PERSPECTIVE

Replication across space and time must be weak in the social and environmental sciences

Michael F. Goodchild[a,1] and Wenwen Li[b,1]

Edited by Alan T. Murray, University of California, Santa Barbara, CA, and accepted by Editorial Board Member B. L. Turner June 24, 2021 (received for review November 30, 2020)

Replicability takes on special meaning when researching phenomena that are embedded in space and time, including phenomena distributed on the surface and near surface of the Earth. Two principles, spatial dependence and spatial heterogeneity, are generally characteristic of such phenomena. Various practices have evolved in dealing with spatial heterogeneity, including the use of place-based models. We review the rapidly emerging applications of artificial intelligence to phenomena distributed in space and time and speculate on how the principle of spatial heterogeneity might be addressed. We introduce a concept of weak replicability and discuss possible approaches to its measurement.

replicability | artificial intelligence | spatial heterogeneity | place-based analysis

Recently replicability and reproducibility have received attention across virtually all of the sciences, because of failures to replicate certain previously published results. Some have identified a "replicability crisis," and comprehensive discussions have appeared in leading journals (1–3) and in reports of prestigious academies (4). Unfortunately the two terms are used differently by different disciplines. In this paper we define reproducibility as the ability to obtain the same results using the same data and methods, and replicability as the ability to obtain similar results (within acceptable bounds of uncertainty) using similar data (e.g., different samples obtained randomly from the same population) and similar methods (e.g., computer codes that have been designed to implement the same general procedures, but perhaps using different algorithms and running on different machines). In this paper we focus on replicability and use this definition.

While replicability is an accepted requirement in experimental psychology, physics, chemistry, and many other disciplines, its relevance in the social and environmental sciences is more nuanced. Our primary purpose in this paper is to explore replicability for studies of phenomena on or near the Earth's surface, that is, "replicability over the globe." We term this the geographic domain, which we define as extending from tens of kilometers below the surface to tens of kilometers above, and at spatial resolutions from millimeters to tens of kilometers. While this domain and these limits define the scope of our discussion, much of it may also be relevant to other fields.

Much of this concern stems from efforts to assess established results by closely repeating the original experiments. But in a recent paper, Nichols et al. (5) argue that the replicability of a previously published result should not be regarded as a binary outcome of a test, but as the focus of an evolving process. As more positive evidence becomes available, the degree of belief in the initial result increases; while accumulating negative evidence casts increasing doubt on the original research. In this paper we echo some aspects of this argument in addressing replicability across the geographic domain.

This introduction next focuses on spatial heterogeneity, a principle that forms the foundation for this discussion and its relevance to replicability. We then place studies of social, economic, demographic, and environmental phenomena in the geographic domain into two categories, which we term within-area and between-area studies, a distinction that helps to ground the subsequent discussion. Following this introduction we review relevant methodological arguments and approaches in the social and environmental sciences and include a section on what have been termed place-based methods. We introduce relevant methods of machine

[a]Department of Geography, University of California, Santa Barbara, CA 93106-4060; and [b]School of Geographical Sciences and Urban Planning, Arizona State University, Tempe, AZ 85287-5302

Author contributions: M.F.G. and W.L. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission. A.T.M. is a guest editor invited by the Editorial Board.

[1]To whom correspondence may be addressed. Email: good@geog.ucsb.edu or wenwen@asu.edu.

Published August 20, 2021.

learning using the example of convolutional neural networks (CNNs) and examine replicability in this context in the light of spatial heterogeneity. In the final and prospective section we argue for what we propose to term weak replicability, and for the development of associated metrics.

Our arguments in this paper are grounded in the principle of spatial heterogeneity. In a seminal paper titled "What is special about spatial data?", Anselin (6) identified two general properties of information about the geographic domain: spatial dependence and spatial heterogeneity. The principle of spatial dependence was succinctly expressed by Tobler (7) as "nearby things are more similar than distant things," and Sui (8) provides an introduction to a collection of papers that explore many implications of the statement. This simple expression of positive spatial autocorrelation is the conceptual foundation of many spatial methods. Spatial interpolation (9), for example, is a widely used family of techniques that estimates unknown or missing data values, based on known values at nearby locations. Maps showing contours also employ the same principle, and positive spatial autocorrelation is the foundation of the science of regionalized variables, commonly known as geostatistics (10).

In advancing spatial heterogeneity as another fundamental principle, Anselin (6) argued that expectation (the expected value of a discrete random variable) varies across the Earth's surface, as do many other statistical properties. Numerous forms of spatial organization and pattern exist in the geographic domain, and in focusing on spatial heterogeneity we do not wish to imply that spatial variation is necessarily random. Teleconnections and telecouplings (11), for example, can induce strong correlations between locations that are far apart, and two well-separated cities or agricultural regions may nevertheless possess many forms of similarity. Such studies are well outside the focus of this paper.

Rather, our concern in this paper is for the implications of spatial heterogeneity for scientific discovery. Many processes are expected to operate uniformly over the globe, and the objective of sciences such as physics and chemistry can been seen as the isolation of such uniform processes from any spatial (and temporal) variation that may obscure them. In the social and environmental sciences, however, there may be many reasons not only for an appearance of spatial heterogeneity in the relevant variables, but also for a confounding of the process of discovery. Models may fit better to data in some areas than in others, and the fitted values of parameters may also vary. It may be difficult if not impossible to exclude the variable effects of context, and spatial heterogeneity will also appear when models are incompletely specified, since the missing variables will almost certainly exhibit spatial variation. Below we explore some of the literature of the social and environmental sciences that recognizes this issue and some of the approaches that have been used to address it.

Studies of phenomena in the geographic domain fall into two broad categories, which we term within area and between area. Within-area studies examine the spatial variation of some phenomenon of interest within a frame or study area. The terms "extent" and "scale" are often used in this context, and we sometimes refer to studies of large areas as large-scale studies (though scale is also used to refer rather confusingly to spatial resolution and in cartography, to the related property of the representative fraction; ref. 12). Spatial heterogeneity leads us to expect both variation within the study area and dependence of results on the choice of study area. If the area covers less than the entire surface of the Earth we also expect difficulties in replicating the results of the study within other areas that may or may not overlap with the original area.

Such studies use methods of spatial analysis to make inferences and conclusions. Spatial analysis (see, for example, ref. 9, p. 291) can be defined informally as analysis that relies directly or indirectly on the locations of the features or samples that are being analyzed, or more formally as any analysis whose results are not invariant under changes in those locations. This definition is extremely broad: in addition to the kinds of inferences from spatial analysis that advance scientific knowledge and are the subject of this paper, it also covers techniques that are prescriptive or normative in character, seeking to design part of the geographic domain to achieve certain objectives (see, for example, ref. 13).

Methods of spatial analysis are today implemented in geographic information systems (GISs) and in many packages of cartographic or statistical software, and issues of replicability and reproducibility in spatial analysis have been addressed in recent papers (14, 15). However, most of these methods assume that the processes affecting the phenomena within the study area are spatially homogeneous; we focus here on the implications of spatial heterogeneity for replicability. We discuss below how some methods of spatial analysis have been adapted to address specifically the concept of spatial heterogeneity.

Between-area studies compare phenomena at a selection of locations, without necessarily being concerned with how phenomena vary within areas; thus they do not necessarily utilize methods of spatial analysis. They are nevertheless impacted by spatial heterogeneity, so we include them here.

Replicability in Practice

Replicability, or lack of it, has been a long-standing problem in disciplines that focus on the geographic domain, such as regional science, spatial ecology, geography, epidemiology, sustainability science, or spatial econometrics. As complex and adaptive, social–ecological systems (SESs) often embrace observations collected from within a specific spatial and temporal setting, under what may be unique weather and climate conditions, or shifting political regimes, making replicability difficult to achieve (16). Ostrom (17) argues that in SES studies it is unlikely that a simple, universally applicable, predictive model exists due to the complexity of places (see also refs. 18, 19). Sustainability science, a field that studies interactions between nature and society, requires an understanding of the interactions between global processes and the local characteristics of particular places (20), which may differ in socioeconomic, political, and ecological conditions. At the center of these studies is the spatial context that situates the data collection, research design, and findings (21). When data related to a geographic location are collected, they must be set in a specific spatial and temporal frame. This practice of frame-dependent geographic measurement (22) creates contextual uncertainty, causing results to differ from location to location, and this will negatively affect a study's replicability even when all other experimental settings remain the same (23). The data representation, be it a point, line, or area, may also be a factor contributing to the generation of nonreplicable results despite the use of the same methodological workflow (24).

One approach to the problem of replicability across a spatially heterogeneous globe has been to reduce it to a dichotomy: simply, some findings are replicable in the style of physics and chemistry and others are not. The Polish–Dutch geographer Varenius, writing in the 17th century in books that were later annotated by Isaac Newton, distinguished between special geography, that is, the distinct nature of places, and general geography, the general and presumably replicable principles on which the geographic

domain is constructed (25, 26). In the 1950s this issue came to a head in a lengthy debate in the discipline of geography between these two positions, which we term idiographic and nomothetic, respectively. Idiographic science, or "areal differentiation," consists in recording the distinguishing and therefore unreplicable properties of places; to paraphrase Hartshorne (27), the only possible general principle of the geographic domain is that all places are unique. Schaefer (28) argued the opposing case for nomothetic geography, or what later became known as scientific or theoretical geography (29), as the search for general laws and principles that apply and are replicable everywhere, and presumably at all times.

However, the idiographic position nevertheless implies a certain degree of homogeneity within the places or areas being studied, and thus a geographic scale, introducing one form of nuance in what otherwise would be a simple dichotomy. We might study the differences between continents, nations, states, counties, cities, or neighborhoods, while implicitly assuming that within each of these areas the distinguishing characteristics are near uniform. Thus the difference between the nomothetic and idiographic positions is in practice not absolute, but a matter of geographic scale, from the nomothetic global to the yet-to-be-specified idiographic local.

Several place-based methods attempt to address the issue of spatial heterogeneity directly, by incorporating it into the modeling process. This could be done (in a within-area approach to spatial heterogeneity) by partitioning the study area into regions and calibrating models independently in each region. However, there are several obvious objections to this approach, including the lack of an objective basis on which to define the regions, and the sharp breaks in coefficients that would occur at region boundaries. Casetti (30) developed what he termed the expansion method by allowing the coefficients in a model to vary with location. Each coefficient is allowed to be a linear or quadratic function of geographic coordinates, thus incorporating spatial heterogeneity directly into the model. Fotheringham et al. (31) generalized this approach in their technique of geographically weighted regression (GWR). In GWR, the model's structure and the independent variables are defined generally, but the parameters of the model are allowed to vary spatially and perhaps through time. Unlike the expansion method, coefficients are calibrated for each location and can be mapped. Both of these techniques introduce additional degrees of freedom, so it is not appropriate to use goodness of fit as a criterion for choosing between them or for rejecting a simple regression model (32).

In a simple application, we might build a GWR model based on the assumption that the presence of a swimming pool will always affect the price of a house. However, we also acknowledge that the effect of the variable could be very different across space (e.g., in Arizona compared with Minnesota). This is done by estimating the parameters of the model point by point, at each point borrowing support only from nearby observations, and using weights that decrease with distance according to a function parameterized by a preset bandwidth value. In the recently developed multiscale version of GWR (MGWR; 33) the bandwidth is separately calibrated for each independent variable in the model, in effect allowing each variable's influence to vary at its own geographic scale. We might interpret these functions as distinguishing the scales of the various processes that operate on the landscape.

Similar to GWR and MGWR, the spatially varying coefficient (SVC) (34, 35) model and its multiscale extension (36) treat the bandwidth parameters in a Bayesian framework, resulting in explicit measures of uncertainty on each bandwidth estimate. Griffith (37) has proposed a spatial-filter–based local regression (SFLR) as an alternative to GWR; for a comparison of the two techniques see the work of Oshan and Fotheringham (38). There have also been numerous extensions and modifications of the basic GWR format. Wheeler (39), for example, has developed a geographically weighted lasso regression that employs shrinkage to reduce model coefficients, potentially to zero, in some parts of the study area.

All of these approaches appear to find intermediate positions in the longstanding debate between nomothetic and idiographic—between global and local or general and special. Certain aspects of models are held constant while others are allowed to vary. In GWR, the structure of the model and the choice of variables are held constant, but the coefficients and the overall goodness of fit are allowed to vary. We might term this weak replicability, acknowledging that some but not all aspects of a geographic principle or model are replicable across space (and perhaps also through time); we expand on this point below.

This raises an important question, however, since it begs a definition of acceptability: How much weakness is acceptable, and can a result fail to replicate if such a generous approach is taken? We argued earlier that replication must be subject to "acceptable bounds of uncertainty," which allow for different errors in measuring instruments, the effects of taking different samples, and the use of different machines and codes for analysis. With a place-based method such as GWR, we expect also to be able to replicate the geographic variation in the coefficients, but again have no objective basis for determining how much variation between results is to be considered acceptable. We return to this and its implications in the concluding section.

Geospatial Artificial Intelligence

In recent years, geospatial artificial intelligence (GeoAI) has emerged as an addition to the analytic tools that can be used to advance scientific discovery in the geographic domain (40). In this section we focus on the role of spatial heterogeneity in GeoAI, and its implications for replicability and scientific discovery. Concerns have already been expressed about the replicability of methods of artificial intelligence (AI) in general. Hutson (41) argued in a recent paper in *Science* that AI is facing significant technical short-term challenges in replicability due to a lack of code sharing and proper documentation, and the practice of publishing research through nonreviewed platforms such as arXiv, but here our focus is on the more specific concerns of GeoAI. We first review applications of GeoAI, and then focus on one specific technique of machine learning to illustrate the nature of its results, and how it already accommodates the principle of spatial dependence. Finally, we address the difficulties in accommodating the principle of spatial heterogeneity and their implications for replicability.

AI has a longstanding history of notable successes in the geographic domain (42–45). We term this research area GeoAI. From the postwar era to the late 1980s, the dominant AI research paradigm was symbolic AI and expert systems. An expert system is essentially driven by prior knowledge and models; it works by predefining a set of reasoning rules or if-else conditions for a machine to make decisions when different scenarios are present. The creation of these rules relies heavily on domain experts. For instance, researchers at the National Aeronautics and Space Administration (NASA) developed an image-based geological expert system to identify mineral properties in the Earth's surface by analyzing hyperspectral remotely sensed images (46). In this approach a decision tree is codified by a knowledge engineer to

integrate a geologist's 10-step decision-making process into the expert system. To date, decision-tree analyses remain popular approaches for image analysis and data processing. The methods have also been extended from binary classification to probabilistic models, such as random forests (47).

Agent-based modeling (ABM) (48) and cellular automata (CA) (49) are also commonly used methods for building an expert system. They both rely on microscale simulation, with the former simulating the behavior of moving agents in space and the latter simulating process change over a region partitioned into a regular grid (50). When the solution space for a real-world problem becomes too large due to combinatorial complexity, heuristic techniques can be adopted to guide the expert system to achieve an acceptable but perhaps nonoptimal solution (51).

In symbolic expert systems the decision rules are well defined. However, a flaw of such systems is that they are monotonic. As more rules are added, the more knowledge is encoded in the system and the more complex the system becomes. This will inevitably affect its computational efficiency and flexibility because new knowledge cannot overwrite old knowledge. Another stumbling block of an expert system is the lack of generalization in the reasoning rules across space (and time), making them extremely difficult to replicate when a study area changes. This issue is predominantly seen in the analysis of remotely sensed images, in which the spectral, optical, and spatial–contextual properties of an object may vary significantly across spatially heterogeneous landscapes and geographic regions.

Data-driven models attempt to address these issues by having the machine learn, without prior expert knowledge, the correlations between input data and output symbols. A symbol in this context means a high-level concept that is understandable by humans, such as a class label in a classification system. Models such as artificial neural networks (ANNs), support vector machines, and the aforementioned random forest, all belong to this category. They work by supervised machine learning, in which training data are required to feed the model to gain intelligence between the input and the desired output. These models are called shallow machine-learning models because the models do not have the ability to learn from the raw data; instead, the input of such models needs to be a set of independent variables (features) that are known to influence the outcome. Compared to the knowledge-driven models, this kind of model can be easily retrained with new data. While the model stays the same, the model parameters can be adjusted when input data are nonstationary, hence they are better at encoding new knowledge and modifying it. However, the model does not control its transferability or replicability across space, which is indeed determined by the spatial heterogeneity of the spatial processes underlying the data. ANN models that incorporate spatial weights to respect spatial heterogeneity have also been developed in recent years; however, some of the model changes introduce extra computational time without gaining significant improvement in model performance over the GWR model (52, 53).

In recent years, deep-learning models, especially deep convolutional neural network (DCNN) models, have emerged as a breakthrough in AI research because of their outstanding capability in mining from big data and learning representative features (i.e., independent variables) automatically from raw data. These models have shown good performance in specialized tasks, such as image classification and speech recognition (54). In the next subsection we present a short discussion of DCNN and use it to illustrate how this method of GeoAI incorporates the principle of spatial dependence. Later we discuss what replicability might mean for such methods.

Deep learning refers to computational models that are composed of multiple processing layers, creating a data-processing pipeline that can automatically learn and extract prominent features or representations of the data that enhance prediction. The multilayer architecture in deep learning has evolved from an artificial neural network, which purports to mimic how information might be propagated in the human brain in support of decisions. Fig. 1 illustrates an example of a multilayer, feed-forward neural network (Fig. 1*A*), which we might term a shallow learning model, and its extension to a DCNN for geospatial applications (Fig. 1*B*).

A key innovation in the DCNN shown in Fig. 1*B* is its introduction of a convolution module, which applies a moving window to perform convolution, another form of weighted sum, to extract prominent features automatically. The result of this operation is called a feature map. Max pooling is enabled between convolution layers to select the maximal value in a $k \times k$ pixel block to reduce the dimension of the feature map and reveal important features at different spatial scales. Although not explicitly stated, these DCNN models have naturally incorporated the principle of spatial dependence in the methodological design. The moving window idea, which applies convolution on each square subarea containing $k \times k$ adjacent pixels to extract prominent features from data, is based on the underlying assumption that nearby pixels composed together will provide a meaningful way of representing data features. In effect, this assumes positive spatial autocorrelation. Another principle built into the deep-learning technique is scale-based analysis. The max-pooling operation enables scale-dependent feature extraction, as an important strategy to aid understanding of spatial structures and processes.

The concept of replicability which we adopted at the outset asks for "similar" results across the geographic domain; this implies some means of measuring how the results in one area compare with results in another area. In the within-area strategy defined in the *Introduction* it implies that similar results have been obtained across the study area or when the same methods are applied to more than one area. With the between-area strategy, it implies similarity across the results obtained in each area. When traditional methods such as linear regression are used, it is clearly possible to assess similarity by comparing the fitted coefficients and the overall goodness of fit. But how should the results of GeoAI be compared?

The results of an application of DCNN or any other form of neural network consist of a complex set of weights on the links of the network. To assess the degree to which one application replicates another, it would be necessary to compare sets of weights. Moreover these links will not necessarily persist across applications, so it is not always possible to compare weight with weight. Perhaps it would be possible to create a version of DCNN or other GeoAI tool that directly incorporates the principle of spatial heterogeneity, in a similar manner to GWR and other place-based techniques. The original objective function $argmin \sum_{l=1}^{m} \left| Y_l - \widehat{Y}_l \right|^2$ of Fig. 1 would become $argmin \sum_{i=1}^{n} \sum_{l=1}^{m} w_{il} \left| Y_l - \widehat{Y}_l \right|^2$, where m is the number of output variables Y, n is the number of nearby sampled data that would be involved in a localized deep-learning process, and w_{il} is the weight assigned to a sample data point using some spatial weighting scheme. By incorporating this new objective function, the use of the weights will ensure that the model

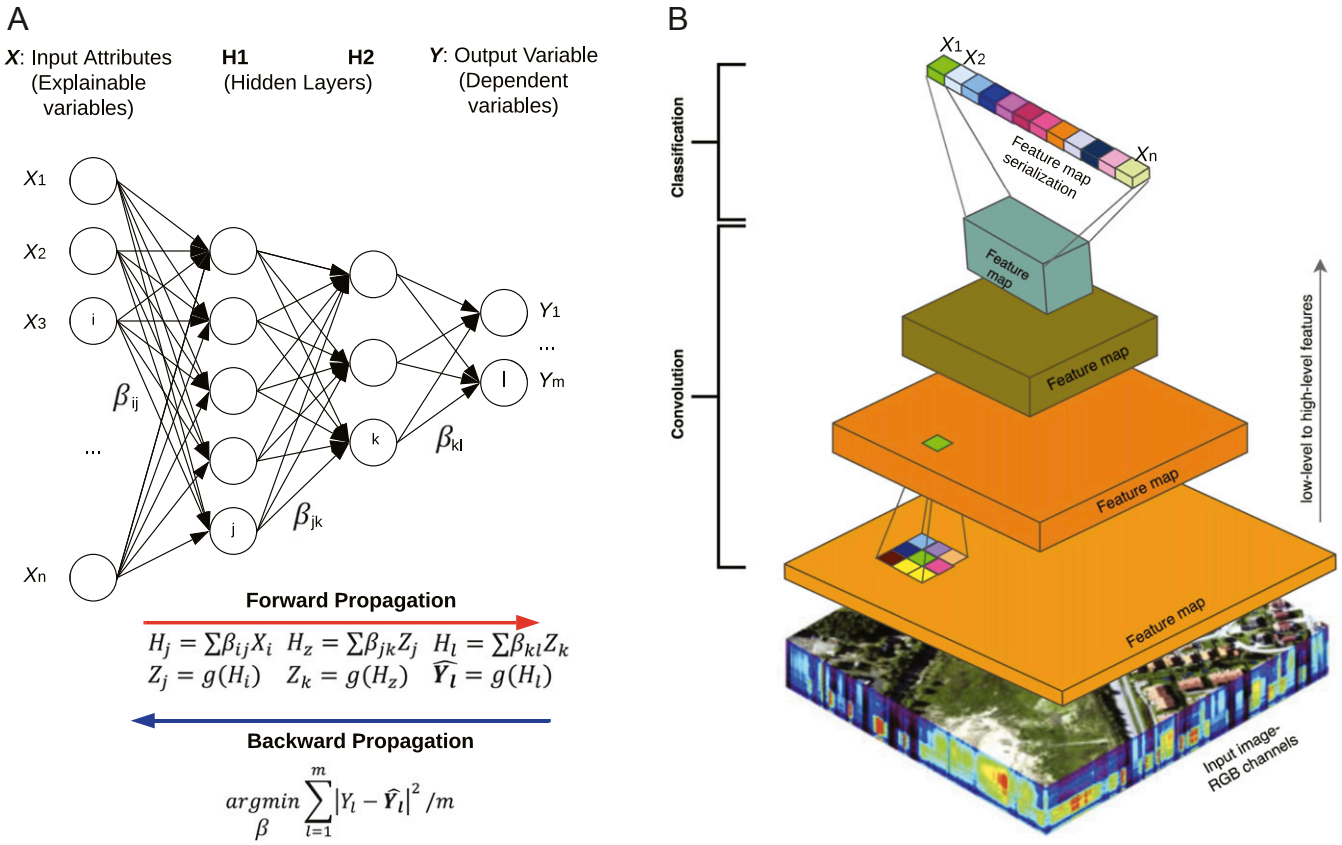

Fig. 1. Examples of a multilayer neural network (*A*), and a DCNN as one might be used in GeoAI for analyzing a remotely sensed image (*B*). Specifically, *A* presents a fully connected, feed-forward neural network, which is composed of multiple layers, and each layer is composed of multiple nodes. The input to the network consists of multiple attributes (or features), represented by X_i, and the output is *Y*, where *Y* could be either a numerical value, suitable for use in a regression scenario, or a categorized value, suitable for use in classifications. The goal of learning is to find a nonlinear mapping between *Y* and *X* that minimizes the mean squared difference between the expected output (*Y*) and predicted output ($\hat{Y}$). This mean squared difference is the most commonly used regression loss function in a machine-learning model. The mapping result is captured in the set of weights (β) that act on the links connecting nodes in the network. *g*() is an activation function that determines the output of a neuron. It also helps to normalize the value into a range between [0,1] or [−1,1]. In *B*, a DCNN is used to extract low- to high-level features (*X* in *A*) in an automated manner such that there is no need for manual feature selection. Once all the features or the independent variables (*X*) are extracted, a DCNN can be linked to a fully connected layer (*A*) for final classification and prediction.

calibrated at location *i* benefits from the support captured from nearby observations and is less dependent on the existence of a study area boundary. This strategy might constitute a new GeoAI model that explicitly addresses spatial heterogeneity in the modeling process. But with GWR the set of independent variables and the algebraic structure of the model remain constant over space, allowing the fitted coefficients to be mapped over space and allowing the researcher to make some level of general statement.

Compared to GWR, the result of the GeoAI model would no longer be a coefficient surface (Fig. 2). Instead, it would be a nonlinear, complex function that varies across space, generating a multidimensional surface that may appear to ignore the principle of spatial dependence and will be even more difficult to interpret than the results of GWR. Since AI models generally have more degrees of freedom than traditional regression models, they will likely achieve better fits. But in the spirit of Occam's razor, if a simpler model works well, and if its mechanics can be easily explained, why use a more elaborate model with millions of parameters that need to be learned?

In summary, there is no obvious way of comparing the results of one GeoAI application, in one study area at one time, to the results obtained independently in another study area or at another time. When GeoAI is used for prediction, it may be acceptable to adopt an idiographic position and regard the results of any application as unique. But we are unable to see how GeoAI can lead to the discovery of nomothetic knowledge of the geographic domain, given the presence of spatial heterogeneity. Despite the compelling performance of AI algorithms in appearing to emulate some of the functions of human minds, these algorithms are often questioned for their opaque learning processes and lack of interpretability. Unlike human brains, which contain a meta-algorithm (55) to explain the rationale for reasoning, either through observations, logical reasoning, or experience based on accumulated scientific knowledge, the deep layers in complex AI algorithms are often not as comprehensible as human intuition. Their black-box nature is highly worrisome, especially when they are adopted to answer scientific questions that require a transparent reasoning process and valid results—key aspects that ensure scientific replicability. Judea Pearl, winner of the Turing Award in 2011, argued that the secret behind deep learning is merely a "curve fitting" (56).

Weak Replicability

The preceding sections have argued for a science of the geographic domain that is neither idiographic nor nomothetic, but

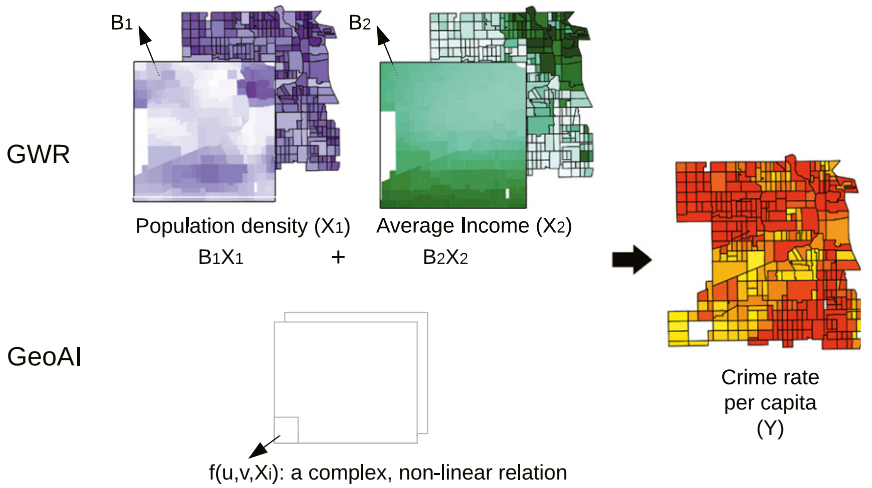

Fig. 2. Hypothetical results generated by the place-based method GWR and a possible GeoAI method in a study of how population and income make an impact on crime. In the GWR result, coefficient values will be determined at each location (*u*, *v*) by borrowing support from nearby data points. These values form a geographically varying coefficient surface for each independent variable X_i. By comparison, the values at each unit location that result from the GeoAI method will be complex, nonlinear functions instead of a single number.

instead occupies an intermediate position between those two extremes. We see this as a logical consequence of the near-ubiquitous presence of spatial heterogeneity and the persistent inability to completely specify models and processes in the social and environmental sciences. Moreover, it is never clear how close a new result must be to a previously published one to qualify as a replication. Thus we concur with Nichols et al. (5) in questioning the traditional approach to replicability, in which a new result must either replicate an existing one, or conflict with it. In the geographic domain and in the ubiquitous presence of spatial heterogeneity, it is difficult to imagine a clear outcome of this traditional approach.

Instead it seems that research in the geographic domain requires a concept of what we choose to term weak replicability, in which some of the requirements of replicability are dropped. This might take the form of a single-model structure but with spatial variation in the model's parameters and goodness of fit, as in GWR, or a single neural network with spatial variation in its weights, as in our suggested spatially heterogeneous version of DCNN. From this perspective the degree of weakness becomes important, because there must be some limit to how much variation is acceptable; without it, all knowledge is implicitly replicable. How much variation should be allowed in a fitted GWR model, for example, and should models with more than a threshold degree of spatial variation in parameters be rejected?

Questions like these are often resolved using the infrastructure of inferential statistics. Variables in a multivariate model are tested to determine the degree to which they contribute to the explanatory power of the model, and if that contribution is within the range of chance contributions from similar but in reality unrelated variables, then the variable is rejected. But consider a simple multiple regression that is applied independently in two study areas, and assume that the variables are all statistically significant in both areas, suggesting that the model is replicable. Yet the values of the model's coefficients will not be identical and may be more different than would be acceptable under inferential tests, given the observed variances and sample sizes. How much variation should be allowed in the fitted coefficients before the model is declared unreplicated? We can find no simple answer to this question.

Moreover the property of spatial dependence creates problems for inferential tests in the geographic domain. While sampling is the basis of inferential tests, in many cases there is no sampling involved in within-region analyses; instead, all of the available data are used. Further, inferential tests commonly assume that individuals are randomly chosen; because of spatial dependence, the independence assumption is often compromised. There is of course a vast literature on this topic, so we merely note its existence here.

If replicability is weak, and if metrics can be devised to measure the differences between models, then we see an opportunity for the development of geographies of replicability, as follows. Given the Tobler principle of positive spatial dependence, it seems reasonable to assume that replicability will be stronger over short distances and decline with distance. However, there may be exceptions to this general principle: similar cities may provide greater replicability than dissimilar ones, independently of their separation in space. As more evidence accumulates, in the form of calibrated models in different geographic areas, it would be fascinating to develop what we might term "replicability maps."

Conclusion

Replicability is a key principle of the scientific method. But its meaning is especially nuanced in those social and environmental sciences that deal with phenomena embedded in space and time. The principle of spatial heterogeneity is by now well established, particularly in disciplines that deal with the surface of the Earth and with change through time. Yet as a principle, it appears to cast doubt on the property of replicability, implying that the results of any study will change when the bounds of the study change, in a direct challenge to replicability over the globe.

In response, we introduced the concept of weak replicability as an intermediate position between the nomothetic and idiographic ideals—between a search for general principles on the one hand, and a belief that all places are unique on the other. Under weak replicability the model specification might be generalizable, for example, but the model's parameters might be allowed to vary spatially and temporally. We used the example of GWR, one of a

number of place-based methods that provide practical ways of implementing this intermediate position.

In line with the concept of weak replicability, we suggested that replicability might be regarded as a variable rather than a binary property in those sciences that deal with environmental and social phenomena embedded in space and time. Perhaps metrics of replicability could be developed, based on the similarity of calibrated parameters. Such metrics would have their own geography, since results are likely to be more replicable in nearby areas than in distant ones, in an echo of the principle of spatial dependence.

Recently there has been much interest in the use of AI techniques in the geospatial sciences, despite the widespread questioning of their value in scientific knowledge discovery. We focused on the analysis of remotely sensed Earth images using deep convolutional neural networks: they incorporate the principle of spatial dependence, but not spatial heterogeneity. We suggested that spatial heterogeneity might be implemented in these models by allowing them to be calibrated locally, in a similar manner to GWR. But we concluded that such models would be very hard to interpret, and their generalizability would be so weak as to add little in the way of useful knowledge.

Nevertheless we should not underrate the value of AI in facilitating advances in science. Karpatne et al. (57) have proposed a theory-guided paradigm for data science in which AI researchers and domain scientists work together to advance scientific understanding and produce novel insights—somewhat similar to our proposal here to incorporate the established principle of spatial heterogeneity, along with spatial dependence, into spatial applications of AI. We should also not underestimate AI's evolving expressive power, which computer scientists have been attempting to improve (58). Open Machine Learning (59) and other open-science frameworks (60) will improve transparency in both AI research and other data-driven scientific research, hopefully leading to a more positive form of scientific replicability in AI.

Data Availability. There are no data underlying this work.

1 A. A. Aarts; Open Science Collaboration, PSYCHOLOGY. Estimating the reproducibility of psychological science. *Science* **349**, aac4716 (2015).
2 M. Baker, Is there a reproducibility crisis? A Nature survey lifts the lid on how researchers view the'crisis rocking science and what they think will help. *Nature* **533**, 452–455 (2016).
3 M. K. McNutt *et al.*, Transparency in authors' contributions and responsibilities to promote integrity in scientific publication. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 2557–2560 10.1073/pnas.1715374115. (2018).
4 National Academies of Sciences, Engineering, and Medicine, *Reproducibility and Replicability in Science* (National Academies Press, Washington, DC, 2019).
5 J. D. Nichols, M. K. Oli, W. L. Kendall, G. S. Boomer, Opinion: A better approach for dealing with reproducibility and replicability in science. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2100769118 10.1073/pnas.2100769118. (2021).
6 L. Anselin, *What is Special About Spatial Data? Alternative Perspectives on Spatial Data Analysis. Technical Report 89-4* (National Center for Geographic Information and Analysis, Santa Barbara, CA, 1989).
7 Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Econ. Geogr.* **46**(supp. 1), 234–240.
8 D. Z. Sui, Tobler's First Law of Geography: A big idea for a small world? *Ann. Assoc. Am. Geogr.* **94**, 269–277 (2004).
9 P. A. Longley, M. F. Goodchild, D. J. Maguire, D. W. Rhind, *Geographic Information Science and Systems* (Wiley, Hoboken, NJ, ed. 4, 2015).
10 P. Goovaerts, *Geostatistics for Natural Resources Evaluation* (Oxford, New York, 1997).
11 V. Hull, J. G. Liu, Telecoupling: A new frontier for global sustainability. *Ecol. Soc.* **23**, 41 (2018).
12 A. H. Robinson, J. L. Morrison, P. C. Muehrcke, A. J. Kimerling, S. C. Guptill, *Elements of Cartography* (Wiley, New York, ed. 6, 1995).
13 R. L. Church, A. T. Murray, *Business Site Selection, Location Analysis, and GIS* (Wiley, Hoboken, NJ, 2009).
14 P. Kedron, A. E. Frazier, A. B. Trgovac, T. Nelson, A. S. Fotheringham, Reproducibility and replicability in geographical analysis. *Geogr. Anal.* **53**, 135–147 (2021).
15 P. Kedron, W. W. Li, A. S. Fotheringham, M. F. Goodchild, Reproducibility and replicability: Opportunities and challenges for geospatial research. *Int. J. Geogr. Inf. Sci.* **35**, 427–445 (2020).
16 S. M. Powers, S. E. Hampton, Open science, reproducibility, and transparency in ecology. *Ecol. Appl.* **29**, e01822 (2019).
17 E. Ostrom, A diagnostic approach for going beyond panaceas. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 15181–15187 (2007).
18 E. Ostrom, H. Nagendra, Insights on linking forests, trees, and people from the air, on the ground, and in the laboratory. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 19224–19231 (2006).
19 E. Ostrom, M. A. Janssen, J. M. Anderies, Going beyond panaceas. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 15176–15178 (2007).
20 R. W. Kates *et al.*, Environment and development. Sustainability science. *Science* **292**, 641–642 (2001).
21 M. F. Goodchild, The quality of geospatial context. In *International Workshop on Quality of Context* (Springer, Berlin, Heidelberg, 2009).pp. 15–24.
22 W. R. Tobler, "Frame independent spatial analysis" in *Accuracy of Spatial Databases*, M. F. Goodchild, S. Gopal, Eds. (Taylor and Francis, Basingstoke, 1989), pp. 115–122.
23 M. P. Kwan, The uncertain geographic context problem. *Ann. Assoc. Am. Geogr.* **102**, 958–968 (2012).
24 R. L. Church, "Location modelling and GIS" in *Geographical Information Systems*, P. A. Longley, M. F. Goodchild, D. J. Maguire, D. W. Rhind, Eds. (Wiley, New York, 1999), pp. 293–303.
25 B. Varenius, *Geographia Generalis* (Elsevier, Amsterdam, 1650).
26 W. Warntz, Newton, the newtonians, and the geographia generalis varenii. *Ann. Assoc. Am. Geogr.* **79**, 165–191 (1989).
27 R. Hartshorne, The nature of geography: A critical survey of current thought in the light of the past. *Ann. Assoc. Am. Geogr.* **29**, 173–412 (1939).
28 F. K. Schaefer, Exceptionalism in geography: A methodological examination. *Ann. Assoc. Am. Geogr.* **43**, 226–249 (1953).
29 W. Bunge, *Theoretical Geography. Second Edition. Lund Studies in Geography Series C: General and Mathematical Geography, No. 1.* (Gleerup, Lund, Sweden, 1966).
30 E. Casetti, The expansion method, mathematical modeling, and spatial econometrics. *Int. Reg. Sci. Rev.* **20**, 9–33 (1997).
31 A. S. Fotheringham, C. Brunsdon, M. Charlton, *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships* (Wiley, Hoboken, NJ, 2002).
32 A. Páez, "Local analysis of spatial relationships: A comparison of GWR and the expansion method." in *International Conference on Computational Science and Its Applications* (Springer, Berlin, Heidelberg, 2005) pp. 162–172.
33 A. S. Fotheringham, W. Yang, W. Kang, Multiscale geographically weighted regression (mgwr). *Ann. Assoc. Am. Geogr.* **107**, 1247–1265 (2017).
34 S. Banerjee, B. P. Carlin, A. E. Gelfand, *Hierarchical Modeling and Analysis for Spatial Data* (CRC Press, Boca Raton, FL, 2014).
35 L. A. Waller, L. Zhu, C. A. Gotway, D. M. Gorman, P. J. Gruenewald, Quantifying geographic variations in associations between alcohol distribution and violence: A comparison of geographically weighted regression and spatially varying coefficient models. *Stochastic Environ. Res. Risk Assess.* **21**, 573–588 (2007).
36 L. J. Wolf, T. M. Oshan, A. S. Fotheringham, Single and multiscale models of process spatial heterogeneity. *Geogr. Anal.* **50**, 223–246 (2018).
37 D. A. Griffith, Spatial-filtering-based contributions to a critique of geographically weighted regression (GWR). *Environ. Plann. A* **40**, 2751–2769 (2008).

38 T. M. Oshan, A. S. Fotheringham, A comparison of spatially varying regression coefficient estimates using geographically weighted and spatial-filter-based techniques. *Geogr. Anal.* **50**, 53–75 (2018).
39 D. C. Wheeler, Simultaneous coefficient penalization and model selection in geographically weighted regression: The geographically weighted lasso. *Environ. Plann. A* **41**, 722–742 (2009).
40 W. Li, GeoAI: Where machine learning and big data converge in GIScience. *Journal of Spatial Information Science* **20**, 71–77 (2020).
41 M. Hutson, Artificial intelligence faces reproducibility crisis. *Science* **359**, 725–726 (2018).
42 M. N. Kamel Boulos, G. Peng, T. VoPham, An overview of GeoAI applications in health and healthcare. *Int. J. Health Geogr.* **18**, 7 (2019).
43 K. Janowicz, S. Gao, G. McKenzie, Y. Hu, B. Bhaduri, GeoAI: Spatially explicit artificial intelligence techniques for geographic knowledge discovery and beyond. *Int. J. Geogr. Inf. Sci.* **34**, 625–636 (2020).
44 S. Openshaw, C. Openshaw, *Artificial Intelligence in Geography* (John Wiley & Sons, Inc, New York, 1997).
45 M. Reichstein *et al.*, Deep learning and process understanding for data-driven Earth system science. *Nature* **566**, 195–204 (2019).
46 W. C. Chiou, Sr, NASA image-based geological expert system development project for hyperspectral image analysis. *Appl. Opt.* **24**, 2085–2091 (1985).
47 M. Belgiu, L. Drăguţ, Random forest in remote sensing: A review of applications and future directions. *ISPRS J. Photogramm. Remote Sens.* **114**, 24–31 (2016).
48 M. Batty, Agents, cells, and cities: New representational models for simulating multiscale urban dynamics. *Environ. Plann. A* **37**, 1373–1394 (2005).
49 W. Li, M. Batty, M. F. Goodchild, Real-time GIS for smart cities. *Int. J. Geogr. Inf. Sci.* **34**, 311–324 (2020).
50 K. C. Clarke, S. Hoppen, L. Gaydos, A self-modifying cellular automaton model of historical urbanization in the San Francisco Bay area. *Environ. Plann. B Plann. Des.* **24**, 247–261 (1997).
51 C. M. Hosage, M. F. Goodchild, Discrete space location-allocation solutions from genetic algorithms. *Ann. Oper. Res.* **6**, 35–46 (1986).
52 S. Georganos *et al.*, Geographical random forests: A spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling. *Geocarto Int.* **36**, 121–136 (2021).
53 J. Hagenauer, M. Helbich, A geographically weighted artificial neural network. *Int. J. Geogr. Inf. Sci.* 10.1080/13658816.2021.1871618. (2021).
54 T. J. Sejnowski, The unreasonable effectiveness of deep learning in artificial intelligence. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 30033–30038 (2020).
55 L. Fletcher, P. Carruthers, Metacognition and reasoning. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **367**, 1366–1378 (2012).
56 J. Pearl, D. Mackenzie, *The Book of Why: The New Science of Cause and Effect* (Basic Books, New York, 2018).
57 A. Karpatne *et al.*, Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Trans. Knowl. Data Eng.* **29**, 2318–2331 (2017).
58 S. Sabour, N. Frosst, G. E. Hinton, "Dynamic routing between capsules" in *Advances in Neural Information Processing Systems*, I. Guyon, Ed. *et al.* (MIT Press, Long Beach, CA, 2017), pp. 3856–3866.
59 J. Vanschoren, J. N. Van Rijn, B. Bischl, L. Torgo, OpenML: Networked science in machine learning. *SIGKDD Explor.* **15**, 49–60 (2014).
60 E. D. Foster, A. Deardorff, Open science framework (OSF). *J. Med. Libr. Assoc.* **105**, 203 (2017).