

The Mathematics of GPS

Richard B. Langley

University of New Brunswick

"Innovation" is a regular column in GPS World featuring discussions on recent advances in GPS technology and its applications as well as on the fundamentals of GPS positioning. This month we look at some of the mathematics involved in determining a position using GPS pseudorange measurements. We also examine some of the ways of gauging the accuracy of GPS positions.

This column is coordinated by Richard Langley and Alfred Kleusberg of the Department of Surveying Engineering at the University of New Brunswick. We very much welcome your comments and suggestions of topics for future columns.

Smart engines, smart ovens, smart cameras. We live in a world of microprocessor-controlled devices that function with a high degree of reliability with little control or thought on the part of the consumer.

Not so many years ago, amateur photographers had to know something about film speeds, f-stop and shutter-speed reciprocity, and the absorption and reflection of light in order to take consistently good pictures. They had to manually adjust the controls on the camera to match the lighting conditions and, of course, focus the lens. With the cameras available today, they just have to point and shoot.

If a good picture isn't possible under the existing lighting conditions, the camera will complain with a beep and may even refuse to take the picture unless switched over to manual operation. Photographers no longer must think carefully about the physical pro-

cess of capturing the image — they can devote all of their talents to the composition of the picture.

Something of the challenge of photography has clearly been lost in this evolution of the machine/operator relationship. Nevertheless, thinking people will continue to be intrigued by the photographic process and will want to have at least a basic understanding of how a camera is able to capture an image on a piece of plastic coated with silver halide crystals — and how that process can be controlled. So it is with GPS.

The GPS receiver is the latest in a lengthening line of smart machines. In the January 1991 issue of *GPS World*, we looked at how a GPS receiver works but glossed over how the measurements are converted into positions by the microprocessor. Let's take a look at that now.

DETERMINING POSITIONS FROM PSEUDORANGES

The basic measurement made by a GPS receiver is the time required for a signal to propagate from a GPS satellite to the receiver. Because the signal travels at the speed of light, c , this time interval can be converted to a distance simply by multiplying it by c .

Let's assume that the clock in the receiver is synchronized with the clock in the satellite, and that the ionosphere and troposphere, which slightly delay the arrival of the signal, do not exist. Furthermore, let's assume there is no measurement noise, that is, no random perturbation to the measurement, something that invariably affects all measurements to a greater or lesser degree. With a single such measurement of the distance or range to the satellite, we can determine something about the position of the receiver: it must lie somewhere on a sphere centered on the satellite with a radius equal to the measured range, as illustrated in the first frame of Figure 1. Let's call that distance p_1 .

If we simultaneously make a range measurement to a second satellite, then our receiver must also lie on a sphere, of radius p_2 , centered on this satellite. The two spheres will intersect, as shown in the second frame of Figure 1, with the locus of intersection points forming a circle. Our receiver must lie somewhere on this circle, which is called a *line of position*. A third simultaneous range measurement, p_3 , gives us a third sphere that intersects the other two at two points only, il-

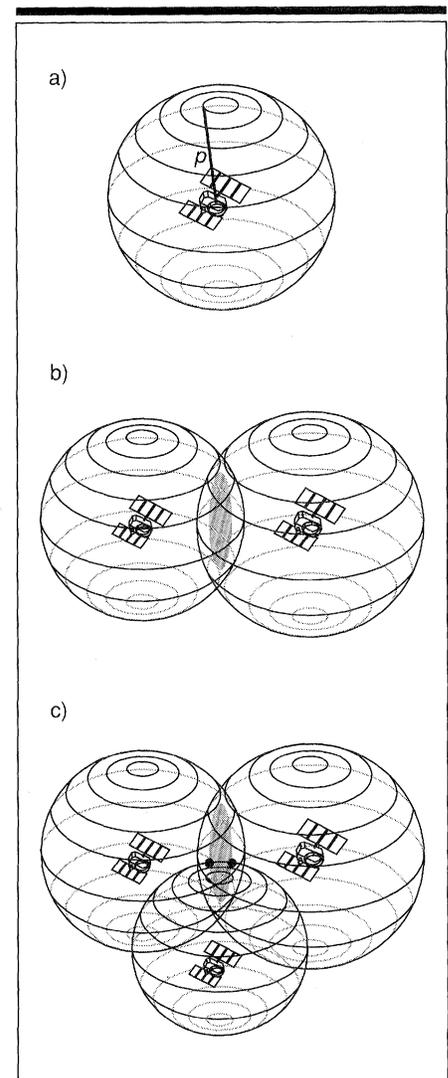


Figure 1. With synchronized clocks, simultaneous range measurements to three GPS satellites produce a determination of a receiver's position. Each range measurement can be portrayed as the radius, p , of a sphere centered on a particular satellite, with the intersections of additional spheres producing ever fewer possible points of receiver location. The line of position is represented in b) and c) by the perimeter of the shaded area.

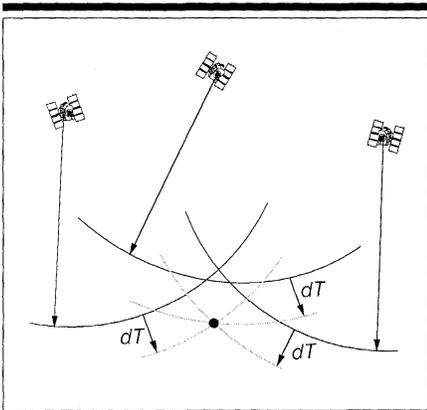


Figure 2. Determination of receiver clock offset (dT) and true user position (intersection of shaded lines) from the intersection of spheres centered on the satellites; pseudoranges are shown by arcs of solid lines

$$\begin{aligned}
 \rho_1 &= \sqrt{(X - x_1)^2 + (Y - y_1)^2 + (Z - z_1)^2} - c dT \\
 \rho_2 &= \sqrt{(X - x_2)^2 + (Y - y_2)^2 + (Z - z_2)^2} - c dT \\
 \rho_3 &= \sqrt{(X - x_3)^2 + (Y - y_3)^2 + (Z - z_3)^2} - c dT \\
 \rho_4 &= \sqrt{(X - x_4)^2 + (Y - y_4)^2 + (Z - z_4)^2} - c dT
 \end{aligned}$$

Figure 3. The set of basic equations for determining user position and receiver clock offset from four pseudorange measurements

illustrated in the third frame of Figure 1. One of these points can be dismissed immediately as being the location of our receiver because it will lie far out in space. So, the simultaneous measurement of the ranges to three satellites provides sufficient information to determine a position fix in three dimensions — at least in principle.

When we started our analysis, we assumed that the clock in the GPS receiver was synchronized with the clocks in the satellites. This assumption, however, is fallacious. When a GPS receiver is switched on, its clock will, in general, be mis-synchronized by an unknown amount with respect to the satellite clocks. Furthermore, the atomic clocks in the satellites are synchronized with each other and to a master time scale — called GPS time — only to within about a millisecond. The range measurements the receiver

makes are biased by the receiver and satellite clock errors and therefore are referred to as *pseudoranges*.

A timing error of a millisecond would result in an error in position of about 300 kilometers, clearly an intolerable amount. System operators conceivably could better synchronize the satellite clocks by frequently sending them adjustment commands from the ground, but atomic clocks have been found to actually keep better time if they are left alone and their readings are corrected. The United States Naval Observatory monitors the GPS satellite clocks and determines the offsets and drifts with respect to GPS time. These parameters subsequently are uploaded to the satellites and transmitted as part of the navigation message broadcast by the satellites. A GPS receiver uses these satellite clock offset values to correct the measured pseudoranges.

Nonetheless, we still have the receiver clock error to deal with. Because of this error, the three spheres with radii equal to the measured pseudoranges corrected by the satellite clock offsets will not intersect at a common point. However, if the receiver clock error, dT , can be determined, then the pseudoranges can be corrected and the position of the receiver determined. The situation, compressed into two dimensions, is illustrated in Figure 2.

So, we actually have four unknown quantities or parameters that we must determine: the three coordinates of our position (say, latitude, longitude, and height) and the receiver clock offset. Now, it is mathematically impossible to uniquely determine the values of four parameters given only three measurements. The way out of this conundrum is to measure simultaneously an additional pseudorange to a fourth satellite.

But just how does the GPS receiver actually extract the position coordinates and the clock offset from the measurements? In the software embedded in the GPS receiver is an algebraic model that describes the geometrical arrangement we've just looked at. For each pseudorange measurement, an equation can be written that relates the measurement to the unknown quantities. The four equations are shown in Figure 3.

The pseudorange measurement made by the receiver, in units of distance, is on the left-hand side of each of the equations. The expression under the square root sign is the true range to the satellite. It is actually a representation of the sphere centered on coordinates x, y, z , the position of the satellite. The satellite coordinates are obtained from the navigation message. The coordinates X, Y, Z

represent the position of the receiver. The term $c dT$ is the contribution to the pseudorange from the receiver clock offset, dT .

The set of four equations must be solved simultaneously to obtain the values for X, Y, Z together with the clock offset, dT . Although the equations are written in terms of geocentric Cartesian coordinates, the resulting X, Y, Z values can easily be converted to latitude, longitude, and height in any geodetic datum or into map grid coordinates.

Linearization of the pseudorange equations. Because of the squares and square roots in the equations, the pseudorange measurements are dependent on the receiver coordinates in a nonlinear way. Consequently, the equations cannot be solved in the usual fashion we all learned in high school. Instead, a procedure known as Newton-Raphson iteration is used. In this procedure, each of the equations is expanded into an infinitely long polynomial based on a set of trial values or guesses for X, Y, Z and dT . Then each series is truncated after the first degree term, resulting in an equation that is linear in incremental corrections to the trial values. The four linearized equations can then be solved simultaneously to determine the values of these increments and the trial values adjusted accordingly.

Because the linearized equations are an approximation of the nonlinear ones, this process, in general, must be iterated, with subsequent iterations yielding smaller and smaller increments. The final solution is the one that satisfies the original nonlinear equations to within an acceptable tolerance. Several iterations may be required to converge to the final solution. However, if the initial position estimate is close to the actual position, the GPS pseudorange equations may be solved in just one iteration.

If one or more of the receiver coordinates is already accurately known, then the remaining coordinates and the receiver clock offset can be determined using fewer than four pseudoranges. For example, say that the height of the GPS receiver is known. Then pseudoranges to three satellites will suffice to determine the two horizontal coordinates and the clock offset. To use GPS to synchronize a clock at a site with known coordinates, only one pseudorange measurement to a single satellite is actually required.

Inconsistent equations. What happens when more than four satellites are above the GPS user's horizon? If the user's receiver can only track four satellites at a time, then the receiver will have to choose which four satellites to track. We'll have something to say about a possible selection method a little

later on. But if the receiver can track five or more satellites simultaneously, then we have a situation in which we have more measurements than unknowns; that is, we have five or more equations, like those in Figure 3, but still have only four unknown parameters.

We cannot solve such a set of equations in the same way as we did for the case of four observations. Why? So far we have neglected the fact that there are other errors in our measurements in addition to the satellite and receiver clock offsets. The presence of these errors means that any subset of four measurements taken from the full set will produce slightly different solutions. In such situations, we say that the system of equations is *inconsistent*.

What do we do? We could discard the extra observations, but, although expedient, that seems wasteful of data. The best approach is to use a method that was devised in the early 1800s by the great German mathematician and father of modern geodesy, Karl Friedrich Gauss — the *method of least squares*. In this method, we obtain a unique solution for the unknown parameters that best fits all of the measurements. This solution is the one that, when substituted into the right-hand side of the pseudorange equations, gives the smallest discrepancies with respect to the measurements on the left-hand side in a summed squares sense. That is, the sum of the squares of the discrepancies is a minimum. Without going into the mathematical reasons for adopting this criterion, we can see qualitatively that it assumes that positive and negative discrepancies are equally likely to occur and that smaller discrepancies are more likely to occur than larger ones.

POSITION ACCURACY MEASURES

As we have mentioned, the pseudorange measurements are contaminated by the satel-

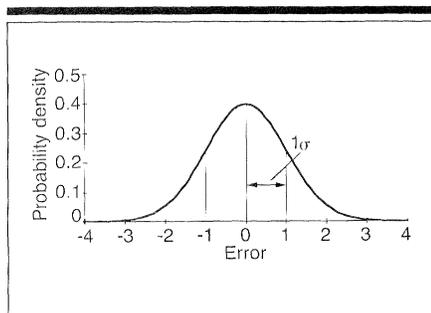


Figure 4. The Gaussian probability distribution function; the shaded area indicates a 68 percent probability of an error having a value between -1σ and $+1\sigma$

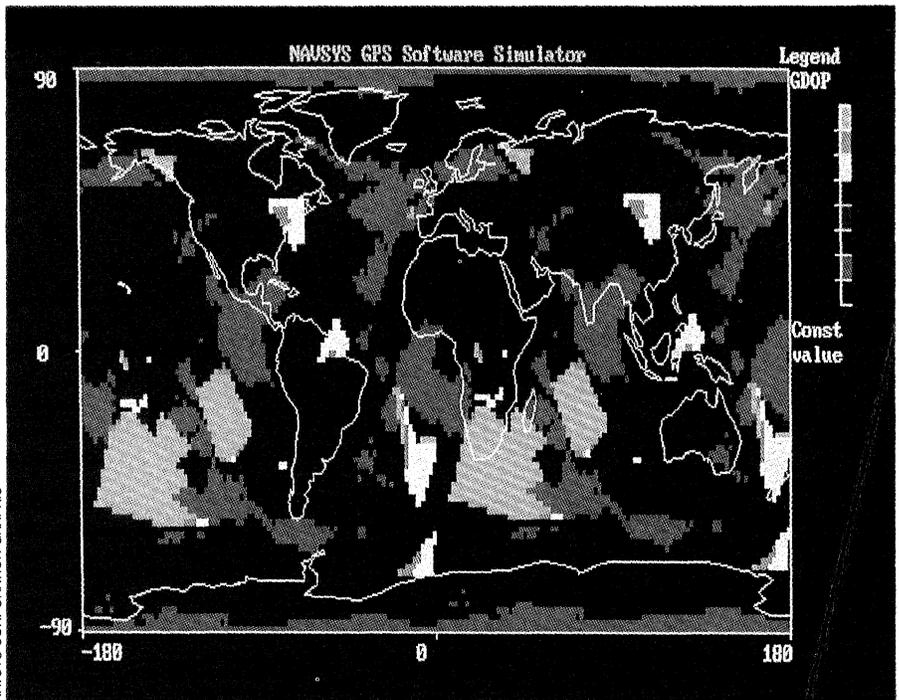


Figure 5. Current global GDOPs; constellation value is the percentage of occurrences for which GDOP is ≤ 6 for a global set of sample points averaged over 24 hours

lite and receiver clock offsets. Even after solving for the receiver clock offset and correcting the pseudoranges for the satellite clock offset using the parameters in the navigation message, errors still remain in the measurements. These errors will, of course, affect the accuracy of the position determination. Because these errors will, in general, change with time, repeated determinations of the position of a fixed location will give slightly different results.

The pseudorange errors come from several sources. The parameters in the navigation message describing the behavior of the satellite clock account for almost all of the clock offset with respect to GPS time. However, because the model used to describe the clock behavior is quite simple and the parameters of the model are predicted ahead of time, there are some small residual clock errors remaining in the pseudoranges. The positions of the satellites as computed from the predicted ephemerides in the navigation messages are also slightly in error. Other errors in the pseudoranges include unmodeled effects of the ionosphere and troposphere, multipath, receiver measurement error, and, for the civilian user, additional clock and orbit errors due to selective availability (SA) when it is in effect.

User equivalent range error. Each of these errors, regardless of its origin, can be ex-

pressed as an error in the range between the user and the satellite. When an error is expressed in this way, it is known as *user equivalent range error* (UERE) or as just *user range error* (URE). Over a sufficiently long period of time, these errors can be considered to be random in nature, with negative and positive errors being about equally probable, giving a mean value of zero. Also, smaller errors are more probable than larger errors.

If a graph is drawn of the frequency of occurrence of an error of a certain size, a curve similar to that shown in Figure 4 will be obtained. Formally, we speak of the curve representing a certain *probability density function*. The probability that an error will occur with a value between specified limits is the area under the curve between these limits on the horizontal axis.

The shape of the probability density curve depends on the particular parameter being measured. However, in science and engineering the probability density curve is often of a particular shape known as a Gaussian or normal distribution. To quantify such a distribution or dispersion of possible errors with a single number, we use the *standard deviation* and usually represent it by the Greek letter sigma (σ). We can determine σ experimentally by making a large number of observations and calculating the square root of the

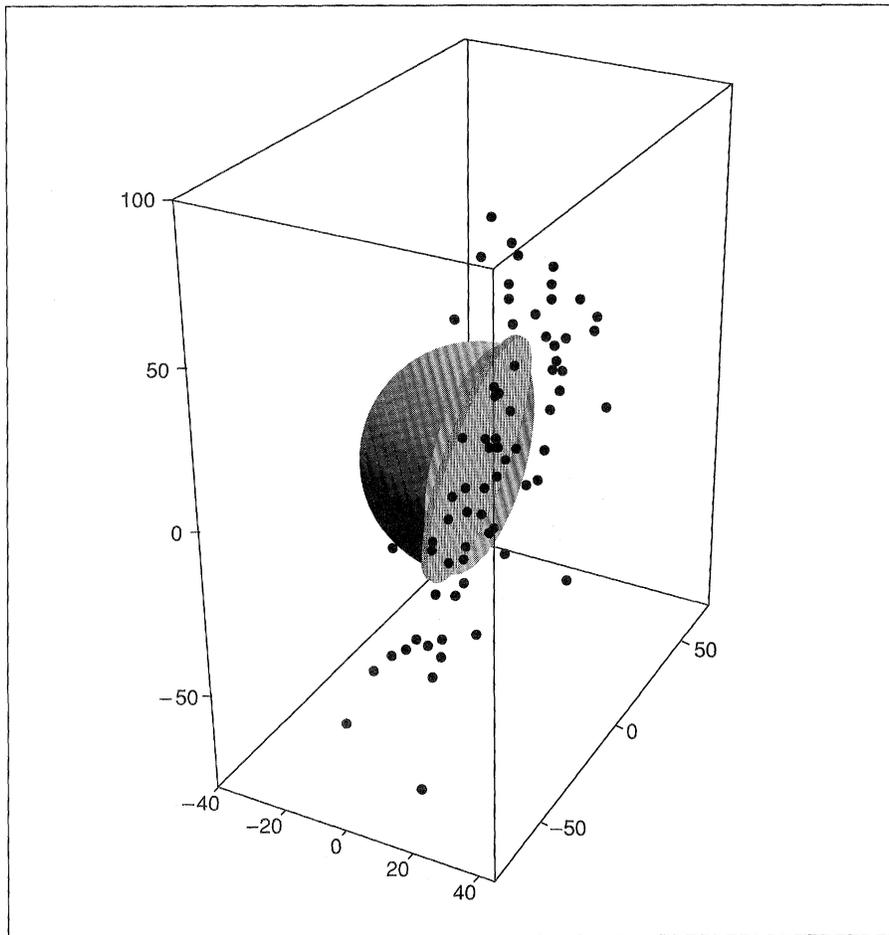


Figure 6. A sample of 100 normally distributed three-dimensional position errors. These artificially generated errors have standard deviations in latitude, longitude, and height of 15.2, 29.1, and 29.0 meters, respectively. The correlation between the errors in latitude and longitude is -0.27 ; between latitude and height, 0.18 ; and between longitude and height, 0.73 . Cross-sectional views of the corresponding error ellipsoid and the sphere with a radius equal to the SEP (34.2 meters) are shown. Data points behind the sphere and some points in its interior are hidden from view.

sum of the squares of the errors in the observations divided by one less than the number of observations made. It is this method of computation that gives σ its alias of root-mean-square (rms) error.

For the Gaussian distribution, there is a 68 percent chance that the magnitude of the error we actually get will be smaller than the standard deviation. There is a 95 percent chance that it will be smaller than twice the standard deviation, and a 99.7 percent chance that it will be smaller than thrice the standard deviation.

UERE errors originate from different sources and thus are independent of each other. However, we can calculate the combined error by taking the square root of the sum of the squares of the individual errors.

This value is the total user equivalent range error.

Dilution of precision. The total UERE is clearly not the error in the position determined by a GPS receiver. UERE is only a measure of the error in the distance to one of the satellites. To determine the three-dimensional position error, we must also take into account where the satellites are in the sky with respect to the receiver. This *satellite geometry* — the spacing of satellites from which GPS signals are received and the resulting angles between the signal paths — results in a larger or smaller uncertainty in the calculation of position.

The contribution of relative satellite geometry to errors in position determination is known as dilution of precision (DOP) and

has a multiplicative effect on UERE. Generally, wider spacing between satellites and the receiver produces smaller errors, for reasons that will be discussed in a moment. Because a GPS receiver user can only obtain GPS satellite signals that are not blocked by the planet, i.e., that are above the horizon, the satellite geometry is already somewhat constrained.

The most common quantification of DOP is through the position dilution of precision (PDOP) parameter. PDOP is the number that, when multiplied by the rms UERE, gives the rms position error (the square root of the sum of the squares of the standard deviations in latitude, longitude, and height).

PDOP is a mathematical function involving the relative coordinates of the receiver and the satellites and can easily be computed for a particular satellite arrangement. PDOP using four satellites can also be visualized geometrically by looking at the tetrahedron formed by the end points of vectors of unit length pointing from the receiver to each of the satellites. PDOP is inversely proportional to the volume of this tetrahedron. The more spread out the satellites are in the sky, the larger the volume of the tetrahedron and the smaller the PDOP, and, hence, the smaller the rms position error. If more than four satellites are in view, a GPS receiver can select the four that give the smallest PDOP.

The minimum value of PDOP is obtained with one satellite at the user's zenith and three satellites with evenly spaced azimuths on the user's horizon. On the other hand, the maximum value of PDOP is, theoretically, infinity. This would occur if the four satellites were situated in the same plane. The final GPS constellation has been designed to provide users anywhere in the world with a PDOP of less than 6 (except for occasional very brief periods of time), assuming four satellites are used with a minimum satellite elevation angle of 5° .

Several other related DOP factors have been defined. HDOP is the dilution of precision in the two horizontal coordinates; VDOP is the dilution of precision in the vertical coordinate; and TDOP is the dilution of precision in the range equivalent of the receiver clock offset. A factor that combines the effects of geometry on both position and clock offset is the geometric dilution of precision, GDOP. Figure 5 shows the global range of GDOP values, averaged over a 24-hour period, for the current 15-satellite constellation. Users can expect the PDOP to be less than 3 most of the time once all the GPS satellites are in place.

Other accuracy measures. In general, the three coordinates of a three-dimensional position fix will have different error probability distributions and, hence, different standard deviations. Also, the errors between any two coordinates may be mutually correlated; that is, an error in one coordinate will have an effect on the other.

If we trace out a contour of equal error probability density in all three coordinates, we get an ellipsoid centered on our position fix. The shape of the ellipsoid is determined by the standard deviations of the coordinates and their correlations. (See Figure 6.) Note that, in general, due to the correlations, the ellipsoid axes are not oriented in the same directions as the coordinate axes. There is a certain probability that the true position lies within the ellipsoid. If this probability is 20 percent, then this ellipsoid is referred to as the *standard error ellipsoid*.

If the standard deviations for the three coordinate directions are identical, the ellipsoid degenerates into a sphere. The radius of such a sphere, inside of which there is a 50 percent probability of the true position fix being located, is called *spherical error probable* (SEP).

The term SEP is used even when the actual error figure is an ellipsoid. If we make a large number of position fixes at a given location, we can say that the SEP is the radius of the sphere containing 50 percent of the individual fixes. The Department of Defense's accuracy goal for GPS is to have a worldwide SEP of 15 meters.

If we forget about the height coordinate for the time being and consider just the horizontal coordinates, we can construct the two-dimensional analogue to the error ellipsoid: the *error ellipse*. It is defined as the contour of equal probability density in the two hori-

zontal dimensions. As with the error ellipsoid, an error ellipse has a certain probability that the true horizontal coordinates of a position lie within it. For the standard error ellipse, this probability is 39 percent. As with the error ellipsoid, the semi-axes of the error ellipse, in general, are not equal to the standard deviations. However, given the standard deviations in the horizontal coordinates and their correlation, the semimajor and semiminor axes of the ellipse can be calculated.

The two-dimensional analogue of SEP is *circular error probable* (CEP). As shown in Figure 7, CEP is the radius of the circle inside of which the true horizontal coordinates of a position have a 50 percent probability of being located.

Another accuracy measure frequently used in navigation is twice the root-mean-square of the horizontal distance error, or *2 drms* for short. It is equal to twice the square root of the sum of the squares of the lengths of the semimajor and semiminor axes of the error ellipse. A circle of radius 2 drms will contain the true horizontal position with a certain probability. Unfortunately, a drawback of 2 drms as a measure of error is that it does not correspond to a fixed value of probability for a given value of error. The probability varies with the eccentricity of the error ellipse, ranging from 95.4 percent (the ellipse collapses to a line) to 98.2 percent (the ellipse becomes a circle). Because of this variation in probability, there is not a constant relationship between values of 2 drms and CEP. The ratio of 2 drms to CEP varies with the eccentricity of the error ellipse from 2.4 to 3.

Because of its wide use in navigation, 2 drms is used to specify the designed level of horizontal positioning accuracy for the GPS Standard Positioning Service (SPS) and Precise Positioning Service (PPS). The latest issue of the Federal Radionavigation Plan (FRP) states that when GPS is declared operational, the horizontal accuracy for SPS is planned to be 100 meters 2 drms at 95 percent probability. This means that 95 percent of all horizontal position fixes should be within 100 meters of the true position.

But what about the other 5 percent? Theoretically, if the position errors due to the various UEREs including SA are from a Gaussian distribution, we could occasionally get extremely large errors. However, the Department of Defense will control SA so that excursions will not exceed 300 meters 99.99 percent of the time. The corresponding designed 2 drms (95 percent) horizontal accuracy for PPS is 17.8 meters.

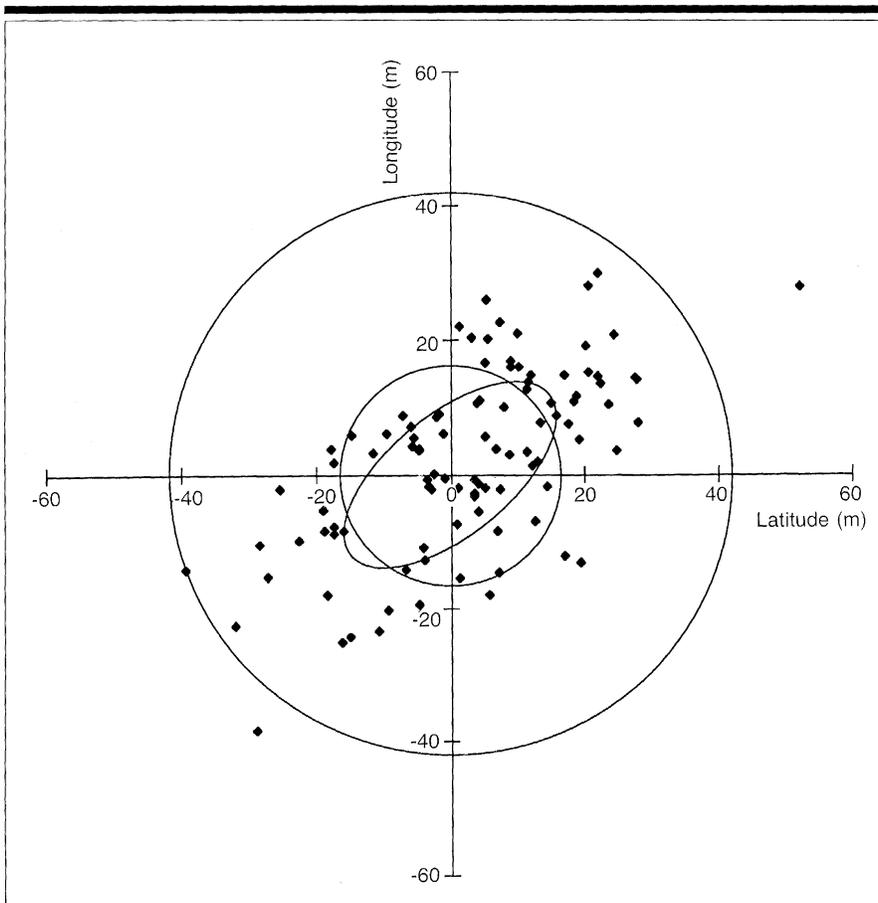


Figure 7. A sample of 100 normally distributed horizontal position fix errors. These artificially generated errors have standard deviations in latitude and longitude of 15.8 meters and 13.6 meters, respectively, with a correlation of 0.64. Shown are the error ellipse, the circle with 2 drms (41.8 meters) radius, and the circle with a radius equal to the CEP (16.4 meters) corresponding to this sample.

The FRP describes the designed accuracy of the vertical component of a GPS-derived position at the 2σ level. As this corresponds to a 95 percent probability level, it is consistent with the accuracy quoted for the horizontal position. For SPS, the designed vertical 2σ is 156 meters; for PPS, it is 27.7 meters. The FRP also gives the designed accuracy of receiver clock synchronization at the 1σ level. For SPS, accuracy is planned to be 167 nanoseconds; for PPS, it is given conservatively as 100 nanoseconds.

It should be pointed out that the stated PPS position and time accuracies are designed estimates of GPS capabilities, and superior results have already been obtained in practice. Significantly greater accuracies can be obtained by both PPS and SPS users by operating in a differential mode with two or more receivers used simultaneously. In fact, almost all of the effects of SA can be removed when operating in this mode.

and statistics involved in determining positions using GPS. Although a GPS receiver can be operated without the user knowing any of this math, a basic understanding of how the receiver determines a position is important for assessing the accuracy and reliability of the numbers presented on a receiver's display screen. ■

CONCLUSION

In this column, we've taken a brief introductory look at the geometry, algebra, calculus,

Reprinted from

GPS WORLD

July/August 1991