Conjunction Data Messages behave as a Poisson Process

Francisco Caldas[1,2] , **Cláudia Soares**[3] , **Cláudia Nunes**[2] , **Marta Guimarães**[1] , **Mariana Filipe**[1] , **Rodrigo Ventura**[2]

[1]Neuraspace

[2] Instituto Superior Técnico and CEMAT

[3] NOVA School of Science and Technology

{francisco.caldas,marta.guimaraes,mariana.filipe}@neuraspace.com, claudia.soares@fct.unl.pt,
cnunes@math.tecnico.ulisboa.pt, rodrigo.ventura@isr.tecnico.ulisboa.pt

Abstract

Space debris is a major problem in space exploration. International bodies continuously monitor a large database of orbiting objects and emit warnings in the form of conjunction data messages. An important question for satellite operators is to estimate when fresh information will arrive so that they can react timely but sparingly with satellite maneuvers. We propose a statistical learning model of the message arrival process, allowing us to answer two important questions: (1) Will there be any new message in the next specified time interval? (2) When exactly and with what uncertainty will the next message arrive? The average prediction error for question (2) of our Bayesian Poisson process model is smaller than the baseline in more than 3 hours in a test set of 50k close encounter events.

1 Introduction

Since the early 1960s, the space debris population has extensively increased [Radtke *et al.*, 2017]. It is estimated that more than 34,000 objects larger than 10 centimetres, and millions of smaller pieces, exist in Earth's orbit [ESA, 2021]. Collisions with debris give rise to more debris, leading to more collisions in a chain reaction known as Kessler syndrome [Krag *et al.*, 2017]. To avoid catastrophic failures, satellite owners/operators (O/Os) need to be aware of the collision risk of their assets [Le May *et al.*, 2018]. Currently, this monitoring process is done via the global Space Surveillance Network (SSN). To assess possible collisions, a physics simulator uses SSN observations to propagate the evolution of the state of the objects over time [Horstmann and Stoll, 2017; Mashiku and Hejduk, 2019]. Each satellite (also referred to as target) is screened against all the objects of the catalogue in order to detect a conjunction, i.e., a close approach. Whenever a conjunction is detected between the target and the other object (usually called chaser), SSN propagated states become accessible and a conjunction data message (CDM) is issued, containing information about the event, such as the time of closest approach (TCA) and the probability of collision. Until the TCA, more CDMs are issued with updated and better information about the conjunction. Roughly in the interval

between two and one day prior to TCA, the O/Os must decide whether to perform a collision avoidance manoeuvre, with the available information. Therefore, the CDM issued at least two days prior to TCA is the only guaranteed information that the O/Os have and, until new information arrives, the best knowledge available.

Several approaches have been explored to predict the collision risk at TCA, using statistics and machine learning [Acciarini *et al.*, 2021], but only a few were developed with the aim of predicting when the next CDM is going to be issued. Very recently, [Pinto *et al.*, 2020] developed a recurrent neural network architecture to model all CDM features, including the time of arrival of future CDMs. GMV is currently developing an autonomous collision avoidance system that decides if the current information is enough for the O/Os to decide, or if they should wait for another CDM to have more information. However, the techniques used are not publicly available. In this work, we propose a novel statistical learning solution for the problem of modeling and predicting the arrival of CDMs based on a homogeneous Poisson Process (PP) model, with Bayesian estimation. We note that standard machine learning and statistical learning solutions are data-hungry and cannot model our problem, suffering from a special type of data scarcity. Further, as our application is high-stakes, we require confidence or credibility information added to a point estimate.

Background. The present work formalizes the problem of modelling and predicting the arrival of a new CDM with a probabilistic generative model. The present formulation has the advantage of providing a full description of the problem, stating clearly the required assumptions. Our proposed model shows high accuracy, decreasing error in more than 3 hours, when compared with the baseline, for predicting the next event. We leverage on a data model depicted in Figure 1. Probabilistic Graphical Models [Koller and Friedman, 2009; Bishop, 2006] is a probabilistic framework that allows for highly expressive models, while keeping computational complexity to a minimum. A few statistical learning developments like the Latent Dirichlet Allocation for topic modelling [Blei *et al.*, 2003], are now the basis for deep learning-powered probabilistic models like the Variational Autoencoder [Kingma *et al.*, 2019] and Normalizing flows [Kobyzev *et al.*, 2020], and as a means of creating modular and interpretable machine learning models [Murdoch *et al.*, 2019]. We
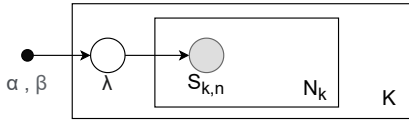
Figure 1: Graphical model representation of the proposed model. The boxes are plates representing replication of the structure. The outer plate represents the dataset of K events, and the inner one the repeated prediction of N_k occurrences in each event.

leverage on the homogeneous Poisson Point Process [Kulkarni, 2016], a counting process of messages occurring over continuous time. Thus, we assume that the inter-CDM times are independent and identically distributed random variables with exponential distribution with a rate λ. The homogeneity property comes from an event rate constant in time. Phenomena like radioactive decay of atoms and website views are well modeled by this process [Ross, 2014]. The homogeneity is, nevertheless, a limitation that will be addressed in future work, as there might be some temporal trend in the data. Not considering it, though, delivers a fast to compute solution that can outperform a standard baseline.

2 Problem Formulation

Our generative probabilistic model of the CDM arrival process will allow answering some questions as (1) what is the probability that, during the decision phase, the roughly one day that the operator has to decide to perform an avoidance manoeuvre, there are going to be received more CDMs with up-to-date information, or (2) what is the best estimate for the next CDM arrival time, and the uncertainty of this prediction.

Definition 1. *Following the literature, we define the stochastic quantities, with k as the id of the object:*

- X_{n+1}^k *time between CDM n and CDM $n + 1$, where $X_{n+1}^k \sim Exp(\lambda_k)$, i.e., X_{n+1} has an exponential distribution of rate λ_k;*

- $N^k(t)$ = *number of CDMs received in the interval (0,t], where $N^k(t) \sim Poisson(\lambda_k t)$, where λ_k is the rate of CDMs issued per day;*

- $S_n^k = \sum_{i=1}^n X_i^k$ *as time of occurrence of the n_{th} CDM;*

- *as $\{X_i^k, i = 1, .., n\}$ are i.i.d., it follows that $S_n^k \sim Gamma(n, \lambda_k)$.*

We are interested in the probability of receiving one or more CDMs in the time interval from the last available observation s_n^k, until a constant security threshold (t_{sec}). In this application, we consider 1.3 days prior to TCA, and the interval will be $(s_n^k, t_{sec}]$.

$$P(N^k(t_{sec} - s_n^k) \geq 1) = 1 - P(N^k(t_{sec} - s_n^k) = 0) \quad (1)$$
$$= 1 - P(\text{Poisson}(\lambda_k(t_{sec} - s_n^k) = 0)$$

To determine the most likely time for the next CDM to be issued, for each event, using the fact that inter-CDM times have Exponential distribution:

$$S_{n+1}^k = S_n^k + X_n^k \qquad X_n^k \sim Exp(\lambda_k)$$
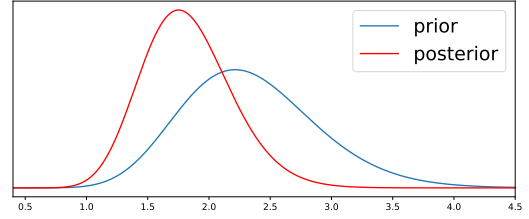$$P(S_{n+1}^k - S_n^k \leq s | S_n^k = s_n^k) = 1 - e^{-\lambda(S_{n+1} - s_n^k)}$$

Figure 2: Prior distribution of λ, and a posterior distribution for a single event in the train data. We can see that while informative, the prior has enough variance that with few updates there is a shift in the distribution of the posterior rate of arrival (λ_k). We can also see that the posterior is narrower, meaning that the credible interval of the posterior is smaller.

With these assumptions, we are left with the need to estimate λ_k, for each event k. In the next section we explain the approach.

3 Maximum a Posteriori

Because of the small number of CDMs in each event, to have a statistically significant estimation of λ_k, and to use information from the other events, we adopt the Bayesian model in Figure 1. More specifically, we use the extensive amount of events in the dataset to get an informative prior, and then we update that distribution with the specific inter-CDMs times of each event. We end up with a posterior distribution for λ_k, of which we can extract a Maximum a Posteriori (MAP), $\hat{\lambda}_k$ [Insua and Ruggeri, 2012].

The posterior λ distribution is defined as such:

1. λ has prior Gamma(α,β) where α and β are hyperparameters estimated using the events in the train data;

2. The likelihood function, considering we have a set of X_i for each event is

$$l(\lambda_k | X_1^k, X_2^k, ..., X_n^k) = \lambda_k^n e^{-\lambda_k \sum_{i=1}^n X_i^k}$$

3. With $T = \sum_{i=1}^n X_i$, the posterior can therefore be written as [Kulkarni, 2016]:

$$f(\lambda_k | n, T) \sim \text{Gamma}(\alpha + n, \beta + T). \quad (2)$$

4 Results

The first result we show is the prediction of the time of the next CDM.

	MAE	MSE
Bayesian PP	0.15170	0.06282
Classic PP	0.17581	0.08789
Baseline	0.26075	0.19105

Table 1: Mean Average Error (MAE) and Mean Squared Error (MSE) for next CDM time prediction in days.

These results were obtained using an unbiased test data of 50000 independent events, that was not seen during the study
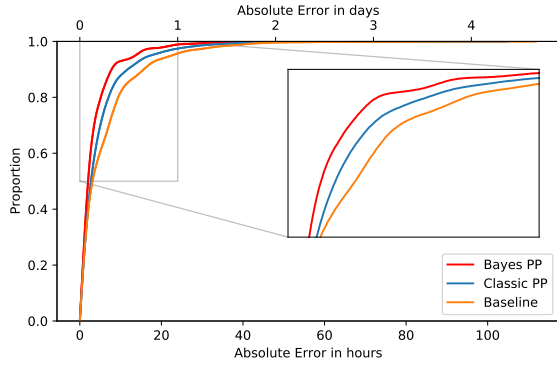
Figure 3: Empirical CDF for each of the models. There is a perceivable difference in the predictive capability of each model. Because the three models use the same formula do predict the expected time of the next CDM, the difference lies in parameter determination, that is, the rate of arrival λ_k.
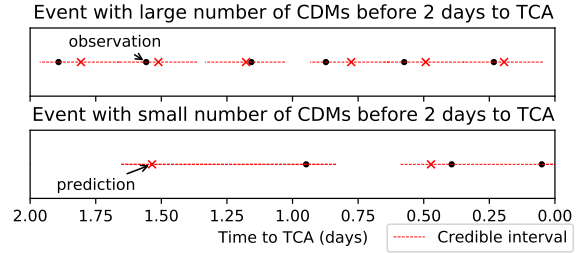

Figure 4: Example of the predictive capability of the model for different examples. Top panel uses 15 observations to update the prior. Bottom panel uses 3 observations. It is of note that for events with fewer observations to update the prior, as is the case of the Bottom event, the credible interval is bigger, to represent a higher uncertainty in the prediction. 90% credible interval in both cases.

and was not used to derive the hyperparameters. The MSE for the proposed Bayesian PP is 0.06282 days, which corresponds to 1h30m. Comparing the Bayesian PP with the classical estimation of the parameter of the PP, we note that we expect better accuracy, as the number of CDMs for each event is small, and thus, the classical approach is more sensitive to extreme values. As can be seen in Figure 3, the proposed model outperforms the classic counterpart and the baseline. The baseline is the assumption that the previous inter-CDM time is the best estimator of the next inter-CDM time.

In order to confirm the good performance of our approach, we compare the estimated probability of receiving a CDM in a decision interval with the experimental probability. To obtain the experimental probability, the events are grouped by the theoretical probability intervals presented in Table 2. Then, analysing the ratio of events that actually received a CDM in that time interval, the probability is obtained. It is important to note that for each group, the probabilities are not uniformly distributed, however, we can still make some observations and the results are indeed promising. We note that there is a small positive drift between the theoretical probability and the experimental one, meaning that the given probability according to equation 1, is conservative when compared to reality. We also note that, as it should be, a lower calculated probability will indeed represent a lower experimental probability, because along the groups, the experimental probability increases.

In Figure 4 we can see two examples of the temporal prediction for two events in our test data, with 90% credible intervals. This figure shows that when the number of observations used for the Bayesian estimation is not loo low (15, in top panel), the observed and estimated times are closed, and more important from a statistical point of view, the observed value is contained within the credibility intervals. In the lower panel case, the credibility intervals have a large range (spanning over almost one day), and the observations and predictions are far apart.

5 Conclusion and Future Work

In conclusion, we found that this real-life problem in space situational awareness can be successfully modeled by a stochastic process, and that by using a Bayesian scheme, we can overcome data scarcity. By modeling this problem as a PP, we can estimate the arrival of CDMs during the decision period, which can be used in practice to aid with expert decision process, helping the operator to confidently delay the manoeuvre decision until new information is received. Our results have shown that our dataset is well modeled by this process. We have used our model to predict the time of the next CDM with accuracy exceeding both the baseline and a simpler predictor using the population average.

The theoretical probability is conservative when compared to the reality, and this might mean that the initial hypothesis of homogeneity might not be completely matched to the data. While the deviation is small, it is consistent, and it might indicate that in the last two days, there is a bias to predict a smaller number of CDMs. However, for the practical case of a high-stakes application like space awareness, it is always safer to be under-confident than over-confident.

Acknowledgments

Theoretical Prob.	Experimental Prob.	Deviation
(0.704, 0.753]	0.7647	0.0117
(0.753, 0.803]	0.83992	0.03692
(0.803, 0.852]	0.90365	0.05165
(0.852, 0.901]	0.9401	0.0391
(0.901, 0.951]	0.96740	0.0164
(0.951, 1.0]	1.0	0

Table 2: Grouping of events by calculated probability, and the experimental results for each group. The Theoretical probability is not over-confident when determining if a CDM is going to be issued during the decision interval.

References

[Acciarini *et al.*, 2021] Giacomo Acciarini, Francesco Pinto, Francesca Letizia, José A Martinez-Heras, Klaus Merz, Christopher Bridges, and Atılım Güneş Baydin. Kessler: A machine learning library for spacecraft collision avoidance. In *8th European Conference on Space Debris*, pages 1–9, 2021.

[Bishop, 2006] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[Blei *et al.*, 2003] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[ESA, 2021] ESA. Space debris by the numbers. https://www.esa.int/Safety_Security/Space_Debris/ Space_debris_by_the_numbers, 2021.

[Horstmann and Stoll, 2017] A Horstmann and E Stoll. Investigation of propagation accuracy effects within the modeling of space debris. In *7th European Conference on Space Debris*, 2017.

[Insua and Ruggeri, 2012] D. Rios Insua and Fabrizio Ruggeri. *Robust Bayesian Analysis*. Springer Verlag, New York, 2012.

[Kingma *et al.*, 2019] Diederik P Kingma, Max Welling, et al. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019.

[Kobyzev *et al.*, 2020] Ivan Kobyzev, Simon Prince, and Marcus Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[Koller and Friedman, 2009] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[Krag *et al.*, 2017] H Krag, M Serrano, V Braun, P Kuchynka, M Catania, Jan Siminski, M Schimmerohn, X Marc, D Kuijper, Ian Shurmer, et al. A 1 cm space debris impact onto the sentinel-1a solar array. *Acta Astronautica*, 137:434–443, 2017.

[Kulkarni, 2016] Vidyadhar G Kulkarni. *Modeling and analysis of stochastic systems*. Crc Press, 2016.

[Le May *et al.*, 2018] S Le May, S Gehly, BA Carter, and S Flegel. Space debris collision probability analysis for proposed global broadband constellations. *Acta Astronautica*, 151:445–455, 2018.

[Mashiku and Hejduk, 2019] Alinda K Mashiku and Matthew D Hejduk. Recommended methods for setting mission conjunction analysis hard body radii. 2019.

[Murdoch *et al.*, 2019] W James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080, 2019.

[Pinto *et al.*, 2020] Francesco Pinto, Giacomo Acciarini, Sascha Metz, Sarah Boufelja, Sylvester Kaczmarek, Klaus Merz, José A. Martinez-Heras, Francesca Letizia, Christopher Bridges, and Atılım Güneş Baydin. Towards automated satellite conjunction management with bayesian deep learning. In *AI for Earth Sciences Workshop at NeurIPS 2020, Vancouver, Canada*, 2020.

[Radtke *et al.*, 2017] Jonas Radtke, Christopher Kebschull, and Enrico Stoll. Interactions of the space debris environment with mega constellations—using the example of the oneweb constellation. *Acta Astronautica*, 131:55–68, 2017.

[Ross, 2014] S.M. Ross. *Introduction to Probability Models*. Elsevier Science, 2014.