Z_{k,t}\right), \quad (6)$$

where $\beta_1, \ldots, \beta_r \in \mathbb{R}$ are the regression coefficients.

Let $\boldsymbol{\theta} = (\phi, d, a_1, a_2, \ldots, a_p, b_1, b_2, \ldots, b_q, \beta_1, \beta_2, \ldots, \beta_r)^{\top}$ and assume that $y_1, \ldots, y_T$ is an observed count time series trajectory. We will use the maximum likelihood estimation method to estimate $\boldsymbol{\theta}$. Under a conditional negative binomial distribution assumption, the likelihood function is given by $L(\boldsymbol{\theta}) = \prod_{t=\max\{p,q\}+1}^{T} \Pr(Y_t = y_t|\mathcal{F}_{t-1})$, where $\Pr(Y_t = y_t|\mathcal{F}_{t-1})$ assumes the form in (4). Under a Poisson assumption, the parameter vector is given by $\boldsymbol{\theta} = (d, a_1, a_2, \ldots, a_p, b_1, b_2, \ldots, b_q, \beta_1, \beta_2, \ldots, \beta_r)^{\top}$, and the likelihood function assumes a similar form as above with $\Pr(Y_t = y_t|\mathcal{F}_{t-1}) = \dfrac{e^{-\lambda_t}\lambda_t^{y_t}}{y_t!}$. The maximum likelihood estimator of $\boldsymbol{\theta}$ is obtained by maximizing the log-likelihood function $\ell(\boldsymbol{\theta}) \equiv \log L(\boldsymbol{\theta})$.

An approximate confidence interval (CI) for each element of $\boldsymbol{\theta}$ can then be readily obtained by using the observed Fisher information matrix, which provides us the

standard errors of the estimates, and then a normal approximation can be used to construct such intervals. Alternately, a parametric bootstrap approach can also be taken up to construct these intervals. In the following empirical analysis in Sect. 3, we present results of both approximate and parametric bootstrap CIs. Note that when carrying out the multiple INGARCH regression analysis in Sect. 3.2, the bootstrap CI will be employed due to the relatively small sample size of the sex trafficking convictions dataset. The model orders p and q of the INGARCH regression models in (5) and (6) should also be determined. The estimation of these quantities will be addressed in Sect. 3.1.

3 Statistical analysis of sex trafficking-related convictions in the United States

In this section, we describe our procedure to identify significant economic factors associated with sex trafficking convictions in the United States. First, the selection of a suitable conditional probability distribution, Poisson or NB, for the response counts is discussed. Second, from the initial list of many economic factors given in Table 1, a variable selection method using the log-INGARCH and sp-INGARCH models is described. Third, a multiple INGARCH regression analysis is conducted wherein multiple covariates (economic factors) are considered simultaneously to determine their joint effect on sex trafficking convictions counts. Finally, model adequacy and out-of-sample forecasting performance results are provided.

The dataset consists of 139 monthly counts of sex trafficking-related convictions, at the federal level, in the United States from March 2011 to September 2022.[4] The data was extracted from a vast volume of relevant news articles collected by DeliverFund, a nonprofit organization founded to fight human trafficking. The complete list of economic factors under consideration is provided in Table 1. The selection of this list of economic variables is partially driven by the discussion in Cho (2015); see Appendix C of that work. To ensure first-order stationarity of the time-dependent covariates, we apply a differencing or a log-differencing to the following time series: monthly consumer price index (CPI), monthly GDP, monthly unemployment rate, monthly labor force participation rate for women, black, and latino, and employment-population ratio of men and women. The list also includes equity market volatility (EMV) trackers which are unofficial monthly indexes developed and produced by *Economic Policy Uncertainty*,[5] an academic research institution, as well as official indexes released by government agencies. These trackers aim to quantify the importance of each variable on the US stock market volatility and are calculated based on news articles. Data on the many economic factors considered are sourced from Federal Reserve Economic Data (FRED).[6]

---

[4] Data source: DeliverFund https://deliverfund.org/.

[5] https://www.policyuncertainty.com/index.html.

[6] Data source: Federal Reserve Economic Data https://fred.stlouisfed.org/.

Table 1 List of the economic factors (covariates) that were considered in the following empirical analysis

Variable name	Description
CPI	Consumer price index for all urban consumers. Seasonally adjusted. (first differenced)
GDP	Gross domestic product - Normalized for the United States. Seasonally adjusted. (first differenced)
UNRATE	Unemployment rate. Seasonally adjusted. (log transformed and first differenced)
LFPR.WOMEN	Labor force participation rate of women. Seasonally adjusted. (first differenced)
LFPR.BLACK	Labor force participation rate of black Americans. Seasonally adjusted. (first differenced)
LFPR.LATINO	Labor force participation rate of Latino. Seasonally adjusted. (first differenced)
EPR.MEN	Employment-population ratio of men. Seasonally adjusted. (first differenced)
EPR.WOMEN	Employment-population ratio of women. Seasonally adjusted. (first differenced)
USEPUINDXD	Economic policy uncertainty index for United States. Not seasonally adjusted
EMV.ELE	EMV tracker: elections and political governance. Not seasonally adjusted
EMV.COM	EMV tracker: competition matters. Not seasonally adjusted
EMV.WEL	EMV tracker: entitlement and welfare programs. Not seasonally adjusted
EMV.FIS	EMV tracker: fiscal policy. Not seasonally adjusted
EMV.NAT	EMV tracker: national security policy. Not seasonally adjusted
EMV.AGR	EMV tracker: agricultural policy. Not seasonally adjusted
EMV.LAB	EMV tracker: labor regulations. Not seasonally adjusted
EMV.IMM	EMV tracker: immigration. Not seasonally adjusted

3.1 Model selection

The model selection procedure for INGARCH regression models is more intricate than that of generalized linear models for independent data. This intricacy arises mainly because variable selection is intertwined with the determination of INGARCH model order (p, q). The inclusion or exclusion of a certain covariate can change the optimal model order, thus impeding the use of the model selection strategies commonly employed in the GLM context, such as the stepwise approach. In the existing literature on INGARCH models, covariates are typically assumed to be already given or are not considered. To the best of our knowledge, any variable selection procedure for INGARCH regression models has not been rigorously investigated.

In the following empirical analysis, we introduce a practical model selection procedure within the INGARCH regression framework. Regarding the choice of a link function, we report results for both log and softplus link functions, instead of selecting

Table 2 AIC of candidate INGARCH(p, q) models without covariates, for each link function

Link function	Response distribution	(p, q)					
		(0,1)	(1,1)	(0,2)	(1,2)	(0,3)	(1,3)
Log	Poisson	798.80	773.14	777.92	773.33	769.02	769.91
	NB	743.01	728.18	730.40	728.98	**727.13**	728.67
Softplus	Poisson	806.03	772.86	780.11	774.35	770.48	771.47
	NB	746.40	727.42	730.19	728.83	**727.38**	729.09

AIC is calculated by excluding the first three observations in all cases

Model with the smallest AIC is marked in bold

one exclusively, so as to assess for the presence of multiplicative and (approximately) additive effects in the conditional mean. Hence, two separate model selection procedures are implemented, one for each link function. The specific procedure is summarized step by step as follows.

> **(Step 1)** Determine the INGARCH model order (p, q) and response distribution (Poisson or NB) using AIC, without considering covariates.
> **(Step 2)** For each covariate in Table 1, fit a simple INGARCH(p, q) regression model (considering one covariate at a time). Test the statistical significance of the regression coefficient at 95% confidence level using approximate standard errors obtained by inverting a numerical hessian matrix.
> **(Step 3)** With only the significant covariates from Step 2, fit a multiple INGARCH regression model. Model order (p, q) is chosen via AIC.

We begin with Step 1 and examine the sample autocorrelations in Fig. 2. The first three sample autocorrelations are approximately 0.35, notably exceeding the upper confidence limit. This visually implies the need to consider the past observations with time lags of at most three months, i.e., the order $q = 3$ in (5) and (6). However, the necessity of including the past mean term, i.e., the order p in (5) and (6), is much less obvious from a visual inspection. The sample autocorrelations do not exhibit the expected rate of decay, which is neither slow enough for the inclusion of the past mean term ($p > 0$) nor fast enough for the INGARCH(0,3) model ($p = 0$). Thus, in Step 1, we consider model orders of $p = 0, 1$ and $q = 1, 2, 3$. Table 2 shows that, using AIC, the negative-binomial INGARCH(0,3) model is selected.

To assess the model adequacy, we utilize the non-randomized probability integral transform (PIT) plot, which has been a popular diagnostic tool for count time series models (Czado et al. 2009). In Fig. 3, the PIT plots corresponding to the NB log-INGARCH(0,3) and NB sp-INGARCH(0,3) models are provided. A blue dashed line in each plot relates to the uniform probability density function. One can witness that both PIT plots exhibit no severe departure from the uniform distribution, implying a decent goodness-of-fit. Consequently, for each of the two link functions, we will implement variable selection by assuming a NB INGARCH(0,3) as the true underlying INGARCH structure.

NB log-INGARCH(0,3)
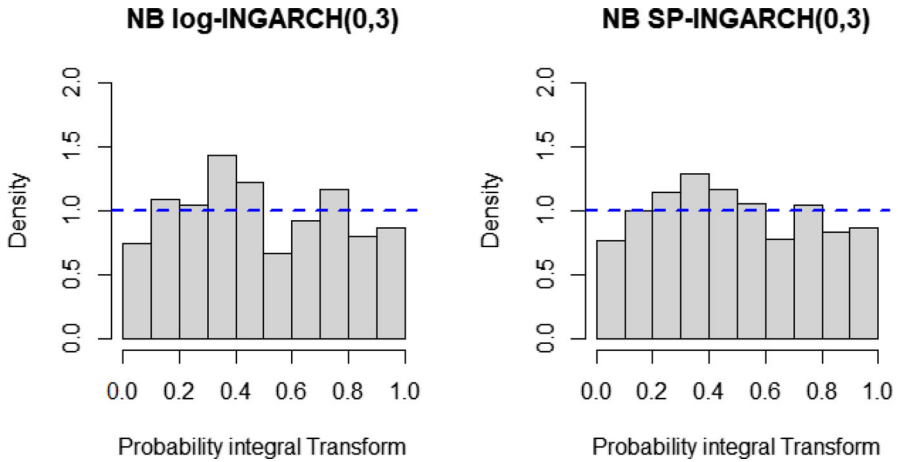
NB SP-INGARCH(0,3)

Fig. 3 PIT plots for NB log- and sp-INGARCH(0,3) models fitted to sex trafficking count time series data without covariates. The blue dashed lines are the probability density of the uniform distribution

Table 3 AICs of simple NB log-INGARCH(0,3) regression model (i.e., one covariate at a time)

Variable	Time lag					
	0	1	2	3	4	5
CPI	–	–	–	–	–	**717.74**
GDP	–	**717.74**	–	–	–	725.15
UNRATE	**723.50**	–	–	–	–	723.66
EMV.AGR	–	–	**722.32**	–	–	–
EMV.LAB	–	–	725.08	721.07	**720.129**	–
EMV.IMM	**723.97**	–	–	–	–	–
LFPR.BLACK	–	–	–	–	–	**724.38**
LFPR.LATINO	–	–	–	–	–	**723.24**
EPR.MEN	–	–	–	–	–	**723.62**
EPR.WOMEN	–	–	–	–	–	**724.38**

For each covariate, the time lag with the smallest AIC (in bold) is selected

'–' means that the corresponding variable is not statistically significant at 95% confidence level

Next, we fit simple NB log- and sp-INGARCH(0,3) regression models with each of the covariates (i.e., one covariate at a time) listed in Table 1, and determine statistical significance at 95% confidence level. For each covariate, time-lagged versions, of up to 5 months, are also considered. If a covariate is significant for multiple time lags, then the time lag with the lowest AIC is selected as illustrated in Tables 3 and 4. The remaining economic factors listed in Table 1 did not turn out to be statistically significant and their results are not presented. Note that the confidence intervals were computed using standard errors obtained by inverting the observed Fisher information matrix.

The EMV tracker for immigration policy was found to be a significant economic factor in terms of AIC. While sex trafficking victims include United States citizens,

Table 4 AICs of simple NB sp-INGARCH(0,3) regression model (i.e., one covariate at a time)

Variable	Time lag					
	0	1	2	3	4	5
CPI	–	–	–	–	–	**718.97**
GDP	726.40	**719.56**	–	–	–	–
UNRATE	**724.72**	–	–	–	–	–
EMV.LAB	–	–	722.69	723.19	**720.07**	–
EMV.IMM	**721.98**	723.74	–	–	723.98	–
LFPR.LATINO	–	724.91	–	–	–	**723.46**
EPR.MEN	**725.51**	726.98	–	–	–	–
EPR.WOMEN	**725.58**	726.97	–	–	–	–

For each covariate, the time lag with the smallest AIC (in bold) is selected

'–' means that the corresponding variable is not statistically significant at 95% confidence level

the statistical significance of this economic factor indicates that foreign nationals are definitely affected by this crime, and changes in the immigration policy is seen to have an influence on sex trafficking convictions. This result, in some ways, agrees with the observation made in Cho (2013) wherein it is written that the majority of trafficking victims are typically foreign nationals. Changes to labor regulations is another significant economic factor, and can be understood as a structural aspect of a country affecting human trafficking. It is known that human trafficking victims' exploitation, in certain situations, begins at domestic labor markets and/or sex industries. A more detailed discourse on how immigration and labor regulations may affect sex trafficking occurrences can also be found in some existing articles; for examples, see Shamir (2012), Avendaño and Fanning (2013). Another significant economic factor is the EMV tracker on agricultural policy. On a related note, it must be noted that the work by Cho (2015) identifies, among several others, the percentage of the workforce in the agricultural sector as a significant economic factor in explaining human trafficking rates. Another similarity with the previous cited work is the significance of the GDP variable in explaining sex trafficking. One dissimilarity with the results in Cho (2015) is that our results indicate that changes in the monthly unemployment rate is statistically significant in determining sex trafficking conviction counts. Our results from Tables 3 and 4 also indicate that changes in the labor force participation of black and Latino Americans and changes in the employment-population ratio of men and women are also statistically significant. A notable advantage of our work over the existing literature is that we treat all the prospective economic factors in Table 1 as time-dependent variables, i.e., as time series data, and also consider time-lagged versions of these economic factors in our regression setup. As an illustration of the previous point, from Tables 3 and 4, one can observe that time-lagged versions of many of the economic factors are statistically significant in explaining sex trafficking-related convictions. Considering time-lagged versions of these economic factors is important since there is often a lag between the time of an initial arrest in a sex trafficking case and the final charging/sentencing date on that case.

Table 5 Model selection results for log and softplus link functions using AIC. The covariates included in the log- and sp-INGARCH multiple regression models are listed in Tables 3 and 4, respectively

Link function	Response distribution	(p, q)		
		(0,1)	(0,2)	(0,3)
Log	Poisson	755.99	738.31	717.67
	NB	732.40	718.99	**705.68**
Softplus	Poisson	757.3	735.77	713.49
	NB	730.1	713.13	**699.89**

Model with the smallest AIC is marked in bold

3.2 Constructing multiple INGARCH regression models

To inspect the joint effect of the marginally significant covariates listed in Tables 3 and 4, we construct multiple INGARCH regression models with each of the two link functions. We first repeat the model order selection procedure in the presence of all of these covariates since their inclusion can possibly change the underlying structure. As indicated in Table 5, NB INGARCH(0,3) model still remains to be the best fit under both log and softplus link function choices, in terms of AIC. Results for the multiple INGARCH regression models with model order $p = 1$ are not reported here because inclusion of the past mean term resulted in estimates that violate the stationarity conditions mentioned in Sect. 2, and these estimates were also found to be highly unstable in the finite-sample setting.

While we do not present the results here, our findings indicate that, when all covariates selected based on Table 3 (and Table 4 for sp-INGARCH) were considered together, only three covariates namely CPI, EMV trackers for labor regulations and immigration turn out to be statistically significant. Finally, we construct multiple INGARCH regression models by only incorporating these three jointly significant covariates and the results are presented in Table 6. We will henceforth refer these models to as final models. The AIC values of the final models, which are similar to those of the full models (i.e., models that include all covariates in Table 3 for log-INGARCH and Table 4 for sp-INGARCH) in Table 5, provide a justification for excluding the jointly insignificant covariates. The regression coefficients show that the sex trafficking conviction counts have positive associations with EMV trackers for labor regulation and immigration, while they are negatively correlated with the monthly CPI variable. Being more precise, we have statistical evidence that each of these three time series Granger causes sex trafficking-related convictions in the United States. For instance, from the results in Table 6, we see that policy changes relating to immigration have a more immediate impact on sex trafficking convictions, whereas policy and regulatory changes in the labor market have a more time-lagged impact on sex trafficking convictions. Note that the statistical significance and the direction of the coefficients convey the essential information in identifying economic factors that are important in explaining sex trafficking-related convictions in the United States.

Due to the small sample size relative to the number of parameters, parametric bootstrapping was employed to compute the standard errors and the corresponding

Table 6 Parameter estimates from the multiple NB log- and sp-INGARCH(0,3) regression models. Standard error and confidence intervals are computed using parametric bootstrap

| | NB log-INGARCH(0,3) | | | | NB sp-INGARCH(0,3) | | | |
| | Coef. | S.E | 95% CI | | Coef. | S.E | 95% CI | |
			Lower	Upper			Lower	Upper
ϕ	11.4370	12.0655	7.5092	43.9751	11.3674	13.0083	7.4785	37.5501
d	0.3811	0.2059	0.1187	0.9120	1.1393	0.6854	0.1560	2.8555
b_1	0.1271	0.0816	−0.0406	0.2723	0.1402	0.0787	−0.0319	0.2806
b_2	0.2468	0.0806	0.0695	0.3821	0.2369	0.0775	0.0583	0.3677
b_3	0.3107	0.0813	0.1395	0.4518	0.2877	0.0796	0.1116	0.4293
CPI_{t-5}	−0.2512	0.0631	−0.3835	−0.1348	−1.3229	0.4022	−2.2180	−0.6705
$EMV.LAB_{t-4}$	0.4481	0.1173	0.2142	0.6870	2.7676	0.8980	1.2997	4.8007
$EMV.IMM_t$	0.5605	0.2115	0.1029	0.9554	4.0222	1.6381	1.0966	7.5573
AIC	701.87				701.30			
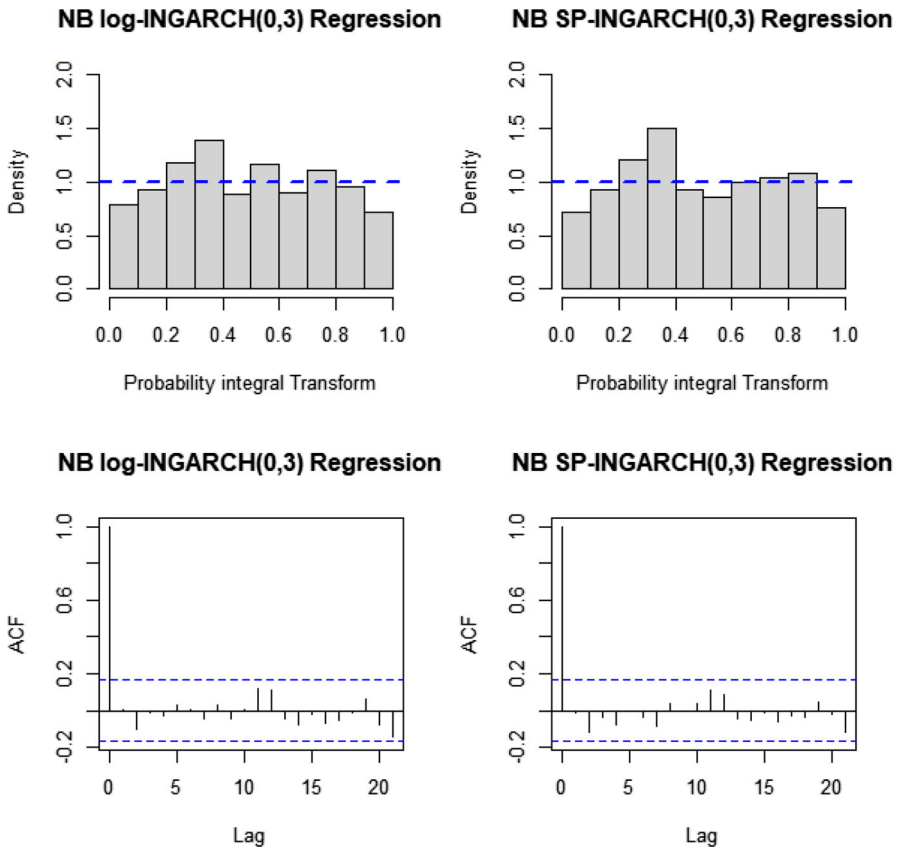
Fig. 4 PIT (top) and ACF (bottom) plots for multiple NB log- and sp-INGARCH(0,3) regression models. The blue dashed lines in PIT plots are the probability density of the uniform distribution

2.5% and 97.5% bootstrap percentiles were taken as confidence intervals. Regarding the model assessment, the ACF plots of Pearson residuals and the PIT plots in Fig. 4 do not reveal any evidence of substantial lack-of-fit, thus affirming the model adequacy.

Finally, we discuss the out-of-sample forecasting performance of the fitted models. With a sample size of 139, the dynamic one-step-ahead prediction is applied to the observations starting from the 101st till 139th. Specifically, the 101st observation is forecasted as the conditional mean estimated from the first 100 observations. Subsequently, prediction for the 102nd observation is made based on the updated conditional mean estimated from the most recent 101 observations. This procedure continues until the last observation. Then, the predictive fit is evaluated with mean squared forecasting error (MSFE), defined as

$$\text{MSFE}_t = \frac{1}{t - t_0} \sum_{s=t_0+1}^{t} (Y_s - \widehat{Y}_s)^2,$$
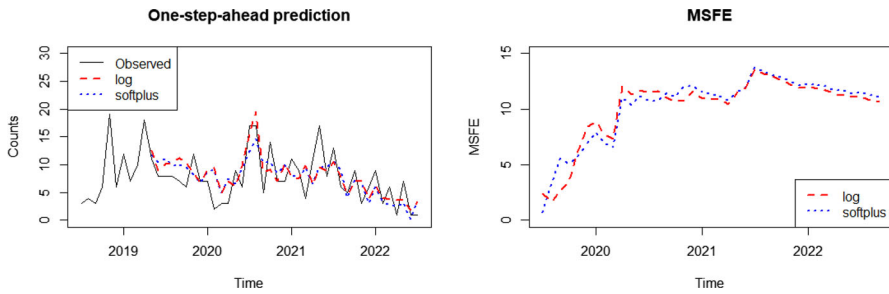
Fig. 5 Trajectories of one-step-ahead predictions (left) and MSFE (right) for NB log- and sp-INGARCH(0,3) regression models. The prediction was recursively updated from the 101st to 139th observation

where $\widehat{Y}_s$ denotes the predicted value for s^{th} observation and $t_0 = 100$. The prediction trajectories given in Fig. 5 appear to generally follow the observed counts without any considerable deviation and both models demonstrate equally decent predictive fits in terms of MSFE.

4 Conclusion

In this article, count time series regression models were considered for modeling monthly counts of sex trafficking-related convictions, at the federal level, in the United States during March 2011 to September 2022. The advantages of the INGARCH models over other count time series models were outlined. With the limited sample size availability in mind, we proposed a model selection procedure to select covariates and orders of the INGARCH processes. The negative binomial log- and sp-INGARCH models were seen to be the best-fitting models for this sex trafficking count time series. Model adequacy checks were done using the probability integral transform (PIT) plots along with a residual analysis. Model performance was also evaluated using out-of-sample forecasting accuracy. Results from our multiple INGARCH regression models indicated that equity market volatility tracker variables related to immigration policy, labor market regulations, along with changes in the consumer price index (CPI) are statistically significant in explaining sex trafficking-related convictions in the United States. To the best of our knowledge, unlike all previous attempts to model human trafficking data wherein aggregating data over time is common practice, our approach treats the response variable (sex trafficking conviction counts) as a time series and also treats the covariates (economic factors) as time-dependent. This enables the uncovering of the true time-dependent relationship between economic factors and sex trafficking convictions. As an important highlight of this time series approach, time-lagged versions of the economic factors can be considered in the regression model, and this is critically important since there is always a lag between the time of an initial arrest in a sex trafficking case and the final charging/sentencing date on that case. As an illustration of this feature, from the results in Table 6, we see that policy changes relating to immigration have a more immediate impact on sex trafficking convictions,

whereas policy and regulatory changes in the labor market have a more time-lagged impact on sex trafficking convictions.

The economic factors in Table 1 is not an exhaustive list and another direction of research is to gather time series data on other socio-economic variables such as crime rates, refugee numbers, immigration levels, and the number of border encounters, to name a few, and assess their effects on sex trafficking. Country of origin information for the victims is needed to call these economic factors as pull factors for sex trafficking. Reliable time series data from other countries is also needed to perform a similar analysis on the push factors of sex trafficking, and also to understand the joint influence of push and pull factors on sex trafficking. Quantitative human trafficking research in the United States continues to remain a challenge with the absence of comprehensive data, and the first, second, and fourth authors of this paper are currently involved in efforts to address some of the data inadequacy challenges. The current work deals with federally prosecuted sex trafficking cases and with the availability of similar data at the state level, our modeling approach can be applied to data on state level prosecutions. Other points deserving future research are (i) analysis of sex trafficking-related convictions through a robust model like the Poisson-inverse Gaussian INGARCH or INAR processes by Silva and Barreto-Souza (2019) and Barreto-Souza (2019), respectively; and (ii) model selection approach based on the prediction performance instead of the AIC measure considered in the paper.

Code Availability The computer code for implementing our method in R is made available on GitHub at https://github.com/STATJANG/HT_Analysis. The monthly sex trafficking convictions dataset considered in this work is based on press releases from the US Department of Justice.

References

Al-Osh M, Alzaid A (1987) First-order integer valued autoregressive INAR(1) process. J Time Ser Anal 8:261–275

Avendaño A, Fanning C (2013) Immigration policy reform in the united states: reframing the enforcement discourse to fight human trafficking and promote shared prosperity. Anti-Traffick Rev 2:97–118

Barreto-Souza W (2019) Mixed poisson INAR(1) processes. Stat Pap 60:2119–2139

Cho S-Y (2013) Integrating equality: globalization, women's rights, and human trafficking. Int Stud Quart 57(4):683–697

Cho S-Y (2015) Modeling for determinants of human trafficking: an empirical analysis. Soc Incl 3(1):2–21

Czado C, Gneiting T, Held L (2009) Predictive model assessment for count data. Biometrics 65(4):1254–1261

Davis R, Holan S, Lund R, Ravishanker N (2016) Handbook of discrete-valued time series, 1st edn. Chapman and Hall/CRC, UK

Davis RA, Fokianos K, Holan SH, Joe H, Livsey J, Lund R, Pipiras V, Ravishanker N (2021) Count time series: a methodological review. J Am Stat Assoc 116(535):1533–1547

Farrell A, Dank M, Kafafian M, Lockwood S, Pfeffer R, Hughes A, Vincent K (2019) Capturing human trafficking victimization through crime reporting. National Institute of Justice, Washington, p 252520

Ferland R, Latour A, Oraichi D (2006) Integer-valued GARCH process. J Time Ser Anal 27(6):923–942

Fokianos K, Tjøstheim D (2011) Log-linear poisson autoregression. J Multivar Anal 102(3):563–578

Hanson EJ, Finklea K (2022) Criminal justice data: human trafficking. Congressional Research Service, Washington, p R47211

IOM (2012) Counter trafficking module. International Organization for Migration

Kangaspunta K (2003) Mapping the inhuman trade: preliminary findings of the database on trafficking in human beings. Forum Crime Soc 3(1):81–103

Kedem B, Fokianos K (2002) Regression models for time series analysis. Wiley, London

Liboschik T, Fokianos K, Fried R (2017) tscount: an R package for analysis of count time series following generalized linear models. J Stat Softw 82:1–51

McKenzie E (1985) Some simple models for discrete variate time series. Water Resour Bull 21:645–650

O'Brien E, Hayes S, Carpenter B (2013) Causes of trafficking. Palgrave Macmillan UK, London, pp 132–165

Shamir H (2012) A labor paradigm for human trafficking. UCLA Law Rev 60:76

Silva R, Barreto-Souza W (2019) Flexible and robust mixed poisson INGARCH models. J Time Ser Anal 40:788–814

Tjøstheim D (2012) Some recent theory for autoregressive count time series. TEST 21:413–438

Weiß C (2018) fAn introduction to discrete-valued time series. Wiley, London

Weiß CH, Zhu F, Hoshiyar A (2022) Softplus INGARCH models. Stat Sin 32:1099–1120

Weiß C (2008) Thinning operations for modeling time series of counts-a survey. AStA Adv Stat Anal 92(3):319–341