The Homogeneity of the Universe

G.F.R.Ellis
Department of Physics, University of Alberta,
Edmonton. Alberta, Canada*.

- An Essay submitted to the Gravity Research Foundation -
March , 1979

Abstract: The observational and philosophical grounds for our belief in the spatial homogeneity of the Universe, are reconsidered. On the one hand, spatial homogeneity cannot be directly verified observationally, and on the other hand, it raises a series of philosophical problems which counter its attractiveness to a considerable degree. It is pointed out that there are several possible alternative approaches to the question of the homogeneity of the universe which may provide at least as attractive a philosophical basis as the current approach, and which merit detailed investigation as possible alternative bases for our cosmological world-view.

* Permanent address: Department of Applied Mathematics, University of Cape Town,
Rondebosch, Cape Town 7700, South Africa.

The homogeneity of the Universe

G.F.R.Ellis
Department of Physics, University of Alberta,
Edmonton, Alberta, Canada.

Introduction

One of the major features of the presently accepted "standard" universe models[1] is their spatial homogeneity. The spatial homogeneity of these models is a direct result of the"Copernican" or "Cosmological" principle, which is introduced right at the beginning and fundamentally shapes the form taken by the theory. Because of the central role played by this assumption, it is essential that it be tested and verified in all ways possible; for should it be incorrect, our understanding of the nature of the Universe could be radically wrong. We shall first consider to what extent this assumption can be observationally tested; and then reconsider its standing as an *a priori* philosophical principle. This leads to consideration of some possible alternative approaches to the question of the homogeneity of the Universe.

The observational status of homogeneity

The observational situation is completely different within, and outside, our past light cone. In the standard models (with vanishing cosmological constant), there exist *particle horizons* , that is, there are regions of the universe outside our past light cone from which we can have received no information at the present time (see e.g. [1]). There may also be such regions in more general universe models. While spatial inhomogeneity in such regions *near* our past light cone might conceivably be detected by "Coulomb" effects on our past light cone (cf.[2]), there is no way we can observationally verify that such regions are spatially homogeneous far out from our past light cone ([3]). When we assume that such regions are spatially homogeneous - as we do in the standard universe models - we can have no observational check whatever of this assumption.

Even within our past light cone, observational verification of homogeneity is difficult. While *isotropy* of our observations can be checked directly, a check of *homogeneity* involves detailed determination of number densities and area distances as a function of redshift, or equivalent observations ([4]); while one can in principle carry out an observational programme that would verify spatial homogeneity within our

[1] The Friedmann-Le Maitre-Robertson-Walker, or *FRW*, universe models; see e.g. [1]. We do not assume here the need for 'exotic' physical processes to explain cosmological observations; so in particular we accept that the space-time geometry is governed by General Relativity.

past light cone, even for relatively small values of the redshift substantial problems arise because of evolution effects, observational selection effects, the "lumpy" nature of the real universe, and so on (cf [4]). At substantial distances, source evolution must be taken into account (for example, the number counts of radio sources contradict the homogeneity assumption unless there is very considerable source evolution); but such evolution makes it difficult to determine the aperture and selection effects which govern actual observations (one has to know the source morphological, brightness, spectral and radial evolution in detail). These problems become so severe that in fact we cannot observationally prove homogeneity at large redshifts; basically, one cannot distinguish a time variation of source properties in a spatially homogeneous universe, from a change in source properties with spatial distance from the observer in a spatially inhomogeneous universe ([4],[5]). Furthermore, on the scales where we can make reasonably unambiguous observations, there are some indications that the Universe does not settle down to a homogeneous state even at the largest scales (see e.g. [6]).

In brief, on the one hand, we can test homogeneity nearby, and find the universe to be rather lumpy; on the other hand, we are unable to observationally prove homogeneity of the universe at large distances from us, or to disprove it; rather the best we can do is to use the observed isotropy of the background radiation and the source counts to deduce space-time isotropy within our past light cone[2], and then adduce an unverified homogeneity assumption ("we are not at the centre of the universe") to deduce spatial homogeneity (cf [1]). The deduction of homogeneity rests on this assumption, rather than on the observations[3].

Philosophical problems arising from homogeneity

The basic problem arising in a spatially homogeneous universe is the question of why the universe should be spatially homogeneous. For this implies that the initial conditions in the universe were very special: all the more so in that while they must not be too strictly homogeneous, or galaxies will not form as required, inhomogeneities must occur with very strictly limited density increments, or the perturbations that are supposed to lead to galaxies at late times will recollapse to form black holes before stars have had time to evolve. Thus an extremely delicate balance must be set up between perturbation densities and expansion rates at very early times, and must be set up the same way at all places in the universe - despite the fact that there can have been

[2] even this step involves unverified assumptions, because we do not have access to all the data needed to verify space-time isotropy: see [4].

[3] See [3 ,4,5]   for further discussion of this point.

no causal communication whatever between most of the places involved, because of the particle horizons ([7],[8]). If one pursues the homogeneity idea back to extremely early times, there may even be substantial self-consistency problems arising from the particle nature of matter and the Coulomb law for charged particles ([9]), which hint at further self-consistency problems for the Gauss law and massive gravitating particles.

Such problems lead to the intriguing "chaotic cosmology" programme initiated by Misner ([10]), in which it was proposed that the early universe was not homogeneous and isotropic, but rather that since dissipative physical processes can act to decrease anisotropy, they could have led to homogeneity as different parts of the universe interacted with each other; such physical processes being potentially able to significantly affect large scales because of the possibility of particle horizons being broken in "mixmaster'-like universes[4]. However this programme does not succeed in the sense that while large amounts of anisotropy can be dissipated by physical processes, the horizon-breaking mechanism is ineffective in many cases ([11]), and in addition very large dissipation of anisotropy and inhomogeneity would lead to generation of much more entropy per baryon than is observed ([12]). Accordingly[4], the situation remains that a homogeneous universe compatible with the currently observed microwave background radiation isotropy and temperature can only develop from a very restricted set of initial conditions at early times. The problem of explaining why such special initial conditions should occur, remains.

The situation is greatly aggravated in homogeneous cosmologies such as the low-density FRW universes. For such universes (with their natural topology) are spatially infinite. Accordingly, any process which occurs with non-zero probability in such a universe, will occur infinitely often in it; this makes it difficult to provide a strong argument why any particular feature of our own world is not replicated infinitely often on other worlds in the universe ([13]). The only ways to avoid this phenomenon in an infinite universe appear to involve breaking the homogeneity assumption. For example, one might try to avoid it by, say, allowing that for distances greater than about 1000 Hubble radii away from our past light cone, one would no longer insist on strict homogeneity- for no conceivable observational test could check conditions there. But once the homogeneity principle is weakened in this way, the essential point of that principle has been lost. If one is willing to weaken it at such large distances, why not rather nearer outside our light cone, say 5 Hubble radii? or 1 ? - or why insist on it even inside our past light cone? Any weakening at all of the homogeneity principle implies a preferred position for our world - which is what the principle was designed to avoid.

4: see e.g. [11] for a review of the programme and its sucess.

The basis of the homogeneity assumption

The homogeneity assumption is an extrapolation, on a grand scale, of the supposition that we do not live in a preferred position in the universe. While the historical roots of this assumption are well-known and based in valid experience, they do not reasonably justify extravagant extrapolations of the idea. In fact, it is manifestly invalid even in the solar system: our life could only have evolved on a planet situated within a very narrow range of distances from our sun, and accordingly does occur in a very special situation within the solar system. A fortiori, life like ours can only have evolved at very special places in the galaxy. Again, in the FRW universes, we live at a special time ([14]).The laudable desire to express the idea that the creation of the universe was not centred on our presence, must not be confused with the incorrect statement that we could equally well live at all places and all times in the universe.

Alternative approaches to spatial homogeneity

The standard assumption may be summarised as follows:

(A) On some scales we live in preferred positions, and on other scales we do not. We assume the second feature is true on the very largest scales (although this is not directly verifiable). This homogeneity is caused by the existence of very special initial conditions at the beginning of the universe, which we cannot explain further.

A second possibility is ([5]),

(B) The universe is spherically symmetric but inhomogeneous, because of special initial conditions which we cannot explain further. We observe it from near the centre because this is the most likely place for life to evolve.

While one can make quite a reasonable picture of such a universe[5], it would also be based on special initial conditions. However if one were to attempt to assign a priori probabilities to each possible situation from the viewpoint of the generality of the initial conditions needed to produce that situation (see e.g. [15]),case (B) would be regarded as more probable than case (A), because case (A) is of zero measure in the set of spherically symmetric models while case (B) is not. However even (B) is clearly of zero measure in the set of all initial conditions. If one cannot delimit in some way the creation mechanisms and the extent to which they (together with subsequent physcial processes) necessarily produce homogeneity and isotropy in the universe models created, suggestions (A) and (B) are both very implausible within the set of all possible universes.

There are at least two other approaches to the question of homogeneity of the universe, which are intrinsically rather more probable.

[5] The static version of this model runs into trouble when the field equations are applied and its predictions compared with observations, but expanding versions are probably compatible with present observations ([5]).

(C) The universe is a completely chaotic infinite (or very large) inhomogeneous universe. Local expansion, homogeneity and isotropy of the universe are to be explained by the Anthropic principle: life would only be likely to evolve at points where these were observed.

In this picture. vast regions of the universe would be expanding, vast regions contracting, some at very high and some at very low densities, some rotating at enormous speeds, and so on. Singularities of various kinds might occur in many places. Observers could only view this universe from very special places and times, because life can only evolve to great complexity under particularly favourable circumstances[6]. It is clear (e.g.[8]) that life would develop in regions of *expansion* (required for suitable galaxy formation) where the radiation background is small enough. The major theoretical problem in this context would be to provide theoretical arguments based on the anthropic principle ([14]; see [16] for a recent discussion)to show that in addition, life would be most likely to evolve to great complexity at points where the universe appears to be *as isotropic* as we see it.

Providing good arguments for such isotropy is a major challenge, largely because of the *depth* from which isotropic radiation reaches us (at least up to a redshift of 10). Nevertheless this viewpoint is worth pursuing further.

On the one hand, suppose one was only able to assign a relatively low probability that a family of observers in such a universe could only evolve in regions from which it appears isotropic. Then the appropriate measure to use in assessing this probability, is the probability of a FRW universe occurring. Essentially,

$$\left(\begin{array}{l}\text{probability of family of}\\ \text{observers seeing an}\\ \text{isotropic universe}\\ \text{about them}\end{array}\right) = \left(\begin{array}{l}\text{probability of the}\\ \text{particular universe}\\ \text{existing, within the}\\ \text{family of all universes}\end{array}\right) \times \left(\begin{array}{l}\text{probability of observers}\\ \text{evolving at a point in that}\\ \text{universe from which it}\\ \text{appears to be isotropic}\end{array}\right)$$

In the FRW approach (A), the second probability is 1 but the first probability is very low. In the "totally chaotic" case (C), the first probability is high,so even if the second probability is low, the *overall* probability of that situation occurring may be at least as high as in the FRW picture.

On the other hand, the chaotic view suggested here has several features strongly in its favour. Firstly, it provides the natural setting for the Anthropic idea ([14],[16]), as considerations of the probability of life occurring at different places within one universe model seems to be philosophically preferable to consideration of such probabilities within ensembles of universes, as in the Everett-Wheeler interpretation

6: We note here that many other kinds of life forms might conceivably exist; nevertheless, complex life would require very particular environments for its development.

(cf [14-16]).Secondly, this view also provides a natural setting in which to examine the "Phoenix" idea ([8]) of collapsing universes changing to re-expanding universes; for even if such a universe were to start our very homogeneously, after one or more collapses and re-expansions it is likely that large anisotropies and inhomogeneities would have developed and that therefore the "chaotic" view would prevail in any case at later stages in such a universe[7]. And thirdly, this outlook is much more in accord with plausible biological (and perhaps even more general philosophic[8]) worldviews than the FRW picture; for the situation in the Earth's biosphere is that there exist many different environments into which seeds of many types are broadcast by many mechanisms, and only in particular places where the environment is suitable does each particular seed survive and its progeny flourish. Approach (C) generalizes this biological world-view - basically, the outcome of the Darwinian revolution - to the cosmological scale (instead of adopting the group-theoretical approach which underlies (A)).

A completely different possibility is,

(D) The universe is a small, finite expanding universe. Apparent homogeneity and isotropy are generated by this finiteness (which allows light to travel round the universe many times).

The point here is that there could exist universes with small numbers of galaxies and finite spatial sections, which could possibly represent our present observations. Consider for example, a FRW universe with k=0 or k=-1, in which identifications have been made (see [19]) so that the space-sections {t=constant} are compact (finite) space sections of rather small volume. Then we could already, due to this compactness, have seen several times round the universe in each direction, and would in fact see each cluster of galaxies in many different directions in the sky. This situation would *not* necessarily be easy to detect, because we would see the same cluster of galaxies at different redshifts (and so at different times in its history) and at different area-distances (so with different selection effects operating, see [4]) as we viewed it in different directions in the sky; and we would effectively see it *from* different directions ([19]).

In general such universe models would have compact space-sections but would be inhomogeneous. A fully satisfactory discussion would have to give some account of the

[7] A relation to Penrose' idea [17] of a minimum entropy initial condition might be established if this could be shown to be necessary for families of observers to develop (or at least to substantially increase the probability of their existence).

[8] See e.g. Olaf Stapledon's book *The Starmaker* . Note that the situation envisaged here is chaotic on a much broader scale than that considered in many other discussions of "chaotic cosmology", see e.g. [18]; the closest to this approach seems to be that in [7].

length-scales involved in these compact space-times (cf[13]), and discussion of the a priori probability of such universes occurring would be difficult[9]. However the attractions of this possibility are at least three-fold.

Firstly, such universes would satisfy the "Machian" reasons for a closed universe originally put forward by Einstein [20] (and later particularly emphasized by Wheeler [21]). Secondly, it would result in a universe where an observer could in principle determine the universe's structure (because there would be no particle horizons in this situation, and "far away" galaxies would also be seen "nearby"), in contrast to all other universes where increasing uncertainty with distance is a basic feature of any observational cosmology programme, and observational ignorance of the situation beyond the horizon is essentially complete ([4],[3]). Thirdly,it would open the way to a quite attractive explanation of the apparent homogeneity and isotropy of the universe. The idea would be to see how few clusters of galaxies could represent our current observations within the context of such a universe[10], in which we have already seen all the galaxies there are. (Could we get by with just one supercluster of galaxies? If not what is the smallest number needed?) Then the universe could be thought of as constructed by repeated duplication of a smallest building block containing these clusters of galaxies. If enough repetitions were included within a"Hubble radius", the universe would necessarily appear homogeneous, because one would indeed be viewing the same galaxies over and over! Further, in general a statistical isotropy would occur in such a universe at large redshifts, for unless local preferred axes lined up along the lines of sight - which would not happen in general [22] - the universe would in the large appear isotropic if viewed from an average viewpoint; in fact in general there would not be observational preferred axes in the space-time. Thus natural reasons for observation of approximate homogeneity and isotropy could be forthcoming in such a universe, even if the basic "building block' (described by one compact spatial section, at any particular time) was quite inhomogeneous.

[9] Surprisingly, FRW universes with such small compact sections might be more likely to occur than those with the 'natural' topology - because there are so many more of them! The 'natural topology' FRW universe is of zero measure in the space of all locally FRW universes, in the case k=-1.

[10] Which one could represent either as a very small compact universe, or as a large (presumably infinite) universe invariant under discrete translations (instead of the continuous translations of the FRW universes).

Conclusion

The real reason for the choice of the FRW picture is philosophical rather than observational; the submission here is that other alternatives, some of which are suggested here, might be as attractive philosophically as the FRW picture; and might provide satisfactory universe models in which there are completely different explanations of the apparent isotropy of the universe.[11]. The FRW model may in fact have triumphed through default of suitable opponents. The contention is that the relative merits of such completely different alternative models (where the homogeneity assumption is not invoked) should be seriously weighed against those of the FRW model; their observational consequences should be determined in detail, and the homogeneity assumption properly assessed relative to some of the alternative possibilities.

[11]: If case (C) were true, the FRW models would not be a satisfactory first-order approximation, except in a (small?) region inside our past light-cone; if case (D) were true, the FRW models could only give a satisfactory approximation if one made suitable identifications to give them small compact space sections. It is not clear whether all universe models of type (D) could be approximated in this way, or not.

References

[1]: Weinberg, S. (1972). Gravitation and Cosmology (Wiley, New York).
Hawking, S.W. and Ellis, G.F.R. (1973). The Large Scale Structure of Space-Time. (Cambridge University Press, Cambridge).

[2]: Ellis, G.F.R. and Sciama, D.W. (1972). "Global and non-global problems in cosmology", in Studies in Relativity, ed. L.ORaiffeartaigh (Oxford University Press, London).

[3]: Ellis, G.F.R. (1975). "Cosmology and verifiability"; Qu. Journ. Roy.Ast.Soc. 16:245-264.

[4]: Ellis, G.F.R. (1979)."Limits to verification in cosmology"; To appear, Ann. New. York. Acad. Sci.

[5]: Ellis, G.F.R. (1978). "Is the universe expanding?" ; Gen.Rel.Grav.9:87-94.
Ellis, G.F.R., Maartens, R. and Nel.,S. (1978). "The expansion of the universe"; Mon. Not. Roy. Ast. Soc. 154:187-195.

[6]: de Vaucouleurs, G. (1970)."The case for a hierarchical cosmology"; Science 167: 1203-1213.

[7]: Peebles, P.J.E. (1972). "Light out of darkness vs order out of chaos";Comments Astrophys.Sp.Sci 4:53-58.

[8]: Dicke, R.H., and P.J.E.Peebles (1979). "The big bang cosmology - enigmas and nostrums"; to appear in Einstein Centenary volume, Ed. S.W.Hawking and W.Israel (Cambridge University Press, Cambridge).

[9]: Penrose, R. (1964). "Conformal treatment of infinity": in Relativity, Groups and Topology, ed, de Witt, B.and de Witt, C.(Gordon and Breach, New York).

[10]: Misner, C.W.(1968). "The isotropy of the universe"; Astrophys. Journ. 151:431-457.
Misner, C.W.(1969). "Mixmaster universe"; Phys. Rev. Lett.22:1071-1074.

[11]: MacCallum, M.A.H.(1979)"Anisotropic and inhomogeneous relativisti cosmologies". To appear in Einstein Centenary Volume, Ed. S.W.Hawking and W.Israel (Cambridge University Press, Cambridge).

[12]: Barrow, J.D. and Matzner, R.A. (1977)."The homogeneity and isotropy of the universe"; Mon. Not. Roy. Ast.Soc. 181:719-727.

[13]: Ellis, G.F.R. and Brundrit, G.B. (1979). "Life in the infinite universe"; to appear, Qu. Journ. Roy. Ast. Soc.

[14]: Dicke, R.H. (1961). "Dirac's cosmology and Mach's Principle"; Nature 192:440-441.
Carter, B. (1974). "Large number coincidences and the Anthropic Principle in cosmology In Confrontation of cosmological theory with observatoon, Ed. M.S.Longair (Reidel, Holland).

[15]: Collins, C.B. and Hawking, S.W. (1973). "Why is the universe isotropic?"; Astrophys. Journ. 180:317-334.

[16]: Carr, B.J. and Rees, M.J. (1979). "The anthropic principle and the structure of the physical world". Preprint, Institute of Astronomy, Cambridge, England.

[17]: Penrose, R. (1979) "Singularities and time asymmetry"; to appear in Einstein Centenary Volume, ed. S.W.Hawking and W.Israel (Cambridge University Press, Cambridge)

[18]: Jones, B.J.T. and Peebles, P.J.E. (1972)."Chaos in cosmology"; Comments Astrophys. Sp. Sci. 4:121-128.
Silk, J. (1973). "The case for a chaotic cosmology"; Comments Astrophys. Sp. Sci. 5:9-14.
Barrow, J.D. (1977). "A chaotic cosmology"; Nature 267:117-120.

[19]: Ellis, G.F.R. (1971). "Topology and cosmology". Gen. Rel. Grav. 2:7-21.

[20] Einstein, A. (1917). "Cosmological considerations on the general theory of Relativity"; in The Principle of Relativity, H.A.Lorentz, A.Einstein, H.Minkowski, H.Weyl (Dover).
Einstein, A. ((1950). The meaning of relativity(Princeton University Press, Princeton), pp. 107-108.

[21] Wheeler, J.A. (1968). Einstein's vision. (Springer, Berlin).

[22] Ellis, G.F.R. and Sciama, D.W. (1966). "On a class of model universes satisfying the Perfect Cosmological Principle"; in Perspectives in Geometry and Relativity Ed. B. Hoffmann (Indiana University Press, Bloomington).
MacCallum, M.A.H. and Ellis, G.F.R. (1970). "A class of homogeneous cosmological models. II: Observations"; Commun. Math. Phys. 19:31-64.

G.F.R.Ellis studied General Relativity and Cosmology at Cambridge under Dr. D.W. Sciama, and was awarded his Ph.D. degree by the University of Cambridge in 1964. He has taught at Cambridge, Texas, Hamburg, Chicago, Boston Universities; is currently on leave from the University of Cape Town (where he is Professor of Applied Mathematics) and is a Visiting Professor at the University of Alberta. He is co-author with Dr. S.W.Hawking of the book The Large Scale Structure of Space-Time.

*   *   *   *