We are IntechOpen, the world’s leading publisher of Open Access books
Built by scientists, for scientists

3,900
Open access books available

116,000
International authors and editors

120M
Downloads

154
Countries delivered to

TOP 1%
Our authors are among the
most cited scientists

12.2%
Contributors from top 500 universities

WEB OF SCIENCE™
Selection of our books indexed in the Book Citation Index
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com
Chapter 1

Introduction to Navigation Systems

Junghyun Lee

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.71047

Abstract

Navigation is the method for determining position, speed, and direction of the object. That is mainly classified into two groups: physical model-based methods (PMMs) and external data-based methods (EDMs). Examples of PMMs are inertial navigation systems (INS) and dead-reckoning navigation. They determine the existing position of an object by measuring various changes in its state, such as velocity and acceleration. Representative EDMs is the global navigation satellite system (GNSS). In the case of spacecraft, auxiliary navigation systems using data compression were proposed. In the case of low earth orbit satellites, the deviations between nominal and real orbit are compressed in the form of Fourier coefficients by using the periodic characteristics of the trajectory. In the event of Deep space explorer, B-spline based orbit compression and transmission was proposed.

Keywords: navigation, GNSS, INS, B-spline, data compression

1. Introduction

Navigation refers to the method of determining aspects such as position, speed, and direction during travel. In the pre-modern era, direction and position were determined using an altazimuth, a compass, and a map; these are now considered primitive forms of navigation. As a result of modern developments in science and technology, exact positions and speeds are determined using equipment such as artificial satellites, global navigation satellite system (GNSS), inertial navigation systems (INS), etc. In the modern sense, navigation is mechanical devices equipped in ground vehicles, ships, and aircraft to determine their positions.

Navigation is classified into two categories in this study: physical model-based methods (PMMs) and external data-based methods (EDMs). Examples of PMMs are inertial navigation systems (INS) and dead-reckoning navigation. They determine the existing position of an object by measuring various changes in its state, such as velocity and acceleration. The global
navigation satellite system (GNSS) is an excellent representative of EDMs. Methods to determine longitude and latitude using polar stars or the sun considered as EDMs, which is utilized in the spacecraft nowadays. PMM and EDM have duality. The accuracy of PMMs is exponentially proportional to the cost, and the error increases over time. However, information can be obtained for three axes (X, Y, and Z); this is a stable configuration as no communication problems are incurred. EDMs are relatively cheap and the accuracy is pre-defined according to sensor. However, there may be some restrictions on communication. The navigation equipment in ground vehicles and the orbit propagator in spacecraft are designed to complement each other, considering the properties of tools such as the INS/GNSS.

In the case of spacecraft, auxiliary navigation systems using data compression were proposed due to the unpredictable space environment and limited communication. The orbital equation used in spacecraft is represented by a nonlinear differential equation containing multiple perturbation terms, which makes the determination of orbit numerically complex. Generally, a precise determination of orbit is made on the ground, and the information is then transmitted to the spacecraft periodically. However, Deep space communication is restrictive and expensive. Therefore, the cost of communication can be reduced by compressing orbital data. For Low Earth Orbit satellites, the deviations between nominal and real orbit are compressed in the form of Fourier coefficients by using the periodic characteristics of the trajectory [1, 2]. Deep space explorer orbit compression and transmission were proposed using B-spline [3].

2. Navigation system

2.1. PMMs

Representative PMMs include dead-reckoning navigation (DR) and the inertial navigation system (INS). They determine the current position by measuring the vehicle’s own velocity and acceleration in addition to initial position data. Due to the nature of PMMs, the error increases with time.

2.1.1. Dead-reckoning navigation (DR)

Dead-reckoning navigation is a method of estimating the current position using the moving direction, velocity, and time. It considers errors according to true north and magnetic north. In the case of ground vehicles, only their own velocity needs to be considered, but aircraft and ships must calculate positions by considering ocean currents, wind, and so on. In fact, all navigation systems currently use this dead-reckoning method. Because the accuracy of this method decreases as time and distance increase, celestial navigation is used to determine the accurate position, and then the dead-reckoning method is used from that point forward. The traditional dead-reckoning method used a plotter (a protractor attached to a straight ruler) or a flight computer to determine position. At present, it is calculated automatically using an electronic flight computer.
2.1.2. Inertial navigation system (INS)

The Inertial Navigation System is a stand-alone navigation system that continuously calculates the position, direction, and velocity of the main body through its own accelerometer, rotation sensor, and arithmetic unit, without receiving any external information [4]. Although GPS offers a precise navigation system, it has limitations in space, deep seas, tunnels, and similar places because the GPS operates only when it can receive signals from the satellite. Furthermore, INS can avoid GPS jamming issues. Because an INS is a PMM, the error increases with time. Moreover, the price increases exponentially as the precision is enhanced. One critical factor in an INS is the accurate entry of the initial position and velocity. After that, the data measured by the accelerometer and rotation sensor are integrated consecutively. The accelerometer provides position data whereas the rotation sensor (gyroscope) provides attitude data. Figure 1 shows an example of a strapdown inertial navigation system. Velocity and position data can be obtained by integrating the acceleration twice. It is strapped with an accelerometer that considers the acceleration of the tangent and vector components.

2.2. EDMs

EDMs include GNSS, which is represented by GPS. The application scope is very broad, and includes ground vehicles, ships, and airplanes. In this chapter, the use of GNSS for satellites and deep space probes is explained. Defining the current position and velocity of a satellite is called “orbit determination.” The orbit determination problem can be largely divided into a system model part, a measurement model part, and an estimation technique part. Each can be explained as follows. First, the system model is a mathematical model that represents the orbital motion and various specific variables. It has to be approximated to some degree because many assumptions are included in the process of deriving the equation of motion. Second, for the measurement model, the GPS navigation solution or the tracking data (line of sight, elevation angle, azimuth angle, etc.) of the ground station is used. Here, the measurement values cannot be the true values due to sensor errors and other reasons, and always include some errors. Third, the estimation technique part estimates the optimum prediction values, that is, the position and velocity of the satellite using the approximated system model.

---

Figure 1. Principle of strapdown inertial navigation systems [5].
and inaccurate measurement values. Among these estimation techniques, the batch mode and the sequential model, such as the Kalman Filter, are widely used [6].

2.2.1. Data estimation method

2.2.1.1. Least squares estimation

Least squares estimation is also known as the “batch filter.” The state equation and measurement equation are as follows:

$$ z_k = h_k^T x + v_k $$

The least squares estimation method estimates the state variable $\theta$, which minimizes the error, that is, the difference between the mathematically predicted value $h_k^T \theta$ and the actual measured value ($z_k$). This can be expressed as follows:

$$ L(x) = \frac{1}{2} E \left[ \frac{1}{2} \sum_{k=1}^{n} (z_k - h_k^T x)^2 \right] = \frac{1}{2n} \left( z_n - H_n X \right)^T (z_n - H_n X) $$

Eq. (2) is a convex shape for the state variable, and the following condition must be met to obtain the minimum value:

$$ \frac{\partial L(x)}{\partial x} = 0 = -H_n^T (Z_n - H_n X) $$

In the above equation, $X$ represents the optimum prediction result. Ultimately, Eq. (3) can be rearranged as follows, which is called the Normal Equation:

$$ \hat{X} = (H_n^T H_n)^{-1} H_n^T Z_n $$

2.2.1.2. Kalman filter algorithm

Unlike the aforementioned batch filter, the Kalman filter is used often as a sequential method. The Kalman filter algorithm estimates the optimum prediction value in real time by appropriately mixing predictions made by a mathematical model with measured values from sensors [7]. In the mathematical propagation, which is the first step for the Kalman filter, it is possible to propagate to the target point through an analytical method using appropriate numerical integration or a state transition matrix, assuming that the initial conditions of the orbit are given by the mean and covariance matrix at a random point. The mathematical model is represented by the following state equation:

$$ x_{v+1} = \Phi x_k + w_k $$
This forms a state-space equation together with Eq. (1).

The covariances of the measurement error and the system error can be expressed as R and Q, respectively. The error was assumed as zero-mean Gaussian white noise. First, the covariances of the estimates and deviations in the system can be obtained by the following:

\[
\mathbf{x}_k^+ = \Phi \mathbf{x}_{k-1}^-
\]

(6)

\[
P_k^- = \Phi P_{k-1}^- \Phi^T + \mathbf{Q}
\]

(7)

Second, in the measurement and processing step, the actual measurement is performed. The measured value is expressed as \(z\) according to Eq. (1). Finally, in the update step, we must determine which value must be given a greater weight depending on the reliability of the mathematical prediction and the actual measurement. For this purpose, the Kalman gain is defined as follows:

\[
K_k = \frac{P_{k-1}^- \mathbf{h}_k}{\mathbf{h}_k^T P_{k-1}^- \mathbf{h}_k + R}
\]

(8)

Accordingly, the weights of the system estimates and measurements are considered. The measurements are updated as follows:

\[
\hat{x}_k = \hat{x}_k^- + K_k (z_k - \mathbf{h}_k \hat{x}_k^-)
\]

(9)

\[
P_k = (I - K_k \mathbf{h}_k) P_{k-1}^-
\]

(10)

The degree of update of the Kalman filter is automatically adjusted according to the reliability of the measurement. If the reliability of measurement is good and \(X\) is very small, the Kalman gain increases and more weight is given to the measurement than the mathematical prediction; otherwise, the mathematical prediction is preferred. In this way, the process of stochastically finding the optimum estimate using the Kalman gain, which is a weighting factor, is repeated in real time. The typical Kalman filter algorithm is shown in Figure 2.

As explained above, the Kalman filter method estimates the value of the state variable in real time by stochastically filtering through a system model after receiving measurements mixed with noise. The estimated value is the orbit.

2.2.2. GNSS

The global navigation satellite system (GNSS) estimates positions through the GNSS dedicated satellite orbiting the earth. The global positioning system (GPS) is open to the public and is frequently used. The satellite has a precise time and sends its own position and time information every moment. As shown in Eq. (11) the GPS receiver calculates the straight distance between the satellite and the receiver considering the incoming velocity of signals [8]. The position is determined by a sphere whose diameter is the distance to the satellite if there
is one satellite, by a circle in space if there are two satellites, and by one point if there are three satellites (Figure 3). Figure 3 illustrates the typical satellite orbit determination method. This orbit determination algorithm can be implemented by a computer installed in the ground station or in the satellite, according to the satellite operation scenario. In other words, an appropriate model and algorithm should be chosen depending on the performing entity. This concept can be expanded to ships, ground vehicles, airplanes, etc.

\[ r = \sqrt{(X - X_s)^2 + (Y - Y_s)^2 + (Z - Z_s)^2 + s} \]  

(11)

Figure 3. Orbit termination scheme via GPS.
The GNSS is comprised of a space part, a ground control part, and a user part, which have an organic relationship with one another (Figure 4). The space part consists of approximately 30 clustered satellites in a space orbit. When satellites reach the end of their life, new satellites are launched to maintain a constant number. The measurement error decreases as the number of satellites increases, and more satellites are also added due to communication interruptions by the Earth. Techniques such as SBAS and DGPS are used to enhance accuracy. The ground control part constitutes control facilities on the ground that monitor and adjust the correct rotation of the satellites around the orbit. The GPS has one main control station and four unmanned monitor stations. They monitor if satellites are moving in a given orbit, and perform orbital maneuvering if any satellite moves out of the orbit. The user part consists of general users and GPS receivers that can be purchased.

2.3. GPS/INS coupling system

The INS and GPS are complementary. The INS detects the navigation information (position, velocity, and attitude) of moving bodies using an accelerometer and gyroscope. The precision and cost of an INS are exponentially proportional and its error increases with time. However, it is more stable than other systems because it can obtain information about three axes, X, Y, and Z, and has no communication problems. It has the advantage of not being affected by the external environment, and providing highly accurate and continuous navigation data for short voyage durations. The GPS is a system that determines the position and velocity of moving
bodies by measuring the pseudo range to moving bodies from at least 4 of the 24 satellites orbiting the Earth. The GPS is relatively low-priced, does not accumulate errors with time, and the error range is fixed. However, its disadvantage is that the performance drops if there is severe jamming, or the number of visible satellites is less than four. Furthermore, visibility and communication may become restricted [9].

If you design a navigation system using only one method, high costs are required to obtain the desired performance. However, high-performing sensors can be obtained at low cost if you consider complementary performance. Thus, the navigation system of ground equipment and the orbiter of a space probe are designed to complement each other. We can design navigation systems more effectively by using the INS and performing periodic calibrations with GPS. Furthermore, this can make the overall navigation system more robust than if using only one method [10]. Current research is attempting to integrate these two systems, using integration methods that can be largely divided into tight and loose INS/GPS integration systems [11]. One characteristic of these integration systems is that the performance can be enhanced by improving the tracking performance of the code and carrier tracking loops of the GPS receiver.

The structure of the tightly coupled INS/GPS integration system is shown in Figure 5. Data measured from the GPS receiver (pseudo range, pseudo range rate) and the INS data (position, velocity, attitude) include many errors. The INS and GPS do not perform separate filtering, but these errors are estimated by one Kalman filter. The loosely coupled INS/GPS integration system is shown in Figure 6. In this system, INS and GPS perform measuring and filtering separately.

2.3.1. Comparison of integration systems

The loosely coupled system can be smaller and faster than the tightly coupled system. However, noises also become amplified. The performance of loosely coupled and tightly coupled integration is the same if the GPS availability is good throughout the test run. When GPS availability is poor, as in an urban canyon, tightly coupled integration performs better than loosely coupled integration. The positional accuracy is best when both code and Doppler measurements are used, and is worst when only Doppler measurements are used. The velocity accuracy is also best when both code and Doppler measurements are used, and worst when code only measurement is used.

![Figure 5. Tightly coupled INS/GPS integration systems.](image-url)
3. Auxiliary navigation system using data compression technique

The analysis of space orbital motion generally means the integration of the nonlinear orbital motion equation. For this analysis, special perturbation [12], which is a numerical method, or general perturbation, which is an analytical method, is used. Special perturbation has a small error due to numerical integration during orbit propagation, but requires a high-performance onboard computer. General perturbation has a small computational load during orbit propagation, but the numerical integration error increases substantially with time. To improve this shortcoming, low-orbit satellites generate a residual, which is the difference between the reference orbit and the true orbit. The residuals, which exhibit periodic characteristics, are approximated using the coefficients of trigonometric and Fourier functions, before being transmitted to the satellites along with the reference orbital elements. The satellite computer can improve the precision of orbit propagation and greatly reduce the computational load by generating a reconstruction orbit [13, 14] using coefficients and reference orbital elements received from the ground station.

However, the approximation method using residuals cannot be applied to deep space probes for the Moon and Mars because they do not rotate around the Earth repeatedly, like satellites. As a solution, an auxiliary navigation system using B-spline data compression has been proposed [3]. The data compression rate must be increased because deep space communication is expensive and limited in communication time. For Earth and lunar orbits where takeoff and landing occur, intensive control and communication are essential due to the danger and unpredictability. However, in the transition segment, which is stable and accounts for the majority of navigation, communication time can be saved, and stable navigation data can be received by sending compressed orbit data calculated on the ground to the probes.

3.1. Satellite orbit compression using the Fourier technique

Satellite orbit compression using the Fourier technique, and development of an auxiliary navigation system using this technique, have mainly been studied using a low-orbit satellite model [1, 2, 12, 15]. The overall operation concept of an onboard orbit propagator is
shown in Figure 7. First, an actual orbit is created through accurate modeling and numerical integration of orbital motions, which is available on the ground. Then, a reference orbit is created that is sufficiently close and has a known solution. After defining the residual, which is the difference between the reference orbit and the actual orbit, a few approximate functions are obtained by reflecting the characteristics of orbital motion. The corresponding coefficients are sent to the satellite. The satellite determines the position and velocity of the satellite by substituting the pre-embedded numbers in the onboard computer (Orbit Reconstruction).

The orbit precision predicted through such orbit reconstruction is closely related to the selection of the reference orbit and residual reproduction function. However, the selection criteria for the reference orbit and residual reproduction function must be based on the calculation power of the satellite onboard computer, the data transmission protocol between the ground station and the satellite, and the required precision of the orbit propagation result.

3.1.1. Creation of reference orbit

Essentially, orbit data is created by numerical integration using the initial time, position, and velocity data on the ECI coordinate system, as well as precise orbit modeling. The data determined in this way is converted to orbit elements, and the reference orbit is established. The general reference orbit can be expressed in first or second order polynomials, as shown below, and the coefficients are interpolated using the least squares method.

\[
n(t) = n_0 + n_1 t + n_2 t^2 \tag{12}
\]

\[
e(t) = e_0 + e_1 t + e_2 t^2 \tag{13}
\]
\[ i(t) = i_0 + i_1 t + i_2 t^2 \]  
(14)

\[ \Omega(t) = \Omega_0 + \Omega_1 t + \Omega_2 t^2 \]  
(15)

\[ \omega(t) = \omega_0 + \omega_1 t + \omega_2 t^2 \]  
(16)

\[ M(t) = M_0 + M_1 t + M_2 t^2 \]  
(17)

where \( n \) is the mean motion, \( e \) is the eccentricity, \( i \) is the inclination angle, \( \Omega \) is the right ascension of ascending node, \( \omega \) is the argument of perigee, and \( M \) is the mean anomaly.

Here, the coefficients are determined using the least squares curve fit or a similar technique based on the precise orbit prediction data of actual orbits created by numerical integration. However, in the case of a near-circular orbit, the argument of perigee cannot be defined or the curve fitting may be inaccurate. Therefore, instead of \( \omega \) and \( M \), \( u = \omega + f \) can be used, which is an argument of latitude of the orbit.

### 3.1.2. Residual reproduction

The residual means the difference between the actual orbit data and the designed reference orbit (Figure 8). The Fourier series coefficient must be determined, and the least squares regression is used for this purpose.

\[ r - r^* = \delta r = b_0 + \sum b_k \sin(kt) + \sum b_k \cos(kt) \]  
(18)

Figure 8. Definition of various orbits.
3.1.3. Orbit reconstruction

The coefficients determined above are uploaded from the ground station to the satellite through communication. Using the reference orbit coefficient and residual coefficient received from the ground station, the onboard computer in the satellite reconstructs the orbit. Eqs. (12)–(18) are stored in the satellite computer, and orbit data is created if the current time is inputted. The velocity data is indirectly calculated from the position data, or can be directly created using the same method. The compressed orbit elements are converted to velocity and position data through the DCM, etc., and vary with the characteristics of the orbit.

3.2. Lunar probe orbit compression using a B-spline

Deep space probes that explore the Moon, Mars, and minor planets stay in the Earth’s orbit after being launched by a launch vehicle. After staying in Earth’s orbit for some time, they escape it by burning the engine to reach the target planet. In the case of lunar exploration satellites, they activate trans-lunar injection (TLI) in the parking orbit of the Earth to enter the lunar transfer orbit. Once a lunar exploration satellite enters the lunar transfer orbit, the satellite is tracked with ground antennae around the world, and orbit determination is performed by processing the obtained tracking data. Communication with a lunar probe corresponds to deep space communication, and a representative example is the Deep Space Network (DSN). Four antennae are in currently in operation at three locations (Goldstone, Madrid, and Canberra) [16–18]. The antennae and communication range of DSN are shown in Figure 9. The orbit data must be compressed as much as possible because communication is very limited in both time and range. Therefore, an algorithm for efficient orbit data compression should be developed.

3.2.1. Ground control station data compression

After orbit determination, the coordinates and velocity data of the calculated orbit are compressed as control points through the proposed procedure based on a B-spline. Compared to the conventional method, this method can stably compress even bulky data at a higher compression rate. This data is sent to the probe in the form of compressed parameters. The

Figure 9. Deep space network facility.
probe reconstructs the orbit based on the received parameters. Methods by which the lunar exploration satellite leaves the earth orbit and enters the lunar orbit include direct transfer, phasing loop transfer, and weak stability boundary (WSB).

The B-spline approximation method creates a three-dimensional curve with position and velocity data over time (4-D), excluding time, and then time is reconstructed by linear interpolation using the B-spline. The initial degree and control points are given, and the parameters are estimated based on the given point data. Control points are obtained from the estimated parameters and corresponding points, and the curve is reconstructed. After performing linear interpolation for the reconstructed data in line with the time, it is compared with the original data, and the control points and degrees are increased until they enter the error range.

Then, the drive command is given with the position and velocity data of the x, y, and z axes for time t. The position data can be expressed by (t, X, Y, Z) and the velocity data by (t, Vx, Vy, Vz). The next step is parameterization, where the parameter values corresponding to the given points are estimated. When a point (=1,2,...,n) is given on a secondary plane, the parameter values can be obtained as follows: \( s_i = s_{i-1} + \Delta i/L_i \) (i = 2,3,...,n), \( \Delta i = |p_i - p_{i-1}|/\alpha \), and \( L = \sum \Delta i \), where s is the parameter and s1 = 0. In these equations, 1/2 or 1 can be used for \( \alpha \).

Next, adjustment points are calculated. The method of obtaining the curve using given points and calculated parameter values is as follows. A k-order B-spline curve with nc adjustment points is given as follows:

\[

c(s) = \sum_{i=1}^{n} b_i N_{ik}(s)
\]

where \( b_i \) is the adjustment point, and \( N_{ik}(s) \) is the k-order B-spline basis function. Here, if nc-3 is rt., the vector becomes \( T = (j = 1,...,rt) \). The k-order B-spline basis function is defined as follows:

\[
N_{ik}(s) = \begin{cases} 
1 & \text{if } u_i \leq s < u_{i+1} \\
0 & \text{otherwise} 
\end{cases}
\]

\[
N_{ik}(s) = \frac{s-u_i}{u_{i+1}-u_i} N_{i-1,k}(s) + \frac{u_{i+k}-s}{u_{i+k+1}-u_i} N_{i+k,k}(s).
\]

where \( u_i \) is the value of the ith knot vector. If the given points are \( p_i (i = 1\ldots n) \) and the parameter value corresponding to each point is \( s_i \), the following equation must be satisfied:

\[
p_i \approx c(s_i) = \sum_{i=1}^{n} b_i N_{ik}(s_i)
\]

This becomes \( n \) simultaneous equations consisting of \( n \) unknown numbers, which can be solved by singular value decomposition [19]. The optimal value for the error between the reconstructed value \( c \) and the original value \( p \) can be derived by adjusting the control points and degrees.

3.2.2. Orbit reconstruction

The optimal solution is calculated using the B-spline method (Figure 10). To apply the B-spline-based approximation method to the orbit of a deep space probe, some changes
need to be made to the system currently in use. The transmission data include the number of calculated coefficients: \( n \); the calculated coefficients: \( b \); and the section beginning and end times: \( t_s, t_e \). Two system changes are required: knot vector creation and basis function evaluation.

For the former, the B-spline method requires a knot vector for calculation. This knot vector can be set in such a way that it will be automatically calculated in the probe without being transmitted from the ground to the satellite. For the latter, the B-spline method uses the basis function introduced in Section 3.2.1. Therefore, the ground control system and satellites must have a function for calculating the B-spline basis function \( N_{i,k} \) and calculating the B-spline curve.

4. Conclusion

In this study, we examined the general operation of satellite navigation systems. Basic navigation methods were defined and described in detail. The navigation methods were classified into PMMs and EDMs, and integration of these two methods was proposed. An auxiliary navigation system using data compression was also described. Regarding the navigation system, we suggested a novel method for designing an organic system, rather than a simple navigation device, according to its purpose, cost, and performance.
Author details

Junghyun Lee
Address all correspondence to: yanggwa82@gmail.com
Defense Agency for Technology and Quality, Jinju, South Korea

References


