Debate on the paper by Gilberto Câmara & Antônio Miguel Vieira Monteiro

Debate sobre o artigo de Gilberto Câmara & Antônio Miguel Vieira Monteiro

Trevor C. Bailey

School of Mathematical Sciences, University of Exeter, Exeter, England.

The authors are to be commended on this interesting and thought-provoking review of the new analytical possibilities offered by "geocomputation" techniques, and I concur with them that there is considerable potential for useful applications of such techniques in the spatial analysis of health data. I agree wholeheartedly that the wider dissemination of such methods along with associated software tools may ultimately benefit many areas of geographical health and environmental research. We do indeed face a "data-rich" future in those fields of study and one where data will be not only voluminous but also complex. I mean both *complex in content* (e.g. in the topographic and geographical detail provided by GIS and remote sensing) and also *complex in structure* (e.g. data from disparate sources relating to different geographical scales and reference frameworks that need to be integrated in the study of many issues of interest in health research). Indeed one suspects that the future may already be with us! Traditional spatial analysis methods are not designed to handle such data complexity (e.g. many make little use of anything more sophisticated than simple Euclidean distance or the contiguity of areal units in order to reflect proximity, many assume some form of stationarity in the processes modeled, and few can handle data sources at different levels of spatial aggregation). We undoubtedly do need new analysis methods that are capable of exploiting more complex concepts. The authors convince me that geocomputational research offers some promising avenues for achieving that, and this paper and the work referenced in it therefore deserves serious and careful consideration by those involved in geographical health research.

However, while generally enthusiastic about the possibilities offered by the techniques discussed and agreeing with much that is said in the paper, there are some issues which I would like to take up from the perspective of an applied statistician with an interest in spatial analysis, and I restrict my remaining remarks to those.

First, I do not consider that we need to think of geocomputational techniques as an alternative to more traditional statistical methods and models, but rather as a complement to them. Modern statistical analysis is itself a broad church and no stranger to computer-intensive methods. To establish a dividing line between many existing forms of descriptive or exploratory statistical analysis and geocomputation may be useful in order to focus attention and promote the use of novel forms of algorithmic approach. However, from the point of view of a practicing statistician, such a distinction is somewhat artificial. Many existing forms of visualization and projection techniques used in statistics, particularly those employed in the analysis of high-dimensional data, have little to do with traditional notions of statistical inference. Statisticians are quite comfortable and familiar with using essentially algorithmic methods where appropriate and have been doing so for many years. What matters in exploring data is that the analyses conducted are careful and thorough, not what type of algorithms are employed to achieve that. So I do not see geocomputation as competing with my current statistical exploratory tool box, but rather as adding to it (in fact I consider two of the methods discussed in this paper, the GAM and local indicators of spatial association, to already be a part of it, although I am happy to see them re-branded as geocomputation if it encourages their use!).

However, I do stress that I see geocomputational techniques as essentially *exploratory*, and that brings me to my second point. Answers to the questions: *Are there any patterns, where are they, and what do they look like?* Are undoubtedly of value, but ultimately they are a preliminary to the more important ones of *Why are they there, will they happen again, and how will they change if we intervene in a particular way?* The answer to this second set of questions requires a scientific explanation of the phenomenon under study, and given the intrinsically stochastic nature of most social, environmental, and health-related phenomena, the best tool for this will, I suspect, remain the statistical model. I am not suggesting that such models will be *true*; the very word model implies simplification and idealization and I fully appreciate that that complex geographical health and environmental systems cannot be exactly described by a few formulae. However, the construction of idealized representations of important aspects of such systems consistent with the existing substantive epidemiological or public health knowledge should remain the ultimate goal. I therefore see the primary value of geocomputation as assisting in the sta-

tistical model-building process and not circumventing it.

This view of the role of geocomputation (which I freely admit may be narrower than that held by the authors) leads me to my third point, which relates to various concerns over the practical use of some kinds of geocomputational algorithms. The process of model-building is ideally both interactive and iterative. The analyst needs to try out ideas on the data, and this requires exploratory tools that can be guided or steered towards particular chosen ends or hypotheses. At present, many geocomputational algorithms appear too much of a "black box" to make this possible. The very nature of the algorithms makes it difficult to provide simple, readily understood control parameters which enable them to be "steered" towards answering particular questions which one might wish to ask of the data. In a sense they provide an answer in the absence of a question. This detracts from their value as exploratory tools for the model builder. In that sense what is often termed "artificial intelligence" might be better referred to as "artificial un-intelligence". There is also the problem of whether such techniques produce *robust* results as opposed to ones which are pure artifacts of the data. I appreciate that traditional notions of statistical significance and standard error cannot and perhaps should not be looked for in relation to these algorithms and that different algorithmic approaches will naturally reveal different aspects of the data. However, the *sensitivity* of the results from any one of them (e.g. to starting conditions or in repeated application to various subsets of the data) needs to be investigated and is often not. If the data are to be mined then we need to establish whether a vein of gold has been found or a vein of fool's gold, and currently the algorithms are weak on the diagnostics that would enable us to measure that.

In summary I do not wish to appear as a dogged defender of existing spatial statistical models and methods. I am well aware how deficient many of those are. For example, traditional spatial models largely involve space in terms of glib abstractions – "distances", "boundaries", and "edge effects". Of course in reality the areas over which analyses are being conducted are vastly complex, criss-crossed with natural boundaries such as forests, rivers, or ranges of hills, or else human constructions such as roads, industrial estates, recreational parks, and so on. Many commonly used spatial statistical methods and models should be viewed in the cold light of their spatial simplicity compared with what we know to exist in geographical reality and upon which data are now available through GIS and remote sensing. Humility would indeed be wise for anyone defending such models, and it is useful to be reminded of that and presented with some novel algorithmic approaches in this paper which may assist to address it. Therefore I welcome new and improved algorithms for exploratory spatial analysis of health data capable of exploiting the complexity of data and of geography. If geocomputation matures to offer that, then I am very comfortable with using it. However, I think we should be cautious about exaggerating its potential. Data analysis in general involves more than methods; it depends on contextual knowledge of the phenomenon under study, the objectives of the analysis, the quality and origins of the data, and the judgment and experience of the analyst. Because of that there is a long-standing resistance among applied statisticians to the suggestion that what they do is just another branch of mathematics. It would not be surprising to find them equally resistant to the suggestion that it should become a branch of computer science. I also doubt that geographical health and environmental research would necessarily benefit if that were to become the case.

David Waltner-Toews

Department of Population Medicine, University of Guelph, Guelph, Canada.

Epidemiologists, after several decades of favoring non-spatial statistical models, are increasingly realizing the importance of understanding socioeconomic and ecological contexts in the interpretation of disease patterns in populations (McMichael, 1999). As the questions we are asking change in both scope and nature, input from scholars in non-health fields with expertise in studying spatial patterns, such as this paper, are a welcome addition to the health literature.

The authors state that their intent is to *"draw the attention of the public health community to the new analytical possibilities offered by geocomputational techniques"*. While the introduction of these techniques to health researchers is laudable in and of itself, I would like to throw out some cautionary notes, based on some experience working with interdisciplinary teams where these techniques have been proposed.