Kalman Filters for Forecasting Open-Ocean White Shipping Location

Alyson R. Grassi and Hayden M. Thomas

ABSTRACT

Merchant vessels travel across the ocean daily to deliver goods and transport cargo or passengers. Understanding the forecasted locations of these vessels is important for many reasons, including collision avoidance. Currently, their captains rely on radar, a global positioning system (GPS) satellite fix, and the Automatic Identification System (AIS) to maintain timely awareness of their surroundings. This article describes a Johns Hopkins University Applied Physics Laboratory (APL) team's research into using a Kalman filter to improve forecasts of vessels' locations. When provided historical geospatial data that contain uncertainties, the Kalman filter algorithm provides a means to estimate future locations of moving objects. The APL team confirmed that when using GPS and AIS data, the Kalman filter forecasting tool can predict the future location of a vessel 90% of the time within 15 nautical miles for 12 h into the future.

INTRODUCTION

Thousands of vessels travel across the ocean every day to deliver goods ranging from fine cheese and wine to luxurious cars. Collecting and analyzing these vessels' location and speed data enables a deeper understanding of their tracks. With a measurable understanding of vessels' movements, analysts can start to form future predictions, with various levels of certainty, on where a ship could be several hours or days into the future. Predicting the future path of a vessel can be an important tool for preventing open-ocean collisions or aiding in the interception of potentially nefarious actors. One way to calculate the future location of a vessel is with a Kalman filter. The Kalman filter provides a dynamic system and estimation approach to predict the future location of a vessel. This article discusses the data analysis method an APL team used to pull the necessary variables for updating the Kalman filter algorithm to predict the location of a vessel several hours into the future and how the team assessed the algorithm's performance and viability.

KALMAN FILTERS

A Kalman filter is a common algorithm for estimating unknown variables that have inherent uncertainties or errors in their measurements and then using those data to predict future states of the system by estimating a joint probability distribution over time. Typically, the system in question involves a moving target, and the analyst has data from the moving object and equations that describe the motion of the object, known as a state space model.

The prediction relies heavily on the understanding of the object's movement (the state equations) and the historical data. It is important to note that the state equations will not perfectly predict the next state and will thus contain errors in the estimate (referred to as the process error or the process noise). It is also important to note that the collected data will contain noisy measurements that will not provide the precise location of the object (referred to as the measurement noise). The measurement noise typically comes from the inaccuracy of the tool used to measure the data. This could be the inaccuracy of a radar for positioning/velocity or the inaccuracy of a global positioning system (GPS). The process noise results from how the dynamic system (the moving object) will not follow the exact movement expected in the state equations. The system will randomly deviate from the expected movement.

In its simplest form, the Kalman filter applies the best known data to the state equations to update the state estimate. From there, the model continues to update the state estimate, considering the measurement and process noise, until all the known data have been used. Then, the prediction portion begins, using the previously tuned state estimate to calculate the next state of the system for each time step thereafter.

Implementation for White Shipping Forecasting

A Kalman filter is a favorable tool to forecast white (nonmilitary commercial vessels) shipping because it does not require or assume uniform periodicity between samples. Automatic Identification System (AIS) signals report oceangoing vessel position every 10–30 s, but satellites and other vessels in the region do not necessarily receive all reported positions, leaving data gaps beyond 5 h at sea. Vessels that receive satellite AIS have the advantage of knowing the location of vessels beyond their radar horizon but still have these data gaps.

Forecasting the future location of a vessel requires data fields for time, latitude, and longitude. Additional fields, such as course and speed, provide a more accurate noise calculation and therefore provide a better prediction for the vessel's location. When course and speed are not available, the values for time, latitude, and longitude are used to derive the course and average speed.

GPS and AIS data are collected and stored by the company MarineTraffic (www.marinetraffic.com). MarineTraffic's global white shipping data can be viewed at no cost via graphics on its website or purchased for personal use. The company receives the satellite AIS data from multiple sources and compiles the data into its own database. The data contain columns of unique ship identifiers, time stamps, locations, courses, and speeds. The APL team used these data in the development of the Kalman filter algorithm described in this article to establish variables and to aid in the prediction of vessels.

Kalman Filter Setup

There are five steps to the Kalman filter process: (1) initialize/update the matrices and vectors, (2) extrapolate the next state, (3) calculate the measurement values, (4) update the next state estimate based on the measurement, and (5) update the estimate uncertainty.

There are three vectors and six matrices to initialize. The state vector ($\mathbf{x}$) is the initial state of the system. The estimate of the state vector ($\hat{\boldsymbol{x}}$) is the estimate of the initial state. The control vector ($\mathbf{u}$) is a measurable input to the system. The observation matrix ($\mathbf{H}$) transforms the state vector values into measurement values. The measurement covariance matrix ($\mathbf{R}$) is the covariance of the measurement noise within the data. The process covariance matrix ($\mathbf{Q}$) is the covariance of the state equations and their relation to each other. The transition matrix ($\mathbf{F}$) defines how much of the next state and uncertainty values are related to each other. The control matrix ($\mathbf{G}$) defines the impact the control vector has on the next state. The estimate uncertainty matrix ($\mathbf{P}$) defines the uncertainties of the estimated state variables. Each value feeds into the calculations of the Kalman filter.[1]

The Kalman filter extrapolates the next state and uncertainty by

$$\hat{\boldsymbol{x}}_{n+1,n} = \boldsymbol{F}\hat{\boldsymbol{x}}_{n,n} + \boldsymbol{G}\boldsymbol{u}$$

$$\boldsymbol{P}_{n+1,n} = \boldsymbol{F}\boldsymbol{P}_{n,n}\boldsymbol{F}^{T} + \boldsymbol{Q},$$

where $\hat{\boldsymbol{x}}_{n+1,n}$ is the uncorrected estimate of the state at time step $n + 1$; $\hat{\boldsymbol{x}}_{n,n}$ is the estimate of the current state at time step n; $\boldsymbol{P}_{n+1,n}$ is the update of the uncertainty matrix; and $\boldsymbol{P}_{n,n}$ is the current uncertainty matrix.

This creates an uncorrected prediction of the next state. To correct the next state prediction, a Kalman gain and the measurement value are required. The Kalman gain seeks to minimize the estimate variance,[1]

$$\boldsymbol{K} = \mathbf{P}_{n,n-1}\boldsymbol{H}^{T}(\boldsymbol{H}\boldsymbol{P}_{n,n-1}\boldsymbol{H}^{T} + \mathbf{R})^{-1},$$

where $\boldsymbol{K}$ is the Kalman gain matrix, and $\mathbf{P}_{n,n-1}$ is the updated uncertainty matrix ($\boldsymbol{P}_{n+1,n}$ calculated above).

The measurement value calculates the estimated output based on the actual or estimated input data,

$$\hat{\boldsymbol{y}}_{n} = \boldsymbol{H}\boldsymbol{x}_{n},$$

where $\hat{\boldsymbol{y}}_{n}$ is the measurement value at the current time step, and $\boldsymbol{x}_{n}$ is the actual state or estimated state at the current time step.

The actual state is the received data at that time step, and the estimated input data is

$$\boldsymbol{x}_{n} = \hat{\boldsymbol{x}}_{n+1,n} + \boldsymbol{N}(\boldsymbol{Q}),$$

where $\mathbf{N}(\mathbf{Q})$ is the Gaussian distribution with variance $\mathbf{Q}$.

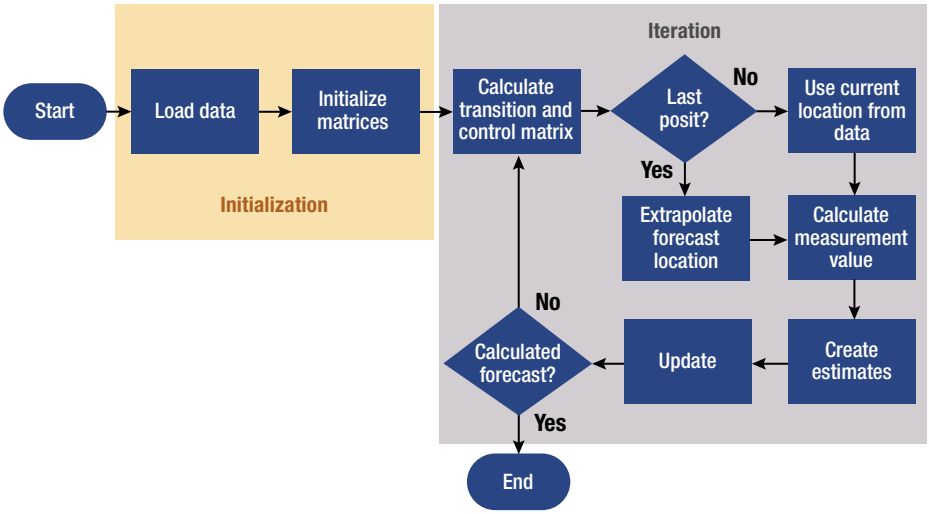


Figure 1. Kalman filter forecasting diagram.

Now it is possible to correct the estimate and the estimate uncertainty by

$$\hat{\boldsymbol{x}}_{n,n} = \hat{\boldsymbol{x}}_{n,n-1} + \boldsymbol{K}(\hat{\boldsymbol{y}}_n - \boldsymbol{H}\hat{\boldsymbol{x}}_{n,n-1})$$

$$\boldsymbol{P}_{n,n} = (\boldsymbol{I} - \boldsymbol{K}_n\boldsymbol{H})\boldsymbol{P}_{n,n-1}(\boldsymbol{I} - \boldsymbol{K}\boldsymbol{H})^{\boldsymbol{T}} + \boldsymbol{K}\boldsymbol{R}\boldsymbol{K}^{\boldsymbol{T}},$$

where $\hat{\boldsymbol{x}}_{n,n-1}$ is the updated uncorrected state estimate ($\hat{\boldsymbol{x}}_{n+1,n}$), and $\boldsymbol{P}_{n,n}$ is the corrected uncertainty estimate.

Variables and Assumptions

The general flow of the white shipping Kalman filter forecast is shown in Figure 1. Some assumptions and calculations are necessary to generate the forecast.

1. The observation matrix (**H**) for the output is the 3 × 3 identity matrix.
2. The control matrix (**G**) and the control vector (**u**) are set to zero.
3. The initialized estimation uncertainty $\boldsymbol{P}_{0,0}$ is a 3 × 3 null matrix. This variable evolves as the Kalman filter updates each iteration.
4. The initialized state vector ($\boldsymbol{x}_{0,0}$) and corrected estimate vector ($\hat{\boldsymbol{x}}_{0,0}$) values are the first longitude, latitude, and speed values of the data set.
5. From calculations and assumptions, the measurement (**R**) and process variance (**Q**) matrices are

$$\boldsymbol{R} = \begin{bmatrix} 3.2 \times 10^{-5}(°)^2 & 0 & 0 \\ 0 & 1.3 \times 10^{-5}(°)^2 & 0 \\ 0 & 0 & 9.7 \times 10^{-4}(°/hr)^2 \end{bmatrix} \begin{bmatrix} Longitude \\ Latitude \\ Speed \end{bmatrix}$$

$$\boldsymbol{Q} = \begin{bmatrix} 3.2 \times 10^{-6}(°)^2 & 0 & 0 \\ 0 & 1.3 \times 10^{-6}(°)^2 & 0 \\ 0 & 0 & 9.7 \times 10^{-5}(°/hr)^2 \end{bmatrix} \begin{bmatrix} Longitude \\ Latitude \\ Speed \end{bmatrix}.$$

The calculations for the noise are described in the next section.

6. Each Kalman filter iteration updates the transition matrix (**F**) based on the most recent reported course (Cse_{n-1}) and the elapsed time (Δt),

$$\boldsymbol{F} = \begin{bmatrix} 1 & 0 & \cos(\theta)\,\Delta t \sec(lat_{n-1}) \\ 0 & 1 & \sin(\theta)\,\Delta t \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} Longitude \\ Latitude \\ Speed \end{bmatrix}$$

$$\theta = \begin{cases} 90 - Cse_{n-1}, 0 \le Cse < 90, 0 \le \theta < 90 \\ 450 - Cse_{n-1}, 270 \le Cse < 360, 90 \le \theta < 180 \\ |Cse_{n-1} - 450|, 180 \le Cse < 270, 180 \le \theta < 270 \\ |Cse_{n-1} - 450|, 90 \le Cse < 180, 270 \le \theta < 360 \end{cases}$$

$$\Delta t = t_n - t_{n-1},$$

where Cse_{n-1} is the course from the previous time stamp where North is 0°, and lat_{n-1} is the previous recorded latitude.

To keep the calculations simple, latitude and longitude were calculated using polar calculations and a basic first-order approximation.

Calculating Uncertainties

As previously mentioned, the Kalman filter prediction algorithm relies on understanding both the measurement noise and the process noise. The measurement noise values determine the uncertainties contained within the MarineTraffic recorded data. The process noise determines the uncertainty in the model. Thus, the first step involves calculating the variance of the error in speed and the variance of the error in location.

Using the data provided and the structure above, these calculated errors account for three noise value inputs to the Kalman filter: speed, longitude, and latitude.

Speed Uncertainty

First, the speed error calculation uses the recorded location and time data. Distance and delta time were calculated from one data point to the next. Then, the expected speed at each location is derived from these distances and times. The error in speed is the absolute value of the stored speeds in the White Shipping data file (spd) minus the calculated expected list of speeds.

It is important to note that this method only applies to $n - 1$ speeds when there are n total posits, individual reported vessel positions, because two sets of locations and times are needed to calculate the approximate arrival speed at the second point. Thus, an error cannot be calculated for the first posit in the list.

This process creates a list of errors for the expected calculated speed and the observed recorded speed. These lists can break down per track, per region, etc. to enable calculation of the overall variance in these errors for specific regional or vessel interests. Ultimately, the variance in the list of speed errors for a given track is the input to the process noise matrix for speed.

The speed measurement noise value is 9.7×10^{-4} (degree/h)^2; this has a standard deviation of 1.9 knots. By trial and error, the selected process noise for speed is 4.8×10^{-5} (degree/h)^2, which is a standard deviation of 2.3 knots.

Latitude and Longitude Uncertainty

One way to calculate locational errors is by calculating the speed when the records are zero in the data. This implies that the vessel is not moving or no speed data exist for this time step. The assumption is that the speed recorded at time t is the speed of the vessel at time t when it arrived at the recorded location. This understanding is necessary because, similarly to the speed error calculations, computed locational errors exist for $n - 1$ recorded locations for n total posits.

The analysis evaluates all latitude and longitude distances separately and identifies when the observed recorded speed is zero. Ultimately, there are two lists that represent the error in latitude and longitude, separately, of a GPS fix. Then, the variance

Table 1. February 1–4, 2019, median distance error

Data Resolution (h)	Total posits	Median Distance Error (Nautical Miles) at Forecast to Future (h)					
		1	2	4	8	12	24
1	23	0.51	0.80	1.16	2.08	2.99	7.77
2	13	0.46	0.80	1.35	2.00	2.97	7.05
4	5	0.54	0.85	1.25	2.24	2.92	7.42
8	3	0.65	0.86	1.19	2.07	2.27	6.72
12	2	0.89	1.06	1.49	2.65	3.57	7.51

Table 2. October 1–4, 2019, median distance error

Data Resolution (h)	Total posits	Median Distance Error (Nautical Miles) at Forecast to Future (h)					
		1	2	4	8	12	24
1	23	0.31	0.51	1.09	2.35	3.89	10.12
2	13	0.35	0.51	1.07	2.44	3.90	10.13
4	5	0.33	0.54	0.97	2.38	3.74	9.45
8	3	0.38	0.66	0.95	2.24	3.55	10.92
12	2	0.47	0.78	1.44	2.41	4.29	10.77

calculated for each list is used as the representative input to the process noise matrix for latitude and longitude.

The latitude and longitude measurement noise values are 1.3×10^{-5} (degree)^2 and 3.2×10^{-5} (degree)^2, respectively. The standard deviations are 0.22 nautical miles and 0.34 nautical miles, respectively.

Another option for calculating the locational error requires projecting forward a proposed location based on known course and speed and comparing this with the recorded location; however, this would incorporate the measurement errors of both course and speed. Thus, by trying to determine the potential error of either a latitude measurement or a longitude measurement, the distance traveled between posits is recorded as an individual error in just longitude or latitude and only incorporates potential errors in recorded speed.

The process noise for the latitude and longitude selected by trial and error is 2×10^{-5} and 4.8×10^{-5}, respectively. The standard deviation is 0.27 nautical miles and 0.42 nautical miles, respectively.

RESULTS

The metric used to determine the accuracy of the forecast for this analysis is the median (50th percentile) and the 90th percentile positional track error. Positional track error is the distance between the forecasted value and the nearest interpolated track point shown as an absolute error. The data used in this analysis are from two different MarineTraffic data sets. The first includes data from February 1 to 4, 2019, and the second is from October 1 to 4, 2019. The data sets have a track sample size of 138 and 118, respectively. Both sets include data for oceangoing vessels that are more than 400 nautical miles from land, with tracks that have a minimum of at least 2 days of data at sea, and have a

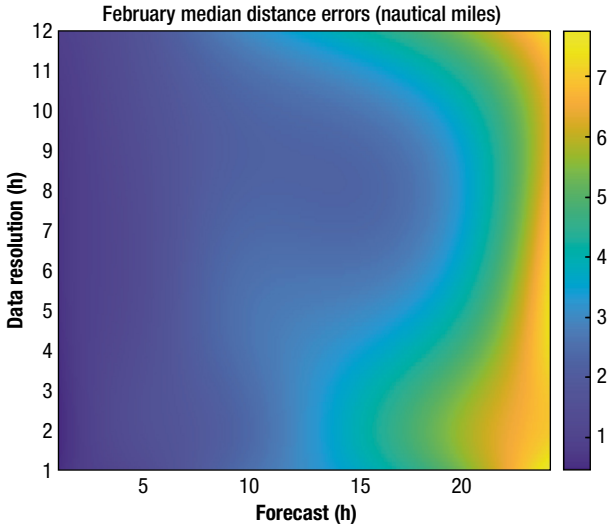


Figure 2. February 1–4, 2019, 50th percentile distance error of track from forecasted position.

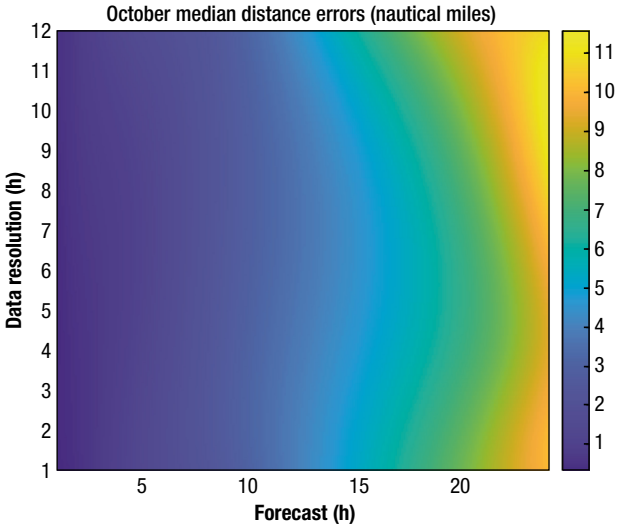


Figure 3. October 1–4, 2019, 50th percentile distance error of track from forecasted position.

Table 3. February 1–4, 2019, 90th percentile error

Data Resolution (h)	Total posits	90% Distance Error (Nautical Miles) at Forecast to Future (h)					
		1	2	4	8	12	24
1	23	1.65	2.10	3.65	9.20	14.00	36.90
2	13	1.35	2.15	3.60	8.45	13.25	29.35
4	5	2.00	2.30	3.65	8.30	13.45	29.95
8	3	2.15	2.55	3.70	8.20	13.85	38.80
12	2	2.45	2.65	4.20	9.00	14.15	31.25

Table 4. October 1–4, 2019, 90th percentile error

Data Resolution (h)	Total posits	90% Distance Error (Nautical Miles) at Forecast to Future (h)					
		1	2	4	8	12	24
1	23	0.75	1.40	2.95	6.25	10.70	37.80
2	13	0.80	1.45	3.10	6.50	10.65	37.40
4	5	0.85	1.60	2.95	6.00	10.75	37.85
8	3	1.05	1.80	2.70	7.40	11.95	39.75
12	2	1.30	2.05	3.45	7.00	11.95	40.50

time difference between all positions of 2 h or less. Only 1 day of data is used for the actual state (**x**), and the forecast is compared with some time into the future within the time frame.

Calculations of the median (50th percentile) positional track error are shown in Tables 1 and 2 and Figures 2 and 3, and calculations of the 90th percentile positional error are shown in Tables 3 and 4 and Figures 4 and 5. The leftmost columns of the tables are the data resolution (time between each time stamp) and the total number of posits available. The top row indicates the number of hours that the filter attempted to forecast into the future. The more available posits in the data set, the more accurate the near-term forecast. As expected, the further into the future the forecast attempted to predict, the greater the error across individual data resolutions. As the Kalman filter forecast projects further into the future, the noisiness of the data and the update methods start to influence the accuracy of the forecast. A surprising outcome from this analysis is that the positional error for forecasts of 12 and 24 h were not always linearly increasing as the number of posits available decreased.

FUTURE WORK

The team identified two potential ways to improve this algorithm by incorporating either a great circle or haversine calculation for the forecast. Either of these may provide a more accurate open-ocean forecast since the great circle route provides the shortest route for a line on a sphere. This differs from the implementation in this article, which assumes that distance globe calculations work within the Cartesian environment with a cosine adjustment for the longitudinal values. Additionally, the team discussed using this data set to perform the Kalman

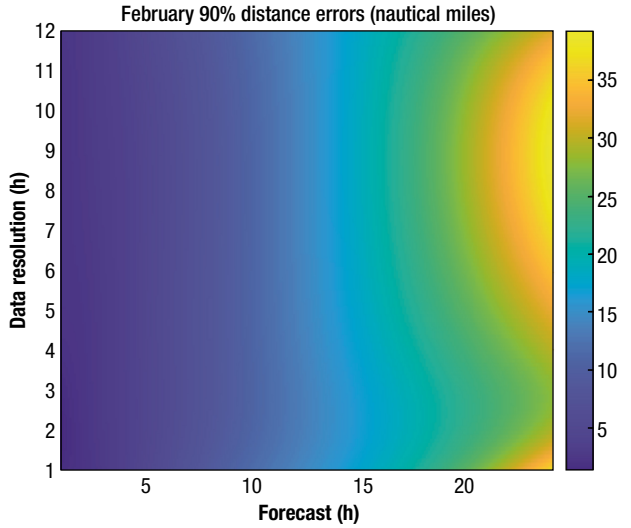


Figure 4. February 1-4, 2019 90th percentile distance error of track from forecasted position.

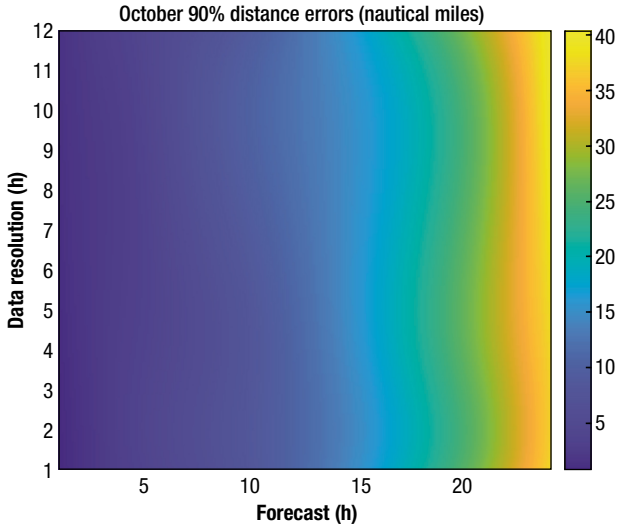


Figure 5. October 1–4, 2019, 90th percentile distance error of track from forecasted position.

filter to calculate the process noise. This involves using the state equation to project forward the positions and compare with the data to determine the process noise. The desired outcome would be finding different "sets" of process noise matrices that represent the different movement patterns expected from various vessel types or indicate whether or not the vessel is moving near/farther from shore. Ultimately, the process noise could be backed out to determine a relative pattern of life for different clusters of vessels.

CONCLUSION

With an increase in data availability, implementation of a Kalman filter can help determine the future location of a vessel 90% of the time within 15 nautical miles 12 h into the future. The Kalman filter algorithm may be improved by using either a great circle or haversine calculation for the forecast. The Kalman filter algorithm forecasts the future location of a vessel using the data of the ship track by extrapolating vessel movement characteristics by using state equations, estimating the uncertainties, and updating the variables in the algorithm to calculate the next state.

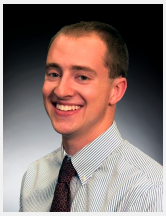
ACKNOWLEDGMENTS: We thank Dr. Bryan M. Gorman for his support, expertise, and encouragement.

REFERENCE

[1]A. Becker, "KalmanFilter.NET," tutorial, 2018, https://www.kalmanfilter.net/default.aspx.

Alyson R. Grassi, Force Projection Sector, Johns Hopkins University Applied Physics Laboratory, Laurel, MD

Alyson R. Grassi is a data analyst and applied mathematician in the Operations and Threat Assessment Group within APL's Force Projection Sector. She earned her BS in mathematics from the University of Delaware and her MS in applied and computational mathematics from Johns Hopkins University. She has extensive experience with modeling and simulation, data analysis in MATLAB and Python, graphical user interface development in Java Python and MATLAB, and testing and evaluation. Since joining APL in 2016, Alyson has successfully led small teams with varying levels of technical skills to solve critical problems for several sponsors. She has facilitated collaboration with outside corporations and led on-site training and testing for sponsors' teams. Her email address is alyson.grassi@jhuapl.edu.



Hayden M. Thomas, Force Projection Sector, Johns Hopkins University Applied Physics Laboratory, Laurel, MD

Hayden M. Thomas is a data visualizationist and operations researcher in the Operations and Threat Assessment Group in APL's Force Projection Sector. He received a BS in electrical engineering from the University of Colorado at Colorado Springs and an MS in electrical engineering from Johns Hopkins University. He has experience in digital signal processing, operations research, and data analysis and visualization in MATLAB and Python. He joined APL in 2017 as a part of the Discovery Program, a 2-year rotational program that allowed him to experience four different groups in the Lab. Much of Hayden's work consists of data simulation, analysis, metric development, and visualization in support of submarines, unmanned surface vessels, and maritime shipping. His email address is hayden.thomas@jhuapl.edu.