Bayesian Methods for Longitudinal Trending in Probabilities of Collision

ai-solutions

J.J. Vallejo[*], M.D. Hejduk[†], and J.D. Stamey[‡]

July 4, 2016

[*]a.i. solutions Inc.,10001 Derekwood Lane, Lanham, MD 20706
[†]Astrorum Consulting LLC, 10006 Willow Bend Drive, Woodway, TX 76712
[‡]Department of Statistics, Baylor University, P.O. Box 97140, Waco, TX 76798

Contents

1 Introduction

Satellites have become an integral part of modern life, supporting phone communication, television and radio broadcasting, internet access, and military activities. Indeed, it is difficult to imagine modern society without many of these technologies, especially in an age when the world is increasingly interconnected via long-distance communications. As of 2013, there were over one thousand operational satellites in orbit about Earth. About half of these active satellites are in Low-Earth Orbit (LEO, meaning an orbital period less than 225 minutes), which is where the International Space Station (ISS) conducts operations, along with other commercial missions such as earth observation and satellite telephone communications. An increasing amount of attention is being placed on protecting satellites in LEO, as the frequency of object launches and satellite fragmentation events has contributed to the proliferation of space debris, resulting in increased congestion. Kelly[32] notes that "the number of space objects has greatly increased in the past 15 years and is currently estimated to be 500,000 objects between 1 and 10 cm and 100 million objects less than 1 cm." She goes on to observe that these objects have three sources: 1) debris from satellites, 2) non-operational or "dead" satellites, and 3) operational satellites that may or may not be able to maneuver. In addition to these smaller objects, other sources suggest that the number of objects greater than 10 cm is roughly 20,000.

As a result of the growing amount of debris in the commonly used orbits, there has been an increased focus on protecting satellites from potential collisions with other objects. For example, NASA produced the first orbit debris mitigation guidelines in 1995 and promulgated a formal conjunction assessment policy in 2007. Though space objects had been cataloged since the late 1950's, these were the first procedural attempts to give guidance on how to manage close approaches between two objects and ultimately how to avoid collisions. In 2005, NASA established agency-wide protocol for performing collision analysis and reactions to close approaches. A project office called Conjunction Assessment Risk Analysis (CARA) was created to perform these analyses for robotic missions, and it currently provides this service to about 65 NASA and civil space satellites.

Although the probability of two space objects colliding is often negligible, collisions do occur; and their impact on future space congestion is often tremendous. Since 1991, eight on-orbit collisions have been reported, the last occurring in 2009 when an Iridium communications satellite was hit by an inactive Russian COSMOS satellite. This collision created two debris clouds of approximately 500 and 1,300 objects that have been subsequently cataloged. In addition to this debris, in 2007 the Fengyun 1C satellite was deliberately destroyed, creating another debris cloud of approximately three thousand cataloged objects. Though only the Iridium reflects the case of a collision that was avoidable by collision mitigation procedures, these two events exemplify how much impact collisions can have on the space debris population. For instance, 40% of detected close approaches in a typical 700km orbit used by NASA involve a debris object from one of these three clouds. In effect, these two collisions doubled the number of close approaches tracked in LEO and consequently increased both the risk of further collisions and the amount of work needed to mitigate this risk.

The Iridium-COSMOS collision was the impetus for a significant increase in breadth and sophistication of conjunction assessment activities, as it represented a worst-case scenario for those attempting to mitigate collisions and the proliferation of space debris, as both satellites were completely intact and collided at hypervelocity ($>$ 6,700 mph). In fact, these two satellites collided at the speed of 26,170 mph, which resulted in the creation of a huge

debris cloud of the number of pieces outlined previously, with perhaps one hundred times that number of pieces too small to be tracked by current radars. The proliferation of additional conjunction events that this debris field generates will only increase the requirements for accurate and meaningful conjunction risk assessment.

For some years, conjunction risk assessment was based only on the closest predicted miss distance between the two conjuncting objects. While this construct has immediate intuitive appeal and is easy to communicate conceptually to decision-makers, because it does not consider the uncertainties of the satellite trajectories, it tends to produce results that are difficult to interpret. Consider Figure 1.0.1, which for 14 reports leading up to the Iridium 33 collision, graphs the closest predicted approach of all tracked space objects, the closest predicted approach for any satellite within the Iridium constellation, as well as for the Iridium 33. The black line indicates the closest predicted approach between the Iridium 33 and the Cosmos 2251, which can be seen to be predicted farther than the closest object. This highlights a major difficulty in assessing risk by using miss distance: often there are multiple serious threats, and deciding which is most imminent is not straightforward, especially when based on a metric that is not actually a measure of conjunction likelihood.

Figure 1.0.2 gives further evidence of the poverty of this particular risk assessment paradigm. This figure shows what rank of risk the Iridium 33/Cosmos 2251 conjunction was of all conjunctions, those for just the Iridium constellation, and for the Iridium 33. Over the 14 reports, the Iridium 33/Cosmos 2251 conjunction varies from a rank of 1,611 of all potential conjunctions on report 3 to a rank of 11 on report 4. Thus, when compared to all possible events, the Iridium 33/Cosmos 2251 varies quite a bit on how comparatively “risky" it is. Even if one were to use its risk rating from the later reports, one would still conclude that it is less serious than 150 or 400 events.

The level of threat is still ambiguous when one considers the comparative risk of the Iridium 33 to only satellites within the Iridium constellation. If one were to use the first four reports, one would conclude that the Iridium 33/Cosmos 2251 conjunction is less serious than about 150 other conjunctions. Even if one were to use later reports, this conjunction is never the most serious of all the satellites in the constellation. Thus, from the available data, it is hard to pinpoint that this is the event which needs most attention.

The situation can be ameliorated with a different type of calculation that considers the uncertainty in the state estimates, which will allow the significance of the miss distance to be assessed. If the uncertainties about both states are much smaller than the miss distance between the two objects, then the conjunction is not particularly worrisome even if the miss distance seems small in an absolute sense—the two objects' states are so well determined that one can have confidence that the two objects really will pass each other with the calculated distance. Conversely, a larger miss distance with uncertainties about the same size as this miss distance could well have a probability of collision large enough to be of concern even though the miss distance might itself seem large. This calculation, called the Probability of Collision, will be discussed in depth in a subsequent section.

The result of these debris-producing events is the potential to beget more debris, a phenomenon known as Kessler syndrome. In 1978, Kessler[34] posited that, due to the increase in objects in space, satellite collisions would be inevitable and create further debris, in turn increasing the risk of future collisions. Kessler predicts that “the result would be an exponential increase in the number of objects with time, creating a belt of debris around the earth." He likened the process for creating this “debris belt" to the creation of the asteroid belt, though at an admittedly faster rate. Interestingly, Kessler predicted that the first satellite

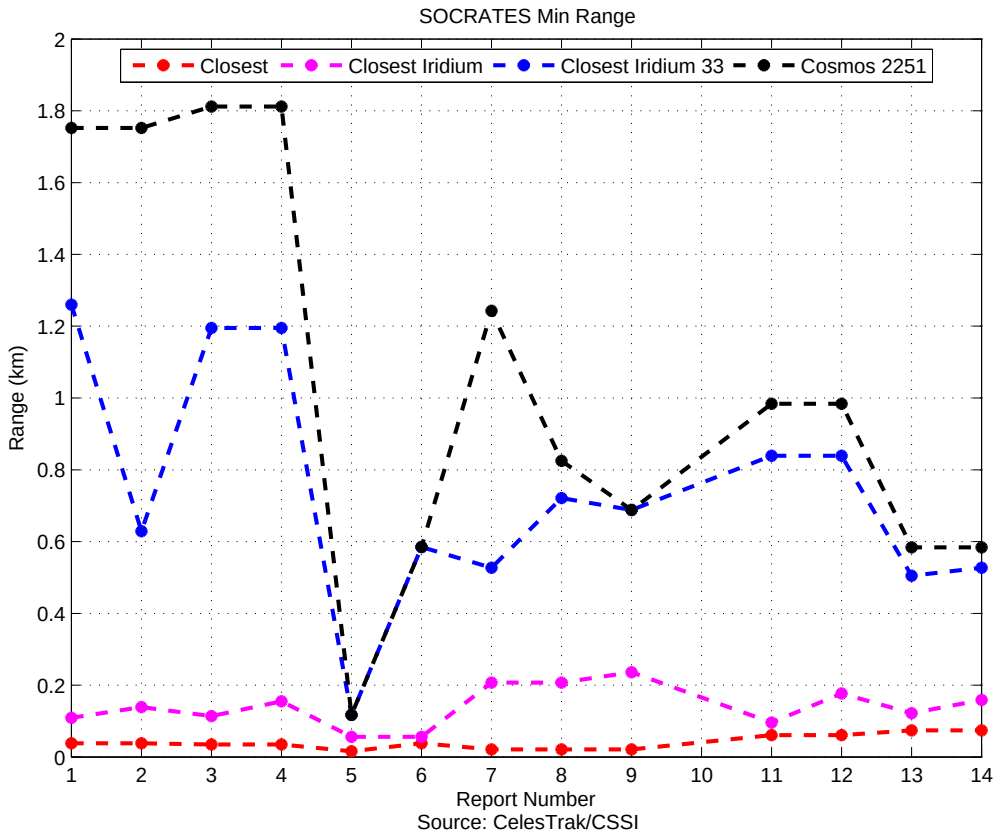

Figure 1.0.1: Closest approach distance distance for all tracked objects (from CelesTrak[1])

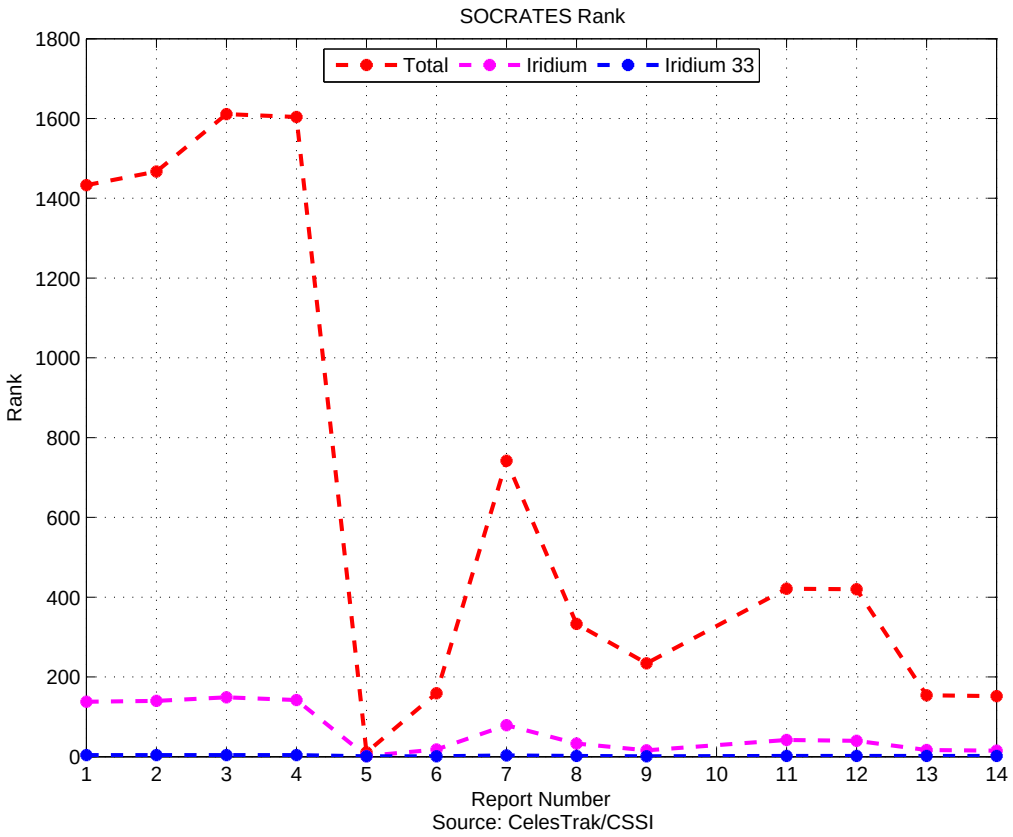

Figure 1.0.2: Closest approach distance for all objects in the Iridium constellation (from CelesTrak[1])

collision could be expected to occur in around 1989 (with a more conservative estimate placing the year at 1997), based on his estimate of an increase of 13%/year growth rate of debris. The first documented satellite collision occurred in 1991 between the Cosmos 1934 and debris from the Cosmos 296. The domino effect described by the Kessler syndrome is most likely in LEO, as this orbit regime contains by far the most space debris. Primack[47] notes that this would not only endanger the International Space Station and Hubble telescope, but also eventually GPS and other communications satellites.

The discussion above outlines the significant risks and ramifications associated with satellite collisions and makes it clear that there is a need for a systematic, sophisticated way of assessing and mitigating these risks. However, there are technical and logistical difficulties in implementing an effective system. The discussion surrounding the Iridium 33 collision gives a broad sense of both, suggesting the technical difficulty in calculating a reliable assessment of risk and the logistical difficulty associated with choosing among hundreds of similarly risky events. We consider both types of difficulty further in order to give a perspective on limitations that add ambiguity to the process and argue for more sophisticated assessment techniques. The subsequent discussion is arranged in sections that address the definition of key terms, the data collection and distribution process, the sources of these data via the orbit determination process, and the calculation of the key parameter presently used in conjunction risk assessment, the probability of collision.

1.1 Definition of Terms

The ensuing discussion will benefit from defining certain key terms more precisely, so we provide such definitions here. The process of describing a satellite's position using various model parameters is called orbit determination (OD); it is a filter estimation process that combines actual sensor observations of satellite positions and other *apriori* information to generate a satellite state estimate (estimate of satellite position and velocity) at a given time, called an epoch time; a robust estimation process will also produce an estimation error covariance matrix, which will specify the expected uncertainties in the estimated parameters and the correlations among them. This information can be used by a satellite propagator to predict satellite positions and velocities at a future time.

When two objects are expected to pass within close proximity of each other, they are said to be in conjunction. A potential collision between two space objects is interchangeably referred to as an event or a conjunction. The primary space object is defined to be the satellite one is attempting to protect (generally speaking, an active satellite). The secondary is the object that is endangering the primary, and this object is usually a piece of debris, although as seen in the Iridium 33 collision, it can be an intact satellite. The time of closest approach (TCA) is the time at which these two objects are predicted to be closest, and the position of closest approach (PCA) is the corresponding position of each object at that time. Conjunction Assessment is the process of determining which spacecraft will be in conjunction with a protected asset over a time period of interest, and Conjunction Risk Analysis is the process of determining the level of collision risk that each of these conjunctions presents. The probability of collision (P_c) is an empirically calculated probability of the two objects colliding, based on a few simplifying assumptions; and it serves as the principal parameter for assessing collision risk. An entire section is dedicated to explaining the calculation of this parameter.

1.2 Conjunction Data Distribution

Because a complete, up-to-date catalogue of the positions and velocities of all known satellites is needed for Conjunction Assessment, this portion of the daily calculation process takes place at the Joint Space Operations Center (JSpOC) at Vandenberg AFB, CA, where the Space Catalogue is actively maintained. Screening runs are executed in which each protected asset, with a volume about it of carefully chosen dimensions, is "flown" several days into the future; and any other catalogued objects that penetrate this volume are identified as conjunctors.

Once a screening run is complete, the results are further processed to generate the orbital information needed to perform Risk Analysis. The precise TCA is determined, and the two satellites' PCAs are calculated. In addition to these position and velocity data that constitute the PCAs, the state error estimates at TCA, represented as covariance matrices, are also provided; these give a statement of the expected variance and covariance of each of the position and velocity components (the estimation process is presumed to be unbiased and therefore produce mean errors of zero), as well as the additional solved-for parameters of atmospheric drag and solar radiation pressure. Finally, information about the force model settings used in the OD is also provided so that, should the user of the data wish to propagate the solution, this can be done with the same model settings enabled. These data are collected into what is called a Conjunction Data Message (CDM) and distributed as a discrete message to the owner or protector of the primary asset.

1.3 Conjunction Data Quality

Before considering methods for quantifying risk, we must first consider limitations that may be imposed due to the quality of the data used for this purpose; and such an investigation has two parts: the quality of the sensor observations that feed the OD process and the quality of the OD modeling itself. We will treat each of these in turn, beginning with the issue of sensor data quality.

Observation data are collected by a variety of space sensors that constitute the Space Surveillance Network (SSN). These include dish and phased-array radars, which provide range-to-target and two angles from the sensor to the spacecraft; optical telescopes, which observe two angles but cannot observe range-to-target; and occasionally other sensor types, such as interferometers or radio-frequency trackers, which typically provide angular data. Radar data are typically reasonably accurate in their range determination but not nearly as reliable in the angular measurements; angle measurements by optical sensors are frequently quite accurate (since they are calculated with reference to the star background, which is accurately known), but there is no range measurement provided. Because it is difficult to track LEO satellites with telescopes and non-LEO satellites with most radars, it is unusual for a satellite to receive both types of tracking and allow the strengths of both sensor phenomenologies to complement each other. So there are errors in observational data due to inherent weaknesses in the different sensor types. Additionally, observations are typically taken in "tracks," or groups of observations all obtained during the same observing session; one might receive a set of, say, six observations in the span of one minute. While groups of data are certainly welcome, if correlation exists among the observations then the basic premise of most estimation techniques—that measurements are uncorrelated—will not be strictly met.

A second issue with tracking data is irregularity of supply. The sensors in the SSN each have different detection and tracking capabilities, meaning that while all sensors can,

for their orbit regime of specialty, track large objects, only a narrow subset can track the smallest objects. Because much of the Space Catalogue consists of debris objects and most debris objects are small, only a few of the SSN sensors are responsible for tracking a good bit of the Space Catalogue; and there is contention for the tracking resources of these sensors. The JSpOC possesses a software-managed sensor tasking paradigm that assigns the tracking of certain objects to certain sensors; but if only a few sensors are responsible for most of the catalogue maintenance, even with the priority scheme that this sensor tasking functionality allows, many debris objects receive far less tracking than one would wish. Furthermore, sensor outages, space weather phenomena, and radar energy misapplications can all conspire to encumber tracking throughput yet further. Lower tracking levels leave the OD process more vulnerable to sensor observation errors and provide a weaker fit overall.

Once tracking data are obtained, they are subjected to a batch minimum-variance estimation process to generate an updated set of orbital parameters for the satellite. The process begins by the appropriate choice of force model parameters, which includes the selection of the proper fidelity of a geopotential model (number of spherical harmonics to solve for in the Laplace equation series expansion that models the irregular Earth's gravity field), the effects of non-Earth gravity (such as the sun and moon), the geopotential irregularity introduced by liquid and solid Earth tides, and parameters that govern the solutions for atmospheric drag and solar radiation pressure. Next, a proper fit-span of observations needs to be chosen, as the batch technique does not correct for each observation sequentially but considers the entire dataset as a "batch." Choosing a group of sensor observations that goes back too far in time (too long a fit-span) tends to weaken the solution for prediction into the future; choosing a group that does not go far back enough (too short a fit-span) tends to produce a poor atmospheric drag solution. Finally, the observation set must be reviewed for "outlier" data that can corrupt the estimated solution. While such an enterprise should of course be conducted with care since there is no *apriori* reason to suspect any particular observation, given the volume of objects and observations most such exclusions must be performed by computer, which is a much less robust process than a trained analyst's data exclusion through the visual review of residual plots. All of these areas are ripe for error that can weaken the correction and therefore the generated state estimate.

Finally, in order to produce data that can be used for conjunction risk analysis, the epoch states of the primary and secondary objects must be propagated to TCA. The principal source of error in this propagation is the inability to model the atmospheric density accuracies over the propagation interval, as the atmospheric drag acceleration on the satellite depends on the local atmospheric density. The atmospheric density is difficult to estimate because it requires estimating the atmospheric temperature, and a variety of factors influence this temperature. Since the temperature is generally governed by the extreme ultraviolet (EUV) heating of atmospheric gases by the sun, temperature is a function of time of day and latitude, as well as the sun's 27-day rotation cycle and 11-year cycle of activity. Acutely, the temperature is also affected by solar ejecta that enter the earth's atmosphere through the polar cusp and heat gases through the manipulation of the earth's magnetic field; this is a product of solar storm activity and is extremely difficult to predict. There can thus be considerable error associated with producing satellite future predicted positions; and given that the model cannot represent these processes well, it is unlikely that the covariance matrix emerging from the fit will model the error robustly. It is often necessary, therefore, to add a consider parameter to the drag variance in the covariance matrix in order to try to represent the atmospheric density error more completely; while this approach is certainly welcome as

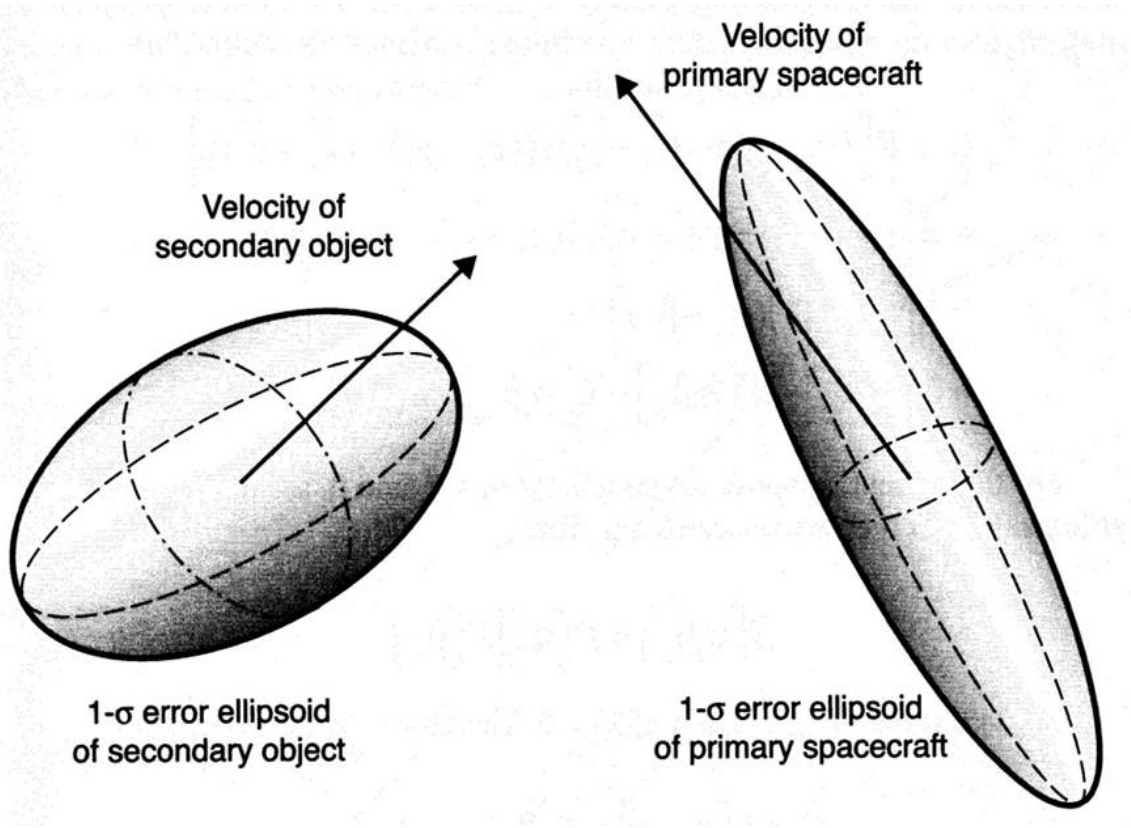

Figure 1.4.1: Visualization of a conjunction (from Chan[17])

an improvement over using the unaltered covariance directly, it is an imperfect compensation method.

For all of these reasons, the information contained in a CDM to describe a conjunction is uncertain and subject to change with future tracking and OD updates; this is the reason that a single estimate of the situation taken several days from the expected event is not adequate for Risk Analysis and that the more elaborate trending approaches explored by this research are warranted. Before turning directly to these methods, however, it is necessary to explain the calculation of the probability of collision (P_c), as it is the parameter used by the industry as the single encapsulation of collision risk. Having just discussed the potential issues with the data, one can observe how these issues work their way through the calculation.

1.4 Calculation of Probability of Collision

As mentioned above, provided in the CDM is a parameter called the probability of collision (P_c), which is generally considered to be the best possible measure for quantifying the risk for an event. Here, we briefly outline the methods used for calculating this value. As noted previously, each conjuncting satellite has an estimated position at TCA, about which a 3-dimensional error covariance is estimated. This covariance is ellipsoidal, and, for near-Earth orbits, usually oriented in such a way that the semi-major axis is close to the direction of the velocity of the object.

An image depicting these assumptions is given in Figure 1.4.1. This ellipse is usually defined in terms of radial, in-track, and cross-track (RIC) coordinates, which is a satellite-centered coordinate system. The radial direction is the direction of the position vector

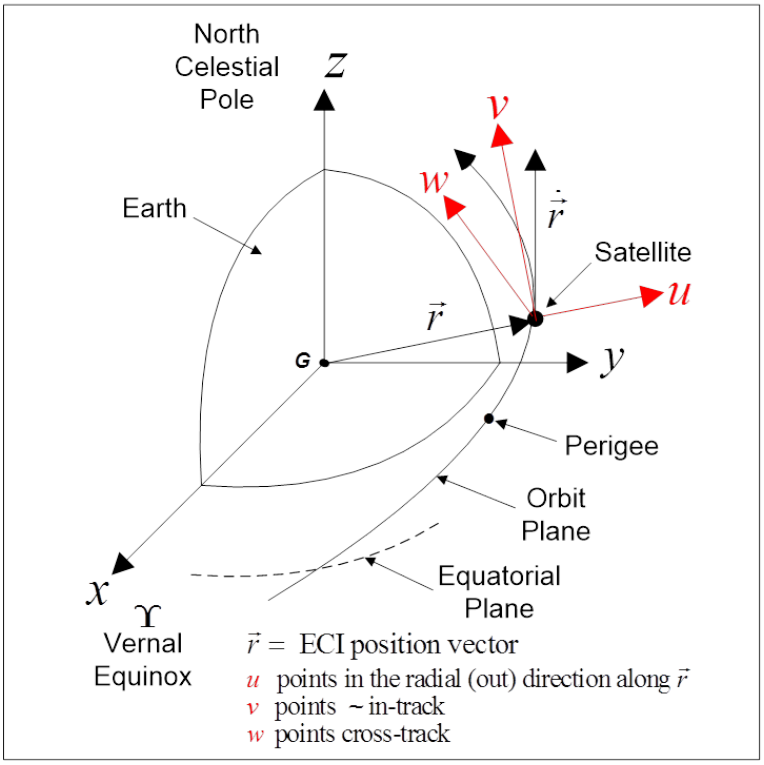

Figure 1.4.2: Visualization of trajectory using UVW coordinates (from Barker[5])

emanating from the earth, the in-track direction is the direction along the trajectory of the object, and the cross-track direction is perpendicular to these two vectors (using the right-hand rule).

These vectors are sometimes given the alternative designation UVW, as seen in figure 1.4.2. It is usually the case that the covariance ellipsoid is longest in the in-track direction, so that it is most difficult to be accurate about where on its trajectory a satellite is when TCA occurs.

This phenomenon is depicted in 1.4.3, where we see the ellipsoid longest in the “V" (in-track) direction. In practice, we further assume that the ellipsoidal errors associated with positional uncertainty are trivariate Gaussian. This implies that the mean of the distribution of each object is taken to be the calculated position at TCA. These covariances are presumed to be uncorrelated, implying that the total positional uncertainty can be calculated simply by the sum of the two covariances (after having been rotated to be expressed in the same coordinate system). Traditionally, we take this combined covariance and center it about the secondary object. Likewise, the radii of circumscribing spheres about each object are summed to create a single combined hard body sphere, which is placed at the location of the primary object. This problem is equivalent to the original problem involving two separate Gaussian densities due to the assumption that the covariances are uncorrelated.

The result of these assumptions is visualized in Figure 1.4.4, though in two dimensions as opposed to three.

One last assumption generally made is that of rectilinear motion near the time of conjunction, so that the dimensionality of the problem may be reduced. If the conjunction between the two satellites takes place at high velocity, then the relative motion in the neighborhood of the conjunction will be rectilinear; and a collision, should it take place, will occur in a plane

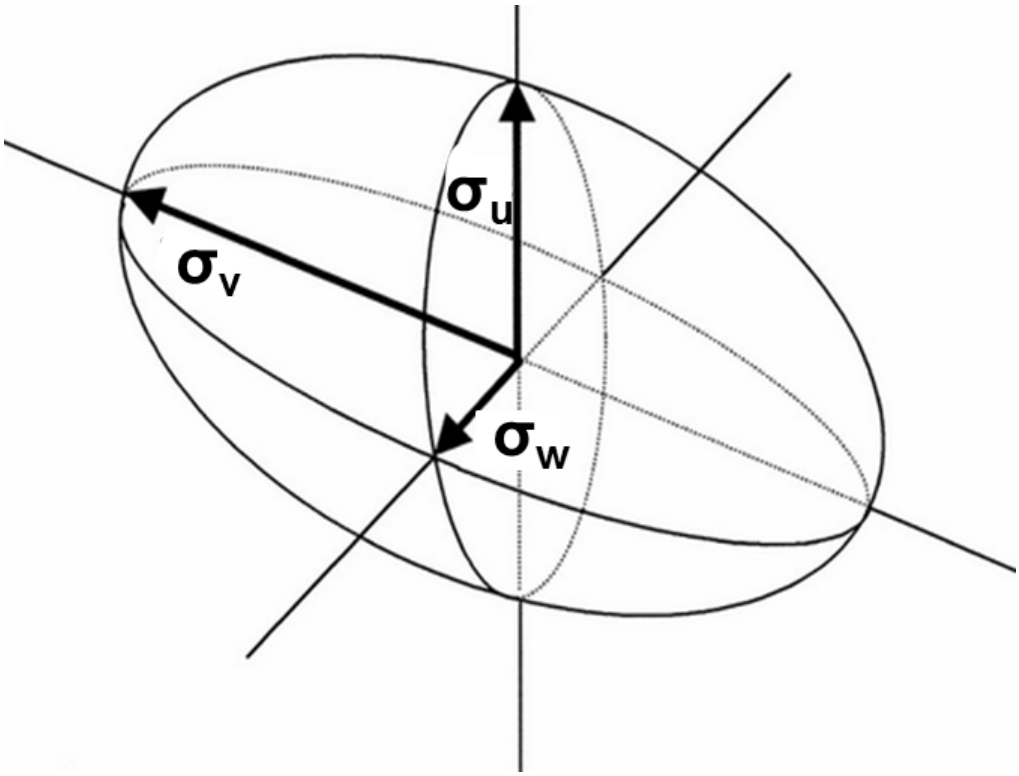

Figure 1.4.3: Positional error covariance ellipsoid defined by UVW coordinates (from Chan[17])

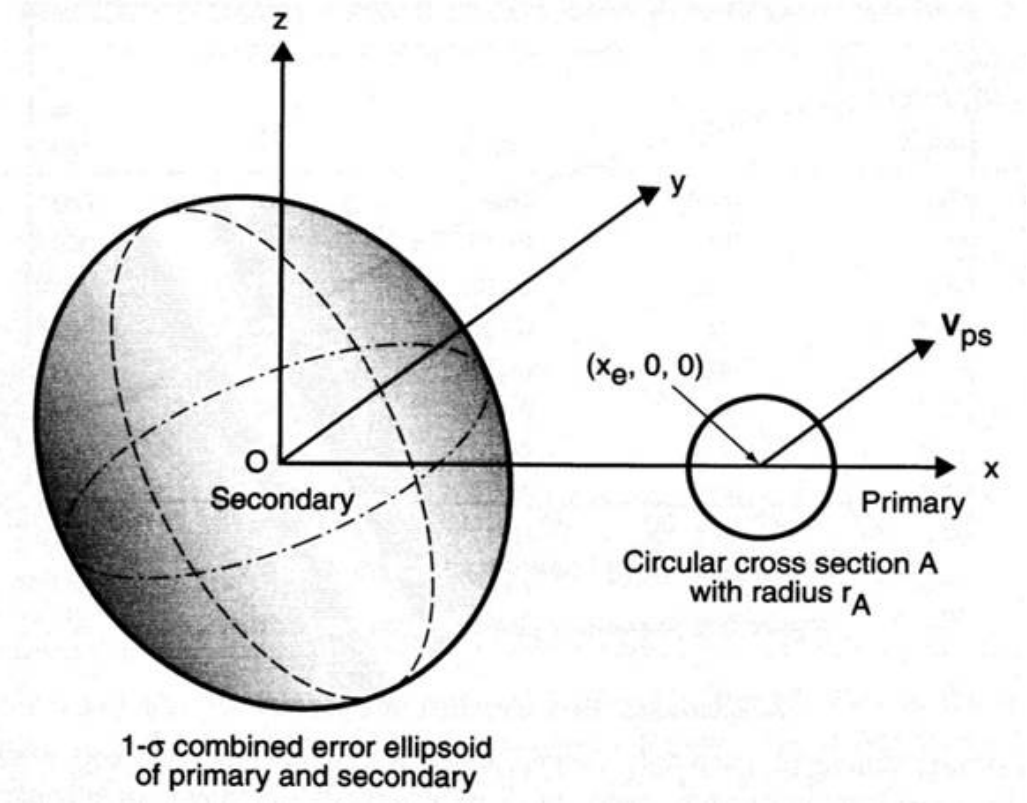

Figure 1.4.4: Projection of conjunction into the conjunction plane (from Chan[17])

normal to the relative velocity vector between the two objects. We can then project the combined covariance and the hard body sphere into this plane and consider the situation as a two-dimensional problem: a circle, resulting from the projection of the hard body sphere, and a covariance ellipse, resulting from the projected combined covariance. We are then interested in the probability of the area swept out by the circle in the probability density formed by the ellipse.[17]

The process for calculating P_c outline above is generally used for events which are approached at a high velocity. If the velocity is sufficiently small (< 10 m/s), many of the assumptions above break down (such as rectilinear motion). In such a case, we use a Monte Carlo approach, simulating millions of trajectories for both space objects and counting the number of times the miss distance is below a pre-specified threshold. This is obviously more time consuming than the above approach and therefore is employed only when necessary.

2 Problem Specification

As made clear by the preceding section the problem of deciding whether to maneuver a satellite which is in conjunction with another space object is often not straightforward, and a serious collision threat often involves the deliberation and cooperation of various parties[25]. Quantifying the risk for any such conjunction is typically accomplished through the use of the calculated probability of collision P_c at TCA. This calculation is generally performed with each received CDM, and such messages typically are received throughout the seven days leading up to TCA. The calculated P_c value is affected by the uncertainty in the positions of the space objects, an uncertainty that generally decreases as one approaches TCA. This decrease in uncertainty typically yields a particular kind of behavior in P_c values, which we shall refer to as the "canonical behavior". We seek to incorporate the shape of this canonical behavior into our understanding of the P_c values, with the goal of making predictions about future P_c values, as well as making inferences about the location of the highest P_c value.

There has been considerable work in calculating the probability of collision, see for example Akella (2000)[2], Patera (2000)[43], Chan (2003)[18], or more recently Xu (2013)[62]. Though methods for improving the P_c are relatively well developed, far less work has focused on detecting trends in repeated measurements of the P_c. Notably, Carpenter and Markley have proposed various implementations of Wald's Sequential Probability Ratio Test (WSPRT) in deciding whether to accept the hypothesis that a new measurement on the P_c is identical in information content to the previous measurement[12][11][13]. Among the advantages of this method are its simplicity and its inherent modeling of false alarms and missed detections. While a considerable advance in P_c predictive methods, this approach is not without limitations. For instance, although the WSPRT tests consecutive measurements, it has no way of directly incorporating the times at which the measurements were taken; it considers measurement time only indirectly through the accumulation of data in forming the total information matrices from which it works. In general, P_c measurements are not taken at equidistant time intervals, suggesting a potential loss of information in the WSPRT approach.

In this manuscript, we propose a simple method to detect the trend in repeatedly-measured P_c values. Our approach has the advantage of directly incorporating the time between observations, which is allowed to be irregular. Additionally, we use the Bayesian paradigm in order to incorporate prior information gathered from past conjunctions. More

sophisticated methods are certainly possible, but we wished to determine how much predictive power could be rendered by a simple and straightforward foundational approach.

2.1 "Dilution" of the P_c

It is clear that the probability of collision depends heavily on the size and shape of the combined covariance ellipsoid. Alfano[3] investigated this relationship for various miss distances, hard body volumes, covariance sizes and shapes. He reported that for a given miss distance, hard body volume, and covariance shape, there is a covariance size which maximizes the probability of collision, with the probability decreasing slowly if uncertainty is increased (that is, the size of the objects' covariances are increased) and decreasing very rapidly if this uncertainty is decreased. We seek to incorporate this known behavior into a statistical model in order to better calibrate each measured probability of collision. In practice, one typically observes a decrease in the size of the covariance as the event moves closer to TCA, producing what we will refer to as a "canonical behavior". We aim to try to recover this behavior beneath all the other "noise" of the problem and ultimately identify the point of maximum probability of collision, in order to make better judgments regarding the degree of continued monitoring that the conjunction merits.

2.2 Canonical Behavior

As noted above, changes in P_c generally follow a canonical behavior with respect to a decreasing state estimate uncertainty; and the the parameter used to illustrate this phenomenon is the ratio of covariance radius to miss distance (for the present we have used a spherical covariance for convenience, but this ratio can be generalized as the Mahalanobis distance and applied to the general case). Figure 2.2.1 depicts what we have called the "canonical behavior" of an event's P_c: an initial increasing change in order of magnitude in P_c as uncertainty decreases, followed by a subsequent drop off when the uncertainty becomes even smaller. The decrease in probability as uncertainty increases is what Alfano[3] referred to as "dilution in probability" because it was caused not by improvements in knowledge of satellite positions but by a lack of positional knowledge that renders any conclusion of high risk more difficult. Note that P_c values are generally particularly small probabilities, and consequently one is usually concerned with changes in orders of magnitude. That is, one is interested in changes in $\log_{10} P_c$ as opposed to simply changes in the P_c value. In the following development, we let y denote the $\log_{10} P_c$ value.

Though informative, using the ratio of covariance size to miss distance as a predictor variable is difficult in practice. Although the size of the combined covariances tend to shrink over time, the rate is not the same for each event. In some cases, the value of this ratio, which appears monotonic before and after the peak point in the figure above, actually increases and decreases several different times before reaching its final value, making modeling a trend even more difficult. Furthermore, the miss distance calculated on the initial CDM is subject to change on subsequent CDMs, and there is often no obvious trend in these updates. As a result, one never knows what the next ratio value will be, even if one knows at which time a CDM would be received. Thus, to use this ratio as a predictor in a statistical model, one would need to regress the ratio on some quantity one could predict, such as time. This is especially difficult because the relationship between the ratio and the $\log_{10} P_c$ value is different for each event, as is the relationship between the ratio and time. To see the

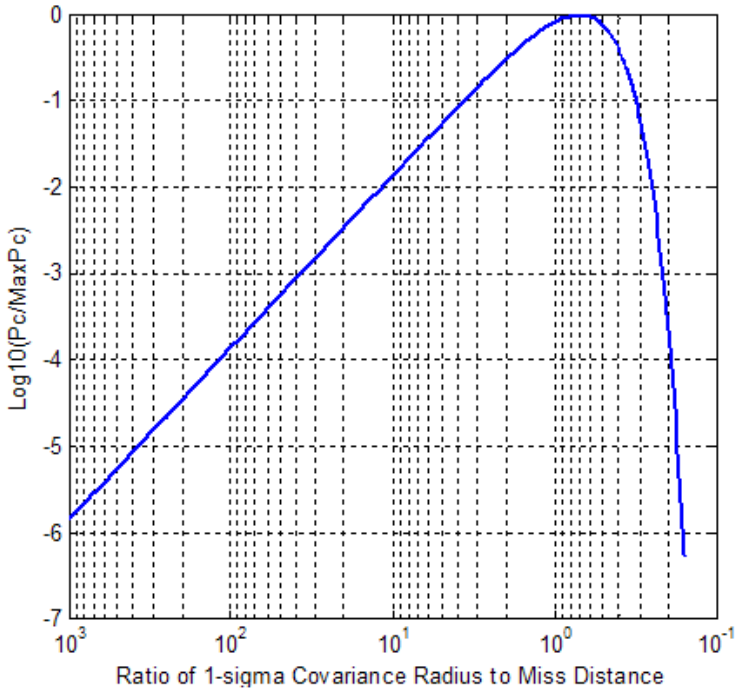

Figure 2.2.1: Theoretical canonical behavior (from Hejduk[29])

difficulty in this kind of modeling, let x_{kt} be the ratio of covariance radius to miss distance for the k^{th} event at time t. We assume that both y_{it} and x_{it} is measured with errors e_{it} and ϵ_{it}, respectively. Then this model is a state-space model[48] and can be written as

$$\begin{aligned} y_{it} &= f(x_{it}) + e_{it} \\ x_{it} &= g(t) + \epsilon_{it} \\ e,\, \epsilon &\sim h(e, \epsilon|\alpha) \end{aligned}$$

where e and ϵ are error terms with a joint distribution $h(\cdot|\alpha)$. It is clear that when attempting to calculate y via the ratio, in practice one needs to specify not only the relationship f between the ratio x_{it} and y_{it} but also specify the relationship g between x_{it} and t.

We leave the exploration of this kind of hierarchical model for later research. In this investigation, we use time as the predictor variable. We assume that the y values still follow a similar canonical behavior with respect to time as they do to with respect to the ratio. Note that this assumes that, as time nears TCA, the ratio decreases. We expect this kind of behavior, as one generally has more accurate information as one approaches TCA. This approach inherently encapsulates the interplay of how the covariances, miss distance, and positional approximations change over time. Because each of these exhibit a high variability across events, we seek to model to overall ensemble progression of these effects over time, as opposed to how each effect impacts the calculated P_c over time.

For simplicity, we attempt to model this behavior with a downward opening parabola, with the aim of correctly predicting the location and, less critically, the magnitude of the peak y value. Though model fit is important, our main goal is to correctly identify the peak y location and, secondarily, value, whether or not the other y values are predicted accurately. For this purpose, the parabola is the simplest curve that can provide a reasonable match to the behavior shown in the previous figure, given the particular attributes of interest here.

2.3 Bayesian Inference

Bayesian inference relies on the posterior distribution of the parameters. To see how the posterior distribution is calculated, let $\boldsymbol{\theta}$ be a vector of unknown parameters, defined on the parameter space Θ. Suppose one has data $\mathbf{y}$, with joint distribution $f(\mathbf{y}|\boldsymbol{\theta})$. Let $\pi(\boldsymbol{\theta})$ be a prior distribution on $\boldsymbol{\theta}$ with CDF $P_{\boldsymbol{\theta}}$. Treated as a function of θ for fixed $\mathbf{y}$, the joint distribution becomes the likelihood, $l(\boldsymbol{\theta}|\mathbf{y})$, defined on Θ. The posterior distribution of $\boldsymbol{\theta}$, given by Bayes' theorem, is

$$\pi(\boldsymbol{\theta}|\mathbf{y}) = \frac{l(\boldsymbol{\theta}|\mathbf{y})\pi(\boldsymbol{\theta})}{\int l(\boldsymbol{\theta}|\mathbf{y})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}.$$

This is the distribution of the parameters after having seen the data vector $\mathbf{y}$. Thus, the prior beliefs about parameters and their distributions are updated after encountering the actual data. The posterior distribution often does not have a closed form and must be approximated using numerical methods such as Markov Chain Monte Carlo (MCMC).

2.3.1 Inference for a new data point

Suppose one has data $\mathbf{y}$ and one wishes to predict $y^* = \log P_c$ at a new data point at time t^*. One can make an inference on y^* by using the predictive distribution

$$g(y^*|x^*, \mathbf{y}) = \int_\Theta g(y^*|\theta, x^*, \mathbf{y})\pi(\theta|\mathbf{y})d\theta,$$

which can be estimated by using the posterior samples from the MCMC draws. In practice, we usually do not know t^* in advance, as it is determined by the exigencies of any particular event. However, for simulation purposes, we use the next time point at which a CDM was received and make a prediction. We construct a 95% credible interval for y^* and check to see if the actual value of y is contained in the interval. The percentage of credible intervals which contain the true y value is known as coverage. If the coverage is close to the nominal value of 95%, we can assume that these predictions are reliable.

3 Methods

3.1 Last Observation Carried Forward

A common technique for imputing missing data in a longitudinal data is Last Observation Carried Forward (LOCF)[46]. In this approach, one replaces any missing value for a given subject with their last observed value. Many have offered criticisms of this approach. For instance, Saha[50] argues that this method induces bias under informative dropout, Kenward[33] lambastes the method, and argues that it is only appropriate under unrealistic special cases. Nevertheless, this method is widely used in missing data, as it is simple and intuitive.

In conjunction risk analysis, LOCF more or less represents the current practice for interpreting P_c values. For instance, the most recent P_c value is generally taken to be the most reliable measurement, and decisions are currently based on these values. In effect, operators "predict" all future P_c values to be the last observed P_c value, though they certainly expect some variability in future values. One of our objectives is to quantify how much variability

operators can expect in future values. In addition, we consider whether the last P_c value really is the "best" prediction, or if other information can be used to improve this prediction. Underlying these objectives are the empirical findings by operators that some events exhibit more variability in P_c than others, and that P_c values seem to follow a general trend over time.

3.2 Look-Up Tables

Look-up tables have been in wide use in statistics since at least 1903, due to Sheppard's definitive tables for the standard normal cdf and pdf[52]. These tables have historically been used to avoid having to repeatedly calculate difficult quantities, such as $\int_0^x \exp(-t^2)dt$, which is involved in the normal cdf and pdf[20]. These look-up tables serve as an important example for our problem, as they are in wide use due to the popularity of the normal distribution as a modeling distribution. In the problem of P_c trending, we consider the distribution $log_{10}P_c$ by time to TCA. Similar to the normal distribution, this distribution is to be referred to repeatedly for inference, so that the idea of a look-up table might be useful in practice.

Creating a look-up table for the distribution $log_{10}P_c$ by time to TCA involves estimating the conditional distribution of $y = log_{10}P_c$ at a given time t to TCA. One of the first important approaches to this problem was given by Stone[57], who suggested neighbor-type estimates. To illustrate ideas, let $(X_1, Y_1), ..., (X_n, Y_n)$ be a random sample from the joint distribution of (X, Y). Stone suggested estimates of $F(y|x)$ of the form

$$\hat{F}(y|x) = n^{-1}\sum_{i=1}^{n} W_i(x)I(Y_i \leq y), \quad -\infty < y < \infty \tag{3.2.1}$$

where $W_i(x) = W_i(x; X_1, ..., X_n)$ weights more heavily Y-values for which X_i is closer to x. Stute[58] developed asymptotic properties for estimators of this type based on kernel weights.

Perrachi[44] notes that, in estimating the distribution of a random variable, one can choose to estimate either the conditional quantile function or conditional distribution function. For simplicity, suppose that Z is a random variable with an absolutely continuous distribution with strictly positive density. Then the cumulative distribution function (CDF) is defined on R by $F(z) = P(Z \leq z)$, while the quantile function is defined on $(0, 1)$ as $Q(u) = \{z \in R : F(z) = u\}$. Thus, Q and F are inverses of each other, so that $Q(F(z)) = z$ and $F(Q(u)) = u$. As a result, either function could be used to estimate the distribution of a random variable.

These estimators can be easily extended to conditional distributions. Suppose instead one observes a random vector (X, Y), where X is k-dimensional and Y is a real-valued continuous random variable with strictly positive density, as before. One may characterize the conditional probability distribution through the conditional distribution function $F(y|x) = P_x(Y \leq y)$, or through the conditional quantile function $Q(u|x) = \{y \in R : F(y|x) = u\}$.

Characterizing a conditional probability distribution through the quantile function has a few notable difficulties. To explore these, note that $Q(u)$ may be characterized as the unique solution to the problem

$$\min_{z \in R} El_u(Z - z) \tag{3.2.2}$$

where l_u denotes the asymmetric loss function

$$l_u(v) = [u - I(v < 0)]v. \tag{3.2.3}$$

To describe a conditional distribution $f(y|x)$, one typically employs *quantile regression*. Quantile regression seeks to estimate the parameters of a function $g(\cdot)$, which is the unique solution to the problem

$$\min_{g \in G} E l_u(Y_i - g(X_i)), \quad 0 < u < 1 \tag{3.2.4}$$

where G is the class of real-valued functions defined on R^k. In practice, $g(\cdot)$ is often taken to be linear, so that $Q(u|x) = x^T\beta(u)$, as described in Koenker and Bassett[37]. Then the estimate of the k-dimensional parameter $\beta(u)$ is any solution to the problem

$$\min_{b \in R^k} n^{-1} \sum_{i=1}^{n} l_u(Y_i - X_i^T b), \quad 0 < u < 1. \tag{3.2.5}$$

The resulting estimate $\hat{\beta}(u)$ can be used to define $\hat{Q}(u|x)$, which in turn can be used to estimate the conditional distribution function

$$\hat{F}(y|x) = \sup \left\{ u \in (0,1) : \hat{Q}(u|x) \leq y \right\}. \tag{3.2.6}$$

Perrachi shows that if the conditional distribution of Y depends on x through both a linear location parameter $\mu(x) = \alpha + x\beta$ and a scale parameter $\sigma(x) > 0$, then the conditional quantiles are no longer linear in x. That is, if the data Y exhibit heterogeneity of variance across varying levels of x, then the quantiles are not necessarily linear. Thus, linear quantile regression may yield poor estimates in this case. Furthermore, this non-constant variance may produce estimates of linear models for conditional quantiles which cross each other, violating a basic assumption about quantiles. Some nonparametric estimators have been proposed, based on kernel or nearest neighbor methods (Antoch and Janssen[4]; Samanta[51]; Truong[59]; Bhattacharya and Gangopadhayay[6]; Chaudhuri[19]), regression splines with a fixed number of knots (Hendricks and Koenker[30]), smoothing splines (Koenker et al[37]) and penalized likelihood.

A difficulty particular to quantile regression methods, mentioned above, is the so-called "no-crossing" condition. This is the condition that, for all values of a covariate x, one should have $\hat{Q}(u_1|x) \leq \hat{Q}(u_2|x)$ when $u_1 < u_2$. That is, the regression line of a lower quantile should not up-cross the regression line of an upper quantile. To see why this must be the case, note that when $u_1 < u_2$, the solutions z_1 and z_2 which satisfy

$$F(z_1) = u_1$$
$$F(z_2) = u_2,$$

must also satisfy

$$z_1 < z_2,$$

since the CDF is a monotonically increasing function. Thus, we must have $Q(u_1) < Q(u_2)$ and $Q(u_1|x) < Q(u_2|x)$ by extension. Thus, we seek estimators of $Q(\cdot|x)$ which adhere to this constraint. Koenker[36] avoided this issue by considering parallel quantile functions.

There have been some solutions proposed for the "no-crossing" condition, notably He[28], Wu and Liu[61], Neocleous and Portnoy[41]. Bondell[8] notes that several authors have proposed to first estimate the conditional cumulative distribution function via local weighting, and then invert it to obtain the quantile curve. He notes that this method is suitable for estimation of the conditional quantile, but that it is not suited for estimation of linear predictor effects. As our concern is only estimation of conditional quantiles, this is not a limitation for our application. Because our data exhibits non-linearity and heterogeneity with respect to the covariates, direct estimation of the quantile function is difficult. We proceed with methods based on the empirical CDF, which circumvents these issues.

3.3 Constrained Bayesian Inference

Gelfand[26] introduced MCMC methods for constrained parameter problems. He notes that if one has the full conditionals for each parameter in the model, producing MCMC draws adhering to the constraint simply involves modifying the full conditional density. For instance, suppose one has data $\mathbf{y}$ which has density $f(\mathbf{y}|\theta)$, where θ is a k-dimensional vector constrained to lie in a subset $S^k_{\mathbf{Y}}$ of R^k. Furthermore, in Bayesian models, we specify a prior distribution for θ, say $p(\theta|\lambda)$. Then, Gelfand shows that the posterior distribution of any element of θ is

$$f(\theta_i|\mathbf{Y}, \lambda, \theta_j, j \neq i) \propto f(\mathbf{Y}|\theta)p(\theta|\lambda), \quad \theta_i \in S^k_i(\theta_j, j \neq i),$$

so that the posterior distribution follows its usual form, only with the specified constraints.

As Gelfand notes, there are a few simple ways to simulate draws from this distribution. One way is to generate the full distribution, not accounting for the constraints, and then to only keep the variates which satisfy the constraints. This can also be accomplished by simulating draws U from a uniform(0,1) distribution and following $\theta_i = F_i^{-1}[F_i(a)+U(F_i(b)-F_i(a))]$ where F_i is the full conditional CDF of θ_i. This produces a draw of θ_i which adheres to the constraints, and is due to Devroye[21].

3.4 Bayesian Beta Regression Models

Generalized linear models were introduced by Nelder[40] in 1972 to solve the problem of regression for responses which have a non-normal distribution. These models often are used for binary and count data, so that the usual models involve the Binomial, Poisson, Negative-Binomial, or Multinomial distributions. Of course, this general class of models includes those which handle continuous data, such as Normally-distributed responses (the usual linear regression case), and Weibull-distributed responses (common in survival analysis). Notice that none of the models handle responses with bounded support. To handle such responses, methods for Beta regression were introduced by Paolino[42], Kieschnick and McCullough[35], and Ferrari and Cribari-Neto[23].

Though Beta regression seems to be the most popular regression method for bounded responses, it should be noted that other methods exist. For instance, one may choose to transform the data using the logit function and use linear regression on the transformed responses. Let $\mathbf{y}$ be a vector of bounded responses such that $y_i \in (a, b)$. Furthermore, let $\mathbf{X}$ be an $n \times p$ design matrix containing the corresponding covariates. Then one may choose to model

$$\log\left(\frac{\boldsymbol{y}}{1-\boldsymbol{y}}\right) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}),$$

so that one may proceed to apply usual linear regression techniques. Define $z_i = \log\left(y_i/(1-y_i)\right)$, so that one may write the response vector as $\boldsymbol{z}$. Suppose y_i is a probability, so that z_i can be interpreted to be the log odds of an event. Thus, this model has the advantage of being able to utilize known techniques in linear regression, but has the disadvantage of forcing one to interpret the results in relation to the response on a different scale.

Other techniques are possible, such as the fractional logistic model, various nonlinear models, and models based on other bounded distributions (e.g. the simplex distribution). Beta regression is attractive because it builds upon the GLM framework and it is based on a familiar distribution. Though these other models are possible, we proceed with Beta regression because it has a recently developed framework for mixed and mixture models, and because it can be fit in a rather straightforward manner with Bayesian techniques. Thus, it is sufficiently flexible and is able to incorporate prior information.

The primary aim of GLMs is to model some transformation of the expected value of the response variable with a linear model. This indirectly specifies a relationship for the mean, and one often treats parameters related to variability, such as the so-called "precision parameter", as a nuisance parameter. This parameter can also be modeled, and such models were explored in the context of Beta regression by Cepeda[16], Cepeda and Gamerman[15], and Simas et. al[53].

The aforementioned developments all extended methods of inference for independently and identically distributed Beta variables. Our data consists of observations which are likely dependent, as they are longitudinal in nature. Furthermore, we are interested in making predictions about individual events, such as the $\log_{10} P_c$ behavior for an event as it nears TCA. When one has longitudinal data and the goal of inference is prediction of individual responses, the most common technique is the mixed model. Mixed models were popularized for longitudinal data by Laird and Ware[38], and a general approach for fitting GLMMs was introduced in Stiratelli, Laird, and Ware[56]. As the Beta GLM was late to be developed, mixed models are fairly new to this setting. For instance, one of the first introductions to Beta mixed models was provided by Verkuilen and Smithson[60]. Bayesian implementations of the Beta mixed model soon followed, described in Figeuroa-Zuniga et. al[24], and Bonat et. al[7].

3.5 Functional Data Analysis

Functional Data Analysis is an extension of longitudinal data analysis, a field which attempts to explain the effect of time on various subjects with multiple measurements. For example, consider a study which follows the blood pressure of a number of individuals over time. Each subject's blood pressure will change depending on each individual's genetic predisposition, thus implying that the analyst must account for effect of the subject as well as time. The case of trending P_c over time is similar in that one observes multiple measurements of P_c for a single event, implying correlation among measurements within the event. Additionally, each event seems to have a slightly different effect on the trend of P_c values over time, reinforcing the need for a longitudinal model. In previous sections, we have introduced parametric models, one of which (downward-opening parabola) is of the form

$$y \sim F(\mu, \phi)$$
$$g^{-1}(\mu) = \boldsymbol{X}\boldsymbol{\beta},$$

where $F()$ is the CDF of a distribution from the exponential family. However, we noted that this parametric model may be too restrictive for the overall trend of P_c values, or a within-event trend in P_c values. Instead, we might consider modeling a trend non-parametrically by specifying a general unknown function $f(x)$:

$$y = f(x) + \epsilon.$$

Specifically, let y_{ij} represent the $\log P_c$ value of the i^{th} event from the j^{th} CDM of that event. Let $X_i(T_{ij})$ be a random function for the i^{th} event at the j^{th} CDM at time T_{ij}. The model we consider is

$$y_{ij} = X_i(T_{ij}) + \epsilon_{ij} = \mu(T_{ij}) + \sum_{k=1}^{\infty} \xi_{ij}\phi_k(T_{ij}) + \epsilon_{ij},$$

where ϕ_k are eigenfunctions and ξ_{ij} are uncorrelated random variables with zero mean and variances λ_k, the eigenvalues corresponding to the eigenfunctions.[63] Utilizing eigenfunctions to express an overall function is often called functional principal component analysis. This method allows us to draw inferences about the overall trend and the within-event trend based on only a few observations. This, in fact, was the reason for our choice in pursuing a functional data approach to the problem. It is well known that longitudinal data analysis models tend to need many observations for each subject in order to estimate their numerous parameters. In contrast, the current method estimates only a few parameters. In practice, the sum of eigenvectors is limited to the first few, those which explain the majority of the variability in the model. The need for only a few parameters and a high level of functional variability makes the functional data analysis ideal for the trending of P_c values.

3.5.1 Principal Component Analysis

The model above is a functional extension of a standard dimension-reduction technique called principal component analysis. Consider a set of n observations of a p-dimensional random variable, i.e. $\mathbf{x_i} = \{x_{i1}, x_{i2}, ..., x_{ip}\}$ for $i = 1, ..., n$. The basic idea behind principal components is to explain the variance-covariance structure for a large number of variables (in this case, the dimensions of $\mathbf{x_i}$) through a few linear combinations of these original variables. This method allows for dimension reduction and for interpretation. Let $\vec{x_i}$ denote the n observations of the i^{th} dimension. Then the idea is to find a new set of vectors $\vec{y_1}, \vec{y_2}, ..., \vec{y_p}$ where

$$\vec{y_i} = \sum_{j=1}^{p} l_i j \vec{x_j},$$

where $var(\vec{y_i}) = \vec{l_i'}\Sigma\vec{l_i}$ and for $cov(\vec{y_i}, \vec{y_k}) = \vec{l_i'}\Sigma\vec{l_k} = 0$ and $var(\vec{y_1}) \geq var(\vec{y_2}) \geq ... \geq var(\vec{y_p})$ for $\vec{l_i'} = (l_{1i}, l_{2i}, ..., l_{pi})$. This problem has the following solution:

1. Suppose that the matrix Σ has associated real eigenvalue-eigenvectors given by $(\lambda_i, \vec{e_i})$ where $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$, then the i^{th} principal component is given by

$$\vec{y_i} = \vec{e_i}^i X = e_{i1}\vec{x_1} + e_{i2}\vec{x_2} + ... + e_{ip}\vec{x_p},$$

and $var(\vec{y_i}) = \lambda_i$ for $i = 1, 2, ..., p$, $cov(\vec{y_i}, \vec{y_k}) = \vec{e_i}'\Sigma\vec{e_k} = 0$ for $i \neq j$. Note, the eigenvalues λ_i are unique, however, the eigenvectors (and hence the vectors $\vec{y_i}$) are not.

2. The total variance for the p dimensions is $tr[\Sigma] = \sum_{i=1}^{p} \lambda_i$. Hence, the proportion of variance explained by the k^{th} principle component is $\lambda_k / \sum_{i=1}^{p} \lambda_i$.

3. If the matrix X is centered and scaled so that Σ is the correlation matrix, then $\sum_{i=1}^{p} \lambda_i = p$.

In practice, we choose a suitable "cutoff" for what percentage of the variance we want the new bases $\vec{y_i}$ to explain (say, 95%). Then we choose to use these $q < p$ new bases to re-express the original data in a smaller dimension. In functional data analysis, we replace the vector observations with functional observations.

3.5.2 Functional Data for Sparse Longitudinal Data

As mentioned earlier, the data we consider is sparse longitudinal data. We have many events, but each event has only a few $\log_{10} P_c$ values. This sparseness presents difficulty in estimating functional data models, as one must estimate the functional principal component scores $\xi_{ik} = \int (X_i(t) - \mu(t))\phi_k(t)dt$, which are usually estimated via numerical integration. When the data is sufficiently sparse, a common estimate for this parameter is $\hat{\xi}_{ik}^S = \sum_{j=1}^{N_i}(Y_{ij} - \hat{\mu}(T_{ij}))\hat{\phi}_k(T_{ij} - T_{i,j-1})$, setting $T_{i0} = 0$. As noted by Yao[63], this estimator will not yield reasonable approximations to ξ_{ik} when the data is sparse.

Yao[63] overcomes this difficulty in estimating ξ_{ik} by assuming that ξ_{ik} and ϵ_{ij} are jointly Gaussian. As a result, the expectation of ξ_{ik} is tractable, and is given by

$$\tilde{\xi}_{ik} = E(\xi_{ik}|\tilde{\boldsymbol{Y}}_i) = \lambda_k \boldsymbol{\phi}_{ik}^T \boldsymbol{\Sigma}_{Y_i}^{-1}(\tilde{\boldsymbol{Y}}_i - \boldsymbol{\mu}_i),$$

where $\boldsymbol{\Sigma}_{Y_i}^{-1} = cov(\tilde{\boldsymbol{Y}}_i, \tilde{\boldsymbol{Y}}_i)$. Yao proceeds with inference using this expected value in lieu of the less reliable estimators based on sums. This procedure, which he calls Principal Component Analysis through Conditional Expectation (PACE), is now a common method for handling sparse functional data.

4 P_c Trending

4.1 Last Observation Carried Forward

We briefly describe our implementation for LOCF. In our simulations, we predict future $\log_{10} P_c$ values with the previously observed value for that event. Thus, for the j^{th} OCM of the i^{th} event, we predict

$$\hat{y}_{i(j+1)} = y_{ij},$$

which ignores all other past values, as well as the time to prediction. We construct prediction intervals for this method using Repeated Cross-Validation, in the same way as discussed below for the Look-Up Method.

4.2 The Look-Up Method

4.2.1 Intuition

Here, we propose a simple method as basis of comparison for our more sophisticated models, which we call the Look-Up Method. The Look-Up Method is based on the common expectation that, when an event is observed with relatively high P_c values, we can suppose this event to behave similarly to other events with other similarly high values. In order to formalize this intuitive approach into an explicit model, we need to establish how high "relatively high" is. A natural way to quantify this notion is in terms of quantiles. That is, we expect events with $\log P_c$ values in the qth quantile to behave similarly to other events with $\log P_c$ values in the qth quantile. The method we describe below is similar to methods involving "look-up tables," where quantiles for various scenarios are used to find the probability of an event within the table.

4.2.2 Method

Let x and y be the time and P_c value from the most recent observation. Furthermore, let x^{new} and y^{new} represent the time of prediction and the true P_c value at this time. The algorithm for the Look-Up Method is as follows

Algorithm 1 Look-Up Method

1: **procedure** LOOK-UP
2: Choose an historical data set $\mathbf{Y}_h$ such that the events contained in $\mathbf{Y}$ are believed to behave similarly to the event of interest.
3: Choose a window w.
4: Calculate the empirical CDF $\hat{F}(y)$ of the $\log P_c$ values in the interval $(x - w, x + w)$.
5: Calculate the sample quantile $\hat{q}$ of y
6: Calculate the empirical CDF $\hat{F}(y)$ of the $\log P_c$ values in the interval $(x^{new} - w, x^{new} + w)$.
7: Predict y^{new} to be $\hat{q}(x^{new})$
8: **end procedure**

We find that $w = 2$ days to be a reasonable window length. This length depends on how much prior data is available, as w may need to be smaller for large datasets, allowing for more precise estimation of the CDF. Note that this method only predicts an estimate of y^{new} and does *not* by default generate a prediction interval or any other confidence information.

The method above is simple: find the sample quantile of the observed P_c value at the given time, and assume that future P_c values will be at the exact same quantile. As the model is simple, it also discards potentially useful information. For instance, the predictions are made based only on the sample quantile of the most recent observation, and makes no use of previous observations other than, of course, the historical P_c behavior information. However, one could argue that the most recent observation is the most (or only) meaningful observation, and thus one should make inferences based on this value rather than more immediate past values.

4.2.3 Prediction Intervals

As noted above, the Look-Up Method does not automatically generate prediction intervals; this is a consequence of the method making no distributional assumptions. However, one may still construct prediction intervals via bootstrapping or cross-validation[55]. Recently, these methods were compared[9], and the results from this comparison indicate that estimators based on Repeated Cross Validation (RCV) tend to outperform other estimators (e.g. bootstrap estimators). As a result, we implement RCV to generate prediction intervals. The method was initially proposed by Burman (1989)[10], which describes the algorithm in detail.

We use RCV to estimate the distribution of prediction errors. This will allow us to construct prediction intervals at any confidence level for the Look-Up method. Our procedure is as follows.

Algorithm 2 RCV algorithm

procedure RCV
  for Each repetition $r \in nRep$ **do**
    for Each fold $i \in nFolds$ **do**
      for Each event $j \in nEvents$ **do**
        for Each OCM $k \in nOcms - 1$ **do**
          Predict y_{k+1} using Look-Up Method
          Estimate prediction error $e_{k+1} = y_{k+1} - \hat{y}_{k+1}$
        end for
        Collect prediction errors across OCMS $pVecj = e_2, ..., e_{nOcms}$
      end for
      Collect prediction errors across events $pStore = (pVec_1, ..., pVec_{nEvents})$
      Calculate estimated percentiles $\hat{p}_i$ using $pStore$ for $i = 1, ..., 99$
    end for
    Calculate mean of estimated percentiles
  end for
  Calculate mean of estimated percentiles
  Return estimated percentiles
end procedure

Notice that for each event, predictions are made for all but the first CDM using the previous CDMs. Recall that CDMs are received at varying intervals, so that the procedure above results in making predictions at varying intervals into the future. Thus, the procedure implicitly assumes that the distribution of prediction errors does not depend on time to prediction. While this assumption is generally untenable, the data is such that the time between consecutive CDMs is generally 2 days or less. As one is generally concerned with making predictions no sooner than one day into the future, these prediction errors are conservative for their operational use, meaning that they predict a more worrisome P_c than the actual value. This is because predictions are generally more variable as time to prediction increases. Thus, this procedure results in prediction intervals which hold reasonably for all prediction times of 2 days or less. The procedure is simple and generally conservative.

4.3 Vertex Model

In order to compute this predicted P_c inverted parabola, we use constrained optimization[39] to enforce a downward-opening behavior. There is also precedent for constrained inference in the Bayesian paradigm, as Gelfand[26] introduced an approach to Gibbs sampling in constrained parameter and truncated data problems. Specifically, Gelfand considers problems with ordered parameters, constrained parameters, and censored data. Considering the general equation for a parabola below, our problem is seen as one involving constrained parameters, as we know that $\beta_2 < 0$ and (as discussed subsequently) $\beta_0 < 0$.

$$y = \beta_0 + \beta_1 t + \beta_2 t^2$$

We also show that this induces a constraint on β_1. Implementing these constraints is another way in which we can "inform" the model. Utilizing these constraints along with an informative prior structure allows us to include a maximal amount of prior information, which we believe to be essential, as many of the events we consider contain only 3 or 4 data points, and we wish a reasonable prediction as early as possible within the event.

To allow our model to incorporate prior information from past events, we use the Bayesian paradigm[27]. Let y_{ij} be the $\log_{10} P_c$ from the j^{th} CDM from the i^{th} event. Similarly, let t_{ij} be the time (in days) until TCA for the j^{th} CDM from the i^{th} event. We assume that the observed P_c values over t follow the relationship

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij}^2 + \epsilon_{ij},$$

where $\epsilon_{ij} \sim N(0, \sigma_{ij}^2)$. Furthermore, we assume this parabola will be downward opening, to attempt to model the expected canonical behavior. Utilizing the Bayesian paradigm will allow us to incorporate information about where the peak y value usually is, and how quickly the y values tend to drop off. This incorporation is accomplished by specifying informative prior distributions for the parameters, which are considered random variables in the Bayesian paradigm. Another consequence of treating parameters as random variables is that one can make probabilistic statements about functions of parameters, a Bayesian feature that is not fully possible with a frequentist approach. Thus, one can make statements about the probability of the peak P_c value occurring at a particular time and magnitude. Finally, using the Bayesian paradigm allows us to make predictions in this four-parameter model even with only two or three observations by utilizing the prior distributions of the parameters to help identify the likely values of the parameters.

4.3.1 Prior Structure

The usual prior structure for regression coefficients in linear regression is an independent normal prior for each regression coefficient[27]. We amend this structure to incorporate the constraints we know to exist in our problem. We know that the parabola must open downwards, so that $\beta_2 < 0$. As a consequence of this constraint and the fact that all y values are less than or equal to 0 by definition (since they represent the base 10 logarithm of values between 0 and 1), we also know that $\beta_0 < 0$.

We show that the parameter β_1 must also be constrained. Because $y_{ij} \leq 0$ for all i, j, it follows that the peak y value should also be less than or equal to zero. It is easy to show that the location of the peak is $h = -\beta_1/2\beta_2$, and that the magnitude of the peak is

$b = \beta_0 - \beta_1^2/4\beta_2$. In order to force the magnitude of the peak b to be less than or equal zero, we must have

$$\begin{aligned} \beta_0 - \beta_1^2/4\beta_2 &\leq 0 \\ 4\beta_2\beta_0 - \beta_1^2 &\geq 0 \\ \beta_1^2 &\leq 4\beta_2\beta_0, \end{aligned}$$

where the second line follows since $\beta_2 < 0$. This implies that $\beta_1 \in [-2\sqrt{\beta_0\beta_2}, 2\sqrt{\beta_0\beta_2}]$. Implementing these constraints in conjunction with the usual prior structure, we have

$$\begin{aligned} y_{ij} &= \beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij}^2 + \epsilon_{ij} \\ \epsilon_{ij} &\sim N(0, \sigma_i^2) \\ \beta_0 &\sim Normal(\mu_0, \sigma_0^2) I_{(-\infty,0)} \\ \beta_1 &\sim Normal(\mu_1, \sigma_1^2) I_{(-2\sqrt{\beta_0\beta_2}, 2\sqrt{\beta_0\beta_2})} \\ \beta_2 &\sim Normal(\mu_2, \sigma_2^2) I_{(-\infty,0)} \\ \sigma^2 &\sim InverseGamma(a, b), \end{aligned}$$

where $I()$ is the indicator function. Thus, we fit a downward opening parabola to the $\log P_c$ values over time for each event. This implies that each event has $\log P_c$ values which will rise and fall over time and that each event is allowed to have its own rate of increase/decrease. Eliciting informative priors on the regression coefficients will allow us to borrow information about what the shape of this parabola is for most events, and how much it is prone to vary. Though there are other ways to borrow information, *e.g.* a mixed model, we find this to be a simple and straightforward way to allow the model to be flexible enough to fit all of the events. On a more technical note, attempting a mixed model in this setting is not particularly straightforward, as any random effects specified in the model would also have to be constrained. Furthermore, at least two random effects would be necessary (a random intercept and a random slope), as we desire a model which can have a different peak location and value for each event. Ultimately, we favor a more simple model that is interpretable and flexible.

We note a few additional attributes of this model here. First, although we know that the regression coefficients are necessarily correlated, we choose not to incorporate this correlation in our prior structure, principally because the prior for β_1 depends on other regression coefficients. Although estimates may be slightly more efficient by including more information, we believe that independent priors are sufficient in this case. Additionally, it is worth noting that because the regression coefficients are defined on half the real line (β_0 and β_1) and a closed interval (β_1), other prior distributions could be chosen. For instance, the Gamma distribution is defined on $(0, \infty)$, so theoretically it could be used as a prior distribution for $-\beta_0$ or $-\beta_2$. Similarly, the Beta distribution could be considered for β_1. However, our testing of these priors showed problems with their use. The sampling generally exhibited a high amount of autocorrelation and/or slow convergence, which is not the case with the truncated normal distributions.

Because P_c values can assume very small values, including the value of 0 to machine precision, using these data in an unbounded way introduces a very large dynamic range in the observed values. Operationally, there is little interest in events with a P_c below 1E-07

and essentially none with a P_c below 1E-10; so it is quite reasonable to truncate (left-censor) the dataset by resetting the values of P_c data $<$ 1E-10 to the 1E-10 value. Of course, in such a case one must accept the cognitive dissonance of the model predicting P_c values less than 1E-10. However, this is acceptable to us for a few reasons. One reason is that we are mainly concerned with the time point at which the peak y value occurs and, to a lesser degree, its predicted value. The other reason is more practical: we are not particularly concerned with prediction for smaller values of y. Because the y values represent orders of magnitude, we are far less worried about prediction error for small values of y than we are for large values of y.

Lastly, we admit that our model cannot capture the rare occurrence that the y values initially decrease and then increase, i.e. an upward opening parabola. We do not concern ourselves with this case, as in such a case our model would fit essentially a horizontal line, indicating no discernible peak value. Though the shape of the data is not preserved, our end goal is: we seek significant statistical evidence of the size and location of the peak, and in this situation its size and location are unclear.

4.3.2 Inference for the Peak Value

The supposed canonical behavior suggests that the order of magnitude of the P_c value increases as the uncertainty decreases and drops off after a certain point. In general, uncertainty tends to decrease with time. Thus, we expect that this relationship holds with reference to time as well. Though some events exhibit this behavior, many events only exhibit the decline in order of magnitude of the P_c value. That is, if we believe the $\log_{10} P_c$ truly increases in time initially, this increase is censored within many events–the earlier small $\log_{10} P_c$ values lie outside of the 7-day screening window or outside of the physical screening volume and were thus not reported. Similarly, because we are only able to observe a few $\log_{10} P_c$ values, we are unlikely to observe the true peak. Thus, it is difficult to measure the accuracy of any prediction of the peak we might make. Because we are not certain of being able to observe the true peak, we take the highest observed $\log_{10} P_c$ value to be the peak.

We can infer the distribution of the location of the peak by utilizing the well-known identity that the peak is located at $x_{max} = -\beta_1/2\beta_2$. We estimate this distribution by collecting the posterior samples of β_1 and β_2 from the MCMC output and transforming them as x_{max} is defined. From the empirical distribution of x_{max}, we can compute a point estimate and a 95% credible interval for x_{max}. For the point estimate, we utilize the posterior mode of x_{max}, which is found by fitting a kernel density to the samples of x_{max} and finding the most likely value. For the credible set, since we define time as time until TCA, we are mainly concerned with the lower bound. Here, the lower bound represents, with 95% probability, the latest time at which our model predicts a peak will occur. This is operationally useful, as we are often interested in whether the peak will occur before 48 hours until TCA. Thus, if we can say that the peak will occur before this time with 95% probability, then the operator may be able to use this information to make a more informed decision regarding the importance of continuing to follow the event. Similarly, we can construct bounds for the magnitude of the peak. The distribution for the magnitude of the peak y_{max} is computed in the same way as for the location but instead using the transformation $y_{max} = \beta_0 - \beta_1^2/4\beta_2$.

4.4 The Bayesian Beta Regression Model

4.4.1 The Distribution of $\log_{10} P_c$ Values

When modeling the trend in P_c values, one is generally concerned with changes in order of magnitude, thus one generally models $\log_{10} P_c$ as opposed to the observed P_c values. This poses an interesting statistical question, namely the distribution of $\log_{10} P_c$ values. Distribution selection is more obvious for the P_c values, as they are bounded between 0 and 1; thus a statistical modeler generally chooses a beta distribution to model these values (although there are a few other less commonly used distributions, such as the simplex distribution, that could be deployed). Theoretically, there is no lower bound on $\log_{10} P_c$ values, as P_c values can be arbitrarily close to zero. Operationally, however, one often considers P_c values below 1E-10 to be effectively 0. To account for this, in a previous section we "floored" the $\log_{10} P_c$ values at -10, so that the large number of small $\log_{10} P_c$ values did not overly influence the model. This allows one to focus inference on the operationally relevant $\log_{10} P_c$ values, which tend to be around -5 and greater. We follow suit here, flooring all $\log_{10} P_c$ values at -10. Therefore, even in modeling the $\log_{10} P_c$ values, we have bounded data (between -10 and 0) that yield a mixture of discrete and continuous outcomes. In this case, the data are -10-inflated, but when the variable is rescaled to fit the Beta distribution, the data are zero-inflated. This can be accommodated by the zero-inflated beta distribution; the form of the equation is given below, with the specific symbology explained in the subsequent sections:

$$f(y|\mu,\phi,p) = (1-p)\frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1} I_{(0,1)}(y) + pI_{[0]}(y). \qquad (4.4.1)$$

Here, $I_A(\cdot)$ is the indicator function, so that the first term corresponds to values falling between 0 and 1 (or $\log P_c$ values falling between -10 and 0) and the second term corresponds to values equal to 0 (or $\log P_c$ values equal to -10). The parameter μ is the mean of the Beta distribution, which will be modeled in the GLM framework, and the parameter ϕ is the corresponding dispersion parameter, which is a measure of variability. The parameter p can be interpreted to be the probability that one observes a 0 (or a $\log P_c$ of -10). To our knowledge, no one has investigated a joint model for an Bayesian inflated Beta regression model.

As noted previously, the P_c values for each event tend ultimately to decrease with time but at a different rate in each conjunction. This suggests approaching the problem within a mixed model framework, allowing random terms for each conjunction. This is a natural approach to take, as the data are longitudinal in nature: one observes an overall trend in time, yet each subject (in this case, each conjunction) deviates somewhat from this trend, and observations within a subject are correlated with each other. In Figure 1, we visualize the longitudinal nature of the data. We plot the $\log_{10} P_c$ values of ten events over time, with each events' values connected by a line; and we also plot these values versus the so-called ratio of combined covariance radius to miss distance. This is to expose the canonical trend in P_c development: as the event moves closer to TCA, the covariance shrinks, bringing this ratio slowly to a peak and then a marked drop-off.

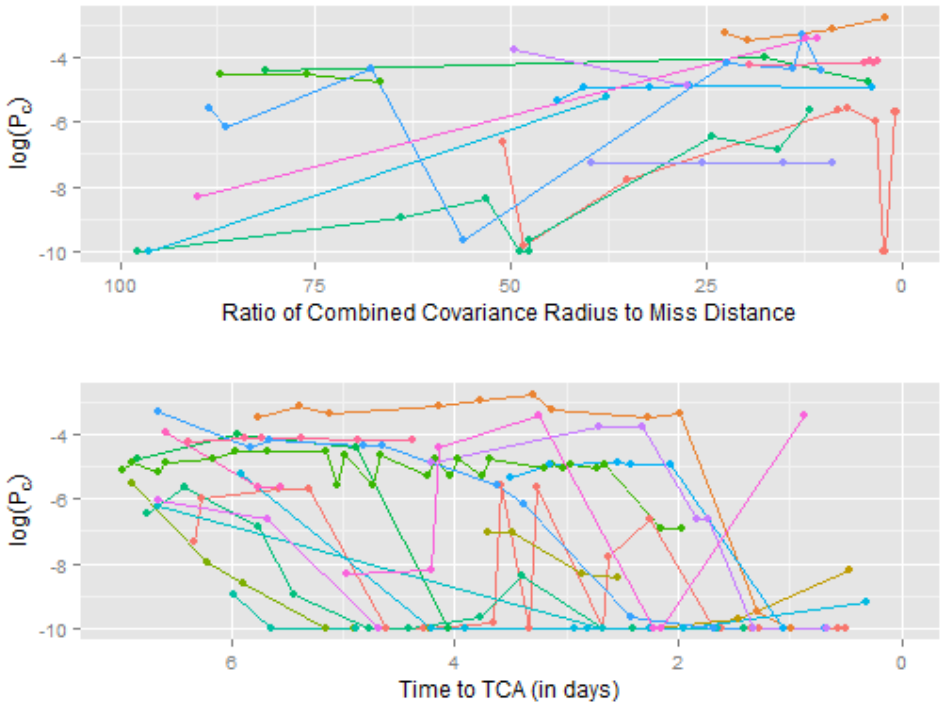

Figure 4.4.1: Plot of $\log_{10} P_c$ vs. ratio of covariance radius to miss distance

It is not clear that the trend in $\log P_c$ values is stronger for ratio of covariance of radius to miss distance than days to TCA. In fact, the $\log P_c$ values exhibit a slight increasing trend with respect to the ratio, while the expected decreasing trend occurs with respect to time. It is possible that the $\log P_c$ values "drop off" at some small value of the ratio, but it is not clear from the data when this might occur, and how frequently one would observe it. Additionally, as was discussed in the introduction, this value is not monotonically increasing or decreasing with time (due to unpredictable changes in the covariance size and the estimate of the mean miss distance between the two satellites); so despite its theoretical linkage to the actual phenomenology of the situation, it is actually a less desirable independent variable for performing trending and prediction. As previously, model construction for P_c trending and prediction will use time to TCA as the independent variable.

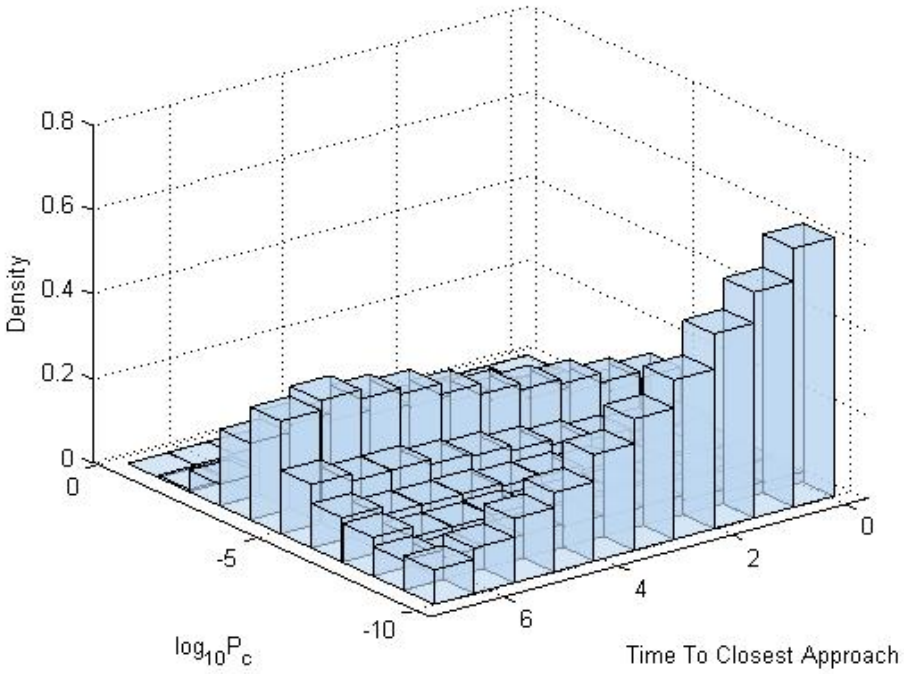

Figure 4.4.2: Two-dimensional histogram of $\log_{10} P_c$ values vs. TTCA

We can visualize the trend of the $\log_{10} P_c$ values over time by considering a two-dimensional histogram, as displayed in Figure 4.4.2.

Recall that we have replaced all $\log_{10} P_c$ values below -10 with -10. Then the figure above indicates that the probability of observing a P_c value of 1E-10 or lower increases as one approaches TCA. In fact, at 2 days to TCA, about 40% of events observed have a P_c of 1E-10 or lower. At 7 days until TCA, the most observed value is about -5, which becomes less frequent over time, as more events observe a $\log_{10} P_c$ of -10. Interestingly, -5 seems to be the most likely value when one does *not* observe a -10, regardless of the time. We can use this information to construct a prior distribution for the model in Equation (1), as the increase of observed -10 values gives us an idea of how p behaves over time, and the observed mode of -5 of the $\log_{10} P_c$ values above -10 gives us some information about the mean of the Beta distribution.

4.4.2 Beta Regression

To model a Beta-distributed random variable with reference to a covariate (such as time), one must use a generalized linear model (GLM). Although GLM's for many other members of the exponential family (Normal, Gamma, etc.) have been developed since 1972[40], the GLM for the beta distribution is relatively new, being introduced in 2001[42]. The reason for this late development is due to the fact that one must reparameterize the beta distribution in order to model the mean sufficiently, an expansion that was not explored until recently. We develop this reparameterization for completeness. The probability density function (pdf) of a random variable X with a beta distribution is generally given as

$$f(x|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1},$$

where $\Gamma(\cdot)$ is the gamma function. The mean of this distribution is $E(X) = \frac{\alpha}{\alpha+\beta}$. GLM's are generally specified by setting some function $g(\mu)$ of the mean equal to a linear combination of covariates. For instance, logistic regression uses the logit link $g(\mu) = \log(\frac{\mu}{1-\mu})$, which is then set equal to a linear combination of covariates, e.g. $\beta_0 + \beta_1 X$, where X is a covariate, such as time. However, as the beta distribution is specified, it is unclear how to model the mean. To facilitate direct modeling of the mean, let $\mu = \frac{\alpha}{\alpha+\beta}$ and $\phi = \alpha + \beta$. Then we can rewrite the beta pdf as

$$f(x|\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} x^{\mu\phi-1}(1-x)^{(1-\mu)\phi-1}.$$

Now we may model the mean μ directly. For instance, we may choose the logit link and model

$$\log\left(\frac{\mu}{1-\mu}\right) = \beta_0 + \beta_1 x + ... + \beta_p x^p,$$

so that the log-odds of the mean has a linear relationship to X. Various link functions are possible, such as the probit link, the complementary log-log link, and the log link. Our simulations have shown that there is no significant advantage in choosing one over the other, so we proceed with the logit link, as it is easy to interpret.

Recall that for each conjunction, one observes a different progression of P_c values. Sometimes the P_c values drop off quickly before TCA, other times they drop off much nearer TCA, and sometimes not at all. To model such a behavior, we may include a random intercept for each event as follows. Let μ_{ij} be the mean of the j^{th} P_c value in the i^{th} event, scaled to be between 0 and 1. Since we have $\log_{10} P_c$ values bounded between -10 and 0, a suitable transformation is $\mu_{ij} = E(Y_{ij})/10 + 1$, where Y_{ij} is the $\log_{10} P_c$ value of the j^{th} P_c value in the i^{th} event. We may consider the model

$$\log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = \beta_0 + \beta_1 t_{ij} + ... + \beta_p t_{ij}^p + b_i$$
$$b_i \sim N(0, \sigma_b^2),$$

where b_i is a random effect allowing for an intercept for the i^{th} event, and t_{ij} is the time until TCA of the j^{th} P_c value within the same event. One may additionally consider a random slope or other random effects for higher order terms, but given the amount of data, these would be difficult to fit and depend strongly on choice of prior distribution.

Recall that in Equation (1) we also introduced the parameter p. This parameter controls what percentage of the time we observe a zero. In our case, since about a third of our data are zeros, p might be close to 1/3. However, we also know that the closer an event approaches TCA, the more likely one is to observe a P_c value that is 0. As a result, we can also let p depend on our covariate. This parameter is also bounded between 0 and 1, so we again use a logit link function here (or any of the other aforementioned link functions). Additionally, we may consider a random term for this model for each event, as the probability of observing a zero is higher for some events than others. Thus, we may consider a regression such as

$$\log\left(\frac{p}{1-p}\right) = \alpha_0 + \alpha_1 t_{ij} + ... + \alpha_p t_{ij}^p + a_i$$
$$a_i \sim N(0, \sigma_a^2)$$

which is similar to the regression for μ above. Again, more random terms could be introduced if necessary.

4.4.3 Model Selection

Given below are some selected results from an exploratory model selection. To evaluate the relative merits of different levels of model complexity, we use the penalized deviance construct[54], where lower values indicate a better fit. Specifically, define $D(\theta)$ to be the "Bayesian Deviance", with form

$$D(\theta) = -2\log p(y|\theta) + 2\log f(y), \tag{4.4.2}$$

where $p(y|\theta)$ is the likelihood of y given θ and $f(y)$ is the saturated model, where $f(y) = p\{y|E(Y) = y\}$. We can rewrite $D(\theta)$ as

$$D(\theta) = -2\left(\log p(y|\theta) - \log f(y)\right), \tag{4.4.3}$$

which shows that $D(\theta)$ is -2 times the difference between the fitted model and the saturated model. Put simply, $D(\theta)$ measures how well a model fits the data relative to a model that fits the data perfectly. We estimate $D(\theta)$ with $\overline{D(\theta)}$, which can be written as

$$\overline{D(\theta)} = D(\bar{\theta}) + p_D, \tag{4.4.4}$$

where $p_D = \overline{D(\theta)} - D(\bar{\theta})$. The estimate $\overline{D(\theta)}$ is known as the *penalized deviance*, as it is computed as the sum of $D(\bar{\theta})$, the mean deviance, and p_D, the penalty term. The term $D(\bar{\theta})$ measures how well one fits the data, with lower values indicating better fit, and the term p_D penalizes this fit for more parameters, where higher values indicate a larger penalty. The penalty term p_D is also known as the *effective number of parameters*, so that one may interpret this term as an estimate of how many parameters the model is actually estimating in order to describe the data. This is to account for models with more parameters fitting the data better, or over-fitting the data.

In Table 4.1 we provide the calculated mean deviance, penalty, and consequent penalized deviance for various models. These values justify how we came to our final model, as we chose the model with the lowest penalized deviance. The variables Y_c and Y_d represent the continuous and discrete parts of the model given in Equation (1), respectively. That is, Y_c are the values produced by the beta distribution, and Y_d are the 0-1 variables that either indicate a zero (1) or a continuous variable (0). All added complexities are in addition to the baseline linear model specified below:

$$Y_{ij} \sim f(y_{ij}|\mu_{ij}, \phi, p_{ij}) \tag{4.4.5}$$

$$\log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = \beta_0 + \beta_1 t_{ij} + b_i \tag{4.4.6}$$

$$b_i \sim N(0, \tau_b) \tag{4.4.7}$$

$$\tau_b \sim Gamma(0.001, 0.001) \tag{4.4.8}$$

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \alpha_0 + \alpha_1 t_{ij} + a_i \tag{4.4.9}$$

$$a_i \sim N(0, \tau_a) \tag{4.4.10}$$

$$\tau_a \sim Gamma(0.001, 0.001), \tag{4.4.11}$$

$$\phi \sim Gamma(0.001, 0.001) \tag{4.4.12}$$

$$\beta_k, \alpha_k \sim Normal(0, 1), \quad k = 0, 1, 2. \tag{4.4.13}$$

Model	Mean Deviance	Penalty	Pen. Deviance
Linear	-17.23	74.93	57.7
Quad Term for Y_c	-23.02	77.32	54.31
Quad Term for Y_d	-26.78	76.45	49.67
Quad Term for Y_c and Y_d	-32.52	79.23	46.71
QuadTerm for both, RanSlope for Y_c	-31.12	81.07	49.95
QuadTerm for both, RanSlope for Y_d	-36.91	85.54	47.63
Cubic Term for Y_c	-31.89	81.1	49.21
Quadratic, linear for phi	-27.76	80.87	53.11

Table 4.1: Model Selection Output

Note that adding a random slope to either Y_c or Y_d did not produce a better fit, nor did specifying a correlation between the random effects.

Based on these results, we propose the following model. Let Y_{ij} be the j^{th} scaled $\log_{10} P_c$ value of the i^{th} event. Also, let t_{ij} be the corresponding time until TCA (in days).

$$Y_{ij} \sim f(y_{ij}|\mu_{ij}, \phi, p_{ij}) \tag{4.4.14}$$

$$\log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = \beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij}^2 + b_i \tag{4.4.15}$$

$$b_i \sim N(0, \tau_b) \tag{4.4.16}$$

$$\tau_b \sim Gamma(0.001, 0.001) \tag{4.4.17}$$

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \alpha_0 + \alpha_1 t_{ij} + \alpha_2 t_{ij}^2 + a_i \tag{4.4.18}$$

$$a_i \sim N(0, \tau_a) \tag{4.4.19}$$

$$\tau_a \sim Gamma(0.001, 0.001), \tag{4.4.20}$$

$$\phi \sim Gamma(0.001, 0.001) \tag{4.4.21}$$

$$\beta_k, \alpha_k \sim Normal(0, 1) \quad k = 0, 1, 2. \tag{4.4.22}$$

For this and all other models proposed in this section, we generate predictions for the next OCM by conditioning on the random effects. This yields the best linear unbiased predictor (BLUP), as discussed in Diggle[22]. Thus, we generate the predictive distribution

$$Y_{i(j+1)} \sim f(y_{i(j+1)}|\mu_{i(j+1)}, \phi, p_{i(j+1)}, a_i, b_i),$$

where $\mu_{i(j+1)}$ and $p_{i(j+1)}$ are produced by using $t_{i(j+1)}$ in their respective models. Using this distribution, we can construct credible sets. As in the other Bayesian models, we use the posterior mode of this predictive distribution as our estimate for $Y_{i(j+1)}$.

4.4.4 Issues of Identifiability

The model proposed in equations (10)-(17) has a total of 7 parameters and 2 random effects, which suggests one must estimate a total of 9 quantities in order to make inferences and hence predictions. However, this issue can be ameliorated by using informative priors in a Bayesian framework. To acquire these informative priors, we run the proposed model on a training

dataset of a large number of events, which are not themselves used for model evaluation. We use the posterior distribution of the parameters as informative prior distributions by matching sample moments of the posterior samples with its prior distribution family. We do this for all of the population-level parameters, which are ϕ, α_k, and β_k for $k = 0, 1, 2$. Then we are left with two parameters to estimate, the random effect variances, τ_a and τ_b. Because we begin making predictions beginning with the second observation, these parameters are identifiable when making inference on a single event.

Motivated by the large number of events in our testing data set, we investigated whether prediction could be improved by making inferences on more than one event at once. In order to test this, we followed the mean squared prediction error (MSPE) when making predictions on one event, 5 events, 10 events, and 25 events. Including more events did not improve the MSPE, likely due to the fact that, in reference to a single event, other events only contribute to the population-level parameters, which are already well-known due to the informative prior distributions. Thus, we make predictions on a single event at a time.

4.4.5 An Aside: Coverage from an Initial Simulation

In an initial exploratory simulation, we found that 97.5% prediction intervals constructed in the Beta model had 86% coverage. Though coverage is often lower than the nominal rate with real data, we found this coverage to be too low to have any meaningful operational use. In exploring this phenomenon, we found that splitting up the dataset into three parts, a high-, medium-, and low-risk group, ameliorated the issue of low coverage. Specifically, if an event had a high (above -4) $\log P_c$ value by 3 days time to TCA (TTCA), we called it high-risk. If an event had a medium (between -7 and -4) $\log P_c$ value by 3 days TTCA, we called it medium-risk. If an event had a low (below -7) $\log P_c$ value by 3 days TTCA, we called it low-risk. We shall refer to the high-, medium-, and low-risk groups as Red, Yellow, and Green hereafter.

Incidentally, the fact that our model performed well when the data were separated into different risk groups supports the notion that the $\log P_c$ value behaves differently depending on the quantile it inhabits. In terms of the Beta regression model, this implies that the population-level trend is different for these different risk groups, which suggests that they ought to be modeled separately. For the simulation presented in this chapter, these definitions worked well and possess the additional advantage of aligning closely with thresholds presently in use operationally for categorizing conjunction event severity.

4.4.6 Checking Assumptions

A variety of assumptions are employed in the model. These include

- A linear model for $g(\mu) = \log(\mu/(1-\mu))$
- A linear model for $h(p) = \log(p/(1-p))$
- The dispersion parameter ϕ is constant across time.
- The logit of the mean values associated with the Beta distribution Y_c are parallel for each event. That is, the logit of the means for each event is separated by a random intercept.
- The random intercepts have a normal distribution.

We check these assumptions graphically below.

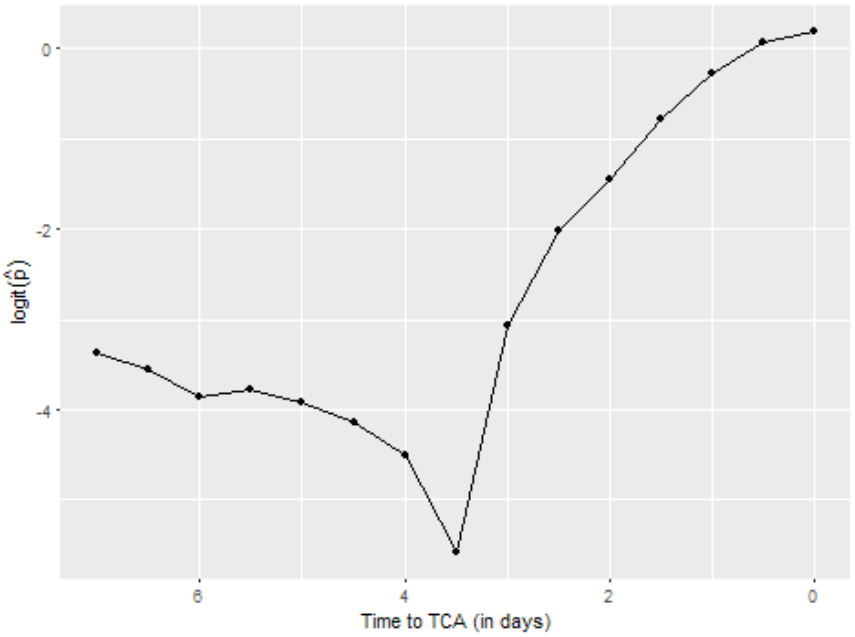

Figure 4.4.3: Plot of $\log(\hat{p}/(1-\hat{p}))$ vs. TTCA

The graph above shows that, though a single second order polynomial may fit $logit(p)$, a piecewise function of two second order polynomials may be more appropriate. The graph suggests that separate models might be appropriate for the time intervals $(0, 3.5)$ and $(3.5, 7)$.

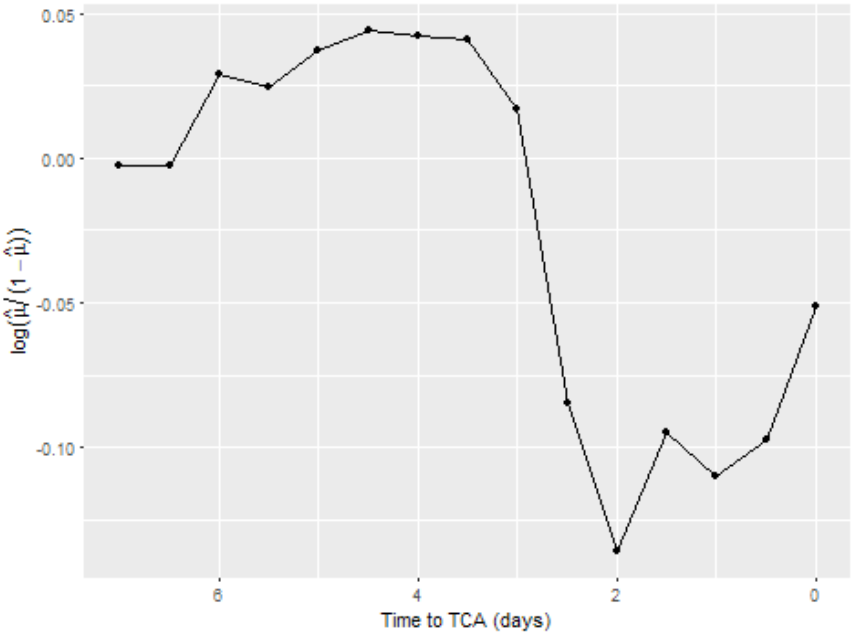

Figure 4.4.4: Plot of $\log(\hat{\mu}/(1-\hat{\mu}))$ vs. TTCA

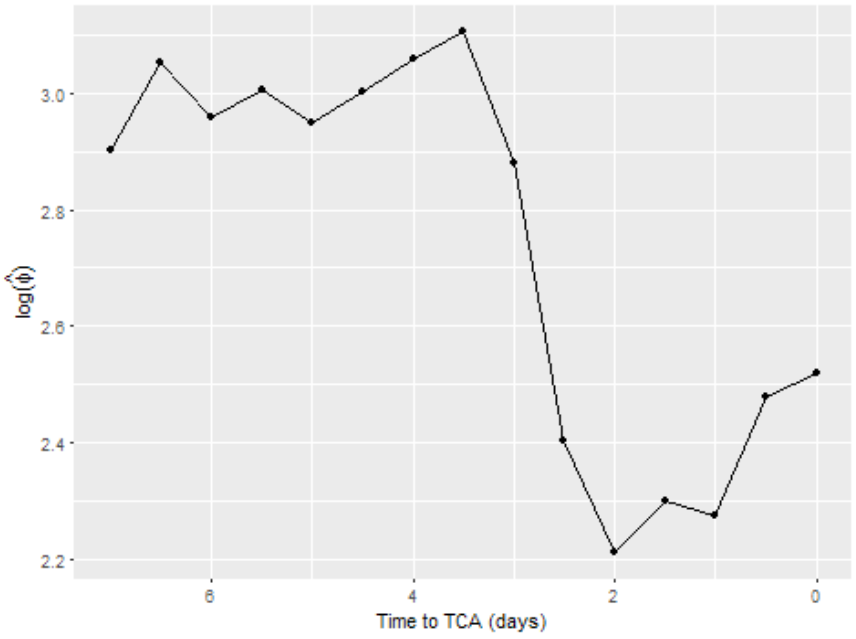

Figure 4.4.5: Plot of $\log(\hat{\phi})$ vs. TTCA

Again, these graphs suggest that it may be appropriate to fit piecewise polynomials on the intervals $(0, 3.5)$ and $(3.5, 7)$.

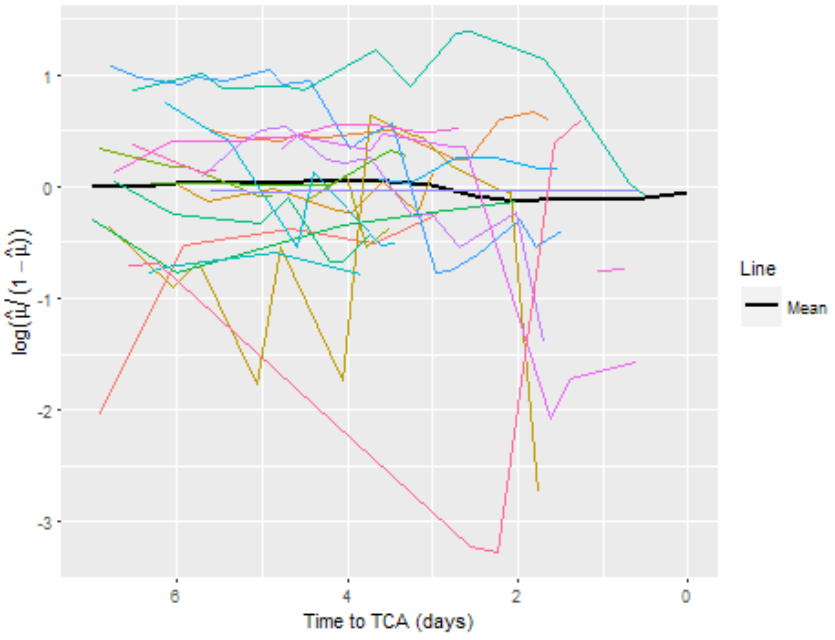

Figure 4.4.6: Spaghetti plot of $\log(y/(1-y))$ vs. TTCA

The graph above depicts the overall mean (in black), $\log(\mu)/(1-\mu)$, as well as this value for 20 random events (color lines). Though there is significant heterogeneity in the paths of the events over time, it appears that a random intercept alone may be sufficient, as most paths seem to be parallel to the overall mean.

4.4.7 Threshold Model

As evidenced in the graphs above, it appears that at 3.5 days until TCA, there is a noticeable change in behavior in both the behavior of μ and p. As a result, modeling this change in behavior may be of interest. To model this behavior, we specify a trend on the interval (7, 3.5) days to TCA, and a different trend on the interval (3.5, 0) days to TCA.

We specify two such models. In the table below, "Threshold" is a model with this threshold modeling for μ only, while "Threshold 2" uses such a structure for p as well. It is clear from the results that a threshold may be appropriate for both parameters, as the DIC is lowest for the "Threshold 2" model.

Model	Mean Deviance	Penalty	Pen. Deviance
Beta Reg	-451	81.66	-369.3
Threshold	-492.9	85.31	-407.6
Threshold 2	-751.5	107.6	-653.9

Table 4.2: Model Selection Output

The resulting model is given below. Notice that it has the same construction as the previous Beta regression model, but with different trends specified on the time intervals (7, 3.5) and (3.5, 0) days to TCA.

$$
\begin{aligned}
Y_{ij} &\sim f(y_{ij}|\mu_{ij}, \phi, p_{ij}) && (4.4.23)\\
\log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) &= (\beta_{01} + \beta_{11}t_{ij} + \beta_{21}t_{ij}^2)I_{[0,3.5]}(t_{ij}) && (4.4.24)\\
&\quad + (\beta_{02} + \beta_{12}t_{ij} + \beta_{22}t_{ij}^2)I_{(3.5,7]}(t_{ij}) + b_i && (4.4.25)\\
b_i &\sim N(0, \tau_b) && (4.4.26)\\
\tau_b &\sim Gamma(0.001, 0.001) && (4.4.27)\\
\log\left(\frac{p_{ij}}{1-p_{ij}}\right) &= (\alpha_{01} + \alpha_{11}t_{ij} + \alpha_{21}t_{ij}^2)I_{[0,3.5]}(t_{ij}) && (4.4.28)\\
&\quad + (\alpha_{02} + \alpha_{12}t_{ij} + \alpha_{22}t_{ij}^2)I_{(3.5,7]}(t_{ij}) && (4.4.29)\\
a_i &\sim N(0, \tau_a) && (4.4.30)\\
\tau_a &\sim Gamma(0.001, 0.001), && (4.4.31)\\
\phi &\sim Gamma(0.001, 0.001) && (4.4.32)\\
\beta_k, \alpha_k &\sim Normal(0, 1) \quad k = 0, 1, 2. && (4.4.33)
\end{aligned}
$$

4.5 New Beta Regression Model

In light of the cubic nature of the means shown above, as well as the observed success of the Look-Up model, we introduce a new model, which we refer to as the New Beta Regression model. This model incorporates a cubic model into the model for μ_{ij}, as well as previous observations. Unlike the Look-Up method, we incorporate the values themselves rather than the quantiles.

$$Y_{ij} \sim f(y_{ij}|\mu_{ij}, \phi, p_{ij}) \tag{4.5.1}$$

$$\log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = \beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij}^2 + \beta_3 t_{ij}^3 + \beta_4 y_{(i-1)j} + b_i \tag{4.5.2}$$

$$b_i \sim N(0, \tau_b) \tag{4.5.3}$$

$$\tau_b \sim Gamma(0.001, 0.001) \tag{4.5.4}$$

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \alpha_0 + \alpha_1 t_{ij} + \alpha_2 y_{(i-1)j} + a_i \tag{4.5.5}$$

$$a_i \sim N(0, \tau_a) \tag{4.5.6}$$

$$\tau_a \sim Gamma(0.001, 0.001), \tag{4.5.7}$$

$$\phi \sim Gamma(0.001, 0.001) \tag{4.5.8}$$

$$\beta_k, \alpha_k \sim Normal(0, 1) \quad k = 0, 1, 2. \tag{4.5.9}$$

4.6 Bayesian Beta Cluster Regression

Though the Beta regression model accounts for more aspects of the data than the more naive models, results (presented in an upcoming section) suggest that this model may not be accurate enough to be operationally useful. In addition, there is an aspect of the data which this model cannot address, which may suggest a better model. In particular, operators often suspect that there are different categories of events: low risk, medium risk, and high risk. Furthermore, operators believe that these categories of events behave slightly differently, so that if one knew with a high degree of certainty which category of event was observed, we would be able to obtain more accurate inference. However, these categories are relatively arbitrary. It may be of interest to know exactly how many significantly different categories of events there are. The Bayesian Beta Cluster Regression model is constructed to tackle this question, and hopefully provide more accurate inference. More technically, we are interested in how many clusters are in the data, what the clusters look like, and how likely we are to be able to identify which cluster is occurring by the time a decision is to be made.

4.6.1 Model

The model is conceptually straightforward. We now assume that there are K different means within the data, representing the means of K clusters. That is, each event is assumed to come from one of the K clusters. This can be expressed as a mixture model, so that

$$Y_{ij} \sim \sum_{k=1}^{K} \pi_k f(y_{ijk}|\mu_{ijk}, \phi_{ijk}, p_k) \tag{4.6.1}$$

where π_k is the probability Y_{ij} is from cluster k. In this general model, each cluster has its own mean μ_{ijk}, dispersion parameter ϕ_{ijk}, and probability of observing a P_c of zero p_k. Model selection will enable us to determine the size of K and which parameters do not vary across clusters.

To form a complete Bayesian specification, priors for all parameters are necessary. The same non-informative prior structure given above is used for the parameters of all clusters.

In addition, we let

$$\pi_1, ..., \pi_K \sim Dirichlet(e_1, ..., e_K) \tag{4.6.2}$$

where the e_i are chosen to be non-informative, and thus all are set to one. Let S_{ik} be a random variable which indicates whether event i is part of cluster k or not, taking value 1 if this is true, and 0 if it is false. Then

$$Y_{ij}|S_{ik} \sim f(y_{ijk}|\mu_{ijk}, \phi_{ijk}, p_k), \tag{4.6.3}$$

where $f(\cdot)$ is a zero-inflated Bayesian Beta regression model, as given above.

4.6.2 Model Selection

Given below is a model selection table similar to that in section 2.3.2. The penalized deviance is given by number of clusters, k, for the above model.

Clusters	Mean Deviance	Penalty	Pen. Deviance
1	-728.4	103.7	-624.7
2	-1462	901	-561
3	-1841	1085	-756
4	-1960	1917	-43.64

Table 4.3: Model Selection Output

As is evidenced above, the model with 3 clusters seems to fit better than the model with 1 cluster. Though it may seem odd that the model with 2 clusters does not also have a lower DIC than the model with 1 cluster, it should be noted that DIC is notoriously difficult to calculate for mixture distributions, and also has problems associated with it[14]. As a result, we also consider an ad-hoc method of selecting the number of components. We set k to the maximum number of possible clusters, and calculate the posterior probability of each component. Specifically, we set $k = 5$, and find that the probability of the first two components is 0.63, and 0.35, respectively. This yields some evidence that the number of clusters necessary is two.

From a practical standpoint, adding more clusters increases the computation time, making fewer clusters more desirable for implementation. In addition, as the number of clusters grows, one encounters other computational issues. For example, a greater number of clusters is generally accompanied by slower convergence, as well as label-switching[49][31], a problem in the posterior samples in which a parameter “switches clusters", due to the invariance of the likelihood to label-switching. This, combined with the above evidence, leads us to consider a model with only 2 components. Results from an initial fit of this model show that, if two clusters indeed do exist, one is a “high variability" cluster and one is a “low variability" cluster. Figure 4.6.1 shows these two clusters.

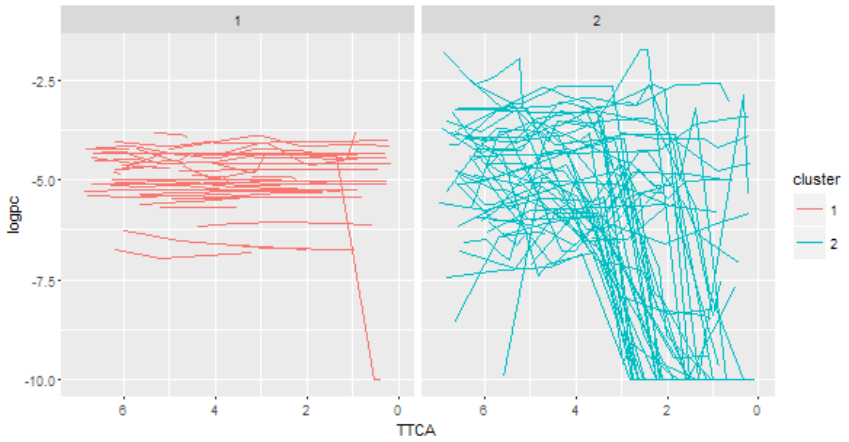

Figure 4.6.1: Clusters found from Beta clustering model

Cluster 1 contains events which show little variability in $\log_{10} P_c$ over time. Interestingly, most of these values are centered around -5. In contrast, cluster 2 contains events which exhibit much more variability over time. Of course, the difficulty with such a model must identify the cluster appropriately to make valid predictions for a given event. Regardless of predictive performance, this model is useful because it gives us new insight into the data. For instance, if such a model could reliably classify events into “high" and “low" variability clusters, operators would have yet another way to temper their expectations about future $\log_{10} P_c$ behavior.

5 Measures of an Effective Model

In this section we discuss the properties on which we will compare our two models. We focus on model fit and decision-making performance.

5.1 Model Fit

The main concern in building predictive models is fitting the data well enough to predict new observations accurately. In order to quantify this, we check the bias, prediction errors, and upper bounds of the proposed models. Specifically, we would like our models to be unbiased, so that the prediction errors are centered at zero. Secondly, we check to see if the prediction intervals are bigger or smaller for different times, predicted values, and times until prediction. Lastly, we check to see that our upper bounds have the correct coverage.

5.1.1 Prediction Coverage

Let t_{ij} be the TTCA of the j^{th} observation i^{th} event. Then our predictions and the associated confidence intervals are made at $t_{i(j+1)}$. Thus, we predict the distribution of $y_{i(j+1)}$. As noted before, we make predictions beginning with the second observation, $j = 2$. We construct 95% confidence intervals, and check whether the true value, $y_{i(j+1)}$, is contained in the interval. The average of the number of true values contained within these predicted intervals is our prediction coverage.

It should be noted that the time until prediction is *not* the same for all predictions. As noted before, the time at which new P_c values are received are random, producing irregular

times between successive observations. The time between observations is usually at most 2 days. Though the time between observations varies, we calculate coverage irrespective of time, so that a prediction 2 days into the future contributes equally to the coverage as a prediction 0.5 days out. Though one would prefer to make predictions at the same number of days into the future for all events, this is not possible due to the nature of the data.

5.2 Decision-Making Efficacy

5.2.1 Framework

In order to assess our models in the framework of making decisions about whether to continue active monitoring of an event, we implement a simple decision-making framework and study its properties in both models. Because the most weighty period for conjunction assessment operational decision-making occurs 2-3 days TTCA, we focus on this region of data. Specifically, we make predictions at 2 days TTCA and make a decision based on this prediction. Let $\hat{y}_2$ be an estimate of the $\log P_c$ predicted to occur at 2 days TTCA. We will make the decision that the $\log P_c$ values will remain above the threshold θ after 2 days TTCA if

$$\tilde{y}_2 > \theta \tag{5.2.1}$$

and will make the decision that the $\log P_c$ values will fall below the threshold θ otherwise. To couch this in the hypothesis testing framework, we write

$$H_0 : \tilde{y}_2 < \theta \quad vs. \quad H_1 : \tilde{y}_2 > \theta, \tag{5.2.2}$$

so that rejecting H_0 is synonymous with claiming the $\log P_c$ will remain high. In our simulations, we set $\theta = -5$ for the Red group and $\theta = -7$ for the Yellow group. Note that while -7 is the lower bound for being in the Yellow group at 3 days TTCA, -5 is below -4, the corresponding lower bound for the Red group. A lower threshold was chosen as these events are generally of much higher concern, thus one prefers an extra order magnitude of certainty before claiming the event is at a lower risk level. In order to explore this trade-off fully, we tried various quantiles of the distribution of $\tilde{y}_2$, which we describe below.

5.2.2 Type I and Type II Errors

As with most decision-making frameworks, our framework can admit Type I and Type II errors. The hypothesis in (22) is framed in terms of the event of a P_c value remaining high, as this is the event we are most concerned with. A Type I error here is the incorrect assertion of a high P_c value (i.e. a false alarm), and a Type II error is a the more worrisome incorrect prediction of a low value (i.e. a missed detection). Thus, while we may find it acceptable to trigger an alarm when the $\log P_c$ value has actually dropped off, it is almost *never* acceptable to miss detecting a high $\log P_c$ value. Of course, we can make our system as powerful against this event as we want, with the trade off of triggering more false alarms. It's worth noting that a false alarm for a high value is the same thing as missed detection for a value which has dropped off. Ideally, we would like to have an alarm that detects high values *and* low values with a high degree of accuracy. However, since we are more concerned with high values, we seek to quantify how often, if ever, can we detect these low values while still maintaining the high accuracy needed for detecting the high values.

6 Results

6.1 Data

Recall that we split our data into two groups. The Red and Yellow groups are defined below.

- If an event had a high (above -4) $\log P_c$ value by 3 days time to TCA (TTCA), it is part of the Red group.
- If an event had a medium (between -7 and -4) $\log P_c$ value by 3 days TTCA, it is part of the Yellow group.

The dataset used for tuning (*i.e.*, setting the parameters for the informative prior distributions) and testing the model was taken from the NASA Conjunction Assessment and Risk Analysis historical Conjunction Data Message (CDM) database. For the Yellow group, five hundred events' worth of data from calendar year 2013 was used for model tuning (the "training" dataset), and the tuned model was evaluated against approximately 2000 events from 2014 (the "validation" dataset); so there was no overlap in terms of time-period or actual data between the two datasets. For the Red group, 82 events were used for training and 70 were used for testing (this data set is far smaller, as these kinds of events are more rare). Data were taken from conjunctions against primaries in the orbital region defined by a perigee height between 500 and 750 km and an eccentricity less than 0.25. As described above, data flooring at a $\log_{10} P_c$ value of -10 was performed on the dataset. To qualify for use in tuning or evaluation, an event must have had at least two CDMs with a $\log_{10} P_c$ greater than -10.

6.2 Simulation Setup

To train our model, we perform a Bayesian analysis on the training data using non-informative priors. We determine the distribution parameters for the informative priors used in the test data by matching the first and second moments of the observed distributions to the hypothesized prior distributions. All MCMC inference is conducted in JAGS[45].

The simulation procedure for a given event is as follows. We attempt to make predictions for the next y value only after the second received CDM. We are interested in estimating the next $\log P_c$ value, which we predict by using the time at which it was observed. The predicted value is taken to be the mode of the posterior predictive distribution. In this context, it is important to use the posterior mode as opposed to the posterior mean, as the posterior predictive distribution is generally bimodal, with some mass at -10 and the remaining density between -10 and 0, inducing another peak. Thus, we choose the "most likely value" as opposed to the mean. We make predictions for five models: the Vertex model, the Beta Regression model, the New Beta Regression model, the Beta Clustering model, and the Look-Up model.

To further assess model fit, we also track a two-sided 95% credible interval for each prediction. We utilize the upper bound from the credible set for checking coverage. This is also done for the Look-Up and LOCF methods, though here the interval is a confidence interval and is calculated using repeated cross-validation. In addition to coverage, we are also interested in how many of these upper bounds are low enough to be "useful". That is, we would like to know how many of these lower bounds are lower than the lower threshold of the Yellow and Red groups.

We present results for the Yellow group only. We found that the Red group had too small of training and testing sets to yield any kind of meaningful conclusions about predictive performance. Further simulation is required to determine just how different these two groups are, and if the results shown below hold for more high risk events like those in the Red group. We suspect this may be the case, as our construction of these two groups was somewhat arbitrary in the first case. As shown by the Beta clustering model, it is likely that when it comes to modeling, there is more of a delineation between “high variability" and “low variability" events than high-risk and low-risk.

6.3 Models

Here, we make a few remarks about the models chosen for simulation. We choose to compare seven of the proposed models: Last Observation Carried Forward (LOCF), the Look-Up model (LKUP), the Vertex model (Vertex), the Beta Regression model (BetaReg), the Beta Clustering model (BetaClust), and the New Beta Regression model (BetaNew). Not all models are shown here to reduce clutter in the graphs and to ease comparison. We do not include the Beta Threshold model because this gave results similar to the Beta Regression model. The functional data model is not considered because it too yielded results similar to the Beta Regression model. Though promising, this model did not have greater success because one of the assumptions of the model was not met. As mentioned earlier, the functional principal component scores are assumed to follow a normal distribution. In practice, we found these to have a bimodal distribution, suggesting clustering. This prompted us to investigate clustering models, such as the Beta clustering model.

Inference for all models is based on prediction at the next OCM. Originally, we investigated the potential use of basing inference on the estimated peak in the Vertex model. We found that this produced poor results, and that better inference resulted from simply making predictions at the next OCM.

6.4 Simulation Results

Our main goal of prediction is to make a decision by 2 days to TCA. As a result, for the Yellow group, we are interested in whether a $\log P_c$ which is observed after 2 days to TCA will be below -7 or not. Additionally, for this group, we may be interested in a “worst case" scenario, which would be observing a $\log P_c$ value above -4 at this time. We also present results for this prediction.

As the goal of our inference is prediction, we present prediction errors for the various models. Figure 6.4.1 shows the density of the prediction errors for all five models. The Look-Up model is more peaked than the other models, suggesting that is produces more prediction errors close to zero. Though this suggests the predictions are more accurate than the other models, a closer inspection suggests that the tails of the prediction errors are nearly as long as those produced by other models. Interestingly, the Beta clustering model is the second most peaked.

Figure 6.4.10 provides a look at the prediction errors of the Look-Up method vs. LOCF. Here we see that the LOCF method produces similar prediction errors to the Look-Up method. In addition, we see an interesting artifact of prediction error density generated by the Look-Up model: the tails are jagged, unlike the LOCF model, which has smooth tails. This suggests that many prediction errors are nearly identical, likely resulting from nearly

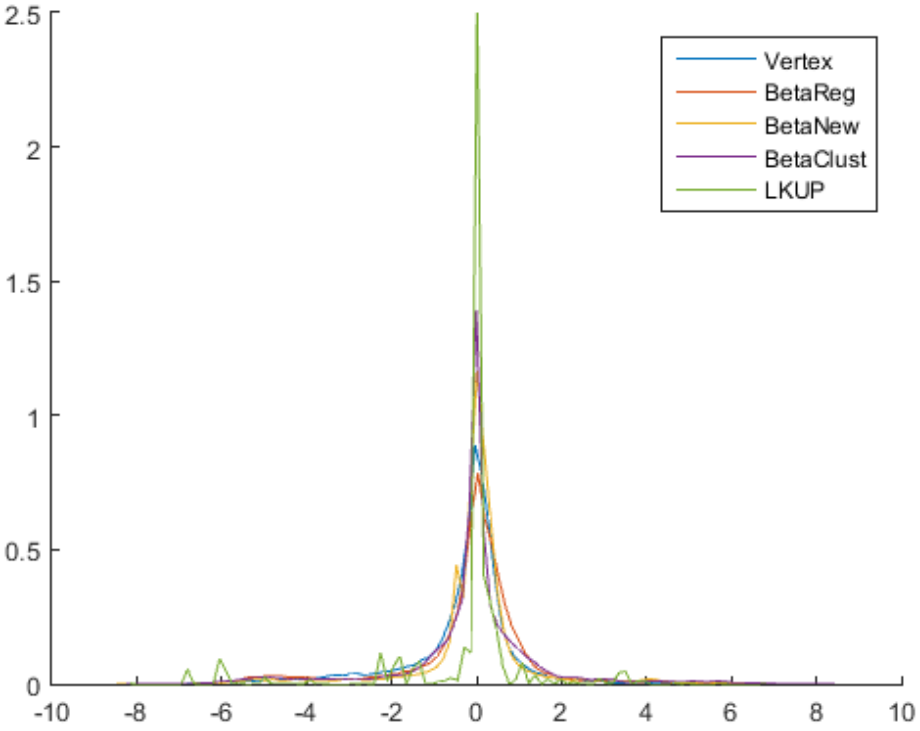

Figure 6.4.1: Density plot of estimated prediction errors for all models

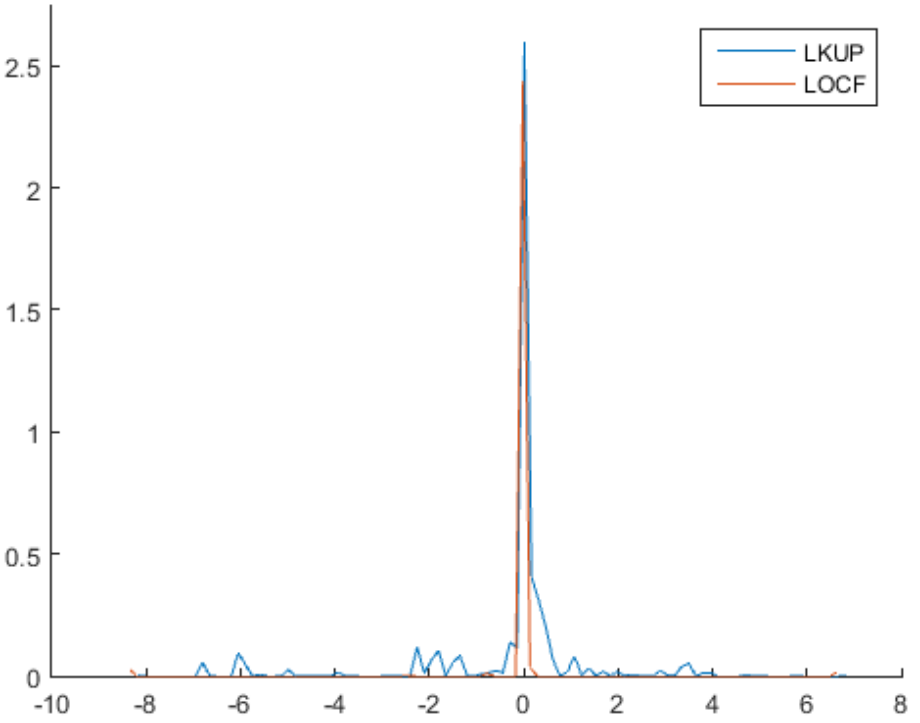

Figure 6.4.2: Density plot of estimated prediction errors for the Look-Up and LOCF methods

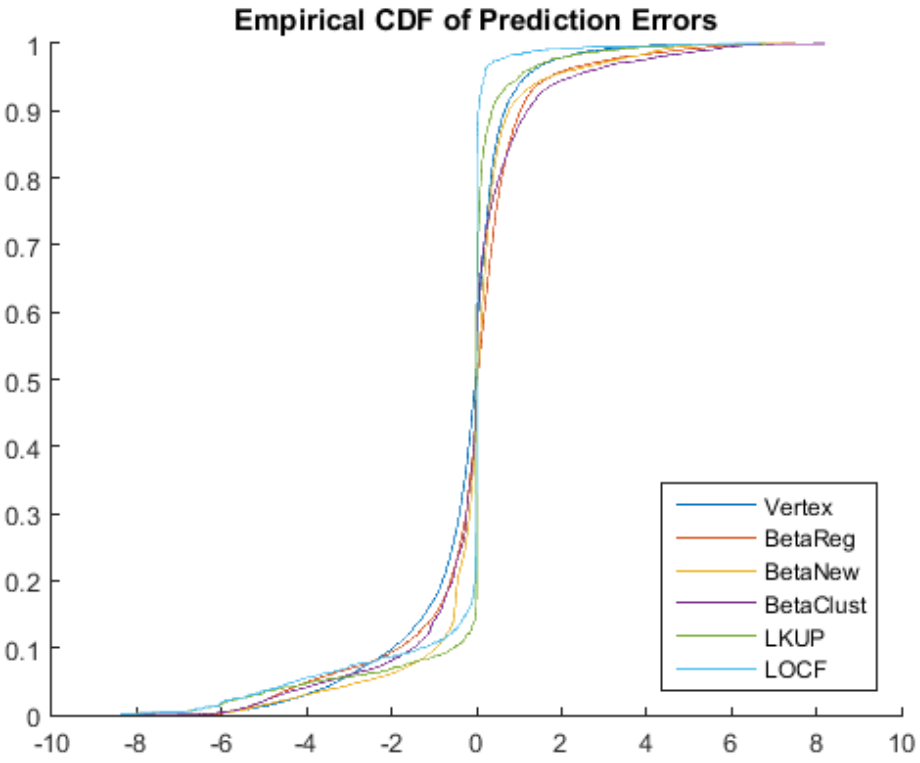

Figure 6.4.3: Empirical CDF of prediction errors for all models

identical predictions in similar situations. For instance, it may be that this model routinely generates a prediction of -7 when the next value is actually -10, resulting in many prediction errors around -3. This is likely due to the fact that the method makes predictions based on percentiles, and many percentiles have nearly identical values, resulting in nearly identical predictions.

To better understand the tail behavior of all of the models, we plot the empirical CDFs for the predictions errors produced by each model. The resulting plot in Figure 6.4.3 suggests that LOCF model has the shortest right-hand tail, resulting in smaller and fewer positive prediction errors. This is ideal, as a positive prediction error means that one predicted a low $\log P_c$ when in fact the next $\log P_c$ was higher, indicating an under-estimation of risk. Note that this implies that the practice of using the previous $\log_{10} P_c$ value for inference is more likely to overstate the risk than to understate it. We visualize exactly how much the risk is overstated later. The Beta Clustering model seems to generate the shortest left-hand tail.

Figures 6.4.4-6.4.10 plot the prediction errors vs. time and vs. actual $\log P_c$ value for each of the models. All of the models have smaller prediction errors from 6 days to TCA to 4 days to TCA than later time points. Prediction errors in this time range are smallest for the Look-Up and LOCF methods, where they are virtually all zero, implying near-perfect prediction. However, prediction in this time range is not particularly of interest, as decisions are usually made at 2 days to TCA (or later). It is interesting to note that the LOCF method produces errors largest in the positive direction, suggesting that in this time frame one may see a significant drop-off, but one almost never sees a jump from, say, a $\log_{10} P_c$ of -10 to -4.

We focus on the time range of 2 days to TCA. Most of the models produce more and larger negative prediction errors than positive prediction errors in this time range, suggesting

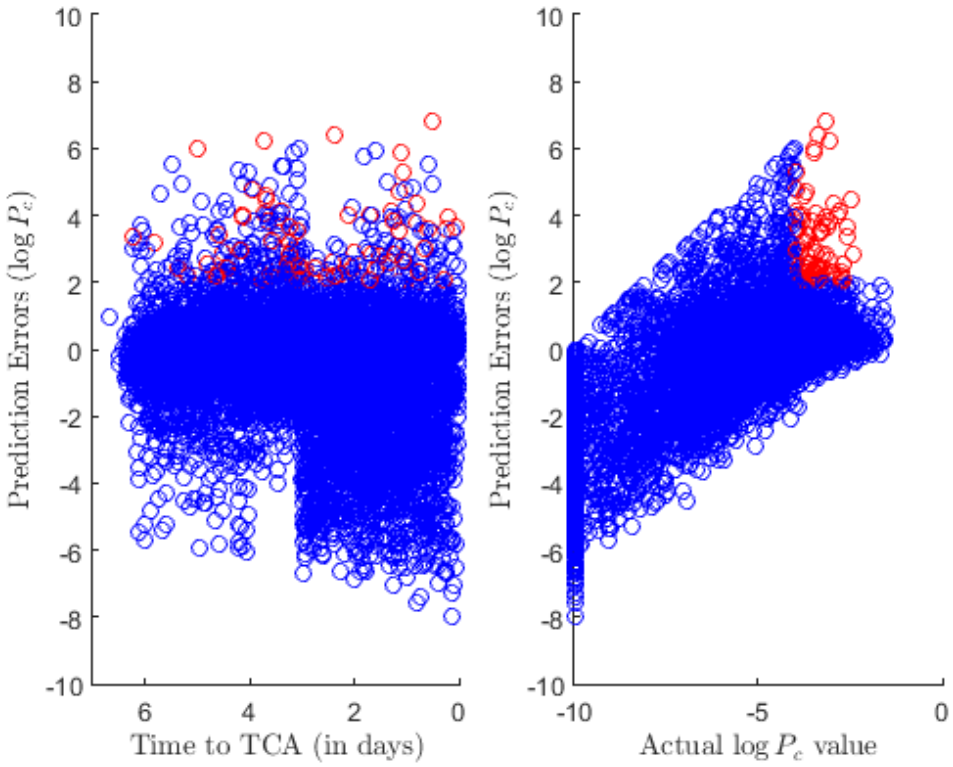

Figure 6.4.4: Vertex Model

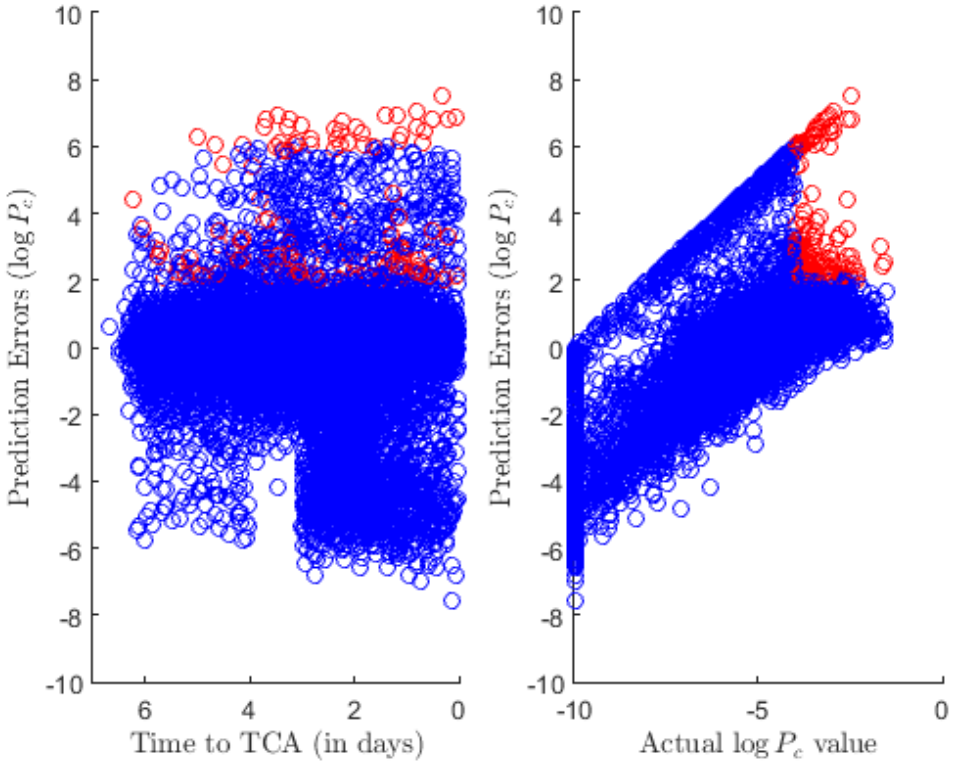

Figure 6.4.5: Beta Regression Model

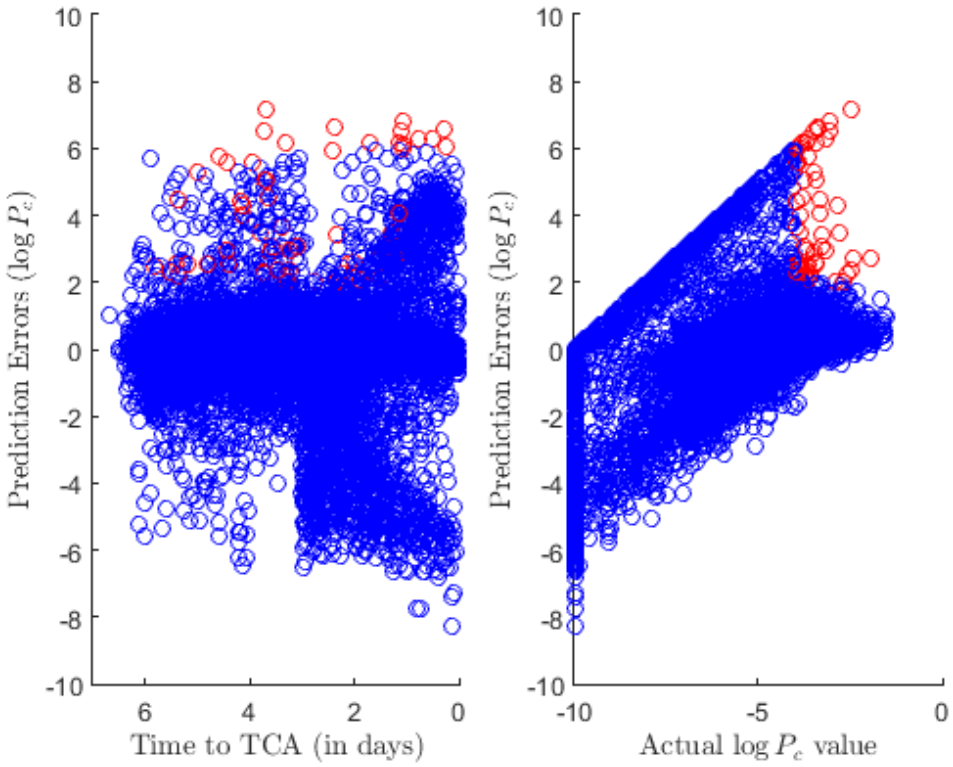

Figure 6.4.6: New Beta Regression Model

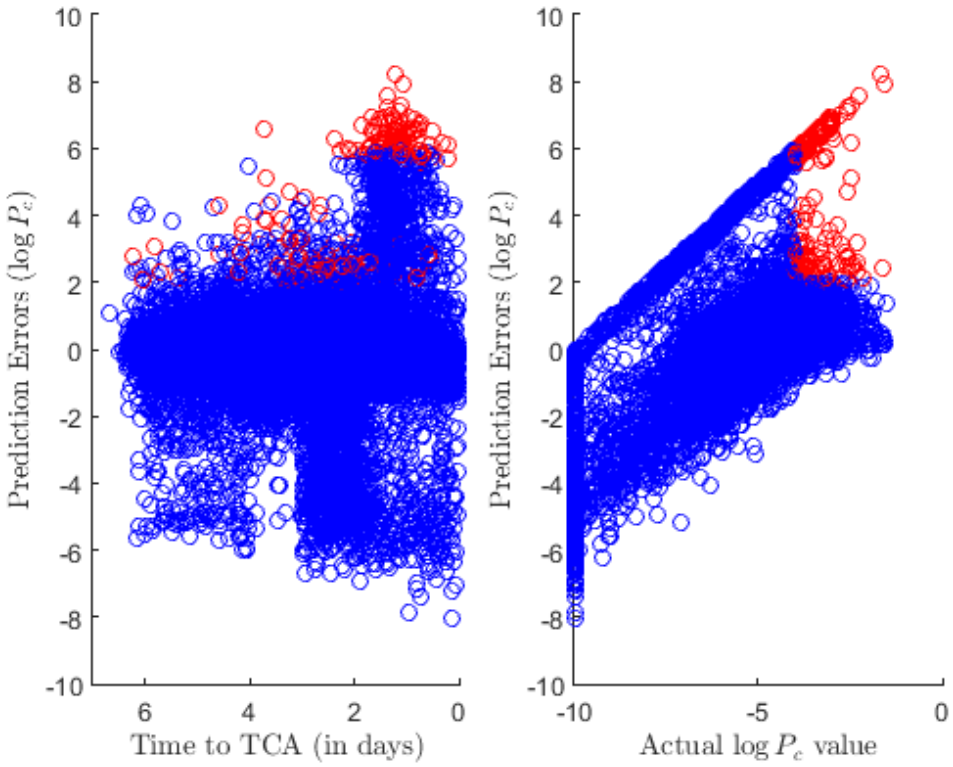

Figure 6.4.7: Beta Clustering Model

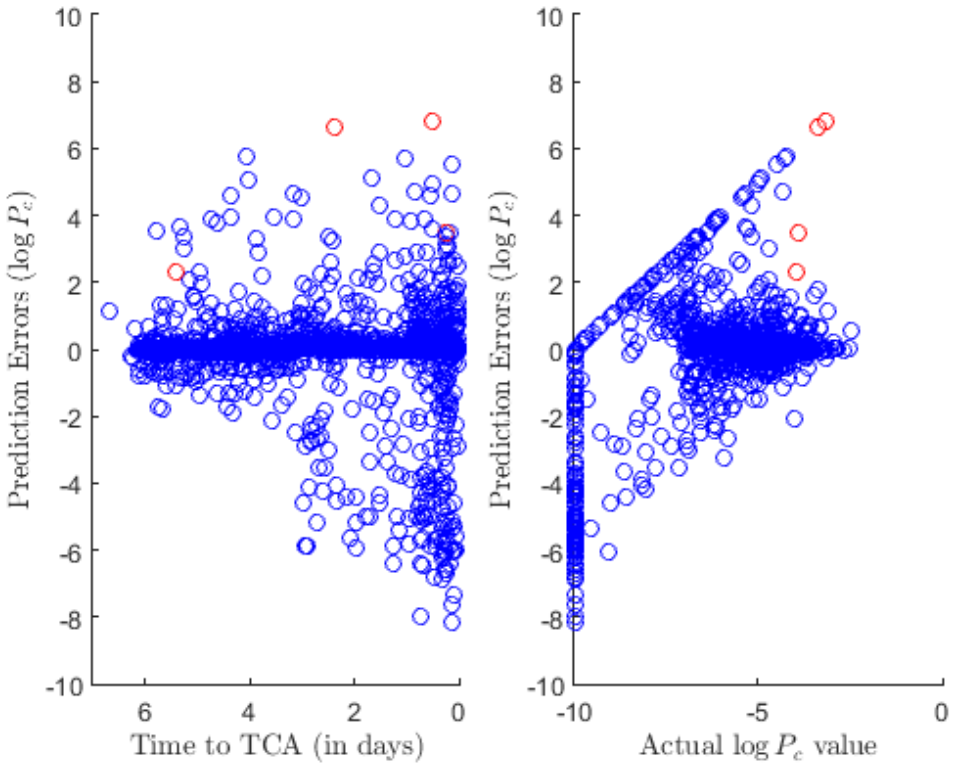

Figure 6.4.8: Look-Up Model

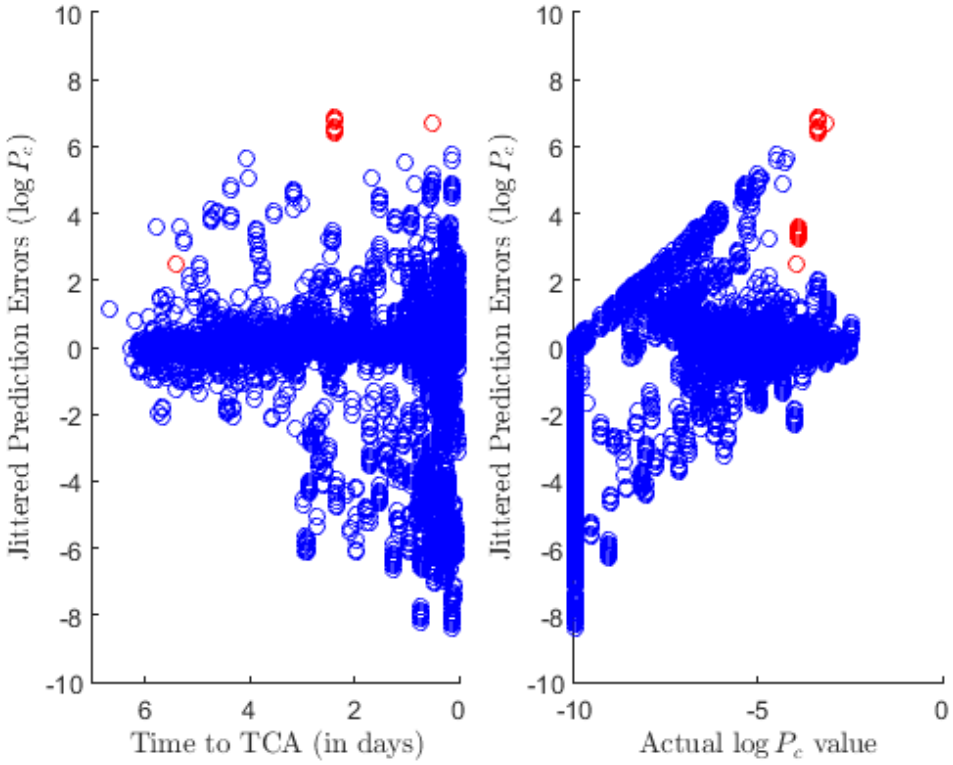

Figure 6.4.9: Look-Up Model (with jitter)

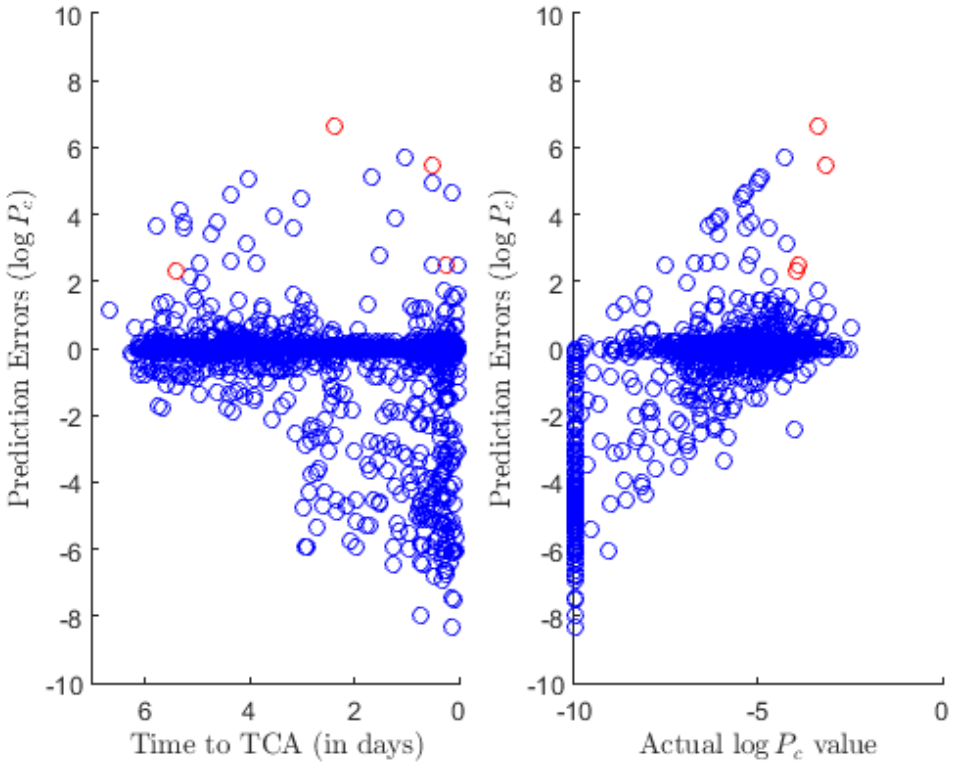

Figure 6.4.10: Last Observation Carried Forward

they err on the side of predicting “too high a risk". In general, this conservativeness is better operationally. This trend is **not** true for the Beta Clustering model, which produces a far greater number of positive prediction errors. This is likely due to predicting a $\log P_c = -10$ far more often than it actually happens. Perhaps this could be ameliorated by exploring some tuning of the priors in the model, though this is likely more onerous than practical for decision-making. In addition, this model takes far longer to run the other models, which only adds to the time one would need to implement it.

To highlight underestimates of high-risk events, we color in red all observations where the prediction error was 2 or greater and the actual $\log P_c$ value is -4 or higher. Interestingly, we see that the Vertex and New Beta Regression models have few of these points, whereas the Beta Regression and Beta Clustering models have noticeably more. In addition, many of these points are within 2 days to TCA for the latter two models. Hence, for erring conservatively in the “worst case scenarios", the Beta Regression and Beta Clustering models fare poorly. The best models in this regard are the Look-Up and LOCF models, which have few red points. In fact, Figures 6.4.4 and 6.4.10 suggest that for these high values, these models almost always produce relatively small prediction errors.

Lastly, Figure 6.4.8 is slightly misleading, as many of the prediction errors overlap. Figure 6.4.9 jitters these prediction errors, so that one may see the phenomenon seen earlier in the densities, where many prediction errors are nearly identical. This same feature holds for the LOCF model as well. Though these models produce prediction errors tightly centered around zero, they still produce a fair number of large prediction errors (many are 4 or greater in magnitude). Generally, predictions which are within one or two orders of magnitude are considered useful, so these models are not necessarily guaranteed to be operationally useful,

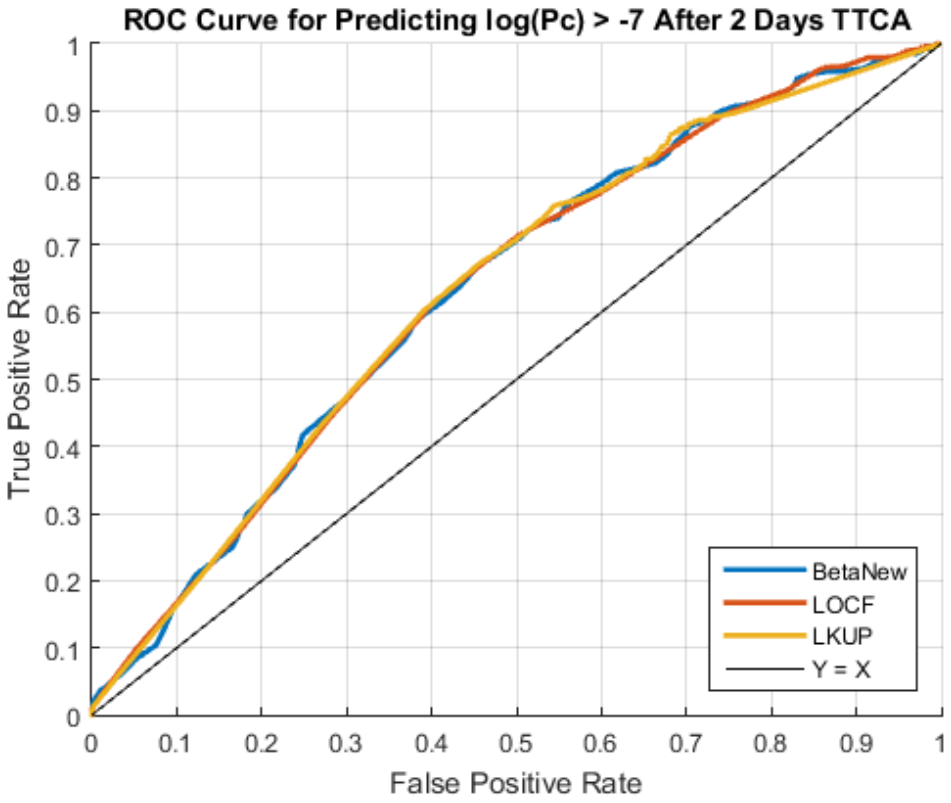

Figure 6.4.11: ROC curve for classifying final $\log_{10} P_c > -7$ (best models)

though they seem to be more promising than the others.

Figures 6.4.11-6.4.14 plot ROC curves for all of the models. These ROC curves present sensitivity and specificity for prediction if the final $\log P_c$ value will be above -7 after 2 days to TCA. Recall that these simulation results are for the Yellow data set, so that one is primarily interested in knowing whether the $\log P_c$ will be above -7, and if it isn't one can "downgrade" the risk to low (the Green group). However, one may also be interested in knowing that this is still a medium-risk event and not a high-risk event, so we also consider prediction of values above -4.

Figures 6.4.11 and 6.4.13 plot the ROC curves for the Look-Up, LOCF, and New Beta regression models, and the results are nearly identical. For clarity, the results for the remaining models are graphed in Figures 6.4.12 and 6.4.14.

Figure 6.4.12 suggests that the Beta Clustering model is best for determining whether the final $\log_{10} P_c$ value will be above -7 if one can accept a true positive rate of 60% or lower, as it generates the fewest false positives. As we are concerned with being highly certain that the value will be above -7, this feature is not particularly useful. For a high true positive rate (90% and above), the models perform quite similarly, though the Beta regression model is somewhat inferior here. Note that, though the Look-Up method has a line in this region, no actual values occurred in this region, as it is simply connecting the point at (0.70, 0.88) with (1, 1). This feature of the Look-Up method is troubling, as it implies that one cannot easily implement a model with a true positive rate of 95% (or higher). This may be due to the way the confidence intervals were constructed. We consider this in future work.

Figures 6.4.11-6.4.14 showed the operating characteristics for the models when determining if the *last* $\log_{10} P_c$ value would be above -7 or -4. Figures 6.4.15-6.4.18 plot these operating

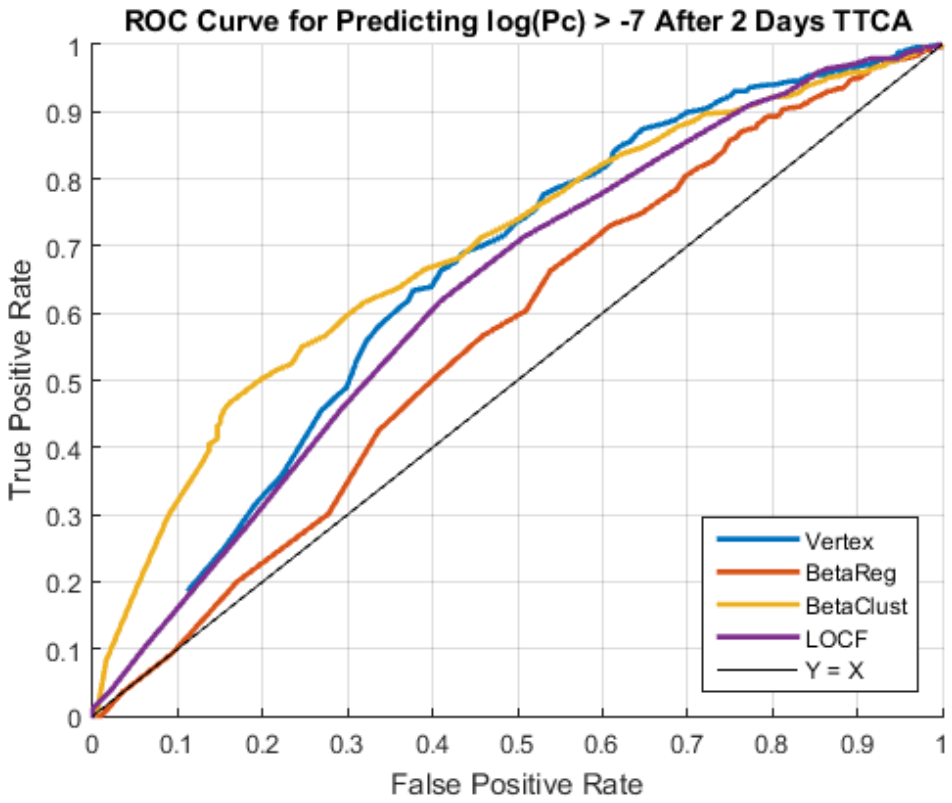

Figure 6.4.12: ROC curve for classifying final $\log_{10} P_c > -7$ (worst models)

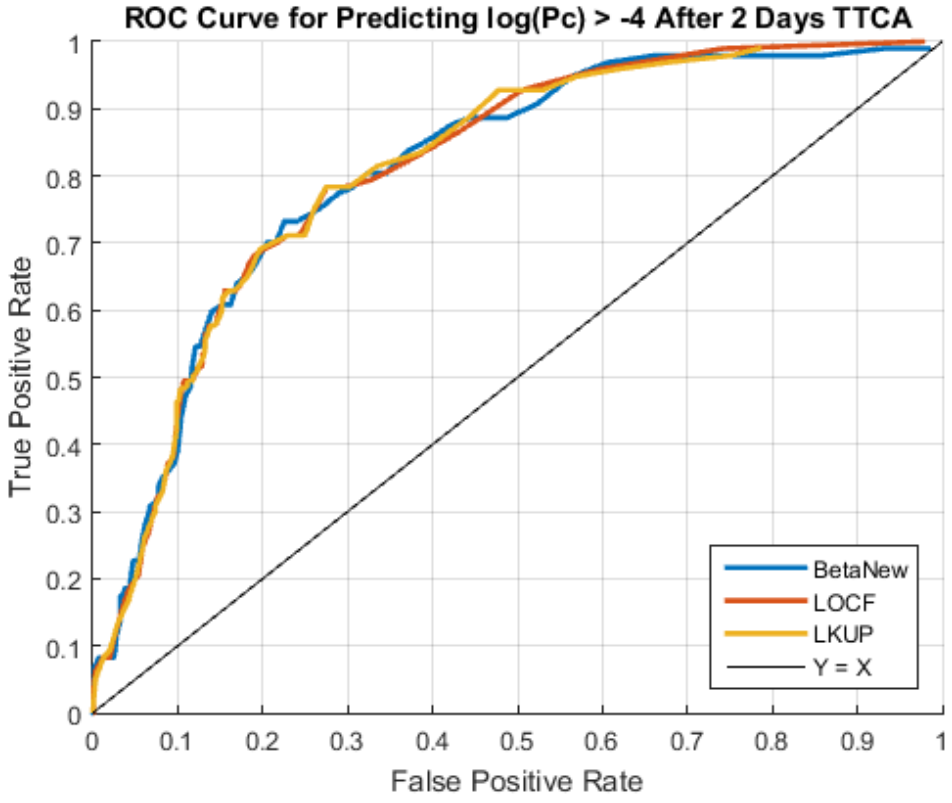

Figure 6.4.13: ROC curve for classifying final $\log_{10} P_c > -4$ (best models)

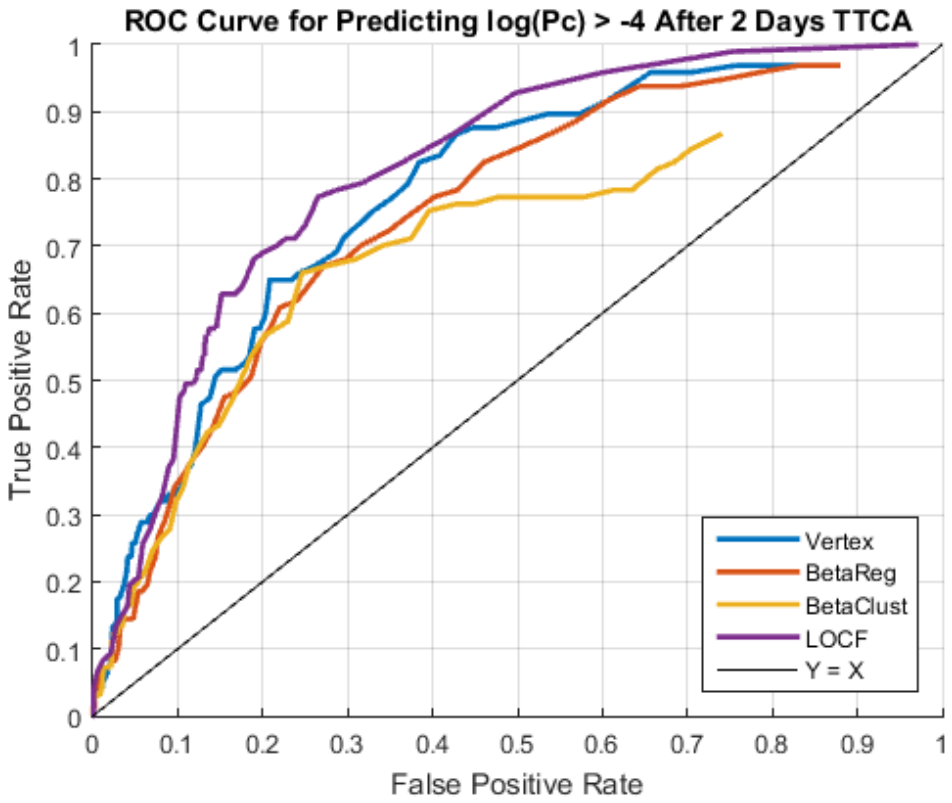

Figure 6.4.14: ROC curve for classifying final $\log_{10} P_c > -4$ (worst models)

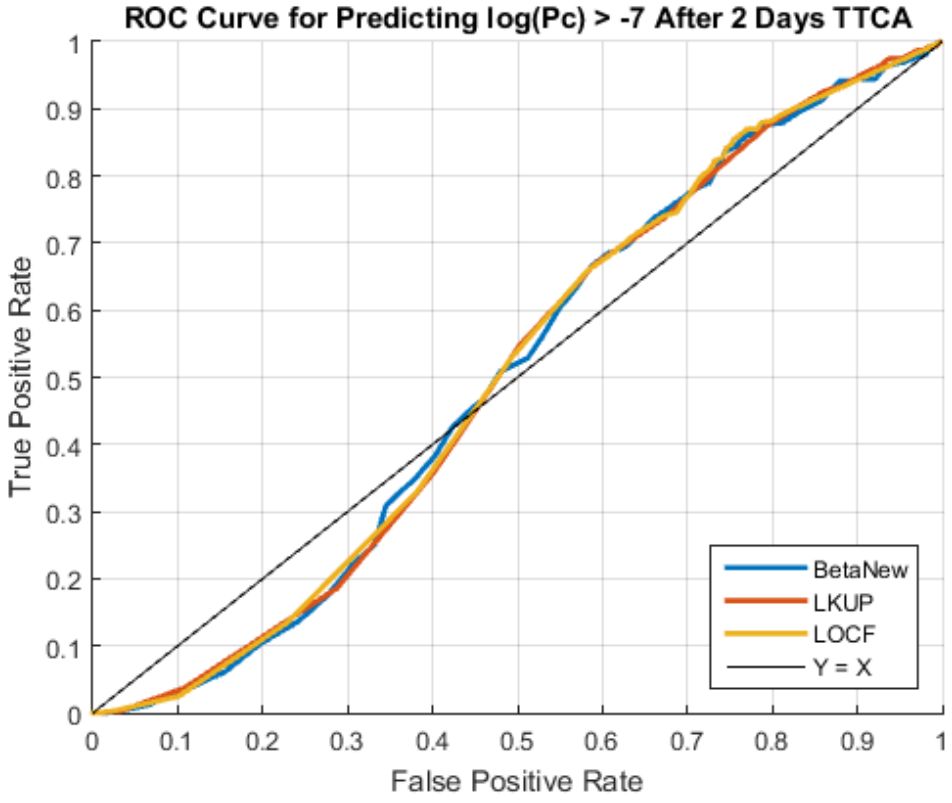

Figure 6.4.15: ROC curve for classifying next $\log_{10} P_c > -7$ (best models)

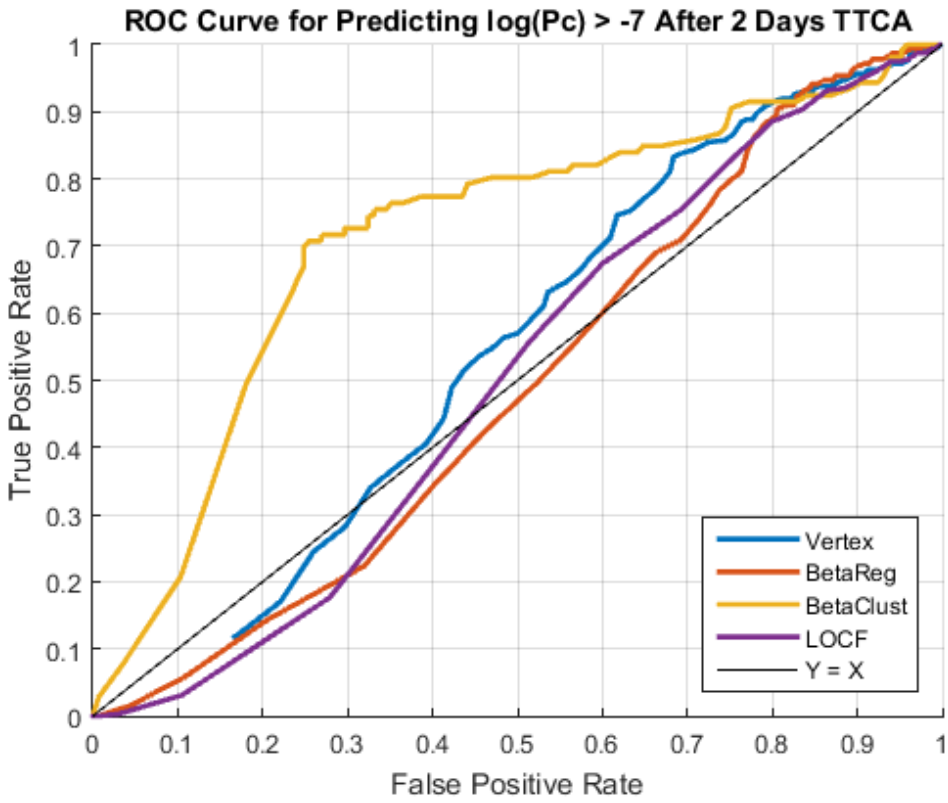

Figure 6.4.16: ROC curve for classifying next $\log_{10} P_c > -7$ (worst models)

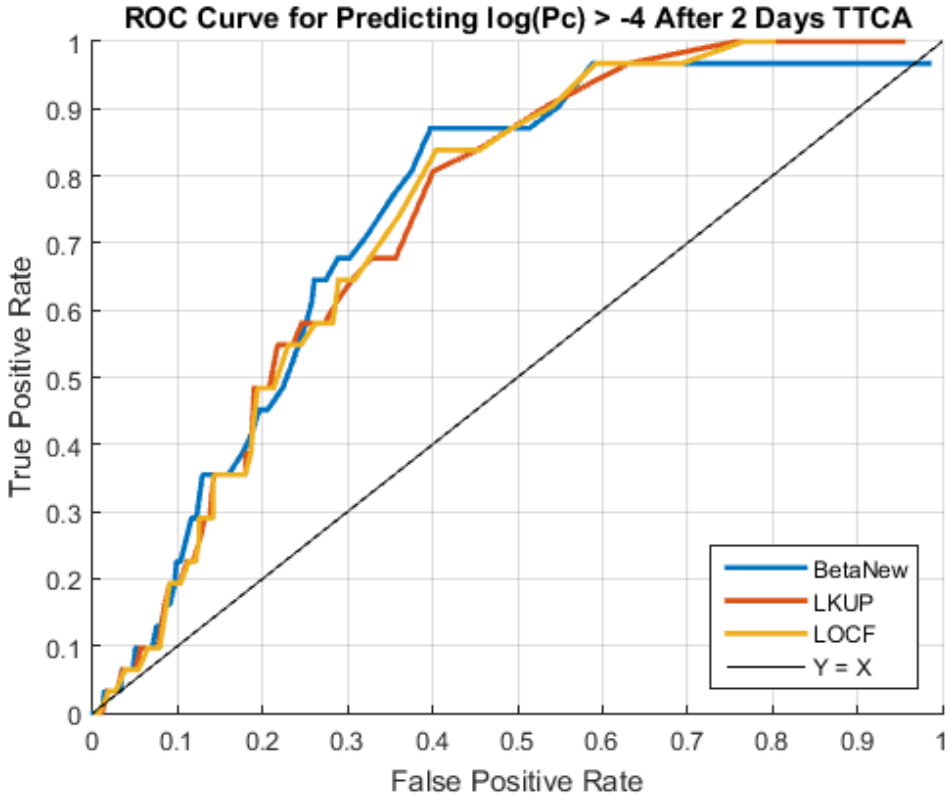

Figure 6.4.17: ROC curve for classifying next $\log_{10} P_c > -4$ (best models)

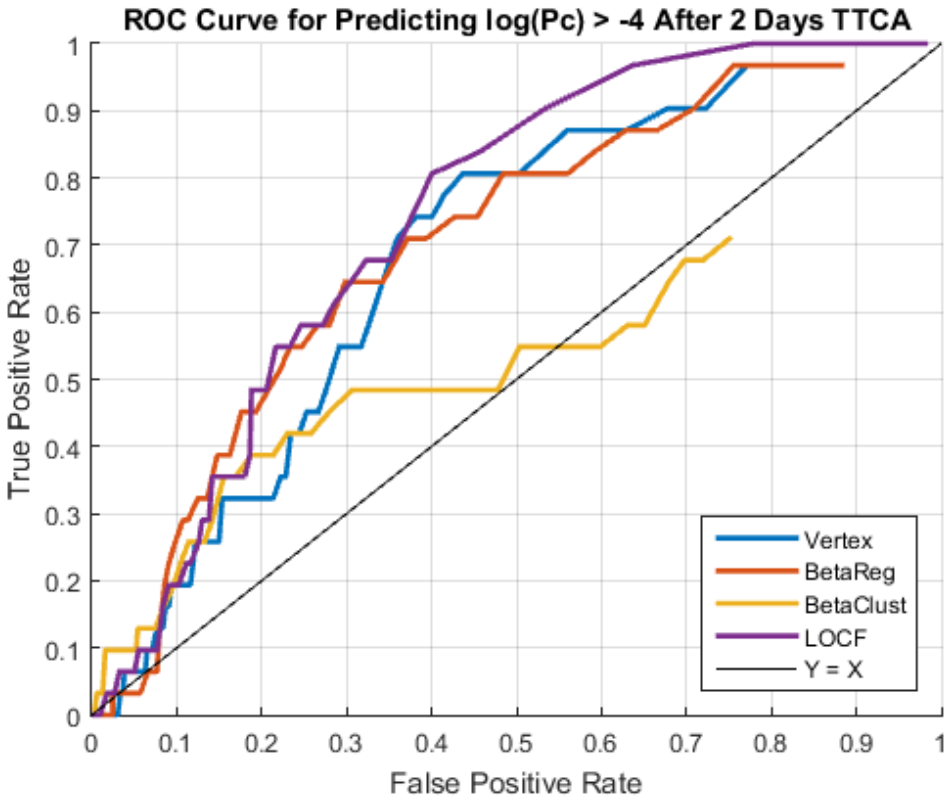

Figure 6.4.18: ROC curve for classifying next $\log_{10} P_c > -4$ (worst models)

characteristics for the models when the objective is to determine if the *next* $\log_{10} P_c$ value will be above -7 or -4 (given the next value occurs after 2 days to TCA). When the threshold is -7, many of these models are only slightly better than guessing, as evidenced by Figures 6.4.15 and 6.4.16. The Beta clustering model performs well here for lower true positive rates, though as mentioned above, this is not operationally useful. This model performs poorly when the threshold is -4. More investigation is needed to ascertain whether a model such as this could ever be operationally useful.

Overall, Figures 6.4.15-6.4.18 support the earlier conclusion that the three best models are the Look-Up, LOCF, and New Beta regression models. Not surprisingly, these models all share the same characteristic: they all utilize the most recent $\log_{10} P_c$ value. Though it's not particularly surprising that this value is useful, it *is* surprising that accounting for the trend over time doesn't seem to provide any noticeable improvement over simply using the previous value and ignoring the trend. This may be due to the fact that the trend is small and an event generally has only a few observations.

7 Conclusions and Future Work

We introduced a number of models for making predictions about future $\log P_c$ values. These models have various advantages and disadvantages, but all of the models produce unbiased predictions over time. Additionally, these predictions are within an order of magnitude at least 60% of the time, with the best models produce predictions within an order of magnitude 85% of the time. This is counter to the conventional wisdom that $\log P_c$ values cannot be

predicted and that we can make no claims about future behavior. Indeed, we can make relatively strong claims about most of their future values, though this is more difficult as one approaches TCA. Still, these predictions are good enough to create decision-making frameworks that are better than guessing, and reliable enough to give us a high degree of confidence in saying whether a future CDM will contain high-risk values or not.

These models all reveal various features about $\log P_c$ values which had not been discussed before. For instance, the Beta Clustering model shows that it is likely that there are two kinds of events, those which have very low variability and those which have very high variability. This information may be used in turn to diagnose what might be different about these events with low variability, and if perhaps the information for these events is simply more accurate. The Look-Up model suggests that the quantile of the $\log P_c$ value tells us a great deal about future values. Furthermore, it reveals that this holds true to a very high degree up to 4 days to TCA. And though the Vertex model is simple and does not utilize a longitudinal framework, it still has many nice properties. Fortunately, many simple models work well here, which eases computation time and interpretation.

It is worth noting that these models are simply an initial exploration into the problem of predicting $\log P_c$ values. Future work may hone these models to be more accurate, and ultimately to be more useful in the decision-making process. For instance, it seems clear from the ROC curves that some more exploration should be done in constructing confidence intervals for the Look-Up method. The Beta Clustering model suggests that perhaps identifying a particular cluster early will make prediction easier. More specifically, it may be worth exploring models which take into account not only the quantiles of the previous observations, but also the change in quantiles, as the low-variability cluster suggests the quantile does not change much over time. The New Beta Regression model suggests that a cubic trend is likely better for describing the mean of the $\log P_c$ values over time, as is supported by the discussion leading to the Beta Threshold model.

Due to the success of the Look-Up and LOCF models, it seems that future work should focus on non-parametric procedures. Though Beta regression is flexible, many of these features (such as the threshold feature) are more easily described through splines or quantiles. A sufficiently flexible non-parametric model which can account for quantile, time, and longitudinal observations may very well improve upon all of these models. Additionally, a non-parametric framework may more easily accommodate the identification of clusters.

In addition to focus on non-parametric models, future work may also consider other features in the data which might help to explain the variability. This may include the positional error covariance matrix, tracking information, or miss distance. In initial exploration, we found that many of these covariates are noisy and may add more noise than the amount of variability in $\log P_c$ they describe. However, there may be a simple way to incorporate one or more of these features into a model so that prediction is improved.

References

[1] Celestrak: Iridium 33/cosmos 2251 collision. http://celestrak.com/events/collision. Accessed: 2016-06-28.

[2] Maruthi R Akella and Kyle T Alfriend. Probability of collision between space objects. *Journal of Guidance, Control, and Dynamics*, 23(5):769–772, 2000.

[3] Salvatore Alfano. Relating position uncertainty to maximum conjunction probability©. 2005.

[4] J Antoch and Paul Janssen. Nonparametric regression m-quantiles. *Statistics & probability letters*, 8(4):355–362, 1989.

[5] William N. Barker. Astrodynamics concepts and terminology, omitron, inc.

[6] Pallab K Bhattacharya and Ashis K Gangopadhyay. Kernel and nearest-neighbor estimation of a conditional quantile. *The Annals of Statistics*, pages 1400–1415, 1990.

[7] Wagner Hugo Bonat, Paulo Justiniano Ribeiro Jr, and Walmes Marques Zeviani. Likelihood analysis for a class of beta mixed models. *Journal of Applied Statistics*, 42(2):252–266, 2015.

[8] Howard D Bondell, Brian J Reich, and Huixia Wang. Noncrossing quantile regression curve estimation. *Biometrika*, 97(4):825–838, 2010.

[9] Simone Borra and Agostino Di Ciaccio. Measuring the prediction error. a comparison of cross-validation, bootstrap and covariance penalty methods. *Computational statistics & data analysis*, 54(12):2976–2989, 2010.

[10] Prabir Burman. A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika*, 76(3):503–514, 1989.

[11] J Russell Carpenter, F Landis Markley, and Dara Gold. Sequential probability ratio test for collision avoidance maneuver decisions. *The Journal of the Astronautical Sciences*, 59(1-2):267–280, 2012.

[12] J Russell Carpenter, F Landis Markley, and Dara Gold. Wald sequential probability ratio test for analysis of orbital conjunction data. In *AIAA Guidance, Navigation, and Control (GNC) Conference*, 2013.

[13] J Russell Carpenter, FL Markley, KT Alfriend, C Wright, and J Arcido. Sequential probability ratio test for collision avoidance maneuver decisions based on a bank of norm-inequality-constrained epoch-state filters. In *AAS/AIAA Astrodynamics Specialist Conference,âĂİ Girdwood, AK*, 2011.

[14] Gilles Celeux, Florence Forbes, Christian P Robert, D Michael Titterington, et al. Deviance information criteria for missing data models. *Bayesian analysis*, 1(4):651–673, 2006.

[15] Edilberto Cepeda and Dani Gamerman. Bayesian methodology for modeling parameters in the two parameter exponential family. *Revista Estadística*, 57(168-169):93–105, 2005.

[16] E Cepeda-Cuervo. Modeling variability in generalized linear models. *Mathematics Institute, Universidade Federal do Rio de Janeiro*, 2001.

[17] F Kenneth Chan. *Spacecraft collision probability*. Aerospace Press El Segundo, CA, 2008.

[18] Ken Chan. Improved analytical expressions for computing spacecraft collision probabilities. *Advances in the Astronautical Sciences*, 114:1197–1216, 2003.

[19] Probal Chaudhuri et al. Nonparametric estimates of regression quantiles and their local bahadur representation. *The Annals of statistics*, 19(2):760–777, 1991.

[20] Herbert A David. Tables related to the normal distribution: A short history. *The American Statistician*, 59(4):309–311, 2005.

[21] Luc Devroye. Sample-based non-uniform random variate generation. In *Proceedings of the 18th conference on Winter simulation*, pages 260–265. ACM, 1986.

[22] Peter Diggle. *Analysis of longitudinal data*. Oxford University Press, 2002.

[23] Silvia Ferrari and Francisco Cribari-Neto. Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 31(7):799–815, 2004.

[24] Jorge I Figueroa-Zúñiga, Reinaldo B Arellano-Valle, and Silvia LP Ferrari. Mixed beta regression: a bayesian perspective. *Computational Statistics & Data Analysis*, 61:137–147, 2013.

[25] Ryan C Frigm, Joshua A Levi, and Dimitrios C Mantziaras. Assessment, planning, and execution considerations for conjunction risk assessment and mitigation operations. In *Proceedings of SpaceOps 2010 Conference: Delivering on the Dream, Huntsville, Alabama*, pages 25–30, 2010.

[26] Alan E Gelfand, Adrian FM Smith, and Tai-Ming Lee. Bayesian analysis of constrained parameter and truncated data problems using gibbs sampling. *Journal of the American Statistical Association*, 87(418):523–532, 1992.

[27] Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*, volume 2. Taylor & Francis, 2014.

[28] Xuming He. Quantile curves without crossing. *The American Statistician*, 51(2):186–192, 1997.

[29] MD Hejduk, D Plakalovic, ME Hametz, LK Newman, JC Ollivierre, BA Beaver, and RC Thompson. Launch cola operations: Examination of data products, procedures, and thresholds. In *Flight Dynamics (FD) Task Order 21 Technical Memorandum*, 2014.

[30] Wallace Hendricks and Roger Koenker. Hierarchical spline models for conditional quantiles and the demand for electricity. *Journal of the American statistical Association*, 87(417):58–68, 1992.

[31] Ajay Jasra, CC Holmes, and DA Stephens. Markov chain monte carlo methods and the label switching problem in bayesian mixture modeling. *Statistical Science*, pages 50–67, 2005.

[32] A Kelly and W Watson. Collision avoidance: Coordination of predicted conjunctions between nasa satellites and satellites of other countries. In *Advanced Maui Optical and Space Surveillance Technologies Conference*, volume 1, page 78, 2014.

[33] Michael G Kenward and Geert Molenberghs. Last observation carried forward: a crystal ball? *Journal of biopharmaceutical statistics*, 19(5):872–888, 2009.

[34] Donald J Kessler and Burton G Cour-Palais. Collision frequency of artificial satellites: The creation of a debris belt. *Journal of Geophysical Research: Space Physics*, 83(A6):2637–2646, 1978.

[35] Robert Kieschnick and Bruce D McCullough. Regression analysis of variates observed on (0, 1): percentages, proportions and fractions. *Statistical modelling*, 3(3):193–213, 2003.

[36] Roger Koenker. A note on l-estimates for linear models. *Statistics & probability letters*, 2(6):323–325, 1984.

[37] Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50, 1978.

[38] Nan M Laird and James H Ware. Random-effects models for longitudinal data. *Biometrics*, pages 963–974, 1982.

[39] Kenneth Lange. *Numerical analysis for statisticians*. Springer Science & Business Media, 2010.

[40] John A Nelder and RJ Baker. Generalized linear models. *Encyclopedia of Statistical Sciences*, 1972.

[41] Tereza Neocleous and Stephen Portnoy. On monotonicity of regression quantile functions. *Statistics & Probability Letters*, 78(10):1226–1229, 2008.

[42] Philip Paolino. Maximum likelihood estimation of models with beta-distributed dependent variables. *Political Analysis*, 9(4):325–346, 2001.

[43] Russell P Patera. General method for calculating satellite collision probability. *Journal of Guidance, Control, and Dynamics*, 24(4):716–722, 2001.

[44] Franco Peracchi. On estimating conditional quantiles and distribution functions. *Computational statistics & data analysis*, 38(4):433–447, 2002.

[45] Martyn Plummer et al. Jags: A program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing*, volume 124, page 125. Vienna, 2003.

[46] Stuart J Pocock. *Clinical trials: a practical approach*. John Wiley & Sons, 2013.

[47] Joel R Primack. Debris and future space activities. In *University of California at Santa Cruz. Presented at the Conference on Future Security in Space, at New Place (Southampton, England) May*, volume 28, page 29, 2002.

[48] Nalini Ravishanker and Dipak K Dey. *A first course in linear model theory*. CRC Press, 2001.

[49] Richard A Redner and Homer F Walker. Mixture densities, maximum likelihood and the em algorithm. *SIAM review*, 26(2):195–239, 1984.

[50] Chandan Saha and Michael P Jones. Bias in the last observation carried forward method under informative dropout. *Journal of Statistical Planning and Inference*, 139(2):246–255, 2009.

[51] M Samanta. Non-parametric estimation of conditional quantiles. *Statistics & Probability Letters*, 7(5):407–412, 1989.

[52] WF Sheppard. New tables of the probability integral. *Biometrika*, 2(2):174–190, 1903.

[53] Alexandre B Simas, Wagner Barreto-Souza, and Andréa V Rocha. Improved estimators for a general class of beta regression models. *Computational Statistics & Data Analysis*, 54(2):348–366, 2010.

[54] David J Spiegelhalter, Nicola G Best, Bradley P Carlin, and Angelika Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639, 2002.

[55] Robert A Stine. Bootstrap prediction intervals for regression. *Journal of the American Statistical Association*, 80(392):1026–1031, 1985.

[56] Robert Stiratelli, Nan Laird, and James H Ware. Random-effects models for serial observations with binary response. *Biometrics*, pages 961–971, 1984.

[57] Charles J Stone. Consistent nonparametric regression. *The annals of statistics*, pages 595–620, 1977.

[58] Winfried Stute. Conditional empirical processes. *The Annals of Statistics*, pages 638–647, 1986.

[59] Young K Truong. Asymptotic properties of kernel estimators based on local medians. *The Annals of Statistics*, pages 606–617, 1989.

[60] Jay Verkuilen and Michael Smithson. Mixed and mixture regression models for continuous bounded responses using the beta distribution. *Journal of Educational and Behavioral Statistics*, 37(1):82–113, 2012.

[61] Yichao Wu and Yufeng Liu. Stepwise multiple quantile regression estimation using non-crossing constraints. *Statistics and its Interface*, 2:299–310, 2009.

[62] Xiaoli Xu and Yongqing Xiong. A method for calculating collision probability between space objects. *arXiv preprint arXiv:1311.7216*, 2013.

[63] Fang Yao, Hans-Georg Müller, and Jane-Ling Wang. Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100(470):577–590, 2005.