7 March 1958, Volume 127, Number 3297 SCIENCE

Distance and Relativity

G. C. McVittie

What does an astronomer mean when he says that the sun and the earth are, on the average, 92,900,000 miles apart? And what bearing can the two theories of relativity have on the matter? It is questions of this kind that I shall try to answer in this article, but I must first warn you that I have been described by my scientific colleagues as an uncompromising empiricist. I daresay that this is true, but it is also a little strange, for all my training and all my research work have lain in theoretical astronomy and not at all in the extremely difficult and fundamental task of making astronomical observations. Perhaps, however, the theoretician does have an advantage over his colleagues who are engaged in observational or experimental work. He can stand slightly to one side and ask himself: What exactly are these men doing, what kind of significance can be attached to the results of their efforts, and in what way are their data really conditioned by theories? To speak in generalities would, I think, be profitless, and it is for this reason that I propose to concentrate on the question of distance in the solar system and to leave aside the equally intricate and fascinating problem of distance in the universe at large.

Large-Scale Distances

It might be thought at first sight that the distance between two objects was such an obvious physical idea that there could be little to say about it. Two marks have been made on a certain bar of platinum that is carefully preserved in France; the separation of the two marks defines the unit of length called the meter. If, then, we want to find the distance from the front door of the J. I. Holcomb Observatory in Indianapolis to the front door of the University of Illinois Observatory in Urbana, all we need do is to lay the meter-bar down successively, starting at the first spot and finishing at the second, and count the number of times we have to repeat the operation. The number so arrived at would give the required distance in meters. But if I were to take any one of you to the Holcomb Observatory, hand you the meter-bar, and tell you to go ahead, I feel sure that you would be at a loss as to how to proceed. You would want to know in what direction to start out, and the specification of a direction means that some angle measured from a standard direction must be furnished. You would also want to know how to allow for the fact that you would at times—admittedly not very often in the flat midwestern plains—be going uphill and at others, down. In other words, how are the accidental errors in the measuring procedure to be dealt with? Again you might ask, if you were of a philosophical turn of mind: Do the two marks on the meter-bar correspond to some immutable degree of separation between them, or is this separation affected by accidental circumstances? Clearly, if in the course of your journey you were forced to pass the meter-bar through the center of a blast furnace in which it was vaporized, you would be hard put to it to know what to do next.

This illustration draws attention to three characteristics of determinations of large-scale distance: that angular measurements are involved, that errors have to be compensated for, and that the persistence of the unit of distance is important. But there are other matters of equal moment. Suppose that we have somehow measured to our satisfaction the distances from the Holcomb Observatory to the University of Illinois Observatory in Urbana and to the Dearborn Observatory in Evanston, respectively. Then is it necessary to repeat the measuring operation between the two last-named observatories in order to find the distance between them? The answer is, unfortunately, yes, if we do not know the geometry of the surface on which these three points lie. But if we do know it, then the rules of the geometry will enable us to calculate the required distance without further appeal to measurements. Conversely, measurements of distances between a sufficient number of points will give us a clue to the geometry that must be used to describe the interrelationships between the distances and the angles that are necessarily involved in them. With geometry, mathematical theory enters into our distance computations and, unfortunately for those who like their science to be simple, there is more than one kind of geometry. For example, the ordinary plane geometry of Euclid that we learn at school is different from the geometry on the surface of a sphere that surveyors use when they are making a survey of the United States. By this I mean that the mathematical formulae that are employed for interrelating distances and angles in the one case are not the same as those used in the other.

If you feel dismayed by the complexity of these problems, I shall now increase your dismay by introducing some further factors that are of the greatest importance in astronomy. The three observatories I have referred to are, to the best of our knowledge and belief, at mutual relative rest, their distances apart remaining unchanged as time goes on. This very simple situation almost never

Dr. McVittie is the head of the University of Illinois Observatory, Urbana. This article is adapted from a paper presented 26 December 1957 at the Indianapolis meeting of the AAAS.

arises in astronomy, nearly every astronomical object being in motion relative to ourselves on earth. Clearly, a distance determination for a moving object is meaningless unless we are also told the time to which the determination refers. Thus, two more elements are involved when the distances of moving objects are concerned—that is, we must know what is meant by the time of occurrence of an event and we must also know how the motions of bodies are governed, what the theories of dynamics and of gravitation may be.

But even if these complications have been allowed for, there is an additional one that enters into distance determinations in astronomy. Up to the present, it has not proved possible for an astronomer to leave the earth and go to any of the bodies whose distances he would like to find. In this respect, he differs from the terrestrial surveyor who can, at a pinch, actually proceed from one point to another on the earth's surface. The astronomer has to rely almost entirely for his knowledge of the astronomical universe on the information conveyed to him through the medium of electromagnetic radiation, whether in the form of optically visible light or in that of radio waves. The behavior of electromagnetic radiation in the astronomical universe is partly determined by experimental evidence in the laboratory but also very largely by a liberal use of theory. In the application to the solar-system distance problem, it can be established experimentally on the terrestrial laboratory scale that light rays travel in straight lines in Euclid's sense of this term. The belief that this is a property of all electromagnetic radiation and that it continues to hold as the radiation travels over the vast spaces between astronomical objects is based on theory, and a theory is something which is constructed by the human mind and which therefore need not enshrine some everlasting truth. When we come to discuss the phenomenon of the bending of light rays, we shall see that this laboratory conclusion about the nature of the paths of light rays cannot always be sustained.

Distance in the Solar System

After this brief survey of the main difficulties of principle that confront the astronomer when he tries to determine the distances of the objects in which he is interested, let us examine the procedures that he in fact employs. The famous astronomer Johannes Kepler, who lived from 1571 to 1630, showed how it was possible to set up a scale model of the solar system without having any accurate notion of the distances involved in terms of miles or kilometers. Suppose that Kepler came back to earth today and that we carefully concealed from him all the accumulated astronomical knowledge of the past 327 years. But suppose that we showed him how to work our modern telescopes and clocks and then suggested that he go ahead and, with this greatly improved instrumentation, repeat the investigations that made him famous. Kepler would no doubt begin by measuring, each day, the angular diameter of the sun. By methods which would be familiar to him, even if they are not to all of us today, he would, by measuring certain other angles at noon each day, determine the sun's position relative to the background of the stars. Note the important point that his observations consist entirely of measuring *angles* and not at all of measuring *distances.*

The changes in the angular diameter of the sun from day to day he would interpret by saying that it was the (unknown) distance of the earth from the sun that was varying; the apparent motion of the sun relative to the distant stars, sometimes faster and sometimes slower during the course of the year, would eventually lead him to the conclusion that the earth was moving round the sun in an ellipse, with the sun in one focus. He would draw on a piece of paper an accurate reproduction of this ellipse, with all the distances relatively correct but with not one of them known in terms of miles or kilometers. Having completed this operation he would take each planet in turn and, again observing angles only, find that each planet moved in an ellipse with the sun in one focus, the reproduction of each ellipse being on the same scale as that used for the earth's orbit.

But in all this there is one point of theory that would seem self-evident to Kepler, though it would not be so to a modern mathematician looking over his work. He would take it for granted that the interrelations between the angles he was measuring and the distances involved were those of Euclidean geometry and that, moreover, the light rays reaching him from the objects he was studying traveled along Euclidean straight lines. To this day astronomers, in dealing with the solar system, make, for most purposes, the some geometrical assumptions that Kepler did. And they express the distances in terms of the astronomical unit of distance, which is nearly, though not quite, the same as the earth's mean distance from the sun.

Turning the astronomical unit into miles or kilometers is a separate problem. Since a scale model of the solar system is at our disposal, all that is required is the measurement of one distance in the system in order to fix the scale. To obtain this, the earth must first be surveyed and the distances between astronomical observatories must be determined. The base lines from which the survey starts are measured, not with the standard meter-bar, but with flexible metal tapes stretched to a given tension, and the belief that a consistent system of distance relations between points on the earth's surface can be established on such a basis depends on a good deal of theory. There must be a theory of the nature of a metal and theories of dynamics and of elasticity—here the classical Newtonian theories are employed—before the surveyors can be sure that using the metal tapes in the way they do is likely to provide an internally self-consistent network of distances. It is also taken for granted that the geometry which interrelates the measurements is the Euclidean, the surface of the earth being regarded as a mathematical surface in a three-dimensional Euclidean space.

Suppose then, that the Euclidean straight-line distance between two widely separated observatories A and B has been thus established, partly by measurements and partly by geometrical calculation. A planet P having been selected, the angles between the lines AP and AB and between BP and AB are observed simultaneously at A and B. Again by means of Euclidean geometry, the distances from A and B to the planet can be calculated and the scale of the solar system determined. In all this procedure, the important elements of principle are: firstly, that all measurements are angular measurements, except for the distances between the ends of the base lines on earth that are established by use of flexible metal tapes; secondly, that calculations are performed through the formulae of Euclidean geometry; thirdly, that the paths of light rays are straight lines in the Euclidean sense; and fourthly, that it must be possible to attach a meaning to the statement that two widely separated events are simultaneous, the two events in question being the observations of the planet from the observatories A and B.

Time and Distance

It is not only by way of this question of simultaneity that time has reared its ugly head; our resurrected Kepler had already tacitly incorporated the notion of time intervals into his scale model of the solar system through his procedure of making observations on successive *days* and in successive *years*. He must have known what a day was and what a year was before he could act in the way he did. And, indeed, the astronomers of today introduce the notion of time into distance computations in the solar system in a much more indirect and yet fundamental way. They have at their disposal all the dynamical theory and the theory of gravitation which were first invented by Isaac Newton in the second half of the 17th century and whose detailed consequences have engaged the attention of mathematical astronomers ever since. As G. M. Clemence (*1*) has pointed out, the unit of time employed in these theories is the average value of the mean solar day during the 18th and 19th centuries, and the astronomical unit of distance is defined as the theoretically predicted radius of the circular orbit around the sun of a planet of very small mass moving at an angular rate, expressed by a fraction given to 11 places of decimals. Now the Newtonian theory of planetary motion contains two very important principles which underlie the three laws of motion. These are, firstly, the postulate that all angular and distance measurements in the universe can be manipulated by the rules of Euclidean geometry and, secondly, that there exists an absolute time. The absolute time is pragmatically defined with reference to the mean solar day during the two centuries referred to, and it is one of the merits of the Newtonian theory that distance measurements and time measurements can be manipulated separately—the first, as I have said, by the rules of Euclidean geometry; the second, on the principle that time is a scalar and that time intervals are not affected by the geometry that is being employed for distances.

I would like to put forward here the view that astronomers are not engaged in discovering eternal truths about the dimensions of the solar system relative to the selected base lines on earth, or even relative to an unknown mean distance of the earth from the sun. They are not even achieving approximations to these truths within some limit of experimental error. In my opinion, the combination of measurements, geometrical argument, and dynamical theory serves to establish an internally self-consistent scheme of distance relationships with which time relations are inextricably intermingled. The magnitudes of the errors do not measure the departure of the scheme from some ideal eternal truth laid up in heaven; rather do they establish a measure of the degree of self-consistency. The adequacy of the scheme can be tested by using it to make predictions of future events—for example, to predict in what part of the sky the planet Mars will be at some date in the future. The prediction can eventually be checked by observation, and this observation in turn will be incorporated into the scheme and will modify it. That the modification may be very slight is irrelevant; its acceptance by astronomers as a natural corollary of their work is the interesting point.

Special Relativity

You will be wondering by this time why I included in my title the subject of relativity, since I have so far made no reference to this theory. The reason is that, in my opinion, the theories of special and general relativity cannot be discussed *in vacuo* but must be envisaged in some context of observational data. You will remember that special relativity owed a good part of its origin to the Michelson-Morley experiment, which was intended to measure the velocity of the earth in its orbit by a purely optical experiment performed in a laboratory from which no celestial object could be observed. The expectation of success depended on calculations based on the principles of the Newtonian theory of motion applied both to the motion of the earth and to the motion of light. The experiment revealed that the expectation was unjustified—that the orbital speed of the earth could not be measured in this way and that this was not due to errors of observation but was a genuine null result.

The special relativity interpretation of this conclusion idealized the situation to the extent that the effect of gravitation was omitted and the earth was regarded as moving in a Euclidean straight line, with constant speed, during the very short time interval occupied by the experiment. It was then shown, firstly, that the hypothesis of the validity of Euclidean geometry could be retained but that, secondly, the assumption of a single absolute time would have to be rejected. On the basis of these two postulates, a new theory of motion could be worked out from which the null result of the Michelson-Morley experiment could be predicted. More precisely, if there are two frames of reference for distance measurements, distances in each being established by identical procedures, and if the frames of reference are in constant relative motion as judged from either frame, then to each frame there must be associated a system of timekeeping of such a kind that the velocity of light has the same value relative to either frame.

The important points that are brought out by special relativity are, firstly, that there is no absolute theory of motion and of the associated dynamics; secondly, that time and distance measurements can be much more closely interwoven than they are in the Newtonian theory of motion; and thirdly, that as a result of this closer interweaving, two events can be simultaneous in one frame of reference but not in another and, concomitantly, that the distance between two events is not an absolute but depends on the frame of reference relative to which it is measured. The device that is employed in relativity theory for interlocking closely time and distance measurements is to use what may be called a four-dimensional geometrical representation for all the events under contemplation. Each event is "plotted," so to speak, as a "point" having four, instead of the usual three, coordinates; three of these fix the position in space of the event relative to some frame of reference, the fourth is a coordinate specifying the instant of occurrence of the event in the same frame of reference.

My own personal view of this procedure is that a geometrical manifold used for such a representation is a mathematical device that is very valuable for purposes of calculation, but I am not convinced that it is anything more than this. A convenient name for the four-dimensional manifold is "space-time," but as for its "reality" or its "physical existence," I would not like to commit myself because I do not know what these terms mean. What I do know, as a theoretician, is that I can make calculations and predict the circumstances of the motions of bodies by using the mathematical technique of the space-time representation.

Let us return to the astronomers in the observatories *A* and *B* who are engaged in determining the distance of the planet *P*. You will remember that an

essential part of the procedure was that their observations had, in principle, to be made at the same instant. But this simultaneity is now dependent on the selection of one of the reference frames of special relativity; there is no absolute definition available. Nor is there any absolute definition of the distance between the two observatories. We might think that we could select the frame of reference relative to which A and B were at rest, which I will call the zero frame, and assume that the time associated with the zero frame was the time kept by the observatory clocks. But if we imagine a second frame of reference moving with uniform speed u with respect to the zero frame, then, in terms of the time appropriate to this new frame, the astronomers at A and at B would not be carrying out their observations simultaneously and the distance from A to B would have a value that differed from that in the zero frame. This difference would depend on the ratio of the square of u to the square of the velocity of light, c.

Now it might be thought that we could eliminate these difficulties by simply asserting that the zero frame had priority over all the other possible ones, but this cutting of the Gordian knot will not do, for there is another principle inherent in relativity theory to which I have not yet referred. It is this: Reference frames, even of the combined space and time variety used in relativity, are elements introduced into the description of the physical situation by the investigator and are therefore alterable at will. Nothing essential must therefore be made to depend on the choice of one reference frame rather than another; all reference frames must be on an equal footing, and we should fix attention on those conclusions that can be demonstrated to be equally valid in every reference frame. It happens that the numerical value of a distance between two events is not an invariant quantity in special relativity but depends on the selected reference frame, and the same applies to the time interval between any pair of events. This is not to say that measures of distances and of time intervals are valueless but that the conditions of measurement must be carefully stated and the relativity of the results must be firmly kept in mind.

If you have followed me so far, you will no doubt have noticed that there is an apparent contradiction between what I have just said about special relativity and what I said earlier about the procedure actually employed by astronomers in establishing distances in the solar system. I have emphasized that, in doing so, these scientists take it for granted that distance is an absolute in the Euclidean sense and that time is also an absolute in the sense of Newtonian dynamical theory.

Since astronomers are acquainted with special relativity, why do they continue to act as if they were not? The answer is that the differences inherent in the use of one frame of reference rather than another depend on the ratio of the square of the relative velocity of the frames to the square of the velocity of light. Now the velocity of light is 186,000 miles per second, and the relative velocities of objects in the solar system rarely attain, say, 70 miles per second. Thus, the ratio of the squares of these velocities will be found to introduce changes in the distances that amount at most to one part in ten million. But the astronomical unit of distance is determined to within the very much larger error of one part in five thousand. Thus, if this is the kind of accuracy we are working to, or, as I would prefer to put it, if this is the measure of the internal self-consistency we aim at, special relativity can be disregarded in direct calculations of distances. But it is not self-evident that the effects of relative velocity on our distance and time computations may not give rise to observable differences from the predictions of Newtonian theory when these effects can be shown to be cumulative in time. To consider this type of question we must pass on to general relativity.

General Relativity and the Solar System

Expositions of general relativity very often give the impression that the only difference between this theory and special relativity lies in the character of the relative motion of the "time and space" frames of reference employed. Whereas in special relativity the frames are in uniform relative motion, in general relativity mutually accelerated frames are considered. This is certainly one aspect of general relativity, but, by itself, the introduction of mutually accelerated frames is not sufficient. The essence of the difference between the two theories is more subtle; it lies in a change of character of the four-dimensional geometrical representation of the events under contemplation. Whereas in special relativity this representation has the geometrical characteristic known technically as "flatness," in general relativity the representations are "curved."

It is not easy to explain this essential difference to nonmathematicians, especially as the terms *flat* and *curved* have so many connotations that their very use in this technical connection is misleading. One can but try to use an analogy drawn from geometrical manifolds of two, instead of four, dimensions. The Euclidean plane is one such two-dimensional manifold, the surface of a sphere is another; the former possesses the geometrical attribute of flatness, the latter, that of curvature.

This difference manifests itself in the geometries appropriate to the two surfaces. Thus, in the Euclidean plane, straight lines are of infinite length, the three angles of a triangle have a sum equal to π, the sum of the squares of the two sides of a right-angled triangle is equal to the square of the hypotenuse, and so on. On the surface of a sphere, great circles are the analogs of straight lines, and they are of finite length; the angles of a spherical triangle do not have a sum equal to π; and Pythagoras' theorem in the form I have stated it is no longer true.

The four-dimensional space-times of general relativity differ from one another and from that employed in special relativity in analogous but far more complicated ways. This is not only because four, as against two, dimensions give rise to greater intricacy but also because one of the dimensions is interpreted physically as referring to the time relations between events. The time and distance measurements made by astronomers are thus interwoven in a more elaborate way than they are in special relativity, and they lead to a new theory of motion and of dynamics. In general relativity it is postulated that each physical situation has its own appropriate space-time, the connection between the two being made through Einstein's gravitational equations. Moreover, the only effects that are taken into account are those of relative velocity (as in special relativity), gravitational attractions, and a peculiar force which is either one of repulsion or of attraction, according as the so-called cosmical constant is a positive or negative number. The physical situation, consisting of a very massive sun around which move planets of relatively infinitesimal mass, has one space-time representation; the whole astronomical universe of galaxies has another, with quite different properties of

curvature. Thus, once again, I would regard these space-times as devices for calculation purposes, rather than as entities having any kind of "physical existence."

But to return to our astronomers studying the solar system. The space-time representation appropriate to this physical situation was worked out almost immediately after Einstein had propounded, in 1917, the general theory of relativity. The geometry of space, by which the relations between the angles and distances that the astronomers measure are calculated, can be shown to be so nearly Euclidean that this classical geometry can be safely employed for most purposes. Similarly, there is a kind of time appropriate to the physical situation which possesses nearly, but not quite, the properties of absolute Newtonian time.

But do not let us be hasty and jump to the conclusion that, if the departures from the Newtonian theories of dynamics and gravitation are so small, we can dispense with general relativity. For these departures can have cumulative effects when we are concerned with predicting how a body in the solar system will move. Observations of the planet Mercury, which has an orbit lying fairly close to the sun, where the gravitational field is strongest, and which also moves with a relatively high orbital speed, had revealed, long before the advent of general relativity, an unexplained motion in space of the planet's point of closest approach to the sun. The amount of this motion was some 43 seconds of arc per terrestrial century, and it was found that general relativity could interpret it in terms of the mass of the sun and the dimensions of the orbit of Mercury.

You might be interested to know what the corresponding figure for an artificial satellite of the earth would be. Let us suppose that the earth is perfectly spherical—which it is not—and that the maximum and minimum distances of the satellite from the surface of the earth are 200 and 600 miles, respectively; and let us ignore the frictional resistance of the earth's atmosphere. Then the point of closest approach of the satellite's orbit would move in space through 1324 seconds of arc in a century. The figure is large compared with that for Mercury in spite of the fact that the earth's gravitational field is incomparably weaker than the sun's and that the satellite's average speed in its orbit is only about one-sixth of Mercury's. The cumulative gain lies in the number of revolutions: in 100 years Mercury goes round the sun about 414 times, whereas in the same period the satellite would circumnavigate the earth nearly 540,000 times.

But to return to the sun. Another striking effect of the non-Euclidean character of the space-time appropriate to the solar system is the phenomenon of the bending of light rays. When the stars beyond the sun are photographed during a total solar eclipse and when the picture is compared with one taken at night, it is observed that the stars appear to be relatively further apart in the eclipse picture than in the other. An effect of this kind is predicted by general relativity, because the paths of light rays in the space-time can be regarded as Euclidean straight lines only in regions remote from the sun, where its gravitational effect is weak. Near the sun itself the geometry departs sufficiently from the Euclidean to give an observable difference. The theoretical prediction is that the distortion of position for a star seen at grazing incidence beyond the sun's disc would be 1.75 seconds of arc. As far as I am able to judge from the results of observations made during six total eclipses since 1919, this figure corresponds to that found from observation.

Conclusions

Let me now try to summarize, in conclusion. Distance determinations in the solar system have always, in the main, consisted of the measurements of angles and of time intervals. Determinations of the positions occupied by a celestial body at successive instants of time have necessarily involved some theory of dynamics and of gravitation. The theories of relativity have shown that time measurements can be interconnected with the distances deduced from the angular measurements in a more intimate way than that postulated by Newton. This has had as a consequence the development of dynamical and gravitational theories different from his. In addition, each physical situation has its own system of time and distance interconnections, obtained through the mathematical device of four-dimensional geometrical representations in which time is regarded as a fourth coordinate interwoven with the three spatial coordinates. General relativity has achieved an interpretation of phenomena in the motions of the planets and of light which had escaped the Newtonian net. I wish I had the space to explain, and that you had the endurance required to absorb, the further applications of these ideas to the universe at large, where distances are measured in millions of light-years instead of in millions of miles.

Reference

1. G. M. Clemence, *Science* 123, 567 (1956).