32nd European Modeling & Simulation Symposium
17th International Multidisciplinary Modeling & Simulation Multiconference

ISSN 2724-0029 ISBN 978-88-85741-44-7 © 2020 The Authors.
DOI: 10.46354/i3m.2020.emss.033

Using Hybrid Bayesian Networks to Detect Audience Behaviour Changes in Youtube

Ezequiel Leonardo Castaño[1,*] and Guillermo Leale[1]

[1]Universidad Tecnológica Nacional Facultad Regional Rosario, Zeballos 1341, Rosario, C2000, Argentina

[*]Corresponding author. Email address: ecastano@frro.utn.edu.ar

Abstract

The nowadays volatile fame one could gain from the Internet led to a disruption in the media industry, with important repercussions in platforms such as the educational channels from YouTube. In this work, one of such channels, which was shut down after more than 2 years of activity, is studied through the use of a Hybrid Bayesian Network with Markov Chain Monte Carlo based sampling. With the application of our model, the behaviour of users can be inferred and thus find whether it changed at some point. As a result, it was indeed possible not only to identify two specific moments in time when that changed but also to provide a transition zone between the steady states before and after the change.

Keywords: Markov Chain Monte Carlo; Hybrid Bayesian Networks; Youtube;

1. Introduction

Nowadays, Youtube is the *de facto* platform to share video content. However, its purpose shifted from only sharing media to becoming a social network, where the video publishers motivate their viewers to perform certain actions within the platform, such as leave a comment or send them an approval signal (usually in the form of a "like" or "share"). This context makes the content strongly dependent not only on the quality but also on how the publishers interact with their audiences. Nevertheless, unexpected events often occur and may have a significant impact in the audience behaviour, the most common of which is the so-called "Viral Videos", which are videos that for some reason received disproportionately more attention than the rest of the content, another possible example could be controversial claims in the video or its comments. In this work some of the publicly available data of one video publisher will be analysed to determine whether it is possible to detect such changes by means of a statistical based method. This publisher was carefully chosen because it experimented great variation which is thought to be the cause of its posterior shutdown. This work uses data from a single publisher as a test for applicability since the results were promising future research will include determining the exact conditions required to apply the method, in principle, there are no particular restrictions known, so applying the method to similar data from other publishers would be possible, however, that hypothesis should be tested, such test is outside of the scope of the current study.

Therefore, the questions to be answered are the following: a) Is there any statistically identifiable point where the behaviour of the audience change? b) If so, was the change positive or negative? c) Is it possible to identify more than one of such changes?

This work first presents the data collection process. Then, basic exploratory analysis is performed and a model is proposed. This model is then compared with some alternative formulations. Finally, the results and conclusions are discussed in detail.

2. State of the Art

Monte Carlo Methods were first proposed to solve physics problems through simulation (Metropolis, 1987), one of the applications of these methods was called Markov Chain Monte Carlo (MCMC) (Gelfand and Smith, 1990), whose aim is to approximate the stationary distribution of a Markov Chain. On the other hand, Bayesian Networks (BN) (Pearl, 1988) are a type of Probabilistic Graphical Model (PGM) represented as a directed acyclic graph (DAG), where each of the nodes represents a random variable, being only some of the nodes observable. The objective is to infer the distribution of the connected nodes for which there are no observations. This process is usually called "Bayesian Inference" and although it could be solved analytically, complex models representing high dimensional and complicated relationships involve intractable integrals. Therefore, in these cases the analytical approach is unfeasible.

However, these two methods can be combined such that the posterior distribution of the variables in a BN corresponds to the stationary distribution of a Markov Chain (Gamerman and Lopes, 2006). In spite of providing simulation-based solutions, the applications of BNs were rather limited for a long time because of the computation costs involved. Nevertheless, in recent years new and more efficient sampling algorithms such as the No-U-Turn Sampler (NUTS) were developed (Hoffman and Gelman, 2014).

The variables modeled through BNs are often continuous. This is actually one of the conditions of many modern sampler methods with the exception of the simpler and poorly performing ones, for example, the well-known Metropolis-Hastings Hastings (1970). Therefore if both continuous and discrete variables are needed, the resulting model is called Hybrid Bayesian Network, which requires much more computational resources than the continuous BN (Salmerón et al., 2018).

Some of the common applications of BNs nowadays are (but not limited to) Control Engineering (Bapin and Zarikas, 2019), Transportation (Corman and Kecman, 2018), Reliability Evaluation(Cai et al., 2018), Agriculture (Drury et al., 2017) among others (Pourret et al., 2008). Although some previous works included time series data and also change point detection using BNs (Aminikhanghahi and Cook, 2017), little has been done in the context of social networking sites.

This work studies a particular scenario in the context of video streaming platforms, using BNs as a proof for the applicability in the field and to evaluate whether the results are insightful for retrospective analysis. The most common platform, Youtube, was chosen as a source for the data because of its popularity and the fact that it was proven to be widely used in academic research in the past (Arthurs et al., 2018). Generally, to answer the questions at hand, either classical statistics (time series analysis and regression) or more recently Deep Learning techniques are used, both being (without some exceptions) discriminative models, as opposed to BNs which are generative and, as such, more suitable when the number of observations is small (Ng and Jordan, 2002). Moreover, generative models could be used afterward to answer different questions than the ones initially addressed, which is a major advantage for the stakeholders in this industry.

A similar approach in this direction was done as an example model in an introductory book of Bayesian Statistics (Davidson-Pilon, 2015), however, the application context was different. Only a single likelihood was used and the incorporation of different transformations was not covered, hindering potential remarks as pointed in other works on count data (O'hara and Kotze, 2010).

3. Materials and Methods

When dealing with Bayesian Inference, the quality of the data is crucial to properly fit the proposed models (Dose, 2003; Gamerman and Lopes, 2006). The data should not only be collected in a careful way but also an exploratory analysis has to be conduted to detect possible outliers or patterns. The next step is to assess whether the models converge and finally, as a sanity check, some alternative models should also be proposed to test the initial findings. This section will cover all of the previously mentioned topics.

3.1. Data Collection

We selected the channel *PBS Infinite Series* on YouTube, which is no longer publishing videos. Data is available on their site publicly. All the videos can be found on the main page of the channel, available at https://www.youtube.com/channel/UCs4aHmggTfFrpkPcWSaBN9g/videos. Although many variables can be considered, in this case, the following set of variables were recorded:

- Number of Views
- Number of Likes
- Number of Dislikes
- Number of Comments

In addition to the previous ones, some meta-data was collected to better identify each of the points, such as Title, URL, Date, and Channel name.

This data, however, presents two challenges:

1. **Atemporality**: The data will be read from a particular moment in time but this number will likely change in the future.
2. **Correlation**: The data is highly correlated because each video will influence the variables of other videos. Although this correlation is typically in a forward direction (old videos affect new ones), it is also possible to have a back-propagation effect (new videos influencing older ones), such as when new people find the channel and want to see previous content.

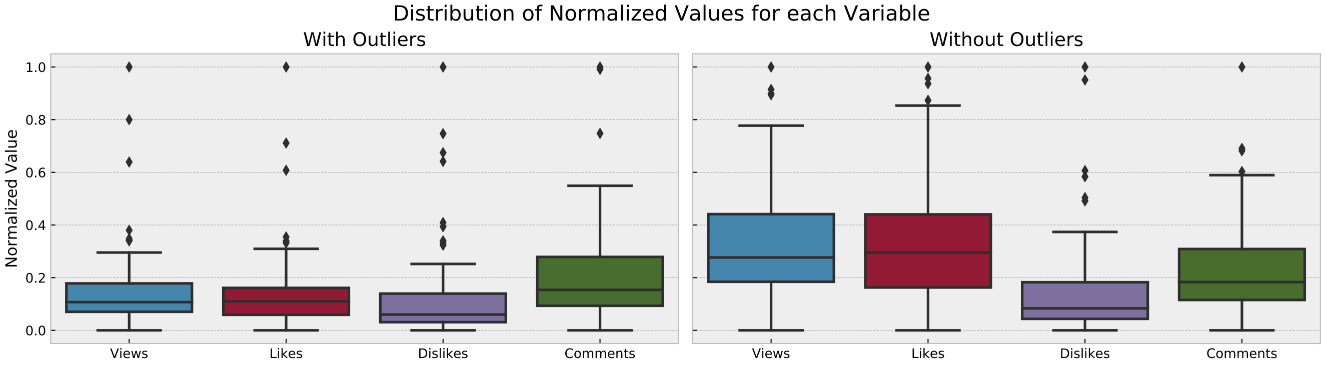

Figure 1. Distribution of the normalized values with outliers (left) and without (right). Whiskers represents 1.5 IQR

One of the reasons for choosing an already shut down channel was precisely to mitigate these effects. The data will undergo changes in the future but this variation could be neglected if the difference in time is short enough. Regarding the correlation, it is impossible to track all the interactions since each of the variables represents a time series of data for each video, of which only the last value is publicly accessible, but since the amount of videos is small ($N = 65$) it should be fair to assume that and the channel has a niche audience. This phenomenon will have also less impact as it could have been under other conditions.

The data was collected through scraping (a technique consisting of analysing the HTML of the website and extract specific segments) on the 5th of May 2020. The data collection was done with the Selenium Web Driver. A summary of the data is presented in the Table 1. The total number of data points is 65 and there are no missing values (Although Youtube provides the data publicly there are some cases where the user can block the comment section or, in the case of YouTube premiums, there might be no count for likes, neither of these occurs in this scenario). Moreover, the specific dataset used is made publicly available at https://doi.org/10.5281/zenodo.3929778 (Castaño, 2020)

Table 1. Summary statistics of the collected data

	Views	Likes	Dislikes	Comments
Mean	161298	4493	152	628
STD	145217	3735	186	464
Min	23392	1082	23	141
25%	83144	2414	53	351
50%	114477	3569	81	487
75%	174787	4724	158	769
Max	873302	23686	991	2394

3.2. Exploratory Analysis

From the table, one can see that the maximum of every variable is quite far from the 75% percentile, which might be an indication of outliers. In order to properly identify them, a box plot for each variable is used but, since the range of values is orders of magnitude different between variables, the values will be scaled to the $[0, 1]$ interval so that proportions are kept and the same scale can be used.

In the boxplots shown in the Fig. 1, there are three points that lie further from the rest, although it is expected that some points lie outside the whiskers, this case is rather extreme. The main reason for these outliers is the so-called "Viral Videos" phenomenon, which represents videos that received such an uncommon amount of attention that they could be easily differentiated from the rest. Since such effect is outside the scope of this work, these points can simply be omitted in the analysis.

With the outliers removed, it is important to identify whether there are relationships between the variables. This could be done through a correlation matrix, which is shown in the Table 2. The variables are highly correlated with each other (Pearson's r is always greater than 0.5), which could introduce multi-colinearity to the model. Therefore, instead of using all of the variables only the View Count (from now on x) will be used. A causal study of these correlations is left for a future investigation.

Table 2. Pearson's r correlation between the variables. Only half of the correlation matrix is shown to improve readability.

	Views	Likes	Dislikes	Comments
Views	1.0			
Likes	0.94	1.0		
Dislikes	0.54	0.53	1.0	
Comments	0.78	0.79	0.65	1.0

Another important aspect to check is whether there is seasonality and/or auto-correlation. Deep analysis in this direction is beyond the scope of this study but the auto-correlation and the seasonal decomposition were checked using black box models and no significant anomalies were found.

Since the aim of this study is to identify some change-point among the videos, the number of videos published each month along with the views for each

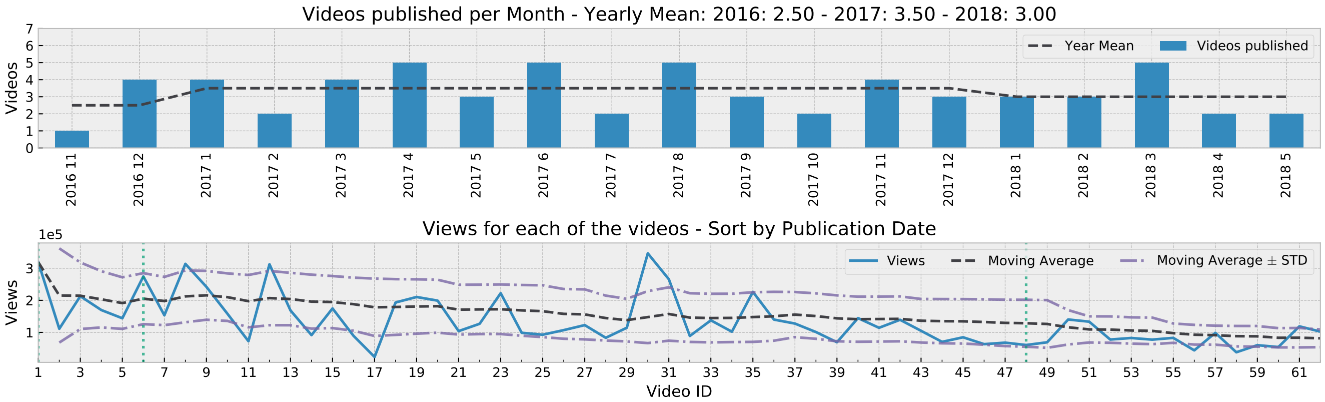

Figure 2. Videos Published per Month (Above) and views per video (below). The moving average is calculated using a windows size of 20 and the STD is calculated using the corresponding moving average with the same windows size. The vertical dotted lines at 6 and 48 denote the change in the year.

video is shown in the Fig. 2. It can be seen that there is a significant yet weak decreasing trend and a reduction in the spread starting from the 50th video. Most of the videos were published in 2017.

Only analysing these basic plots, there is no clearly visible change-point that could be identified. In the next section, Markov Chain Monte Carlo models will be used to check whether this initial assumption is true.

3.3. Model Formulation

The data represented by the Views could be interpreted as count data, therefore, discrete distributions over Natural Numbers are appropriate. In this case, two candidate likelihood functions will be used to model the data, a Poisson Distribution and a Negative Binomial Distribution (Although the Negative Binomial is also used to model number of failures before a success, it has been proven adequate to model count data too.). The former will present a good fit if the mean and the variance are similar and the latter will in contrast provide a better result if they differ significantly.

In this section, the formulation will be expressed for the general case. However, in the results section, two different approaches are compared, the analysis with the outliers and without them. That means that the total number of videos (n), will have two possible values, $n = 65$ when outliers are considered and $n = 62$ otherwise.

3.3.1. *Model Description*

First, it is assumed that there is a change-point (τ), which denotes a moment in time where the basic behaviour of the audience, represented as the number of views (x), changed. Here τ is the index number (dimensionless quantity) of the video from the list of all the videos sorted by date. Since there are no prior assumptions about where this τ may be, a non-informative prior will be used, in this case, a Discrete Uniform (U_{disc}) bounded between the possible values of the Videos $[1, n]$, namely:

$$\tau \sim U_{disc}(1, n) \tag{1}$$

The data before and after τ is assumed to be generated from different candidate distributions (from the same family), if this were not the case, the two distributions might be indistinguishable from each other, i.e. the null hypothesis of their parameters before and after τ being different would be rejected. In the case of the Poisson model it would be expressed as $H_0 : \mu_1 \neq \mu_2$ and $P(reject\ H_0) \approx 1$.

Since the number of views represents count data two discrete distributions will be used, a Poisson and a Negative Binomial. In the case of the Poisson distribution there will be one parameter for each period (μ_1 and μ_2) and in the case of the Negative Binomial there will two parameters for each period ($\{\mu_1, \alpha_1\}$ and $\{\mu_2, \alpha_2\}$). The particular parametrization of the distributions is as follows:

$$Poisson(x \mid \mu) = \frac{e^{-\mu}\mu^x}{x!}$$

$$NB(x \mid \mu, \alpha) = \binom{x + \alpha - 1}{x} \left(\frac{\alpha}{\mu + \alpha}\right)^{\alpha} \left(\frac{\mu}{\mu + \alpha}\right)^{x}$$

Having such configuration results in the following model:

$$x \sim Poisson(\mu_1)\ I_{x\leq\tau} + Poisson(\mu_2)\ I_{x>\tau} \tag{2}$$

$$x \sim NB(\mu_1, \alpha_1)\ I_{x\leq\tau} + NB(\mu_2, \alpha_2)\ I_{x>\tau} \tag{3}$$

Where I_A is the indicator function, namely:

$$I_A = \begin{cases} 1 & \text{if event } A \text{ occurs} \\ 0 & \text{otherwise.} \end{cases}$$

This way of parametrizing the Negative Binomial makes it easier to use the mean and the variance from the data to determine proper informative priors. Moreover, the μ parameter has the same interpretation in both distributions when parametrized this way. In the case of the Negative Binomial the α parameter is a shape parameter associated with the variance (σ^2) of the distribution, namely:

$$\mu = \overline{x} \qquad ; \qquad \alpha = \frac{\mu^2}{\sigma^2 - \mu}$$

However, the expressions shown above only represent the expected value of these parameters, in order to add another degree of freedom, a prior for each will be defined. Since all are positive numbers, an Exponential distribution will be used, the mean value of the distribution will be calculated according to the segment of the data being modeled. The parametrization will be as follows:

$$E(\mu_1) = \overline{\mu_1} = \frac{1}{\tau}\sum_{i=0}^{\tau} x_i \qquad E(\mu_2) = \overline{\mu_2} = \frac{1}{n-\tau}\sum_{i=\tau}^{n} x_i$$

$$\sigma_1^2 = \frac{1}{\tau}\sum_{i=0}^{\tau}(x_i - \overline{\mu_1})^2 \qquad \sigma_2^2 = \frac{1}{n-\tau}\sum_{i=\tau}^{n}(x_i - \overline{\mu_2})^2$$

$$E(\alpha_1) = \overline{\alpha_1} = \frac{\mu_1^2}{\sigma_1^2 - \mu_1} \qquad E(\alpha_2) = \overline{\alpha_2} = \frac{\mu_2^2}{\sigma_2^2 - \mu_2}$$

Then, the prior distributions for μ and α are:

$$\mu_1 \sim Exp(1/\overline{\mu_1}) \qquad \mu_2 \sim Exp(1/\overline{\mu_2}) \tag{4}$$

$$\alpha_1 \sim Exp(1/\overline{\alpha_1}) \qquad \alpha_2 \sim Exp(1/\overline{\alpha_2}) \tag{5}$$

To summarize (1), (2), (3), (4) and (5), two probabilistic graphical models (PRGMs) are shown in the Fig.3.

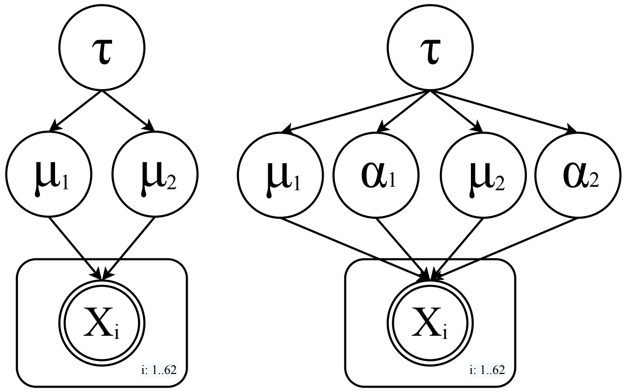

Figure 3. Probabilistic Graphical Model for the two alternatives, one with the Poisson Likelihood (left) and the other with the Negative Binomial Likelihood (right). The double circle in X_i denotes that this is the observed variable.

If there was an actual change-point in x, it will be reasonable to expect that the posterior distribution of τ is not uniform and that the model parameters before and after τ differ by a significant amount. Determining when a difference is significant depends on the stakeholders and the specific problem but, mathematically it can be stated that if $P(\mu_1 > \mu_2) \approx 1 \vee P(\mu_1 < \mu_2) \approx 1$ the difference is significant beyond any reasonable doubt and if $P(\mu_1 - \mu_2 > 0) \approx 0.5 \vee P(\mu_2 - \mu_1 > 0) \approx 0.5$ the difference is imperceptible and the two distributions should be consider equal (within a small enough ϵ). In order to incorporate the vision from the stakeholders, an appropriate approach would be to consider the difference significant when $P(|\mu_1 - \mu_2| > \epsilon) \approx 1$ (absolute), or $P(\frac{max(\mu_1,\mu_2)}{min(\mu_1,\mu_2)} > \epsilon) \approx 1$ (relative). In both cases the value of ϵ should be chosen by the stakeholders.

The actual values of the parameters besides τ are not relevant since this model will not be used for any sort of prediction.

3.3.2. *Model Assesment*

Since the values from x are quite big, this could make the Poisson model inadequate, since it assumes that the variance and the mean are equal. To mitigate this effect and at the same time test the robustness of the solution several models were fit. In each case, 20 chains were used to test for convergence and the number of iterations was increased until the Gelman-Rubin diagnostic (Gelman et al., 1992) was less than 1.02 (Therefore, the number of iterations was not the same for each model).

3.4. Alternative Comparison

In total 8 different models were fit using the Programming Language Python and the PyMC3 Framework (Salvatier et al., 2016), these models represent the incorporation or exclusion of outliers, the likelihood candidate and the use of transformation. In order to compare the models the Root Mean Squared Error (RMSE) is used. The results are shown in the Table 3. For the two likelihood candidates, excluding the outliers produced much better results.

Table 3. Root Mean Squared Error (RMSE) of the residuals for each model, divided by the global minimum (absolute value of 62866)

	With Outliers		Witout Outliers	
	Poisson	Negative Binomial	Poisson	Negative Binomial
No Transform	2.18	2.12	1.01	1.00
Square Root	2.15	2.14	1.02	1.01

All the transformations were done in the explanatory variable x. In the case of the Square Root, the model was fit using $y = \sqrt{x}$. A Log transform was also tested but quickly discarded, when applied to the explanatory variable, the apparent difference before and after the τ became imperceptible, and when applied to the re-

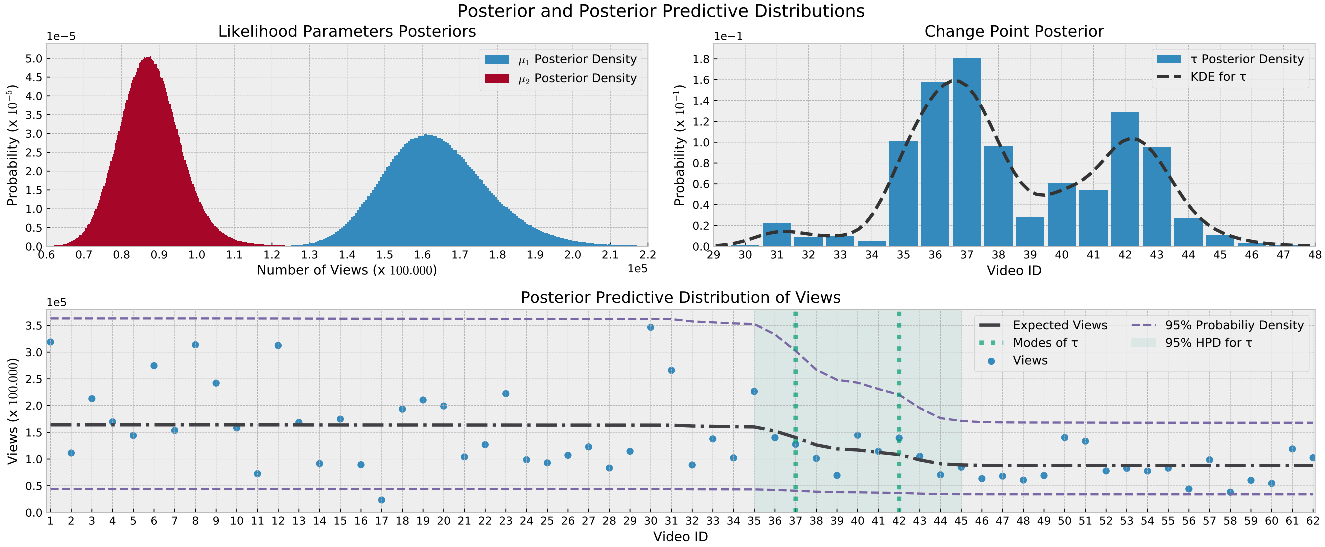

Figure 4. Posterior densities (Top Left) of μ_1 (blue) in contrast with μ_2 (red). Posterior density (Top Right) of τ with Kernel Density Estimate (KDE) to clearly show bi-modality. Posterior Predictive (Bottom) of the model showing the Highest Probability Density (HPD) up to 95 % for the views and also for τ

sponse variable, overflow issues arose when dealing with e^x for bigger values of x.

When analysing the normality of the residuals with respect to the expected value of each model, only three out of the eight models passed the test (with $\alpha = 0.5\%$), these three models are the Poisson without outliers (both with and without transformation) and the Negative Binomial without the transformation and without outliers.

4. Results

Regarding the explanatory power of the models fit, a clear pattern emerged, the ones using the Poisson likelihood had all the posterior density of τ concentrated on a single point ($\tau = 37$), regardless of transformations, while the ones using the Negative Binomial have more spread in the posterior estimates and exhibit a bi-modal behaviour, having one mode in $\tau = 37$ and the other in $\tau = 42$.

In all cases, when comparing the posteriors for μ, it was found that $P(\mu_1 > \mu_2) \approx 1$, indicating that the change was a reduction in the number of views as suspected from the Fig. 2. As the trend was decreasing, this is confirmed by having $\mu_2 < \mu_1$. The posteriors as well as the posterior predictive distributions for the Best Performing model are shown in the Fig. 4 with a sample size of 2.4 times 10^6.

5. Discussion

Incorporating extra information such as sentiment analysis of the comments could give a better interpretation of the results obtained using only the number of views. Another important consideration is the fact that the sample size is somewhat small ($n = 62$ without outliers). Replicating this study with other YouTube channels with bigger samples could reveal further limitations in the approach taken.

6. Future Work

A natural future step in this research will be to improve the method so that it could be used with active channels, estimating the minimum sample size to detect audience changes and provide insights to the owners to take countermeasures.

7. Conclusions

It was indeed possible to identify change points in the data, suggesting that there was an actual moment in time when the behaviour of the users changed, in all the models (although the values differ) the change represented a reduction in the views, in the context of YouTube videos such effect is considered negative.

Using Poisson likelihood only one change-point was detected, (Video ID=37) while using a Negative Binomial likelihood, two points were identified, both 37 and 42. However, the Posterior Predictive from the Negative Binomial represents a better fit for the data, since the Poisson posterior predictive left a lot of points outside the 95% Highest Density Interval (HDI). There are two steady-states zones and a transition zone could be identified in the interval [35 – 45] which is the one that accumulates 95% of the probability density. Being 37 and 42 the most likely options.

These results are consistent with the reality because, although there was no particular event (known) around the time of video 37, there was indeed a big change in the video 42, which was a Host change in the channel.

References

Aminikhanghahi, S. and Cook, D. J. (2017). A survey of methods for time series change point detection. *Knowledge and information systems*, 51(2):339–367.

Arthurs, J., Drakopoulou, S., and Gandini, A. (2018). Researching youtube.

Bapin, Y. and Zarikas, V. (2019). Smart building's elevator with intelligent control algorithm based on bayesian networks. *Int. J. Adv. Comput. Sci. Appl.(IJACSA)*, 10(2):16–24.

Cai, B., Kong, X., Liu, Y., Lin, J., Yuan, X., Xu, H., and Ji, R. (2018). Application of bayesian networks in reliability evaluation. *IEEE Transactions on Industrial Informatics*, 15(4):2146–2157.

Castaño, E. L. (2020). Public Dataset for the Paper "Using Hybrid Bayesian Networks to Detect Audience Behaviour Changes in Youtube".

Corman, F. and Kecman, P. (2018). Stochastic prediction of train delays in real-time using bayesian networks. *Transportation Research Part C: Emerging Technologies*, 95:599–615.

Davidson-Pilon, C. (2015). *Bayesian methods for hackers: probabilistic programming and Bayesian inference*. Addison-Wesley Professional.

Dose, V. (2003). Bayesian inference in physics: case studies. *Reports on Progress in Physics*, 66(9):1421.

Drury, B., Valverde-Rebaza, J., Moura, M.-F., and de Andrade Lopes, A. (2017). A survey of the applications of bayesian networks in agriculture. *Engineering Applications of Artificial Intelligence*, 65:29–42.

Gamerman, D. and Lopes, H. F. (2006). *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. CRC Press.

Gelfand, A. E. and Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410):398–409.

Gelman, A., Rubin, D. B., et al. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications.

Hoffman, M. D. and Gelman, A. (2014). The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15(1):1593–1623.

Metropolis, N. (1987). The beginning of the. *Los Alamos Science*, 15:125–30.

Ng, A. Y. and Jordan, M. I. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in neural information processing systems*, pages 841–848.

O'hara, R. B. and Kotze, D. J. (2010). Do not log-transform count data. *Methods in ecology and Evolution*, 1(2):118–122.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

Pourret, O., Naïm, P., and Marcot, B. (2008). *Bayesian networks: a practical guide to applications*. John Wiley & Sons.

Salmerón, A., Rumí, R., Langseth, H., Nielsen, T. D., and Madsen, A. L. (2018). A review of inference algorithms for hybrid bayesian networks. *Journal of Artificial Intelligence Research*, 62:799–828.

Salvatier, J., Wiecki, T. V., and Fonnesbeck, C. (2016). Probabilistic programming in python using PyMC3. *PeerJ Computer Science*, 2:e55.