ROBERT R. NEWTON

THE NEAR-TERM POTENTIAL OF DOPPLER LOCATION

In this paper we explore the precision that can be obtained in locating a point on the earth's surface by analyzing the Doppler shift in the signals from a near-earth satellite. When we limit the discussion to the use of techniques that have been demonstrated in the laboratory but that may not have been introduced into field use, we find that the precision obtained by using the data from a single pass of a satellite should be about 18 centimeters. It should be possible to improve the precision by using data from more than one pass in the usual statistical fashion.

BACKGROUND

Suppose that a source of sound waves is located at the point S_1 in Fig. 1, and suppose that a listener is located at the point R_1. In some number of seconds, t, say, the listener receives some total number of waves or cycles of sound, W say. The frequency that he hears is (W/t) hertz and this is the same as the frequency f_T sent out by the source. That is, $f_T = W/t$.

Now suppose that the source moves from the point S_1 to the point S_2 during these t seconds, so that the source increases its range r from the receiver by the amount Δr. The wave front that just reached R_1 in the first case now reaches only to the point R_2, which is the distance Δr from R_1. The listener, in this case, does not receive the waves that lie in the distance Δr during the time t. If each wave has a wavelength λ, the number of waves that he fails to receive is $\Delta r/\lambda$, the number of waves that he does receive is $W - (\Delta r/\lambda)$, and the frequency f_R that he hears is this number divided by t. Let $\dot{r}$ denote the rate at which the total range r is changing, so that $\dot{r} = (\Delta r/t)$. Then

$$f_R = f_T - (\dot{r}/\lambda). \qquad (1)$$

This is the basic equation of the Doppler shift. If a source of sound is moving away from the listener, the frequency that he hears is shifted downward by an amount proportional to the velocity $\dot{r}$ of the source. If the source is moving toward the listener, $\dot{r}$ is negative, and the frequency is shifted upward.

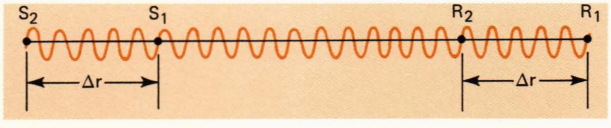

Fig. 1—The origin of the Doppler shift. A source of sound located at S_1 sends out a wave train that is received at R_1, and a certain number of waves are received at R_1 in time t. If the source moves from S_1 to S_2 during the time t, moving a distance Δr away from R_1, the wave train that formerly extended to R_1 now extends only to R_2, and the listener fails to receive all the waves contained in the distance Δr in the time t. Thus he hears a lower frequency if the source is moving away from him.

Little has been written about the history of the Doppler principle, but I believe that the following is basically correct: The principle had long been recognized in its application to sound. In 1842, the Austrian physicist Christian Johann Doppler published a paper in which he pointed out that the same principle should apply to light, and this is the contribution that led to naming the effect after him. It is possible that he also changed the principle from a qualitative one to a quantitative one by deriving Eq. 1 or its equivalent.

In the case of sound, the effect perceived is a shift to a lower pitch if the source is receding and to a higher pitch if the source is approaching. In the case of light, the perceived effect is a shift of color toward the red if the source is receding and toward the blue if the source is approaching. In the case of radio transmissions, which we can neither see nor hear, we do not have the physiological sensations of pitch and color. We simply say that the received frequency f_R is decreased or increased, as the case may be.

Light and radio transmissions are different examples of electromagnetic radiation, and the same relations apply to both. Equation 1 is not correct for light, because of quantum and relativistic effects. The important quantum and relativistic effects are two in number:

1. If the source and the receiver are not at the same gravitational potential, a quantum of radiation changes its energy as it passes from one to the other. This is usually called the "gravitational red shift," because in astronomy we usually deal with the light emitted from the surface of a star. As the quantum climbs out of the gravitational potential well of the star, it loses kinetic energy; losing kinetic energy for a quantum means that its frequency decreases and its color shifts toward the red. However, the shift is toward the blue if the receiver is

deeper in a potential well than the source is. In precise work with the radio signals received on the surface of the earth from an artificial satellite, this "blue shift" effect must be taken into account. That is, the signals increase in frequency as they fall from satellite altitudes to the surface.

2. Even if the source and receiver are at the same gravitational potential, the exact form of Eq. 1 is not correct because it is not consistent with the way that we must combine velocities in relativity theory.

These effects are the same general size for near-earth satellites and they are of opposite sign. For satellites in the orbits that have been used in the Doppler navigation system, the combination of the two effects changes the received frequencies by about 2 parts in 10^{10} of the transmitted frequency. It is thus necessary to take the effects into account when we do Doppler work of high precision. However, it is not necessary to consider them in order to understand the principles and capabilities of Doppler location. In almost all of the remaining discussion, then, I shall assume that Eq. 1 is correct.

When we derived Eq. 1, we assumed that the source was travelling directly away from the receiver. However, since the wave fronts are spherical as they spread away from the source, the number of waves that do not reach the receiver depends only on the amount by which the range changes in the time t. Hence the quantity $\dot{r}$ that appears in Eq. 1 is to be interpreted as the range rate, regardless of the details of the motion, which cause the range to be changing.

In some studies, however, it is convenient to look at the details of the motion. In order to do this, we use Fig. 2. Here $\boldsymbol{r}$ is the range vector from the receiver R to the source S at some instant, and $\boldsymbol{v}$ is the velocity vector of S at the same instant. The angle α is the angle between the two vectors, as drawn. The range rate $\dot{r}$ is obviously equal to $v \cos \alpha$. We let f_D denote the amount of the Doppler shift, in the sense of received minus transmitted frequency. At the same time, we replace the wavelength λ by c/f_T, in which c is the velocity of light. This gives us

$$f_D = f_R - f_T = -(v/c) f_T \cos \alpha$$
$$= -(\dot{r}/c) f_T . \qquad (2)$$

THE PRINCIPLE OF DOPPLER LOCATION

There are several ways of looking at the process of locating a position by means of the Doppler shift. Different ways of looking at the process lead to understanding different aspects of it.

Suppose that an artificial satellite in a near-earth orbit comes over the observer's horizon. At this time the range vector $\boldsymbol{r}$ is almost in the opposite direction to the velocity vector $\mathbf{v}$, so that the angle α is near 180°. From Eq. 2, we see that the Doppler shift f_D is large and positive. As the satellite comes closer, α decreases. It becomes 90° when the satellite is at the point of closest approach, where $\dot{r} = 0$, and finally, as the satellite goes over the horizon, the angle α is nearly zero. The interval from the time the satellite appears over the horizon to the time when it disappears is called a "pass". The variation of f_D with time during a pass is shown schematically in Fig. 3.

We suppose that the position of the satellite is known as a function of time, and therefore we know where it is at the time when $f_D = 0$ in Fig. 3. Since $\alpha = 90°$ at this time, the observer is in the plane that passes through this position and that is perpendicular to the velocity vector $\mathbf{v}$ at the same time. In the usual case, which is the only one we shall consider, the observer is also on the surface of the earth. Thus he lies on the curve that the plane cuts from the earth's surface.

Imagine for the moment that the satellite passes directly through the observer's position. In this case, α remains 180° until the satellite reaches this position

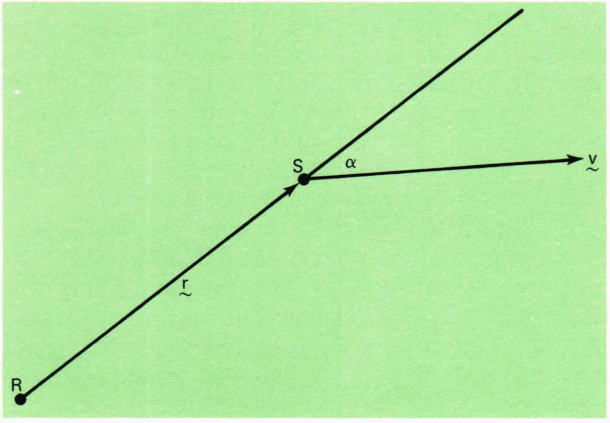

Fig. 2—Relations between range, velocity, and range rate. The vector $\boldsymbol{r}$ points from the receiver R to the source S, and the source S has the vector velocity $\boldsymbol{v}$. Then the range rate $\dot{r}$ equals $v \cos \alpha$. In some studies, it is convenient to write the Doppler shift in terms of v and α rather than of $\dot{r}$.

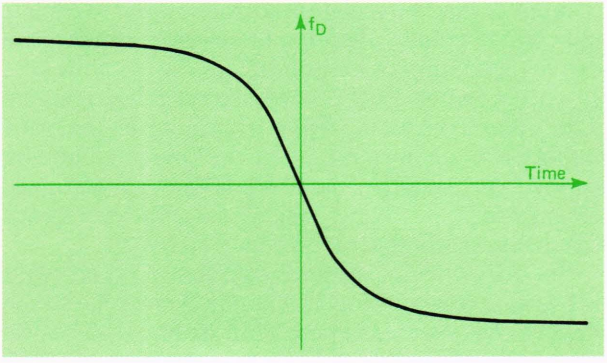

Fig. 3—The schematic variation of the Doppler shift f_D with time during a pass of a satellite above the observer's horizon. An observer on the surface of the earth can find his position from the position of the satellite at the time when $f_D = 0$ and from the derivative of f_D at the same time. At this time, usually called the time of closest approach, the range r is a minimum.

and it then changes abruptly to 0°. The slope of the curve where $f_D = 0$ in Fig. 3 is thus infinity if the "miss distance" is zero. The slope is finite for any real miss distance, and it becomes steadily smaller as the miss distance increases. Thus there is a one-to-one relation between the slope of the f_D curve at closest approach and the miss distance. Since we already know that the observer is on a particular curve in space, knowing the miss distance locates him at one of two points. Typically, the points are separated by thousands of kilometers, so the observer immediately knows which of the two points applies to him. Thus, by using the time of closest approach, and the slope or derivative of the Doppler curve at that same time, the observer can locate himself.

We can view the process of location in a different way by looking at Fig. 2 again. If we measure f_D at some instant, we can calculate α from Eq. 2. The observer therefore lies on a cone whose vertex is at S and whose axis is the direction of $\mathbf{v}$. This cone intersects the earth's surface in some curve. Measuring f_D at a different time gives another curve cut from the earth's surface, and the observer lies at the intersection of these two curves. Thus we see that the observer can locate himself by measuring only two points on the Doppler curve of Fig. 3. Measuring more than two points on the curve provides redundancy and therefore increased accuracy.

Still a third way to look at the process of location is to use what is often called "integrated Doppler". In order to measure the Doppler frequency, the observer must have an oscillator of known frequency. For the moment let us suppose that its frequency is exactly equal to f_T. The observer beats his local oscillator against the received frequency, and the difference is f_D. Instead of attempting to measure f_D at two or more instants, the observer counts the cycles of the beat note and measures the time required for the beat note to go through some number N of full cycles; N is a number that the observer chooses for convenience.

Suppose that the count of N cycles begins at time t_1 and ends at time t_2. Between these times, the satellite moves from point S_1 to point S_2 in Fig. 4. Now N is the number of cycles that the observer fails to receive if the source is moving away from him, or the extra number that he does receive if the source is moving toward him. In either case, $N\lambda$ is the change in the distance to the satellite; that is, it is the difference between the distances to the points S_1 and S_2.

The difference in distances defines a hyperboloid of revolution whose foci are points S_1 and S_2, so the observer lies on this hyperboloid in Fig. 4. Thus he lies on the curve of intersection between the surface of the earth and the hyperboloid. If he repeats the process, he generates a second curve, and he lies at the intersection of the two curves.

This way of looking at Doppler location brings out the relation between Doppler location and location by ranging. In location by ranging, a satellite emits time signals at times t_1 and t_2, say, as measured on its own clock. The observer receives the signals at later times, say at $t_1 + \delta t_1$ and $t_2 + \delta t_2$, as meaured on his own clock. The differences δt_1 and δt_2 place the observer on the surfaces of two spheres whose centers are the corresponding satellite positions and, if the observer is on the surface of the earth, these spheres determine his location.

Suppose that we want an accuracy of 10 centimeters in position. Since the time signals travel 3×10^8 meters per second, both clocks must have errors of less than one third of a nanosecond. This exceeds the present capabilities of measuring time on an absolute basis; therefore ranging by this method cannot be done. Instead, the observer must time at least three signals sent from the satellite and use the information to determine the offset between his clock and the satellite clock at the same time that he determines his position. Thus the observer does not actually measure the range at any time. Instead, he measures the amount by which the range changes between two times, and he repeats this measurement as often as he needs to.

Thus, contrary to a widespread belief, Doppler and ranging systems measure the same thing, which is the change in range between two times. Hence both systems supply exactly the same kind of information; the ranging systems do not have an innate superiority in the kind of information that the user obtains.[1] If

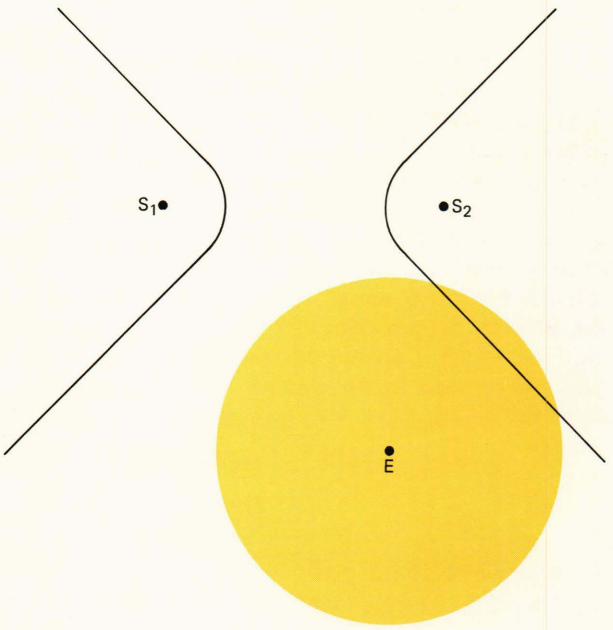

Fig. 4—Location by integrated Doppler. The observer counts the total number of cycles in the Doppler shift while the satellite moves from S_1 to S_2. This number, times the wavelength, is the difference between the distances from the observer to S_1 and S_2; the observer is therefore on a hyperboloid of revolution whose foci are S_1 and S_2. At the same time, he is on the surface of the earth, which is represented by the circle whose center is E. The observer is therefore on the curve of intersection of the hyperboloid with the earth. A second count of cycles then locates the observer.

the kind of information is the same, Doppler systems have an innate engineering superiority.

If we use "information" in the technical sense that it has in information theory, we may say that a measurement of a range difference amounts to a transfer of a certain number of bits of information from the satellite to the observer. The precision of the measurement is directly related to the number of bits transferred, which in turn is proportional to the amount of energy transferred. Hence, if a Doppler system and a ranging system provide the same precision of location from a single pass, the energy transferred is the same, the time is the same, and the average power transmitted is the same. In the Doppler system, the average power and the peak power are the same. In the ranging system, the peak power exceeds the average power by orders of magnitude, and the efficiency is necessarily lower.

In other words, if the same power and the same levels of technology are used in both systems, the Doppler system is inherently more precise than a ranging system.

THE PRACTICE OF DOPPLER LOCATION

A short while ago, we assumed that the observer uses a local frequency equal to the transmitted frequency, and that he measures f_D by beating the received frequency against his local frequency. He would encounter severe difficulties if he actually did this, because the beat frequency would pass through zero at the time of closest approach in Fig. 3. It is difficult to maintain precision if this is allowed to happen, because the dynamic range required of the measuring apparatus becomes infinite.

In all Doppler satellite systems that have been used to date, the observer chooses his local frequency so that it differs from f_T by more than the possible amount of the Doppler shift f_D. A satellite in a near-earth orbit has a speed of about 7000 meters per second, or slightly more, so that v/c is about 2.4×10^{-5}, or about 24 parts per million. Hence, according to Eq. 2, the maximum value of f_D is about $2.4 f_T \times 10^{-5}$, so that the received frequency f_R cannot differ from f_T by more than this amount. The observer therefore sets his local frequency, say f_S, so that it differs from f_T by more than 24 parts per million. An offset of about 50 parts per million, or perhaps somewhat more, has proved to be convenient.

When this is done, the beat frequency, which is what the observer directly measures, can never go through zero. After he measures the beat frequency (or counts its cycles in the integrated Doppler method), the observer calculates the Doppler frequency (or the count of its cycles) by using the difference $f_S - f_T$. I shall use f_A (from the German *abstand*) to denote this difference.

In order to find his position with high precision, the observer needs to know the difference f_A with high precision, but he does not need to know either f_T or f_S with high precision. It is sufficient to know only nominal values of the two individual frequencies.

In finding his position, the observer does not actually use the simple approaches described in the preceding section, except perhaps in simple demonstrations or expositions. Unless he uses the method of integrated Doppler, he starts by measuring the beat frequency at a set of times t_i. I shall denote the resulting set of measured frequencies by f_i.

Let us assume that the observer has a map which gives his distance from the center of the earth as a function of his latitude η and his longitude λ. If so, he knows fully his position with respect to the center of the earth once he finds the two coordinates η and λ. Since he knows the orbit of the satellite, he could calculate the Doppler frequency f_D at the time t_i if he knew η and λ. In other words, f_D at the time t_i is a function of η, λ, and t_i. Further, the frequency f_i measured at the time t_i equals f_D plus f_A, except for experimental error. That is,

$$f_A + f_D(\eta, \lambda, t_i) = f_i \qquad (3)$$

for each value of i. The observer then uses statistical procedures to find the values of η and λ that give the best fit in Eq. 3, using any definition of "best fit" that he chooses.

If he uses the method of integrating the Doppler frequency, the observer starts to count cycles at some measured time and counts continuously until the end of the pass. He then reads the times t_i at which the count takes on a convenient set of values. As in the frequency method, he can calculate what the count should be as a function of f_A, η, λ, and t_i, and he proceeds as before to find the values of η and λ that make the measured values best fit the calculated ones. For simplicity, most of the discussion of this paper will be based upon the frequency method of Eq. 3, but the reader should remember that all the discussion applies equally to the integrated Doppler method, if appropriate modifications are made to the terminology.

Only cost limits the number of parameters that can be found in this way. For example, the observer may have a local oscillator whose frequency is not known accurately. In this case, he may simply take f_A, along with η and λ, to be an unknown parameter that he finds from Eq. 3. The parameter f_A is strongly determined by the measurements.

In a more complicated example, the observer may not know his distance from the center of the earth, which we may take as equivalent to not knowing his altitude. The altitude must be taken as another unknown parameter. The difficulty in doing this comes from the fact that the time of closest approach and the range at closest approach are the geometric parameters most strongly determined by the data. Suppose that there are two observers, both at the same range at closest approach. Suppose further that they are at different altitudes, so that they are at different horizontal positions if they are at the same

range. Under these conditions, there is almost no difference in their measured frequencies (for the same f_A). Hence it is not possible to solve accurately for both the altitude and the horizontal position by using the data from a single pass.

All that the observer needs to do in order to find all three coordinates is to use two passes. These can be two different passes of a single satellite, or they can be single passes of two different satellites. If his position is, say, on the port side of the satellite in the first pass, he should choose the second pass to be one in which he is on the starboard side. The range lines at the two closest approaches then cross at a strong angle, and all coordinates are strongly determined.

A still more complex problem comes from the need to know the orbit of the satellite. In order to find the orbit, we have a network of tracking stations, suitably distributed at known locations over the surface of the earth, and we measure the Doppler shift at each station for each pass over some convenient interval, say two days. We then chose the six orbital parameters that give the best fit to the resulting set of data. If necessary, we also find a separate value of f_A for each pass observed at each station. However, if we have atomic frequency standards at the stations, and if we have a crystal oscillator of high quality in the satellite, we may assume that all station frequencies are known and that the transmitted frequency varies with time in some simple way, such as quadratically.

We find the most complex problem when we consider the other things we must know in order to calculate the satellite orbit; I shall now leave the relatively trivial frequency problems to one side. We assumed in the preceding paragraph that we know the coordinates of all the tracking stations. Further, we tacitly assumed that we know all the parameters that enter into determining the force field acting on the satellite. Actually, of course, there are uncertainties in our knowledge of station coordinates and force-field parameters. If we trace out the effect of these uncertainties upon the location of points, we find that there is a limit to the accuracy of location. In the present state of affairs, this limit is a moderate number of meters.

In making this statement, I deliberately said "location" instead of "Doppler location." We find the same limit, in the sense in which the word has just been used, with all uses of satellites. In finding this limit, it does not matter whether we measure the Doppler shift, the range, or the optical position. It does not matter whether we use the one-way Doppler system, the two-way radio systems such as radar or the NASA range/range-rate system, the most precise camera systems, or laser systems. Further, the same limits apply to surveying systems that are purely groundbased or that use ranging between aircraft and ground sites.

The reason for this comes from the connection between the shape of the earth and its gravity field. By the shape of the earth, we mean the shape of the surface called "mean sea level." This surface is the shape of the actual ocean surface, after we average out the waves and tides, when we are at sea. On land, we extrapolate the ocean surface inland by means of bubble tubes on our levels; the bubble tubes always give us the direction that the water surface would have if the water could percolate freely into the continents.

Except for small effects that we must account for in practice, but that we can ignore in this general discussion, mean sea level is a surface on which the potential energy (including the centrifugal potential due to the earth's rotation) of a kilogram mass is a constant. Thus, if we knew the parameters of the gravitational force field exactly, we would know the shape of the surface at sea level. Conversely, if we knew the shape of the surface, we could calculate the parameters of the force field.

We specify the shape of the sea level surface by giving the local earth radius at any point. That is, at a given latitude and longitude we give the number of meters from the center of mass of the earth to the surface at that point. At present, the uncertainty in the local earth radius is a moderate number of meters. If there is an uncertainty in the shape of the surface, there must also be an uncertainty of the same general amount in locating points on the surface. This uncertainty is a characteristic of the surface and of our knowledge of it, and it does not depend upon the method we use to measure position.

The limit on positioning that is imposed by our ignorance of sea level will be called the geodetic limit. In the next section, I shall try to assess the near-term effect of the geodetic limit. This means trying to answer the question: Can we decrease this limit by further research, or are we too near some minimum limit that is imposed by basic geophysics? The geodetic limit is common to all systems of location. In later sections, I shall try to assess the near-term level of all factors that are known to limit the accuracy of location by Doppler methods using artificial earth satellites.

THE GEODETIC LIMIT

In discussing the geodetic limit on the accuracy of location, we must distinguish between the global problem and the local problem. In the preceding section, I was tacitly talking about the global problem, which is: For any given latitude and longitude, what is the uncertainty in the earth radius? As I have said, this is a moderate number of meters. I refrain from stating a specific number, because different research centers make different estimates. All current estimates, however, are of the order of 5 or 10 meters.

Now pick a specific latitude η_o and longitude λ_o. Let r_o be our current estimate of the earth radius at this point. Suppose that further research ultimately shows that r_o is, say, too large by 5 meters. We express this by saying that the error is +5 meters.

Next let us move away from this point by 0°.01 in some direction; 0°.01 is about 1 kilometer (km). If the error in the earth radius is +5 meters at the first point, can we put limits for the error at the second point? The answer is provided by the direction of the sea-level surface.

The same information that lets us calculate the gravitational force field and the sea-level surface also lets us calculate the direction of the sea-level surface at any point. At present, the uncertainty in this direction is a few tens of seconds of arc. For the sake of illustration, let me use 20″ as the uncertainty at the point with coordinates η_o and λ_o; this is about 10^{-4} radians. If we predict the radius at a point 1 km away, the uncertainty is about 10^{-4} km, which is about 10 centimeters (cm). Thus if the error in the earth radius is +5 meters at the first point, it must be between +4.9 and +5.1 meters at the second point. In other words, we can find the relative position of two neighboring points much more accurately than we find the position of either point individually.

Finding the relative positions of neighboring points is the local problem and finding the position of any point with respect to the center of mass of the earth is the global problem. The geodetic limit is much greater for the global problem than for the local problem at present, and it has also been much greater throughout the past.

The first artificial satellite was launched on 4 October 1957. At that time, the global accuracy of location was a few hundred meters. Only a few areas had been the subjects of intensive local surveying. They included Europe, the contiguous United States plus southern Canada, Australia, and some other smaller areas. Within one of those areas, the local accuracy was a few tens of meters, but the accuracy of locating the areas with respect to each other was hundreds of meters. Now the global accuracy has been improved by an amount that is close to two orders of magnitude. Several programs have contributed to this progress, but we can safely say that the dominant new contributions have been made by satellite programs.

Large extrapolations are always dangerous, but a moderate extrapolation is usually safe. On the basis of present knowledge, it is safe to say that the geodetic limit can almost certainly be decreased by another order of magnitude, to the point that is a few tens of centimeters on the global basis and less on a local basis. Perhaps it is better to say that improving geodetic accuracy to this level is more a political problem than a technical problem. The improvement mentioned is within reach of present technology, but it will take money. Whether the necessary programs are to be funded is a matter of national priorities and political decisions.

We can be rather sure of our ability to improve the geodetic limit because there are no physical effects that might interfere whose existence has even been suspected. If there were any effects that might interfere, it is almost certain that they would have given us some premonitions of their existence. However, in order to achieve an improvement by an order of magnitude, we shall have to revise our ideas and approaches in two important ways.

The Adoption of a New Reference Surface

Until the present time, we have assumed that the physical surface of the ocean is a surface of fundamental physical importance, namely a surface of constant potential. If the waters of the oceans were at rest, this assumption would be correct. Actually, the waters are not at rest, and the assumption is incorrect. The waves and tides quickly average to zero, and they do not affect the validity of the assumption. However, there are well known currents that are permanent or nearly so, and they must be driven by physical forces. Since the forces cannot exist unless there are differences in potential, the ocean surface cannot be a surface of constant potential. Estimates of the forces required to drive the currents are difficult to make, but those estimates that exist indicate that the mean ocean surface departs from a surface of constant potential by amounts that may range up to a meter. Hence we shall probably have to abandon the ocean surface as our reference surface and learn how to use a constant potential surface in its place, if we are to improve the geodetic limit by an order of magnitude.

The Introduction of Time-Dependent Coordinates

Even at the present level of accuracy, it is no longer possible to speak of the latitude and longitude of a place as if they were constant. Both the North and South Poles move rather irregularly over areas about the size of a baseball infield, and each polar motion changes both latitudes and longitudes. Doppler location already provides a standard method of following the poles.[2] The earth tides also change the coordinates of points by amounts that will have to be included if we are to speak of accuracy exceeding a meter. Most importantly, we know that portions of the earth's crust move with respect to other portions at rates of a few centimeters per year. In order to cope with this phenomenon, we must abandon the idea of defining a coordinate system by the use of one or more "fixed" points such as the Greenwich Observatory. We must learn to define an "earth-fixed" coordinate system even if all the points in the system are in motion.[3] The coordinates of an identifiable point such as a brass marker must then be given as functions of time in this system.

ATMOSPHERIC DRAG ON A SATELLITE

Two forces that are not gravitational in origin affect the motion of a satellite. One is the drag produced by the residual atmosphere that is still

found at satellite altitudes; the other is the force produced by the pressure of solar radiation.

The structure of the upper atmosphere is highly complex; it varies drastically with latitude, longitude, altitude, and time.[4] To describe the matter loosely, solar heating "boils" matter up from the lower atmosphere, and dissociates many of the molecules found in the lower atmosphere into their constituent atoms. The atoms, being lighter and perhaps more heated, then diffuse upward, so that much of the atmosphere at satellite altitudes is atomic rather than molecular. This process of producing upper atmosphere is continually being opposed by the tendency of the atoms to recombine into molecules.

From this fact, we expect the density of the daytime atmosphere to be greater than that of the nighttime atmosphere, and measurements show that the ratio of day to night densities at 1000 km altitude or above may be ten or more. However, the "hot spot" is not directly under the sun; because of time lag in the diffusion process, the hot spot is somewhat east of the subsolar point.

Most of the solar spectrum is not effective in boiling up the upper atmosphere. Because a key effect is molecular dissociation, the far ultraviolet is most effective. While most of the solar spectrum is extremely stable, the intensity of the far ultraviolet varies over a large range. It varies with sunspot cycle, with a period of about 11 years, but it also varies sporadically from day to day. The far ultraviolet is completely absorbed by the upper atmosphere, so its strength cannot be monitored at ground level. However, it is highly correlated with the solar spectrum in the microwave region (wavelengths of 10 to 20 cm) and with the magnetic activity of the sun, which can both be monitored at ground level. Thus we can monitor the temporal variations of the upper atmosphere with reasonable accuracy by ground-based observations.

Under typical conditions, experience with the satellites in the Navy Navigation Satellite System shows that the upper atmosphere may give them a drag acceleration of about 10^{-6} centimeters per second per second. Over a period of 14 orbital revolutions, about one day, this displaces the satellite in the direction of its orbital motion by about 100 meters; I shall use the latter as a standard figure to which all other drag results will be normalized. In spite of the size of this displacement, the drag limit on the accuracy of location is certainly no more than 10 cm on a global basis, as I shall now show.

To start with, we can determine the parameters of the satellite orbit by using the data over an interval of one day. The orbital period found this way is the average period over a day, and the simple fact of letting the period be fitted to the data cuts the maximum drag error by a factor of six. However, we are not restricted to this action. On the basis of much experience with the satellites in the navigation system, we have calibrated Jacchia's theory[4] so that it yields the actual drag experienced by the satellites, with considerable accuracy. When the satellites are used for real-time navigation, it is necessary to use a predicted value of drag in the orbital calculations. When the satellites are used for locating places in a research program, however, we can analyze the data after the fact and use the monitored values of solar activity.

More simply, and probably more accurately, we can take the calibration constant for Jacchia's theory, or some equivalent parameter, as an unknown to be determined from the tracking data at the same time that we determine the other orbital parameters. When we do this, the orbit is determined as accurately as the data and our geodetic knowledge allow, and the drag does not impose any limit to the accuracy of the process.

In writing this, I have assumed that the atmospheric density does not change markedly with time during the tracking interval. If the density does change with time in a way that we cannot follow in detail, we must ask what limiting accuracy may result.

Suppose that the density function changes linearly by a factor of two within the tracking interval of a day, and that we find the orbit by using an average density that gives the best fit. The change in density puts a cubic function into the displacement of the satellite, and the cubic is antisymmetric about the middle of the interval. Thus it is effective for only half of the interval, since the average is removed by the tracking process, and the average is the value at the center of the span. Since the variation is as the cube of the time, it is small during most of the interval. Further, the orbital period found for the satellite automatically adjusts to give the best straight line fit to the cubic, and this reduces the residual effect even farther. The algebra to find the residual effect is trivial, and I shall give no details.

The result is the following: If the average density has the value needed to give a total displacement of 100 meters in a day, if the density varies by a factor of two within the day, and if the average density is found by a fitting process, the standard deviation of the residual error is 43 cm.

This does not represent the limit, however, because it is the error that is present when we make no attempt to eliminate it. We can reduce the error in either of two ways:

1. Since we can monitor the solar activity, we can take the time derivative of the density from the monitored activity, and fit out the average in the process of finding the orbital parameters. This would surely leave no more than about a fourth of the error, or about 10 cm.
2. If we cannot satisfactorily remove the error, for reasons that we cannot foresee, we can still take advantage of the fact that changes by a factor of two in one day are rare. Since we do not have to use all of the data in a research program, we can simply omit data gathered during a day when solar activity is changing rapidly.

In conclusion, it seems safe to say that the global limit imposed by drag acting on the satellites is less than 10 cm.

RADIATION PRESSURE

At an altitude of 1000 km, the electromagnetic pressure exerted by the solar radiation is about ten times the drag pressure exerted by the residual atmosphere. If the forces acted in the same direction, the resulting perturbation would be 1000 meters in a day; a perturbation of this size might pose a serious problem. Luckily radiation pressure does not act in the same direction as drag, as we can see with the aid of Fig. 5.

The solar radiation arrives from the right in Fig. 5. If the satellite is symmetrical about this direction, the resulting force acting on it is directed toward the left, away from the sun. If the satellite is not symmetrical about the direction to the sun, the resulting force may also have a component perpendicular to the direction of the sun.[5] If the radiation pressure is to be calculated accurately, one of three conditions must be satisfied: The satellite must be symmetric. If not, it must rotate so rapidly that it seems symmetric on the average over a reasonable time. If it meets neither of these conditions, we must know its configuration and its orientation in space at all times. Under any circumstances, the force component away from the sun is considerably larger than the perpendicular component. In this error analysis, we can consider only the component away from the sun.

The component of force that is tangential to the orbit is largest when the sun lies in the plane of the orbit, and this is the condition shown in Fig. 5. The smaller of the two circles whose center is E represents the earth, and the earth's shadow is the lightly hatched region to the left of the earth. Within this region, there is no radiation pressure.[6] The satellite emerges from the shadow at point A, and from there to point B the tangential component of force is opposed to the velocity. From B around to C, where the satellite enters the shadow again, the tangential component is in the same direction as the velocity. If the orbit is circular, the net displacement in a full revolution is zero.

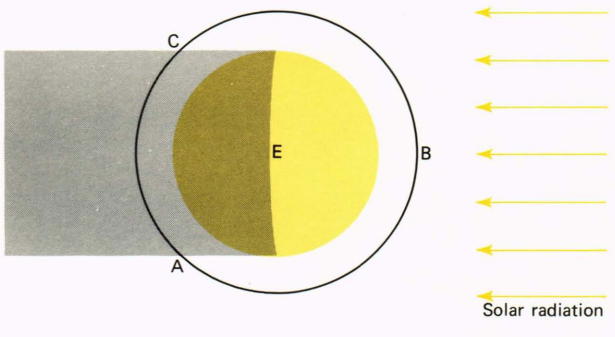

Fig. 5—Radiation pressure acting on a satellite. The smaller circle whose center is *E* represents the earth, and a satellite travels around it on the circle *ABC*. Solar radiation impinges from the right. The earth's shadow is shown by the shaded region, and there is no radiation pressure within this region.

The distance from A to B is about a third of a revolution, but the tangential component is small over much of this distance. If we say that the radiation force opposes the velocity for a fourth of a revolution, we shall have the right size of the effect. In 14 revolutions, the perturbation would be 1000 meters if the force were actually tangential. In a quarter revolution, the displacement is only $(1/56)^2 \times 1000$ meters, and this is about 40 cm. That is, the displacement is about 40 cm when the satellite reaches B, but this is cancelled by the time it reaches C. We can surely calculate this high frequency effect with reasonable accuracy, and therefore it imposes no appreciable limit upon the accuracy of location, even on a global basis.

If the orbit is not circular, the time spent between A and B need not be the same as the time between B and C, and there may be a net displacement during a full revolution and over a day. This displacement is of the order of 1000 meters times the eccentricity of the satellite orbit. For satellites used in navigation or other processes of location, we deliberately make the eccentricity small; the eccentricities of the satellites in the navigation system are about 0.01. Hence the cumulative effect of radiation pressure over a day is about 10 meters, a tenth of the effect of drag. By taking suitable precautions, we expect to cut the residual effect of drag to less than 10 cm. Since the radiation pressure, being almost constant in time, is more tractable, we can expect to cut its residual effect to less than 1 cm.

ELECTRICAL NOISE

There are several sources of electrical noise in a system of Doppler location. We have phase jitter in the satellite's oscillator and in the observer's oscillator, we have ambient radio frequency noise at the observer's antenna, and we have internal noise in the observer's receiver.

The oscillators used so far in the navigation satellite system are quartz crystals in multiple-walled Dewar flasks, with active temperature control in the space between the outer pair of walls. The oscillators used in both the satellite and the ground system are basically the same. Because the frequency of a crystal oscillator is a function of temperature, the operating temperature is chosen to be the temperature at which the derivative of frequency with respect to temperature is a minimum. For the crystal cuts used, the operating temperature has been about 50°C.

The phase jitter of the oscillators has been measured by comparing two oscillators with each other. For example, we can start counting cycles of both oscillators at some instant and end the count of both at some later instant. The interval over which the count is measured is called the averaging interval. We then divide the difference between the counts by

the time interval to get a frequency difference, and we further divide this by the operating frequency to get the relative frequency difference. Finally, we find the standard deviation of the relative difference from zero.

For the oscillators used in the navigation satellite system, the noise level found in this way[7] is typically 60 parts in 10^{12}. Since two oscillators are involved in the measurement, the noise contribution of each oscillator is this number divided by $\sqrt{2}$. I shall round this quotient to 40 parts in 10^{12}. These numbers are for an averaging interval of 1 second.

We routinely measure the noise level of the data obtained by the Doppler tracking stations of which there are two types. One type uses a simple whip antenna, while the other uses a helical antenna with a gain of about 10 dB. In a recent sample, for an averaging interval of 1 second, the first kind showed a noise level of 81 parts in 10^{12} while the second kind showed a level of 57 parts in 10^{12}. Since the noise measured this way includes the effects of two oscillators, one in the satellite and one in the station, the noise measured with the helical antenna is almost exactly that expected from the oscillators alone. If this is so, it means that the receiver noise is small compared with the oscillator noise; it further means that the helical antenna reduces the ambient noise to the point that it is also negligible compared with the oscillator noise.

The noise generated in the receiver itself is probably about the same in both types of station, so the difference between the two noise levels should be mostly the result of ambient radio noise. It seems that the helical antenna provides about as much gain as is useful with the present type of oscillator.

With the aid of computer simulations, Guier and Weiffenbach[8] have developed the following semi-empirical formula for the position error E caused by noise:

$$E = 2.5 \times 10^{13} \; N^{-1/2} \mu \text{ centimeters.} \qquad (4)$$

In this, μ is the noise level for one-second averaging, and N is the number of seconds for which data are obtained. It is explicitly assumed that Eq. 4 applies to the data obtained during a single pass of a satellite and that we attempt to find only two coordinates from that pass. However, we do not need to make any substantial change if we derive all three coordinates, provided it is understood that at least two passes must be used if we do this. It is also assumed that the satellites have an orbital altitude of about 1000 km.

Under these conditions, N is about 625 for a single pass. If we use this value,

$$E = 10^{12} \mu \text{ centimeters.} \qquad (5)$$

Thus a noise level of 60 parts in 10^{12} produces a location error of about 60 cm for a single pass.

There have not been many chances to test Eq. 5 experimentally because other sources of error outweigh the noise under most conditions of use. However, for a few days during the spring of 1971, we did have the opportunity to operate two Doppler receivers simultaneously at the Applied Physics Laboratory. We were able to obtain data from nine satellite passes, using two receivers operating from a common antenna but with separate local oscillators. We were also able to obtain data from eleven passes in which the receivers had a common antenna and also a common local oscillator. The results of these 20 passes are summarized in Table 1.

In order to understand the table, we must know what is meant by the number of deleted points in the second column of the table. The sets were of the integrated Doppler type, in which the count of cycles was read about every 25 seconds, with about 30 counts being obtained during a single satellite pass. In a time series of measurements, especially those obtained by radio apparatus, there will frequently be some readings that are not valid measurements. A burst of external noise, a line transient, or the like, may cause a reading that is characteristic of the noise rather than of the phenomenon being studied. In order to find such readings, we first find the parameters that give the best fit to the series of readings, find the standard deviation of the residual, and try eliminating all residuals that are more than three standard deviations. We repeat the process with the remaining series, and continue in this way until the process stabilizes. If no more than, say, 10% of the series are deleted in this way, we usually accept the remaining readings and use them.

It is not clear that this acceptance is valid. If a reading is deleted, it indicates some sort of malfunction, using the term in a very general way. It may be that the malfunction was confined to the time interval in which the deleted reading was made, but it may well have had some existence outside the interval that was at a level too low to be detected. Hence, the best choice may be to ignore an entire series if even one reading or point is deleted.

We have tested this idea in Table 1. The table gives statistics on both the measurements obtained with

Table 1

COMPARISON OF TWO DOPPLER RECEIVERS AT THE SAME LOCATION OPERATING SIMULTANEOUSLY

Oscillator	*Number of Points Deleted*	*Number of Passes*	*Measured Separation (cm)*	*σ (cm)*
Separate	0	4	32	95
	1-3	5	177	205
Common	0	4	15	39
	1-2	7	69	113

separate oscillators in the two receivers and on the measurements obtained using a common oscillator for both sets. For each of these conditions, we first exhibit those passes in which no point was deleted from the series obtained with either receiver. We then exhibit those passes in which some points were deleted, perhaps from only one set or perhaps from both. In this sample, out of about 60 readings obtained by both receivers during a pass, a total number of deleted points ranges from 1 to 3 when the oscillators were separate and from 1 to 2 when the oscillators were common. This difference must surely be an accident; there is no fundamental reason why more points should be deleted in the first case.

From each pass, we infer a position for each receiver and subtract the positions to find the measured separation. The separation is a vector, and the table gives the magnitude of the average vector separation. It also gives the standard deviation of the magnitude, taken about zero rather than about the mean.

The difference in performance between the passes with no deletions and those with one or more deletions is quite striking. From the table, we may tentatively draw two conclusions:

1. In precise work, we should delete an entire pass if we delete even one point from either receiver; doing so pays off more rapidly than using all passes and relying upon statistical improvement.
2. When we use only passes in which no points are deleted, the error produced by noise is some tens of centimeters, in accordance with Eq. 5.

As we expect, the noise is greater with separate oscillators than it is when we have a common oscillator in both ground sets. In view of the small sample, attempts at more detailed analysis of the data are probably not warranted.

In summary, theoretical considerations indicate that the noise contribution to Doppler location from a single pass in which no points are deleted is of the order of 50 cm, even when we use only a whip antenna. The small amount of experimental data that is available confirms this estimate reasonably well, but it is desirable to obtain more data.

If we accept this figure, we should ask how many passes are needed in order to reduce the error to 10 cm. We need to use 25 passes in order to obtain this reduction, and we can use about half of the total passes. Thus about 50 passes will be needed. This is the number of passes obtained in 4 or 5 days.

The preceding discussion applies to the oscillators that are now in service. During the past year, we have been doing laboratory tests on a new type of oscillator called NP4. The tests performed so far indicate[9] that the noise level of these oscillators is an order of magnitude below the level of the oscillators now in service. If this laboratory improvement can be carried over into field equipment and satellites, the noise contributions will fall to about 5 cm for a single pass.

OSCILLATOR DRIFT AND TIMING ERRORS

We saw in the section on the practice of Doppler location that the user may determine f_A, the difference between his local frequency standard and that in the satellite, at the same time that he determines his position. If necessary, he can make a separate determination of f_A for each pass that he uses. We must now ask how sensitive his inferred position is to errors in f_A. For example, suppose that he uses a satellite frequency supplied by the operators of the satellite system, that he measures his local frequency by some means independent of the satellite system, and that he combines the two frequencies to find f_A, instead of inferring it from the Doppler data. What error in location will he make?

The error depends somewhat upon the exact relative geometry of the observer to the satellite orbit, but we can calculate a representative error by averaging over all geometries. The result is the following rule of thumb:

$$1 \text{ part in } 10^{13} \text{ in frequency} = 1 \text{ centimeter in position.} \quad (6)$$

Equation 6 is not new. On the contrary, it has been considered standard for so long that I do not know where it originated.

The requirement in Eq. 6 is so stringent that the user who wants precise results must probably infer f_A from the satellite data, even if he is a user who can afford an atomic frequency standard.

Since the user does not need an absolute knowledge of either his frequency or that in the satellite, the next limitation comes from a drift in the frequency of either oscillator. When the Doppler navigation system was designed, atomic standards were still rather exotic, so the system was designed to use crystal standards both for ground observers and in the satellites. The worst oscillator that has been used in the system[7] has a drift rate of about 1 part in 10^{10} per day. A pass lasts about 15 minutes, or about 0.01 days, so that the frequency change during a pass is about 1 part in 10^{12}. According to Eq. 6, this gives a position error of about 10 cm if no attempt is made to counteract the error. This is in the worst case. The error is about 1 cm in the best case.

It is quite simple to counteract the drift error. The frequencies of the oscillators in both the satellites and the ground equipment can be monitored and the average drift rate can be determined. If necessary, the measured drift rates can be used in the calculations of position. The question then is not the size of the drift but the amount that the rate over an interval of 15 minutes can depart from the average. We have not attempted to make measurements of this sort, but there can be little question that the deviations should be considerably smaller than the average. Thus the drift contribution to location error should probably be measured in millimeters rather than centimeters.

Let us now look at the requirements on the accuracy of measuring time. The satellites in the Doppler navigation system move somewhat less than 1 cm in 1 microsecond; for simplicity, let us say that the rate is exactly 1 cm per microsecond. Obviously, then, a timing error of 1 microsecond means a position error of 1 cm.

The ground stations in the Doppler navigation system maintain their own atomic time and frequency standards. At appropriate intervals, a portable cesium clock is transported from the Naval Observatory to each ground station in turn and back again; this serves to keep all the clocks in the system accurately set. At the time of the resetting provided by the visits of the portable clocks, the errors are typically 10 microseconds. Thus an observer who kept exactly the time of the Naval Observatory would make an error of 10 cm in his position on a single pass. This error would average to zero almost immediately, however, for the reason that will now be explained.

We remember from the section on the principle of Doppler location that the observer first locates himself in a plane that is normal to the satellite orbit and that passes through the position of the satellite at the time of closest approach. He then finds his distance from the satellite at the same time by using the slope of the Doppler curve at closest approach. If his clock is correct while the one used by the satellite system is in error, he makes an error in the position of the plane but not in his distance from the satellite. In order to simplify the discussion, let us suppose that the observer is at the equator. A mistake in the position of the plane means an error in latitude; there is no error in longitude under the assumptions made.

Now let us suppose, for example, that the observer at the equator uses the satellite on a pass when it is going north. Let us further suppose that the timing errors are such that he puts himself 10 cm too far north. Later he observes the satellite when it is going south. On this pass, he puts himself 10 cm too far south. His average position is correct in spite of the timing error. The same conclusion holds for any position of the observer, but the language involved is more complicated if he is not at the equator.

This conclusion should not be taken to mean that timing errors are unimportant and that they can be allowed to take on any size. There are second order effects of timing that may not average out and that can be serious if the errors are too large. However, if the timing errors are 10 microseconds or less, the second order effects are less than 1 cm.

We have seen that the error in the time used within the Doppler navigation system is of the order of 10 microseconds. It is this large because there has been no requirement to keep it smaller. If there were a requirement, there is little question that the time error could be held to 1 microsecond or less.

Any user can maintain the same error if he wishes to take the trouble. However, we must consider the user who has only a crystal frequency standard and clock. If the reliability of his frequency standard is 1 part in 10^{10} per day, his time can drift by 10 microseconds per day. If he is going to maintain a timing accuracy of 10 microseconds, he must calibrate his clock almost daily.

There are various ways in which he can do this. If he is going to use the satellites in the Navy Navigation Satellite System (the Doppler system), his simplest procedure is to use the satellites themselves. Each satellite in this system transmits a timing signal every 2 minutes. The timing signals are controlled by a clock that in turn is controlled by the same oscillator that controls the transmitted frequency. The timing signals are monitored by four stations in the ground system that are equipped with cesium time and frequency standards, and the clock in each satellite is reset every 12 hours on the basis of the monitoring data. Thus the errors in the satelilte clocks are held to a few microseconds.

We must now ask about the precision with which a ground station can compare a timing signal with its own local clock. Ten years ago, this precision was 21 microseconds for the timing signals obtained during a single pass.[7] Since 10 passes or more can be received in a day, the timing error can be reduced to about 7 microseconds. This leads to a location error that is actually somewhat less than 7 cm for a single pass. Because of the automatic averaging that results from the satellite motion, this quickly reduces to 1 cm or less, as we explained above.

REFRACTION

We now turn to the processes that affect the radio signals as they travel between a satellite and a user on the ground. Since the space between a satellite and the ground is never a vacuum, the signals interact continuously with the matter that is encountered along the signal path. At altitudes above, say, 100 km, the dominant interaction is with the electrons that make up the ionosphere. The positive ions found there affect the signals much less than the electrons because they are much more massive, while the neutral material is too rarefied to have an appreciable effect. In the troposphere, however, the neutral molecules provide the dominant interaction.

Hence we need to consider refraction in the ionosphere and refraction in the troposphere. The nature of the refraction is quite different in the two regions, but there is a certain general principle that applies to both; that principle is the subject of this section.

We let n denote the index of refraction of the material that is found at any point between the satellite and the ground. The index n is a function of the properties of the material and, since these properties (such as density) change continuously with position, n is also a continuous function of position. Further, since the properties at a specific point in space obviously change with time, n is also a function of time. However, the time scale of the temporal changes in n is usually long compared with the dura-

tion of a satellite pass. Hence we shall neglect the time dependence of n.

Suppose that the satellite is in a certain position. We assume that the radiation that it emits from this position is propagated according to Fermat's principle, and we construct the wave fronts that result from that principle. At any point, we say that the ray direction is the direction that is perpendicular to the wave front at that point. We then trace the ray path that leads continuously from the satellite to the user on the ground.

Now let s be the coordinate that measures length along the (curvilinear) ray path. Since the index of refraction is a function of position, it can now be regarded as a function $n(s)$ along the ray path. The quantity called the optical path length is the integral of $n(s)$ with respect to s. When we review the discussion in the section on the principle of Doppler location, we see that the quantity called the range is actually the optical path length. That is, all the considerations of the first section remain valid provided that we define r as

$$r = \int n(s)\, ds. \tag{7}$$

The index of refraction between a satellite and the ground is always close to unity, and its maximum deviation from unity is between 10^{-3} and 10^{-4}. Thus the ray path is always close to the straight line between the satellite and the receiver. This means that the difference Δr between the optical path length and the straight line distance is, to first order,

$$\Delta r = \int (n(\ell) - 1]\, d\ell, \tag{8}$$

in which $d\ell$ is an element of length along the straight line.

Since the optical path actually follows a curved path rather than the straight line, there is a contribution to Δr that is proportional to the difference in length between the curved path and the straight line; this contribution is of order $[n(s) - 1]^2$. Still other contributions involve higher powers of the parameter $n(s) - 1$. Thus the quantity r can be expanded as a power series in this parameter.

REFRACTION IN THE IONOSPHERE

At any point in the ionosphere, there is a certain density N of electrons. For a given value of N, the index of refraction n can be written in the form

$$n = 1 + \sum_{i=2}^{\infty} (\alpha_i / {f_T}^i)\ . \tag{9}$$

Note specifically that there is no term proportional to ${f_T}^{-1}$, although all other negative powers of f are present. The coefficients α_i do not depend upon the frequency, but they do depend upon several other things. They depend upon the electron density N, and they therefore depend implicitly upon position for this reason. Some of them depend upon the magnitude of the magnetic field, and these depend implicitly upon position for this reason also. Those which depend upon the magnetic field also depend upon the angle between the magnetic field vector and the ray direction, and they further depend upon the polarization of the radiation.[10]

The two independent components of polarization are the circular components, namely the right-hand and the left-hand components. If a signal is transmitted with only a right-hand component, say, it will have only a right-hand component all the way to the point of reception. At any point along its path, it will have an index of refraction n_r. If the signal starts with only a left-hand component, it will arrive with only a left-hand component, and it will have an index n_ℓ.

Now suppose that the signal is transmitted with linear polarization. If there were no ionosphere, the polarization would remain linear and the observer could receive it readily with a linear antenna such as a whip. However, the linear polarization consists of right-hand and left-hand components of circular polarization with equal strengths. Since these components have different indices of refraction, they travel with different velocities. When the components are recombined at the receiver, the result is that the direction of the resulting linear polarization is constantly rotating, so that the polarization is sometimes aligned with a whip and is sometimes normal to it. The signal strength falls to zero at these times.

Thus, if an observer uses a simple linear antenna, the signal strength will vary from a maximum all the way down to zero unless the transmitted signal has only a single component of circular polarization. Therefore the satellite antennas should be designed so that they transmit circularly polarized radiation. Since they cannot do so and simultaneously transmit equally in all directions, they must have favored directions of transmission. This in turn means that their orientation with respect to the earth must be controlled with a moderate amount of accuracy, in order that the favored direction of transmission may be pointed at the earth.

Now we suppose that only a single component of circular polarization is transmitted, so that we are concerned with only a single index of refraction n; this single index is, however, a function of position. We saw in the preceding section that the optical path length equals the straight line distance, ℓ, plus a power series in the parameter $n - 1$. We see from Eq. 9 that this power series becomes a series in inverse powers of f_T that starts with ${f_T}^{-2}$. That is,

$$r = \ell + \sum_{i=2}^{\infty} (\beta_i / {f_T}^i).$$

By Eq. 2, the Doppler shift f_D equals $-(\dot{r}/c)f_T$. Hence f_D has the form

$$f_D = -(\dot{\ell}/c)f_T + \sum_{i=1}^{\infty} (\alpha_i / {f_T}^i). \tag{10}$$

If we knew the electron density at all points between the satellite and the ground for all times during a pass, we could evaluate theoretically the coefficients α_i in Eq. 10. On the basis of present knowledge, we cannot hope to calculate the α_i with sufficient accuracy by using measured values of the electron density. However, by using various measurements of density as a function of position, we can reach an important conclusion: At the frequencies that interest us, the next largest term[11] after α_1/f_T is the term α_3/f_T^3. Specifically, the term α_2/f_T^2 is small compared with the cubic term.

The frequencies used in the Doppler navigation system are 150 and 400 megahertz. If an observer uses the higher frequency, and makes no attempt to correct for refraction in the ionosphere, he may easily make an error that is in the range of hundreds of meters. In order to reduce this to 10 cm, the observer would have to know the electron density at all relevant points with an accuracy better than 1 part in 10^3, and this is not possible with methods known at present.

In order to reduce the ionospheric refraction error to a tolerable level, we make use of two coherent frequencies, which are 150 and 400 megahertz in the navigation system, as we have just said. Then we assume that we can neglect all terms in Eq. 10 above the term α_1/f_T. The equation then contains two unknown parameters, namely $\dot{\ell}/c$ and α_1. Let f_1 and f_2 denote the Doppler shifts measured with frequencies 150 and 400 megahertz, respectively, and let the megahertz be the unit of frequency. Then

$$f_1 = -150\,(\dot{\ell}/c) + (\alpha_1/150),$$

$$f_2 = -400\,(\dot{\ell}/c) + (\alpha_1/400).$$

From these, we find

$$f_2 - (3/8)f_1 = -343.75\,(\dot{\ell}/c). \qquad (11)$$

The left member of this relation is a measured quantity, and we calculate $\dot{\ell}/c$ at each instant during a pass from it. The quantity $\dot{\ell}/c$ in turn is the quantity that we use in inferring position, by any of the methods described in the section on the principle of Doppler location.

In order not to degrade the accuracy when we find the difference $f_2 - (3/8)f_1$, we must take two important precautions:

1. Both frequencies transmitted from the satellite must be controlled by the same oscillator, and all frequencies involved in the ground equipment must be controlled by a common oscillator.
2. Instead of measuring f_1 and f_2 independently and forming $f_2 - (3/8)f_1$ by calculation, the frequency f_1 must be multiplied by 3/8 by means of phase-locked multiplying circuits, and the result must be beat against f_2 in order to form the difference. The beat frequency is then directly measured.

If the electron density were a function of altitude only, with no gradients in the horizontal direction, the effect of refraction upon r would be symmetrical about the point of closest approach. This means that the refraction would not affect the value that we measure for the time of closest approach and that it would affect only the value of the range at closest approach. However, the ionosphere varies systematically with latitude, and the electron density therefore does have a horizontal gradient. In order to study this problem as well as other ionospheric problems, we equipped several early satellites with three or more coherent frequencies. This allows us to solve for coefficients beyond α_1 in Eq. 10. Because of limitations in accuracy, it has been possible to find only one coefficient in this way. Since theoretical studies show that α_3 is the most important coefficient after α_1, work with multiple frequencies done so far is based upon the assumption that Eq. 10 contains only the terms in α_1 and α_3, with other coefficients being set equal to zero.

Much of this work has been done by Willman and Doyle.[11] The question studied in their work is the following: If we eliminate the coefficient α_1 by using Eq. 11, what is the remaining error in position because of higher coefficients? There are two main conclusions:

1. The standard deviation of the remaining error is 2 meters, although one instance was found in which the error was 20 meters.
2. Because of the latitude dependence of the electron density, the error parallel to the satellite motion is often as large as the error in the range at closest approach. Further, the error parallel to the satellite motion is not necessarily equal and opposite for northbound and southbound motions of the satellite. Instead, the error tends to introduce a bias in the latitude of the observing station.

In spite of these results, ionospheric refraction does not impose an important limit upon the accuracy of Doppler location. The satellites in the Doppler navigation system were basically designed in 1960, when it was not possible to contemplate a frequency higher than 400 megahertz in a satellite system that had to achieve routine operational status within a few years. Now there would be no difficulty in going to frequencies at least three times as high. Since most of the error in Eq. 11 comes from the term α_3/f_T^3, the error in location when we use Eq. 11 varies as f_T^{-4}. If we triple the operating frequencies, we divide the error by 81. This reduces the standard deviation of the error from 2 meters to about 2 cm.

REFRACTION IN THE TROPOSPHERE

In the part of the troposphere that is in line of sight of a particular ground observer, the index of refraction is usually a function of altitude only, at a particular time. If there is a weather front within line of sight, this condition may be violated; we have sometimes detected the presence of weather fronts by

analyzing the Doppler data from a satellite. However, we can tell independently when a front is close to a station. Such occasions are moderately rare, and thus we can afford to ignore data obtained under conditions when the weather is not suitable. With this understanding, then, we can say that the index of refraction in the troposphere is independent of horizontal position and depends only on altitude.

Refraction is severe when the satellite is near the horizon, so we avoid the use of data obtained at low elevation angles. In most location work that we have done at this Laboratory, we have adopted the following rules:

1. We do not use any data from a pass unless the satellite attains an elevation angle of at least 15°;
2. We discard all data when the elevation angle is less than 10°, and we use this "cut-off" for all passes that are retained under rule 1.

When we adopt these rules, the worst error that arises from tropospheric refraction[7] is about 30 meters, and the standard deviation of the error is about 20 meters. The error never falls below about 12 meters.

The errors just quoted are those found when we make no attempt to eliminate the effects of tropospheric refraction, other than eliminating data obtained at low elevation angles. Since the index of refraction in the troposphere depends but little upon the frequency at radio frequencies, we cannot proceed as we did with ionospheric refraction. In the present state of knowledge, we can attempt to eliminate tropospheric refraction only by calculating it theoretically.

Hopfield[12] has made the most intensive study of the effects of tropospheric refraction on radio signals from satellites, and the following discussion is based upon her work. We must start by dividing the refraction into a "wet" component, resulting from the water vapor in the troposphere, and a "dry" component, resulting from all other constituents. The dry component, except perhaps under extreme conditions, is many times the wet component. For purposes of illustration, we may say that the effect of the dry component is 20 meters and the effect of the wet component is 50 cm.

To high accuracy, the index of refraction at any point in the troposphere is a function of the pressure, temperature, and relative humidity. The goal of Hopfield's work has been to measure these quantities on the ground at the observer's location and to see how the refraction effect can be calculated from the measured quantities, using thermodynamic principles to calculate the variation of the index of refraction with altitude. She then compares the integrated effect of the index with the effect calculated from detailed measurements of tropospheric properties made by balloons. The balloon data are obtained from the National Climatic Center, a part of the National Oceanic and Atmospheric Administration. Hopfield has made these studies at sites as diverse as Samoa, Dulles Airport, and a weather ship in the North Atlantic.

The dry component proves to be quite amenable to theoretical treatment. Since the total effect of the dry refraction depends upon an integral taken through the entire troposphere, it turns out that the dry component depends upon the surface pressure only, being independent of surface temperature. Hopfield's theory, based upon a value of surface pressure measured at the time of each pass, gives a refraction effect that is correct within about 1 part in 500. Thus the remaining error resulting from the dry component is about 4 cm. Since the error seems to be random from pass to pass, so far as we can tell from the available data, the error can be further reduced by using multiple passes. It seems safe to take 1 cm as the limit on Doppler location imposed by the dry component, on the basis of present knowledge.

The wet component is not as amenable to theoretical treatment on a relative basis, but we can afford a larger relative error since the wet component is smaller to start with. Hopfield's current methods give an error of about 25% in dealing with the wet component. This means that the residual error is about 12 cm for a single pass. Again the error seems to be random. If we use six passes, which is about the number of passes that we can obtain in a day, we would apparently reduce the error to $12/\sqrt{6} = 5$ cm, but this calculation is probably illusory. It is plausible that the error is a function of the weather, and the weather is not likely to change in a day. Thus we probably have about the same error for all the passes obtained during a single day, and we can reduce the error statistically only by using passes for which the weather is uncorrelated. This probably dictates operations over several days.

The estimates of the refraction error are those that apply on a global basis. On a local basis, it is only the difference in refraction between two neighboring points that matters. We do not have detailed information about the local variation of the refraction effect. However, for points that are within 100 km of each other, say, it seems implausible that the difference should be more than 10% of the total. Tentatively, then, we shall say that the local limit is a tenth of the global limit.

SUMMARY AND DISCUSSION

In the preceding sections, we have studied all the known factors that affect the accuracy of locating a point by measuring the Doppler shift in the radio transmissions from a near-earth satellite. Table 2 summarizes the limiting accuracy imposed by each source, provided that we take full advantage of present knowledge and techniques. The table does not include the limit imposed by fundamental knowledge of the earth's shape and gravity field; that limit will be discussed separately.

In preparing the table, I have assumed that the observer uses satellite orbits that have been determined from orbital data taken over a period of a day.

Table 2

LIMITS ON THE ACCURACY OF DOPPLER LOCATION FROM A SINGLE PASS, RESULTING FROM ALL FACTORS EXCEPT GEODETIC ONES.

Source	*Limit (cm)*
*Satellite Motion**	
Atmospheric drag	10
Radiation pressure	1
Instrumentation	
Noise	5
Oscillator drift	1
Timing	7
Propagation Effects	
Ionosphere	2
Dry troposphere	4
Water vapor	12
Resultant	18

*It is assumed that the satellite orbits are determined daily.

I also assume that he uses data covering a span of a day in his clock calibration. Otherwise, I have assumed that he uses only data gathered from a single pass of a satellite.

The largest single error in the table is 12 cm from the unpredictable part of the refraction due to humidity in the troposphere. The next largest error is 10 cm from the unpredictable part of the drag acting upon a satellite. The resultant of all the errors listed in the table is 18 cm.

Although the table is based upon current knowledge and techniques, it is not always based upon current practice. The only two errors for which this point is important are those resulting from noise and ionospheric refraction. The limit imposed by noise in the present Doppler navigation system is about 50 cm rather than 5 cm, and we believe that the noise level is set by the oscillators used in the system. The design currently used for the oscillators is about 15 years old. Oscillators that have been tested in the laboratory but not used in the field are better by an order of magnitude. However, until we have actually used these oscillators in both satellites and ground equipment, we cannot know for sure that the improvement assumed in Table 2 can be achieved in field use.

The current limit imposed by ionospheric refraction, on the basis of a single pass, is about 200 cm rather than 2 cm as it is listed in the table. However, this limit varies inversely with the fourth power of the operating frequencies, and it is now feasible to use frequencies at least three times as high as those we are using. Thus it seems safe to say that the effect of ionospheric refraction can readily be reduced by two orders of magnitude simply by increasing the operating frequencies.

With one possible exception, all of the errors listed in Table 2 are random. The possible exception is the error produced by ionospheric refraction. The electron density is a function of latitude[7] and the residual refraction that is not removed by the two-frequency technique tends to produce a bias in the latitude of the observer. However, it should be possible to learn the average latitude dependence of the electron density and hence the average bias left by the two-frequency method. Subtracting the average bias should then leave only a random error.

Thus it should be possible to reduce the errors far below the limits shown in Table 2 simply by using many passes. It is probably not safe to say that the errors can be made as small as we like by using enough passes, because we do not know what correlations and biases there may be at error levels far below those in the table. However, it seems safe to say that the errors could be reduced to 5 cm by using multiple passes. This requires using 13 passes, if ordinary statistics apply, which is about the number of passes received in a day from the satellites in the Navy Navigation Satellite System.

Next, we turn to the geodetic limit. We concluded in the section on the geodetic limit that the acquisition of additional data, without requiring any new observing techniques, would allow us to lower the geodetic limit to a few tens of centimeters on a global basis and considerably less than this on a local basis. Let us use 50 cm as the geodetic limit on a global basis that we can achieve in the foreseeable future, and let us ask what the limit is in locating two neighboring points relative to each other.

The geodetic errors become uncorrelated for two points that are separated by about 90° on a great circle; this is 10,000 km in distance. For two observers separated by 10,000 km, then, the geodetically induced error in their relative position is 70 cm. This is approximately the product of 50 cm, which is our estimate of the global limit in locating a single point, multiplied by $\sqrt{2}$, since two points are involved in a relative location. For two observers separated by 0 km, the geodetically induced error is 0 cm. It should be a reasonable approximation to say that the error grows as the square root of the separation. Hence, for two points separated by L km, the geodetic limit G in finding their relative location is approximately

$$G = 0.7\sqrt{L} \quad \text{centimeters.} \qquad (12)$$

The limit G is the same for all methods of location, whether by Doppler techniques or not, as we saw in the section on the geodetic limit.

Finally, let us estimate the total error involved in the relative location of two points, if we use the Doppler observations that are obtained in one day. We estimate that the non-geodetic factors contribute 18 cm error in the location of a single point if we use

only the data from a single pass. If we obtain 13 passes in a day, we expect to decrease this error to 5 cm. However, since two points are involved in a relative location, we multiply this by $\sqrt{2}$, obtaining 7 cm, approximately.

The total error in a relative location is then the square root of the sum of the squares of 7 cm and of G cm from Eq. 12. The total error E is plotted as a function of L in Fig. 6. For separations less than 100 km, E is dominated by the non-geodetic contribution of 7 cm, which is characteristic of the Doppler method. For greater separations, E is dominated by the geodetic contribution G, which is the same for all methods of measurement.

In many applications, however, the error E is not the error that interests us. Often we are interested only in relative motion. For example, we may be interested in the relative motion of points separated by the San Andreas fault or by other major fault lines in the earth's crust. For another, we may be interested in the relative motion of points in a region where there has been extensive removal of water, oil, or other materials from below the earth's surface. For such applications, we want to know the minimum change in separation that we can detect over an interval of, say, one year.

The geodetic limit G for any two points is a bias that has the same value for measurements made now and a year later. Thus it does not contribute to the minimun detectable separation, which depends only upon the nongeodetic factors. We saw in Table 2 that the nongeodetic factors contain a contribution of 2 cm from ionospheric refraction, and that this may be in part a bias in latitude. The bias part is independent of time and thus it does not affect the minimum detectable motion. However, it contributes a negligible amount whether it is a bias or not, and we can ignore it. The other nongeodetic factors are random, so far as we know. If they are random, their contribution should be $18/\sqrt{N}$, in which N is the number of passes used. Thus the use of 324 passes, which can be obtained in about 30 days, should leave a net contribution of 1 cm to the minimum detectable motion. However, only experimentation can determine whether the minimum can be reduced to this level by averaging over many observations, or whether there are biases that we have so far been unable to determine.

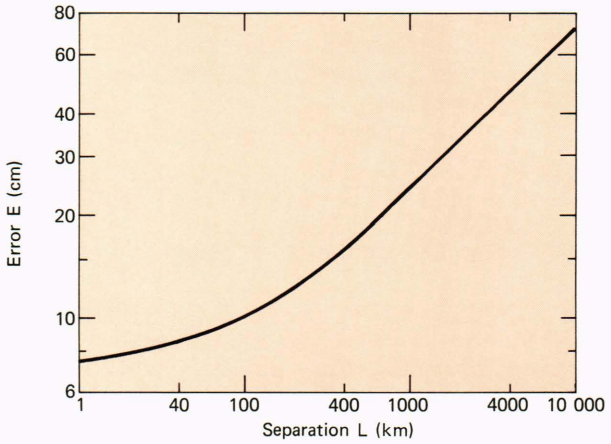

Fig. 6—The estimated error *E* in the relative location of two points as a function of their separation L. The geodetic error is estimated to be $0.7\sqrt{L}$ cm and is independent of the method of measurement. The other error component, which is estimated to be 7 cm for the observations obtained in a single day, is a characteristic of the Doppler method of location. The error probably does not increase for separations greater than 10,000 km, although the figure suggests that it does.

REFERENCES and NOTES

[1] If the observer can use several satellites at once, instead of being restricted to one satellite, there may be a difference in the kind of information supplied by Doppler and ranging systems. Multiple satellites can provide a small amount of extra information if their clocks are synchronized to the necessary precision. It should be clear that a ranging system in this context means a one-way system. It does not include a two-way system such as radar.

[2] R. J. Anderle, "Determination of Polar Motion from Satellite Observations," *Geophys. Surv.* **1**, pp. 147-161 (1973).

[3] R. R. Newton, "Coordinates Used in Range or Range-Rate Systems and Their Extension to a Dynamic Earth," *Reference Coordinate Systems for Earth Dynamics, Proc. International Astronomical Union Colloq, No. 26*, pp. 181-200 (1974).

[4] L. G. Jacchia, "Static Diffusion Models of the Upper Atmosphere with Empirical Temperature Profiles," *Smithson. Contrib. Astrophys.* **8**, No. 9 (1965).

[5] R. R. Newton, *Geophys. J. R. Astron. Soc.* **14**, p. 505 (1968) gives an extensive study of the force due to radiation pressure.

[6] In highly precise work, we must also include the pressure of the thermal radiation emitted by the earth; see Ref. 5.

[7] R. R. Newton, "The U.S. Navy Doppler Geodetic System and Its Observational Accuracy," *Philos. Trans. R. Soc. London* **A262**, pp. 50-66 (1967).

[8] W. H. Guier and G. C. Weiffenbach, "A Satellite Doppler Navigation System," *Proc. Inst. Radio Eng.* **48**, pp. 507-516 (1960).

[9] L. J. Rueger, private communication, 2 Feb 1975.

[10] See K. G. Budden, *Radio Waves in the Ionosphere* (Cambridge University Press, Cambridge, 1961), for an extensive survey of ionospheric refraction.

[11] J. F. Willman and J. F. Doyle, Research Report No. 491, Defense Research Laboratories (Univ. Texas) (1963).

[12] H. S. Hopfield, *Radio Sci.* **6**, p. 357 (1971).

ACKNOWLEDGMENTS—I thank Mrs. H. S. Hopfield, of the Applied Physics Laboratory of The Johns Hopkins University, for a critical reading of the manuscript and for valuable discussions about refraction. This work was supported by the Department of the Navy under Contract N00017-42-C-4401 with The Johns Hopkins University Applied Physics Laboratory.