A Bayesian time series model of Coca leaf production in Colombia

Sergio H. Garrido M. [1] **Emiliano Isaza** [2]

Abstract

We propose a Bayesian model to predict Coca production in Colombia so that policy makers can take uncertainty into account in their decision making process. Given the high dimensionality of the problem and the relatively low quantity of data points, we propose a Hierarchical time series approach. Even though our model performs better compared to three other specifications using Leave Future Out Cross Validation, some future modeling approaches are suggested to improve the model's performance.

1. Introduction

According to official sources, the Colombian government spent, including U.S. military aids, 1.3 billion dollars in the 2000's in Colombia's drug war (Mejia & Restrepo, 2016). Moreover, the effects of illegal crops have vast consequences, ranging from deforestation to child labour. Prediction of illegal crops is paramount in assigning resources efficiently to combat drug trafficking as well as estimating the future growth of this illegal enterprise.

There are many statistical models that have been studied in the context of illegal crops and drug production, mostly in economics; for example, Leoncini & Rentocchini (2012) estimated Cocaine production using an econometric model including solid and liquid precursors as predictors. Econometric models are common estimation techniques used by policy makers but they are usually concerned with theory validation and not with predictive performance. Additionally, there are many general equilibrium models that are used to estimate and forecast these illegal economies (Chumacero, 2008). However, these models are often devoid of reality and perform poorly when asked to predict. Bayesian models are rare within this literature and, to the best of our knowledge, there is no research aimed at predicting illegal crop growth. Bayesian models can be powerful in this context given that they allow to quantify and disaggregate both aleatoric and epistemic uncertainty.

The contributions of this paper are:

- A Bayesian estimation of sq. hectares Coca production in Colombia at a desegregated level.
- The novel Pareto Smoothed Importance Sampling (PSIS) Leave Future Out (LFO) Cross Validation (Bürkner et al., 2019) technique is used to compare alternative models and assess thoroughly their validity.

2. Method

2.1. Data

The used data comes from the United Nations Observatory for Drugs in Colombia (O.D.C), it represents this agency's yearly estimation of coca hectares grown in Colombia at a desegregated level (U.S. equivalent to counties) (Oficina de las Naciones Unidas contra la Droga y el Delito (UNODC), 2017). Furthermore, the data was aggregated into *Departamentos* which are the equivalent of states for the US. We estimate the model in differences and normalize the individual series with their standard deviation. The data has $T = 20$ time periods (20 years) and $C = 24$ series (one for each *Departamento*).

2.2. Model

We use a Hierarchical Bayesian Vector Autoregression of order 1 (HVAR). An HVAR is an extension of an uni variate autoregressive model where the series depend on past observations of the rest of the series. We justify the use of a hierarchical model by acknowledging that there is not a copious amount of data to estimate the model and the hierarchical formalism in Bayesian estimation allows to learn faster (using less data) at higher levels of abstraction. A term coined as *the blessing of abstraction* in (Goodman et al., 2011).

In our model, the observed standardized differences of Coca production in time t are defined as y_t, a vector of c columns; β, a matrix of shape (c, c), contains the parameters that represent the autoregressive relations. Moreover, α_c are the individual intercepts for each of the areas (*Departamentos*).

[1]Department of Transport, Denmark Technical University, Copenhagen, Denmark [2]Independent researcher. Correspondence to: Sergio Garrido <shgm@dtu.dk>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.

We assume each α_c is drawn from a hierarchical distribution with mean μ_α and standard deviation σ_α. Likewise, each row of the matrix β comes from a hierarchical distribution with mean $\mu_{\beta,c}$ and standard deviation $\sigma_{\beta,c}$. The underlying assumption is that the effect of previous observations of each *Departamento* come from a common distribution. Finally, $\mu_\alpha, \sigma_\alpha, \mu_{\beta,c}, \sigma_{\beta,c}$ have weakly informative priors and y_t is assumed to be normally distributed with mean $\alpha + \beta \times y_{t-1}$. Since we standardized the data, y_t has a standard deviation of 1.

The priors and likelihood of the model are summarised as follows:

$$\begin{aligned} \mu_\alpha, \mu_{\beta,c} &\sim \mathcal{N}(0, 1) \\ \sigma_\alpha, \sigma_{\beta,c} &\sim \text{HalfCauchy}(0, 1) \\ \alpha_c &\sim \mathcal{N}(\mu_\alpha, \sigma_\alpha) \\ \beta_{c,i} &\sim \mathcal{N}(\mu_{\beta,c}, \sigma_{\beta,c}) \\ y_t &\sim \mathcal{N}(\alpha + y_{t-1} \times \beta, 1) \end{aligned}$$

We estimated the model in STAN (Lee et al., 2017) using Hamiltonian Monte Carlo and a non-centered parametrization version of the model to improve sampling performance. All the code and data used to train and evaluate this model can be found in a github repository [1].

2.3. Model Validation and selection

Since our data has a relatively small amount of observations, we can estimate the exact Leave Future Out (LFO) Cross Validation (CV). In this paper, $M = 1$ is used as a step ahead cross validation because the data arrives on a yearly basis which, for practical purposes, gives enough time for policy makers to change their decisions accordingly. In order to reduce the computational burden, we use the Pareto Smoothed Importance Sampling version (PSIS) of LFO-CV (Bürkner et al., 2019).

The essential steps of PSIS LFO-CV is: 1) estimate the model with L observations, 2) find the log-probability density function (lpdf) of M steps ahead and 3) for $M > 1$ steps approximate the lpdf using PSIS. 4) If PSIS is not reliable, re-estimate the model with L now set to $L + M_{fail}$. M_{fail} is the M step ahead where the PSIS approximation failed, 5) finally, the lpdfs are summed to get a single value which is known as the expected log probability density (Gelman et al., 2013). Pareto Smoothed Importance Sampling was proposed in (Vehtari et al., 2015) and follows the same principle of importance sampling. The main difference is that in PSIS the longer tails of the distribution of weights are used to fit a generalized Pareto distribution. PSIS returns a variable k which is used to assess the validity of the approximation and returns more stable weights for importance sampling. In the case of time series, if $k \geq 0.6$, the lpdf is overestimated, therefore repeating the estimation is necessary -as explained on step 4) above.

[1] https://github.com/Chechgm/bayesian_ts_coca

3. Results

We compared our proposed model with other Bayesian models, specifically: an autoregressive model of order 1 AR(1), a Vector autoregression model of order 1 VAR(1) and a hierarchical Bayesian vector autoregression model of order 1 HVAR(1) -where the hierarchical distribution is defined for the intercepts only. Our proposed approach, as defined in the previous section, is a HVAR(1) with hierarchical distribution on weights and intercepts. The results of the LFO CV using the Expected Log Probability Density (ELPD) are summarised in Table 1.

Table 1. Expected Log Probability Density for the compared models. Higher is better.

	ELPD
AR(1)	-115.36
VAR(1)	-192.00
HVAR(1) (intercept)	-178.29
HVAR(1) (intercept and coefficients)	-110.94

Figure 1 depicts the results of our proposed model shown for a single Departamento [2]. Qualitatively, it is possible to observe that our model is able to capture a "softened" version of the observed data; Nevertheless, a lot of the variation is still to be captured. We hypothesize that the use of latent variables or exogenous variables can be used to improve the performance of the model.

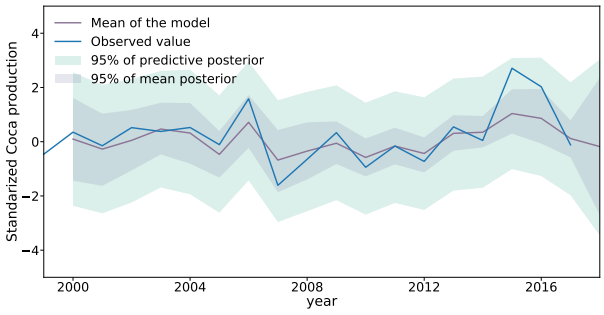

Figure 1. Results for a single Departamento in Colombia. The narrower posterior intervals correspond to the intervals of the mean, while the wider ones correspond to the posterior predictive intervals.

[2] We plot only one Departamento to avoid cluttering. The rest of the figures can be found in the link provided above.

4. Discussion and future work

As stated earlier, even though our model is able to capture a large portion of the variation in the data, we believe that enriching the model with exogenous variables could potentially improve the model quality. These variables could come from economic or criminal theory. Additionally, we propose a latent variable model in the spirit of a Hidden Markov Model for future estimations. The latent variables could capture some dynamics of criminal groups and improve the predicted variation. The great benefit of using Bayesian methods, besides the quantification of epistemic uncertainty, is that we can mix the models in order to come up with more powerful representations of reality.

References

Bürkner, P.-C., Gabry, J., and Vehtari, A. Approximate leave-future-out cross-validation for time series models. *arXiv preprint arXiv:1902.06281*, 2019.

Chumacero, R. A. *Evo, Pablo, Tony, Diego, and Sonny-general equilibrium analysis of the illegal drugs market*. The World Bank, 2008.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. *Bayesian data analysis*. CRC press, 2013.

Goodman, N. D., Ullman, T. D., and Tenenbaum, J. B. Learning a theory of causality. *Psychological review*, 118 (1):110, 2011.

Lee, D., Carpenter, B., Li, P., Morris, M., Betancourt, M., maverickg, Brubaker, M., Trangucci, R., Inacio, M., Kucukelbir, A., buildbot, S., bgoodri, seantalts, Arnold, J., Tran, D., Hoffman, M., Margossian, C., Modrák, M., Adler, A., Sakrejda, K., Stukalov, A., Lawrence, M., Goedman, R. J., Horn, K. S. V., Vehtari, A., Gabry, J., Casallas, J. S., and Bales, B. stan-dev/stan: v2.17.1, December 2017. URL https://doi.org/10.5281/zenodo.1101116.

Leoncini, R. and Rentocchini, F. Let it snow! let it snow! let it snow! estimating cocaine production using a novel dataset based on reported seizures of laboratories in colombia. *International Journal of Drug Policy*, 23(6): 449–457, 2012.

Mejia, D. and Restrepo, P. The economics of the war on illegal drug production and trafficking. *Journal of Economic Behavior & Organization*, 126:255–275, 2016.

Oficina de las Naciones Unidas contra la Droga y el Delito (UNODC). *Monitoreo de territorios afectados por cultivos ilícitos 2016*. 2017.

Vehtari, A., Simpson, D., Gelman, A., Yao, Y., and Gabry, J. Pareto smoothed importance sampling. *arXiv preprint arXiv:1507.02646*, 2015.