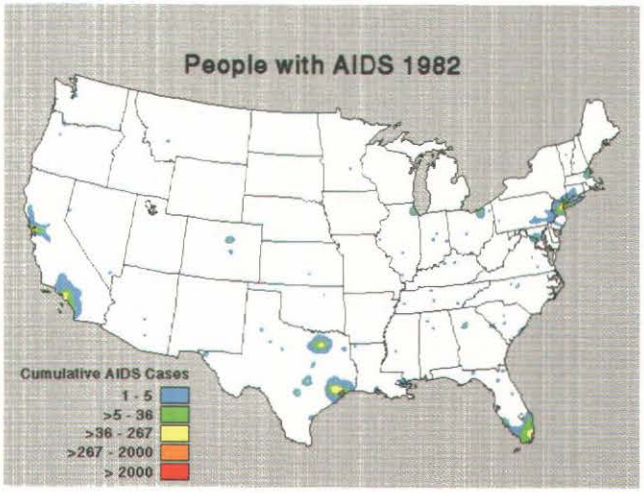
Sources of Error in a Map Series, or Science as a Socially Negotiated Enterprise*

Peter Gould

Peter Gould is the Evan Pugh Professor of Geography at Penn State University, 306 Walker, University Park, PA 16802

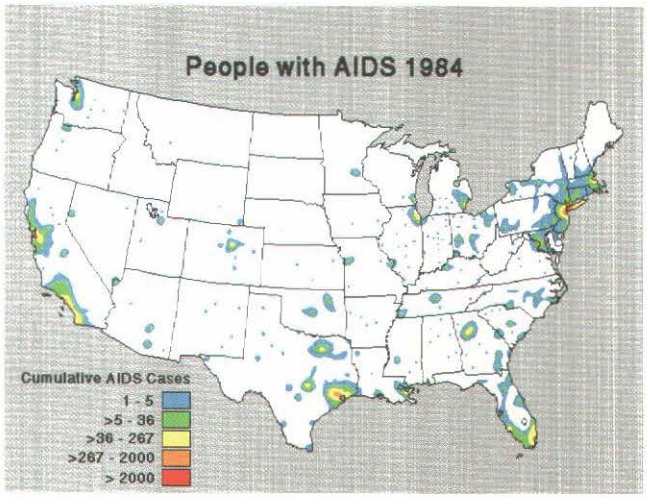
Temporal, definitional, and spatial errors may be present in maps, as well as errors of underreporting and estimation. These are illustrated in a series showing the diffusion of AIDS in the United States, and constitute an example of science as a socially negotiated and hermeneutic enterprise.

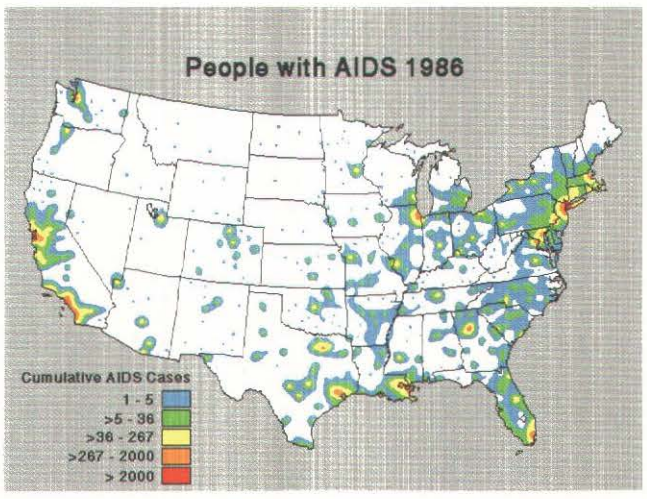


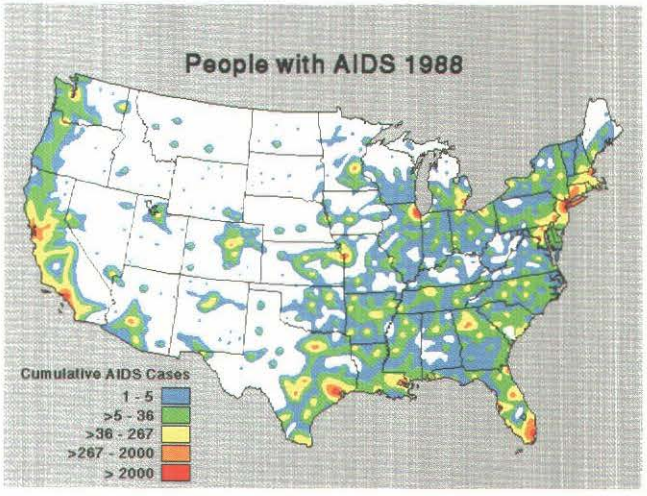
Although terminating in 1990, the five maps illustrating this article still constitute the most detailed graphic expression ever presented of the geographic diffusion of the AIDS epidemic in the continental United States.[1] Based on approximately 2,500 spatially varying units, most of them counties, they have raised questions about the neglect of the spatial perspective by public health authorities during the first decade of the epidemic,[2] and they have shocked an American audience in such public forums as *Time*, *Forbes*, and *Playboy* magazines.[3] They are also available in animated form for educational television directed at young people, who now form the cohort of the population most at risk.[4] It is worth emphasizing immediately that the contour-color interval is geometric, each change up the natural spectrum from blue to red multiplies by 7.5 the previous value, with the red areas simply "over 2000." However, if you made a three-dimensional map in 1995, and used 0.25 millimeters to represent each person dead or dying from AIDS in the five boroughs of New York City, you would have a spike fifteen meters high, and only slightly smaller ones at Los Angeles, San Francisco, Miami, etc.

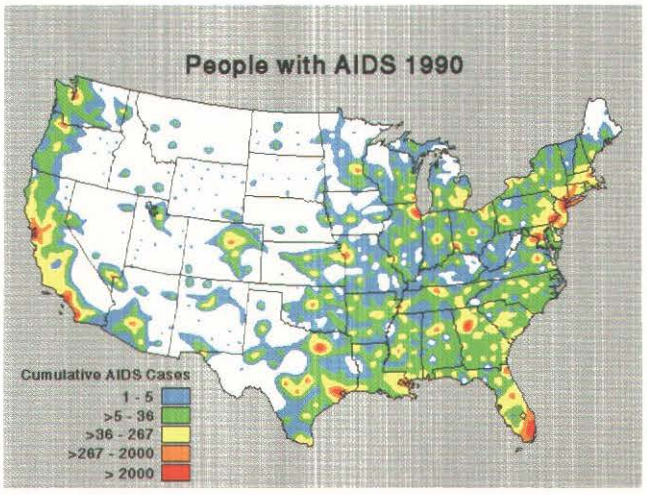
Taken as a whole, the sequence constitutes a powerful rhetorical statement, using the word "rhetorical" in its old and honorable sense as the "art of persuasion."[5] Upon reviewing the "AIDS explosion" in the carto-geographic domain, many people are persuaded for the first time that AIDS is not something "out there," remote and far removed from them, but may well be all around them. To a geographer, the sequence is a classic case of spatial diffusion, with strong evidence of both hierarchical diffusion, controlled by relations of interaction in the urban system or hierarchy,[6] and spatially contagious diffusion from regional epicenters–the "wine stain on the tablecloth" effect.

*Respectfully dedicated to the memory of Brian Harley, University of Wisconsin, Milwaukee. This article is a very slightly revised and updated version of "Sources d'erreur dans une série de cartes, ou: la démarche scientifique, objet de négociations," MappeMonde, 2, 1993, pp. 22-27.









The carto-geographic perspective challenges the totally aspatial view of traditional epidemiological modeling confined exclusively to the time domain.[7] Since the construction of such a series illuminates aspects previously hidden by bureaucratic obtuseness parading in apparently impeccable ethical concern for confidentiality, and since actually seeing the explosive diffusion in the first decade may be politically delicate, especially in an election year in the United States,[8] we may expect attempts to throw doubt on the series, attempts emphasizing that errors are present. It is important to deal with such efforts to denigrate in a direct, firm, scientific and philosophically aware manner.

In any scientific statement, there is error since science is a mortal, rather than a divine, enterprise. However, after examining a large literature, it becomes apparent that little has been added to any theory of error since Karl Friedrich Gauss made maps for the Duke of Hanover during the eighteenth century. Those who boldly pronounced judgment about the amount of error always retreat behind a cloud of assumptions after realizing that one cannot specify any quantity or degree of error without actually knowing the truth. Even such a correspondence theory of truth (and, therefore, of error) has been in disarray since the days of Kant, and no one living in these hermeneutic days could take such an approach seriously. What we can do is make some quite open judgments about the types and sources of error, and then argue that these would not alter any major conclusions about the geographic processes at work and the carto-geographic representations these produce. Whether a reader finds such scientific rhetoric persuasive or not depends on the hermeneutic or interpretative stance he or she is prepared to take. In the end, scientific truth is always socially negotiated, including the construction and interpretation of maps, as Brian Harley was able to teach us before his tragically early death in 1991.[9]

What are the sources and types of error that take this research from the error-free realm of the immortal gods to the foothills of Mount Olympus where ordinary geographers live? There are essentially five, none of which can be cleanly separated except for purposes of exposition. First, there is the problem of underreporting, particularly in the early years of the epidemic. Less was known about the various ways an infected person could convert to symptoms diagnostic of AIDS; tests were less reliable; and some doctors (perhaps up to 50 percent during the early years in Germany) were prepared to sign death certificates for "pneumonia," "cancer," and so on to spare the feelings of shame that some families expressed. Early maps are likely to reflect such errors of omission rather than commission.

Second, there are temporal errors–usually delays in reporting that make it extremely difficult to monitor the course of the epidemic properly, and so use forecasting techniques which rely on recent information to make appropriate parametric adjustments. No matter how sophisticated the methodology and technology used, it is no use monitoring junk.[10] We may even find ourselves in the curious situation that model predictions, far from over-estimating the course of the epidemic, actually turn out in the future to be closer to some unknown truth than the official figures reported by medical bureaucrats. This constitutes a nice philosophical, not to say political, problem in its own right.[11] Even when AIDS is a legally reportable disease, errors of temporal specification may be gross: in June 1991, for example, 75 percent of the AIDS cases reported in Washington, DC were not days, weeks, or even months late but had been diagnosed in previous years.[12]

Errors over time are clearly related to the third type of errors–definitional errors. After the first decade-and-a-half of the pandemic, as more has been learned about the HIV virus and its effects on the human immune system, we are able to recognize, in lowered T4 cell counts and other diagnostic approaches, the earlier signs of conversion to opportunistic diseases. New definitions in 1993 interrupted the previous time series, inflating cumulative totals to the point today (1995) where the 400,000 mark has long been exceeded and the totals are still growing. In the previous year (1992), scientific advances in diagnostic tests were found to be politically unacceptable, and so were socially negotiated away by centralized power structures. For one more year, a nation breathed a sigh of relief that things were not so bad after all. How many more people are infected today we do not know with any reasonable degree of assurance.

The fourth kind of error is spatial error–a form lying in a domain of thinking familiar to the geographer, but an intellectual arena where doctors of medicine and epidemiologists have little if any experience.

The fourth kind of error is spatial error–a form lying in a domain of thinking familiar to the geographer, but an intellectual arena where doctors of medicine and epidemiologists have little if any experience. Unfortunately, such ignorance does not prevent them from making judgments–some of them catastrophic for our deeper understanding of the epidemic and for our ability to intervene with education and health care planning. Spatial error is simply misplacing in space values reported, and it may be thought of as the geographical equivalent of delayed reporting in the temporal domain (i.e., "misplacing" people by a month or a year). Like any other error, it is unavoidable to some degree. Even if we had a dot map of each person,[13] the individual dots would only stand as a spatial mean for a probabilistic smear or "field of movement" created by individual human lives.

Spatial error is particularly likely to arise in connection with the fifth kind of error–errors of estimation. Many of these arise because of the quite proper and understandable ethical concern to protect the identity of people with AIDS. I wish to make it quite clear that I am in total agree-

ment with this ethical ideal, while noting at the same time that it has been carried to quite absurd extremes.[14] Spatial errors of estimation arise when we want to move to finer levels of spatial resolution with data (numbers of people with AIDS) that have been deliberately aggregated spatially, ostensibly to preserve confidentiality. Notice, however, that even the ability to observe scientifically now becomes socially negotiated, with the negotiations directed and informed by the ethos of a society and the power it is willing to allocate to certain groups of professional "experts". In the United States, most states now report regularly by county: some, like Virginia, by zip or postal code; others, such as, Kentucky by aggregations of counties into regions; one (Nebraska) by three, totally irrelevant economic areas; another (South Dakota) by two regions east and west of the Missouri River, although this physical feature is not known to be an effective barrier to the diffusion of HIV; while a few states (Wyoming, Wisconsin) still report only state totals. Generally, there is a tendency to report by smaller and smaller spatial units as the epidemic, measured by rates of infection, intensifies.[15]

Spatial errors of estimation arise when we want to move to finer levels of spatial resolution with data (numbers of people with AIDS) that have been deliberately aggregated spatially, ostensibly to preserve confidentiality.

I want to illustrate the problem of spatial errors of estimation and the way these problems are produced and convoluted by a mélange of other problems by focusing upon three states in the map sequence–namely Texas, Florida and Iowa. Texas is prepared today to report by county, and from 1986 onwards, the map sequence uses the officially reported, updated, and corrected figures. But, before 1985, there was no consistent database, and even today, the state medical authorities only know the cumulative county totals from 1985 onwards. As a result, we have to estimate the 1982 and 1984 values in this part of the country. This actually requires no sophisticated mathematics or computational model: the annual totals for 1985-1990 plot with classical smoothness and regularity, and they can be extrapolated back with a plastic curve to be anchored at 1981, when what we now call AIDS was first recognized (although not the HIV, which was only "discovered" in early 1983). Thus, we can estimate, with what must be only the very slightest error, the cumulative totals for 1982 and 1984.

The question then is: how do we assign these totals spatially? Since we have the map distributions (i.e., the spatial series) for 1985 onwards, we can simply deflate county values, say for 1984, by the ratio of the cumulative totals 1984/1985. This is a simple linear extrapolation backwards, but the difference between the linear and nonlinear approximation will be minute over this time span. Some informed and educated guesswork is involved: a county may appear on the map in the lowest category (blue) with one or two people with AIDS a year before or after it really did, but recall that we have no idea what the reality was in those days and no better way of capturing it. The reader of this text and map must simply judge whether this is plausible, whether it is reasonable, whether it is persuasive. And note: this judgment will be informed by what the reader brings to these written and graphical texts; in other words, it will be related to the hermeneutical stance one is prepared to adopt. I suggest that an experienced geographer will find such spatial estimations acceptable. I further suggest that the ordinary lay person, viewing Texas in the entire sequence, will accept the 1982 and 1984 maps without comment since they reflect what I can only call a "spatial logic" that is to be found everywhere else on the map (California, Washington, New York, New England, etc.). Each map seems to develop quite `logically' out of the previous one, like a photographic plate developing in the darkroom. On the other hand, bureaucratic epidemiologists, trained to think exclusively in the temporal domain, and slowly realizing that they may have been

. . . this judgment will be informed by what the reader brings to these written and graphical texts; in other words, it will be related to the hermeneutical stance one is prepared to adopt.

sitting on scientifically valuable spatial series without knowing what to do with them, may try to exaggerate the minute errors injected by such a procedure. Science becomes, once again, a socially negotiated endeavor.

Florida presents other human, not to say politically charged, problems. Many polite inquiries to the Florida State Health Authorities for cumulative county values produced replies to the effect that they were quite capable of handling the geographical analysis of the epidemic themselves, that only state totals were available to outsiders, and that they needed no help whatsoever–thank you very much! Fortunately, one state health worker, who clearly must remain anonymous, thought this attitude was unreasonable, defensive, and even unethical since, in a state where the epidemic was rapidly becoming catastrophic, it made a major database the private preserve of a few researchers who had no appropriate geographical and methodological ways of using the data. At that time, no geographic modeling had been undertaken, let alone published. We received a xeroxed packet of county values as they had appeared at the end of each year, and we inflated these by a factor based on corrected state totals after the database had been revised by incorporating late reports. In constructing the Florida sequence from these revised figures, we were also fortunate in having perhaps the only geographical analysis of the diffusion of HIV at the time, an analysis which used Florida as one of four case studies.[16] The result is the only published sequence of AIDS diffusion, a sequence with very small, though still unspecifiable, errors.

Iowa presents another problem, one shared by states like Montana and South Dakota. These states are characterized by rural populations of very low density, interspersed by a few urban centers. Only state totals are available for states like these. Even states, such as Wisconsin, which has a higher population density and a reasonably developed central place structure focusing upon Milwaukee-Chicago, exhibit this problem. In cases like these, the cumulative state totals must be assigned in proportion to the county populations. Such an estimation procedure is based on the perfectly reasonable assumption that AIDS is, in large part, density dependent.[17] Detailed analyses of somewhat similar states, like Ohio and Pennsylvania,[18] confirm this procedure as reasonable in the absence of any other information. In actual fact, other information for Montana, Wyoming, and North and South Dakota was made available to me under the standard and strict ethical conditions governing the scientific reviewing procedure. It is clear that in the early stages of the AIDS epidemic–the so-called "seeding" stages–the appearance of AIDS cases in areas of extremely low rural population densities consisted almost entirely of young homosexual men coming home to die from major urban epicenters on the east and west coasts. I cannot and will not use such information to "correct" the earliest maps, so here spatial error must be knowingly left literally in place.

Iowa presents another problem, one shared by states like Montana and South Dakota. These states are characterized by rural populations of very low density, interspersed by a few urban centers. Only state totals are available for states like these.

In the early years, the cumulative numbers for these sparsely settled states are very small, counted in tens or less, and in later years, as the epidemic takes hold, the maps become more and more reliable (i.e., less prone to error). In a year when Montana had ten cases, the national total was already in the tens of thousands. The overall relative error is minute: the local spatial error may be initially quite large but reduces quickly. Notice it is not the totals in a state that are in dispute (except for the other sources of error discussed above) but the exactitude of the spatial allocations. With the exception of Waldo Tobler's "error ellipses" in the very different area of multidimensional scaling and cartography, I do not think we know much about measuring such errors. And, once again, how can you measure error without knowing beforehand what the truth is?

Turning back to the five map sequence, what effects might such errors have on our belief that the sequence is a reasonably accurate representation of the diffusion of the AIDS epidemic at this scale? I emphasize "at this scale" since we recognize degrees of generalization in any cartographic representation.[19] No one would attempt to use these maps for analytical purposes appropriate to much finer levels of resolution. The thickness of a contour line may well exceed the size of some of the townships reporting in a state like Virginia. Some degree of generalization is inevitable in any scientific statement. Indeed, and perhaps almost by definition, a scientific statement in words, graphics or algebras is a generalization where we can see the forest rather than the individual trees. Notice that in regions where the county database is reasonably fine, and the official values reported considered reasonably reliable (over much of the eastern part of the country, for example), the unfolding sequence generates a high degree of trust and therefore belief. The "spatial logic" appears reasonable and truthful mainly because the information content in such "spatial logic" arises precisely out of the plausible spatial autocorrelative properties in such relatively "local" areas. But why should we believe that Iowa, Texas and Florida, and other states where estimations have been made, are any different? Yet, notice further how words like "trust," "belief," "reason," and "truth" have entered the discussion. Whether you trust the map sequence, whether you believe it to be reasonable and close to some unknown truth, depends upon you and what you bring to the hermeneutic task, a task that faces us as human beings as a condition of possibility. Thus, and as thoughtful scientists, it is necessary to negotiate in a communicative discourse.

Some degree of generalization is inevitable in any scientific statement. Indeed, and perhaps almost by definition, a scientific statement in words, graphics or algebras is a generalization where we can see the forest rather than the individual trees.

NOTES

1. They appear in black and white in P. Gould, *The Slow Plague: A Geography of the AIDS pandemic* (Oxford, UK, and Cambridge, Massachusetts: Blackwell, 1993).

2. They were shown at the International AIDS Conference, Amsterdam, July 1992, by Dr. Mindy Fullilove, a doctor of medicine and an AIDS researcher, who pointed out the almost exact correspondence between the geographic development of the epidemic and the assumed sources of original infection specified by several hundred people interviewed in North Carolina. After the cartographic presentation, one epidemiologist in the audience said, "I believe we have a new paradigm here," only 7,000 years after the maps of Babylon were incised on clay tablets.

3. Inexorable march, *Time*, Vol. 40, No. 9, 1992, p. 20; *Forbes*, Vol. 154, No. 6, 1994, p. 250; *Playboy*, Vol. 41, No. 2, 1994, pp. 44-45.

4. I would like to thank Joseph Kabel, Ralph Heidl, and William Holliday for their expert help in making these maps, from the laborious compiling of the original data base, to the final, high-resolution slides, to the animation sequence for educational television.

5. P. Gould, D. DiBiase and J. Kabel, "Le SIDA: la carte animée comme rhétorique cartographique appliquée," *MappeMonde*, 1, 1990, pp. 21-26.

6. Based upon the largest 102 urban centers, containing half the population of the United States, a simple model of hierarchical diffusion predicts AIDS rates in 1986, 1988, and 1990 (r=0.80), with residuals highlighting well-known social and cultural characteristics (P. Gould, "Spreading HIV Across America with an Air Passenger Operator", paper given at *the International Symposium on Computer Mapping in Epidemiology and Environmental Health*, Tampa, Florida, February, 1995).

7. Particularly when such a map sequence becomes the data input for spatial adaptive filtering and parametric tracking, techniques which search out the information in spatiotemporal sequences and use it to forecast the next maps. See, for example, P. Gould, "Epidémiologie et maladie," Chapter 53 in A. Bailly, R. Ferras and D. Pumain (eds.), *Encyclopédie de Géographie* (Paris: Editions Economica, 1992), pp. 949-969; and J. Kabel, *A*

Geographic Perspective on AIDS in the United States: Past, Present and Future (University Park, PA: Ph.D. Dissertation, Department of Geography, 1992).

8. New definitions of AIDS, based upon impeccable medical criteria, and designed to enhance earlier diagnosing and life-prolonging treatments, were meant to come into effect on January 1, 1992, a presidential election year. They were "delayed" until January 1, 1993.

9. J.B. Harley, Deconstructing the map, in T. Barnes and J. Duncan (eds.), *Writing Worlds: Discourse, Text and Metaphor in the Representation of Landscape* (London: Routledge, 1992), pp. 231-247.

10. P. Gould, J. Kabel, W. Gorr and A. Golub, AIDS: predicting the next map, *Interfaces*, 21, 1991, pp. 80-92.

11. We raised this question while modeling, with spatial adaptive filtering, the course of the AIDS epidemic in the 64 health districts of the Bronx (New York). See Gould, *The Slow Plague*, pp. 130-133.

12. AIDS Surveillance Group, *Monthly Report for Washington*, D.C., June, 1991, p. 2.

13. Such maps have been created for Los Angeles, with the full cooperation of the Los Angeles AIDS Surveillance Group, by Dr. William Bowen and his undergraduate students as a socially meaningful cartographic exercise. The dots are placed randomly within the thousands of census districts, some of them no bigger than 4-5 city blocks. Total confidentiality is preserved. See W. Bowen, et al., AIDS in LA, and AIDS in LA 1983-89, *Occasional Publications in Geography*, Nos. 4 and 6, California State University, Northridge, CA, 1989.

14. An excellent empirically-based study on the confidentiality question, using the more open Italian census, particularly the region of Tuscany, has been carried out by S. Openshaw ("An empirical study of confidentiality crisis in the release of micro census data," *CURDS Publications*, 1989, Center for Urban and Development Studies, Newcastle University).

15. T. Dawson, "Towards a spatial ethic: the question of confidentiality and the geographic aggregation of data," *Proceedings of the Association of American Geographer, American Association of Geographers*, Miami, Florida, April 1991, p. 45.

16. This remarkable paper was based on over 2 million medical exams given to young (17-23) people volunteering for military service. Even today, it is the largest data base in the world, generally denigrated by classically-trained statisticians because it is not a pure random sample. The reader must judge when an N, now exceeding 4 million, is worth considering as a source of information, or whether it should be thrown away. See L. Gardner, J. Brundage, D. Burke, J. McNeil, R. Visintine, and R. Miller, "Spatial diffusion of the human immunodeficiency virus infection epidemic in the United States, 1985-87," *Annals of the Association of American Geographers*, 79, 1989, pp. 25-43.

17. Confirmed by analyses using the expansion method of Casetti, in which the parameters of a cubic temporal equation are made quadratic functions of a spatial variable like population density. The fit of such models is quite tolerable, although not as good as spatial adaptive filtering. See J. Kabel, *A Geographic Perspective on AIDS in the United States*; and R. Wallace, "Transmission on geographically-centered social networks: effects of population density and spatial distribution," New York Psychiatric Institute, July, 1992, pp. 1-7.

18. A. Gorlub, W. Gorr, and P. Gould, "Spatial diffusion of the HIV / AIDS epidemic: modeling implications and case study of AIDS incidence in Ohio," *Geographical Analysis*, 1992, pp. 85-100.

19. P. Gould and R. Wallace, "Spatial structures and scientific paradoxes in the AIDS pandemic," *Geografiska Annaler*, 76B, 1994, pp. 105-116.